BIAS-AWARE INFERENCE IN FUZZY REGRESSION DISCONTINUITY DESIGNS

Claudia Noack Christoph Rothe

Abstract

We propose new confidence sets (CSs) for the regression discontinuity parameter in fuzzy designs. Our CSs are based on local linear regression, and are bias-aware, in the sense that they take possible bias explicitly into account. Their construction shares similarities with that of Anderson-Rubin CSs in exactly identified instrumental variable models, and thereby avoids issues with "delta method" approximations that underlie most commonly used existing inference methods for fuzzy regression discontinuity analysis. Our CSs are asymptotically equivalent to existing procedures in canonical settings with strong identification and a continuous running variable. However, they are also valid under a wide range of other empirically relevant conditions, such as setups with discrete running variables, donut designs, and weak identification.

1. INTRODUCTION

Regression discontinuity designs (Thistlethwaite and Campbell, 1960; Hahn et al., 2001) can deliver credible identification of treatment effects from observational data in settings where the probability of receiving the treatment changes discontinuously with a running variable at some known threshold value. Such designs are called sharp (SRD) if the probability changes from zero to one, and fuzzy (FRD) otherwise. With both types of designs, methods based on local linear regression are widely used in empirical practice for estimation and inference.

The confidence intervals (CIs) typically reported in empirical FRD studies are obtained by applying techniques for handling smoothing bias, such as robust bias correction (Calonico

First Version: June 11, 2019. This Version: October 11, 2023. We would like to thank, Tim Armstrong, Marinho Bertanha, Yingying Dong, Keisuke Hirano, Michal Kolesár, the anonymous referees and numerous seminar participants for their helpful comments and suggestions. The authors gratefully acknowledge financial support by the European Research Council (ERC) through grant SH1-77202. Contact information: Claudia Noack, University of Bonn, email: claudia.noack@uni-bonn.de, 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.

et al., 2014) or bias-aware critical values (Armstrong and Kolesár, 2018), to a delta method (DM) approximation of the FRD estimator. Such DM CIs can be unreliable in practice, however, if the running variable is not continuously distributed with full support around the cutoff, or the jump in treatment probabilities is "small". These limitations are important because empirical researchers often face running variables like test scores or class sizes that take only a limited number of distinct values, "donut designs" (Barreca et al., 2011) that exclude units close to the cutoff to increase the credibility of causal estimates, or weakly identified setups where treatment assignment only has a moderate impact on treatment probabilities.[1]

In this paper, we propose a new class of FRD confidence sets (CSs) that are not subject to these issues. The idea is to apply a bias-aware approach, which takes possible finite sample biases into account, to a particular local linear SRD estimator that is conceptually similar to an Anderson-Rubin statistic in a linear IV model (Staiger and Stock, 1997; Feir et al., 2016). We show that the resulting CSs are "honest" in the sense of Li (1989), meaning that they have correct asymptotic coverage uniformly over a class of conditional expectation functions of outcomes and treatments with bounded second derivatives, irrespective of the distribution of the running variable or the strength of identification. We also show that our CSs are asymptotically equivalent to bias-aware DM CIs in settings with a continuous running variable and strong identification.

Regression discontinuity methods that explicitly take possible bias into account have been shown to have favorable theoretical and practical properties, for instance, by Armstrong and Kolesár (2018, 2020), Kolesár and Rothe (2018) and Imbens and Wager (2019). The CSs in this paper complement these methods, as they allow for reliable FRD inference settings where DM CIs can fail without sacrificing efficiency in the canonical setup. Our approach is related to that of Feir et al. (2016), who also consider Anderson-Rubin-type statistics in FRD designs with potentially small jumps in treatment probabilities, but differs in that it allows for discrete (or otherwise irregularly supported) running variables, takes potential bias explicitly into account, and includes a method for choosing bandwidths in practice.

[1]To illustrate the scope of the issue, we surveyed the articles published between 2015 and 2021 in the "Top 5" economics journals. We found 20 papers that used a fuzzy regression discontinuity design as one of their main empirical specification; 9 of which had an "irregular support" (in the sense of having less than 100 support points within the bandwidth window on each side of the cutoff), and one was potentially affected by weak identification (in the sense that the reported first stage estimate differed by less than three standard errors from zero).

2. SETUP AND PRELIMINARIES

Let $Y_i \in \mathbb{R}$ be the outcome, $T_i \in \{0,1\}$ the actual treatment status, $Z_i \in \{0,1\}$ the assigned treatment, and $X_i \in \mathbb{R}$ the running variable of the ith unit in a random sample of size n from a large population. Treatment is assigned if the running variable falls above a known cutoff that we normalize to zero, so that $Z_i = \mathbf{1}\{X_i \geq 0\}$. The parameter of interest is $\theta = \tau_Y/\tau_T$, where for a generic random variable W_i we write $\mu_W(x) = \mathbb{E}(W_i|X_i = x)$ for is its conditional expectation function given the running variable; $\mu_{W+} = \lim_{x\downarrow 0} \mu_W(x)$ and $\mu_{W-} = \lim_{x\uparrow 0} \mu_W(x)$ for the right and left limit at the cutoff; and $\tau_W = \mu_{W+} - \mu_{W-}$ for the corresponding jump.[2] In a potential outcomes framework with certain continuity and monotonicity conditions (e.g. Hahn et al., 2001), the parameter θ has a causal interpretation as the local average treatment effect among units at the cutoff whose treatment decision is affected by the assignment rule.

Our goal is to construct powerful confidence sets $\mathcal{C}^\alpha$ that cover the parameter θ in large samples with at least probability $1-\alpha$, uniformly over (μ_Y, μ_T) in some function class $\mathcal{F}$ that embodies shape restrictions imposed by the analyst:[3]

$$\liminf_{n\to\infty} \inf_{(\mu_Y,\mu_T)\in\mathcal{F}} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1-\alpha. \tag{2.1}$$

Following Li (1989), we refer to such CSs as *honest with respect to* $\mathcal{F}$. This is a stronger requirement than correct pointwise asymptotic coverage:

$$\liminf_{n\to\infty} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1-\alpha \text{ for all } (\mu_Y, \mu_T) \in \mathcal{F}. \tag{2.2}$$

In particular, under (2.1) we can always find a sample size n such that the coverage probability of $\mathcal{C}^\alpha$ is not below $1-\alpha$ by more than an arbitrarily small amount for every $(\mu_Y, \mu_T) \in \mathcal{F}$. Under (2.2) there is no such guarantee, and even in very large samples the coverage probability of $\mathcal{C}^\alpha$ could be poor for some $(\mu_Y, \mu_T) \in \mathcal{F}$. Since we do not know in advance which function pair is the correct one, honesty as in (2.1) is necessary for good finite sample coverage of $\mathcal{C}^\alpha$ across data generating processes.

As in Imbens and Wager (2019) or Armstrong and Kolesár (2020), we specify $\mathcal{F}$ as a smoothness class. Specifically, let $\mathcal{F}_H(B) = \{f_1(x)\mathbf{1}\{x \geq 0\} - f_0(x)\mathbf{1}\{x < 0\} : \|f_w''\|_\infty \leq B, w = 0,1\}$ be the Hölder-type class of real functions that are potentially discontinuous at

[2] We write $f_+ = \lim_{x\downarrow 0} f(x)$ and $f_- = \lim_{x\uparrow 0} f(x)$ for generic functions f throughout the paper.

[3] Note that we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on μ_Y and μ_T implicit in our notation. Each function pair (μ_Y, μ_T) corresponds to a single distribution of $(Y, T, X, Z) = (\mu_Y(X) + \epsilon_M, \mathbf{1}\{\mu_T(X) \geq \epsilon_T\}, X, Z)$, where (ϵ_M, ϵ_T) is some fixed random vector.

zero, are twice differentiable on either side of the threshold, and whose second derivatives are uniformly bounded by some constant $B > 0$; and let $\mathcal{F}_H^\delta(B) = \{f \in \mathcal{F}_H(B) : |f_+ - f_-| > \delta\}$ be a similar class of functions whose jump at zero is larger than some $\delta \geq 0$. We then assume that there are constants B_Y and B_T, whose choice we discuss in Section 6.3, such that

$$(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^0(B_T) \equiv \mathcal{F}. \tag{2.3}$$

As $\mathcal{F}$ is a Cartesian product, this rules out cross-restrictions between μ_Y and μ_T. Note that we impose $\mu_T \in \mathcal{F}_H^0(B_T)$, and thus $\tau_T \neq 0$, only to ensure that $\theta = \tau_Y/\tau_T$ is well-defined. We explicitly allow τ_T to be arbitrarily close to zero.

If the running variable is discrete, or more generally such that there are gaps in its support, condition (2.3) is understood to mean that there exists a function pair $(\mu_Y, \mu_T) \in \mathcal{F}$ such that $(\mu_Y(X_i), \mu_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i))$ with probability one (cf. Kolesár and Rothe, 2018). With this interpretation, the parameter θ is generally partially identified:

$$\theta \in \Theta_I \equiv \left\{ \frac{m_{Y+} - m_{Y-}}{m_{T+} - m_{T-}} : (m_Y, m_T) \in \mathcal{F},\ (m_Y(X_i), m_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i)) \right\},$$

where the identified set Θ_I is either (i) a closed interval $[a_1, a_2]$ with $a_1 \leq a_2$; (ii) the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$ with $a_1 < 0 < a_2$; (iii) the entire real line; or, as a knife-edge case (iv) a half-line $[a_1, \infty)$ or $(-\infty, -a_1]$, with $a_1 > 0$.[4] The classical point identification result when the support of X_i contains an open neighborhood around the cutoff is then simply a special case of (i) with Θ_I a singleton. Our goal is to construct CSs for θ that have correct uniform asymptotic coverage under both point and partially identification (cf. Imbens and Manski, 2004), without applied researchers having to decide which of the two notions of identification more accurately applies to their specific setting.

3. BIAS-AWARE ANDERSON-RUBIN-TYPE CONFIDENCE SETS

We argue in Section 5 that conventional CIs, based on local linear regressions (Fan and Gijbels, 1996) and "delta method" (DM) arguments, can potentially break down in a number of practically relevant setups, including ones with discrete running variables or weak identification. Our proposed approach is still based on local linear regression, but avoids these issues through a construction similar to that of Anderson and Rubin (1949) for inference in exactly identified linear IV models. It also takes possible bias from local linear smoothing

[4]This holds because the range of $(m_{Y+} - m_{Y-}, m_{T+} - m_{T-})$ over $(m_Y, m_T) \in \mathcal{F}$ is a Cartesian product of two intervals $I_Y \times I_T$. The four cases then obtain depending on which of these two intervals contain zero, possibly as a boundary value.

explicitly into account. We hence refer to our CSs as *bias-aware AR CSs*.

To describe the approach, we write $\widehat{\tau}_W(h)$ for the local linear estimator of the jump τ_W in the conditional expectation of some generic random variable W_i given the running variable X_i at the cutoff:

$$\widehat{\tau}_W(h) = e_1^\top \operatorname*{argmin}_{\beta\in\mathbb{R}^4} \sum_{i=1}^n K(X_i/h)(W_i - \beta^\top(Z_i, X_i, Z_iX_i, 1))^2 = \sum_{i=1}^n w_i(h)W_i. \tag{3.1}$$

Here $K(\cdot)$ is a kernel function, $h > 0$ is a bandwidth, $e_1 = (1,0,0,0)^\top$ is the first unit vector, and the $w_i(h)$ are weights, given explicitly in Appendix A, that only depend on the data through the realizations $\mathcal{X}_n = (X_1, \ldots, X_n)^\top$ of the running variable. We refer to estimators of the form in (3.1) as *SRD-type estimators of* τ_W in the following.

The natural point estimator of θ is $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$, but we will base inference on different statistics. Define the auxiliary parameters $\tau_M(c) = \tau_Y - c\tau_T$ for $c \in \mathbb{R}$, and note that $\tau_M(c) = \mu_{M+}(c) - \mu_{M-}(c)$, with $\mu_M(x,c) = \mathbb{E}(M_i(c)|X_i = x)$ and $M_i(c) = Y_i - cT_i$. We then consider the SRD-type estimator $\widehat{\tau}_M(h,c) = \sum_{i=1}^n w_i(h)M_i(c)$, and exploit the properties of the regression weights $w_i(h)$ to write its conditional bias $b_M(h,c) = \mathbb{E}(\widehat{\tau}_M(h,c)|\mathcal{X}_n) - \tau_M(c)$ and conditional variance $s_M^2(h,c) = \mathbb{V}(\widehat{\tau}_M(h,c)|\mathcal{X}_n)$ given $\mathcal{X}_n$ as

$$b_M(h,c) = \sum_{i=1}^n w_i(h)\mu_M(X_i,c) - (\mu_{M+}(c) - \mu_{M-}(c)), \quad s_M^2(h,c) = \sum_{i=1}^n w_i(h)^2\sigma_{M,i}^2(c),$$

respectively, with $\sigma_{M,i}^2(c) = \mathbb{V}(M_i(c)|X_i)$ the conditional variance of $M_i(c)$ given X_i.[5]

The bias depends on (μ_Y, μ_T) through the transformation $\mu_M = \mu_Y - c\mu_T$ only, and we have that $\mu_M \in \mathcal{F}_H(B_Y + |c|B_T)$ by (2.3). As in Armstrong and Kolesár (2020), for any value of the bandwidth h we can thus explicitly bound $b_M(h,c)$ in absolute value over $\mathcal{F}$:

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_M(h,c)| \le \overline{b}_M(h,c) \equiv -\frac{B_Y + |c|B_T}{2} \sum_{i=1}^n w_i(h)X_i^2\operatorname{sign}(X_i).$$

The supremum is achieved by the "worst case" pair of piecewise quadratic conditional expectation function with second derivatives equal to $(B_Y\operatorname{sign}(x), -B_T\operatorname{sign}(x))$ over $x \in [-h,h]$.[6]

[5]To keep the notation simple, the estimator $\widehat{\tau}_M(h,c) = \widehat{\tau}_Y(h) - c\widehat{\tau}_T(h)$ uses the same bandwidth on each side of the cutoff, and also the same bandwidth for estimating τ_Y and τ_T. It is straightforward to accommodate more general bandwidth choices; see Online Appendix B for details.

[6]Note that this bound may not be sharp if no such pair of piecewise quadratic functions is a feasible candidate for (μ_Y, μ_T). For example, there is no function μ_T with $\mu_T''(x) = B_T\operatorname{sign}(x)$ and $\mu_T(x) \in [0,1]$ for all $x \in [-h,h]$ if $h > (2/B_T)^{1/2}$. Still, the bias bound is valid in such cases.

For every $c \in \mathbb{R}$ we can then construct an infeasible bias-aware CI for $\tau_M(c)$ as

$$C_M^\alpha(h,c) = \left[\widehat{\tau}_M(h,c) \pm \mathrm{cv}_{1-\alpha}(r_M(h,c))s_M(h,c)\right],$$

where $r_M(h,c) = \overline{b}_M(h,c)/s_M(h,c)$ is the "worst case" bias to standard deviation ratio, and $\mathrm{cv}_{1-\alpha}(r)$ is the $(1-\alpha)$-quantile of "folded" normal distribution $|N(r,1)|$. Armstrong and Kolesár (2018, 2020) show that such CIs are honest with respect to $\mathcal{F}_H(B_W)$ irrespective of the distribution of the running variable, have correct asymptotic coverage $1-\alpha$ at the "worst case" conditional expectations, are valid for wide ranges of bandwidths, and are highly efficient for SRD inference if the running variable is continuous.

The bandwidth that minimizes this CI's asymptotic length is

$$h_M(c) = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}(r_M(h,c))s_M(h,c).$$

We assume that this optimal bandwidth is unique; and the minimization in its definition is understood to be carried out over the set of bandwidths for which the involved objects are well-defined. An efficient but infeasible bias-aware AR CS for θ is then given by the set of all $c \in \mathbb{R}$ for which the auxiliary CI $C_M^\alpha(h_M(c),c)$ contains the value zero:

$$\mathcal{C}_*^\alpha = \left\{c : |\widehat{\tau}_M(h_M(c),c)| \le \mathrm{cv}_{1-\alpha}(r_M(h_M(c),c))s_M(h_M(c),c)\right\}. \tag{3.2}$$

Our proposed bias-aware AR CSs are feasible versions of (3.2) based on a standard error $\widehat{s}_M(h,c) = \sum_{i=1}^n w_i(h)^2\widehat{\sigma}_{M,i}^2(c)$ and some estimate $\widehat{h}_M(c)$ of the optimal bandwidth:

$$\mathcal{C}_{\mathrm{ar}}^\alpha = \left\{c : |\widehat{\tau}_M(\widehat{h}_M(c),c)| \le \mathrm{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c),c))\widehat{s}_M(\widehat{h}_M(c),c))\right\}, \tag{3.3}$$

with $\widehat{r}_M(h,c) = \overline{b}_M(h,c)/\widehat{s}_M(h,c)$. Both the standard error and bandwidth estimator can be implemented in different ways, and our theoretical analysis below therefore only imposes some weak "high level" conditions. We propose a specific standard error based on nearest-neighbor linear regression estimates $\widehat{\sigma}_{M,i}^2(c)$ of $\sigma_{M,i}^2(c)$ in Section 6.1; and a feasible bandwidth that combines a plug-in construction with a safeguard against certain small sample distortions in Section 6.2.

4. THEORETICAL PROPERTIES

4.1. **Coverage.** Our main theoretical result is that $\mathcal{C}_{\mathrm{ar}}^\alpha$ is an honest CS for θ with respect to $\mathcal{F}$, in the sense of (2.1), under the following rather weak conditions.

Assumption 1. *(i) The data* $\{(Y_i, T_i, X_i), i = 1, \ldots, n\}$ *are an i.i.d. sample; (ii)* $\mathbb{E}((Y_i - \mathbb{E}(Y_i|X_i))^q|X_i = x)$ *exists and is bounded uniformly over* $x \in supp(X_i)$ *and* $(\mu_Y, \mu_T) \in \mathcal{F}$ *for some* $q > 2$*; (iii)* $\mathbb{V}(Y_i|X_i = x)$ *is bounded away from zero uniformly over* $x \in supp(X_i)$ *and* $(\mu_Y, \mu_T) \in \mathcal{F}$*; and* $Cov(Y_i, T_i|X_i = x)^2/(\mathbb{V}(Y_i|X_i = x)\mathbb{V}(T_i|X_i = x))$ *is bounded away from one uniformly over* $x \in supp(X_i) \cup \{x : \mathbb{V}(T_i|X_i = x) > 0\}$ *and* $(\mu_Y, \mu_T) \in \mathcal{F}$*; (iv) the kernel function* K *is a continuous, unimodal, symmetric density function that is equal to zero outside some compact set, say* $[-1, 1]$.

Assumption 1 collects mostly standard conditions from the literature on local linear regression. Part (i) could be weakened to allow for certain forms of dependent sampling, such as cluster sampling. Parts (ii) and (iii) ensure that $\mathbb{V}(M_i(c)|X_i = x)$ is bounded away from zero for all $c \in \mathbb{R}$ and allow for the special case of a SRD design. Part (iv) is satisfied by most kernel functions commonly used in applied RD analysis, such as the triangular or the Epanechnikov kernels.

Assumption 2. *The following holds uniformly over* $(\mu_Y, \mu_T) \in \mathcal{F}$*: (i)* $\widehat{h}_M(c) = h_M(c)(1 + o_P(1))$*; and (ii)* $\widehat{s}_M(\widehat{h}_M(c), c) = s_M(h_M(c), c)(1 + o_P(1))$.

Part (i) of Assumption 2 states that the empirical bandwidth is consistent for the infeasible optimal one, and part (ii) states that the empirical standard error is consistent for the true standard deviation at the infeasible optimal bandwidth. We discuss specific implementations in Sections 6.1 and 6.2.

Assumption LL1. *The support of the running variable* X_i *is finite and symmetric, in the sense that it is of the form* $\{\pm x_1, \ldots, \pm x_k\}$*, for positive constants* $(x_1, \ldots, x_k)$ *over some open neighborhood of the cutoff.*

Assumption LL2. *The running variable* X_i *is continuously distributed with continuous density* f_X *that is bounded and bounded away from zero over an open neighborhood of the cutoff.*

Assumptions LL1–LL2 describe RD setups with discrete and continuously distributed running variables, respectively. These settings are meant to be exemplary and are considered because they allow explicit characterization of the bandwidth $h_M(c)$. Note that the symmetry of the support in Assumption LL1 is for notational convenience only. Discrete running variables with asymmetric support can easily be accommodated by using a different bandwidth on each side of the cutoff, as described in Online Appendix B. In Appendix A.1.1 we also consider an alternative asymptotic framework for the discrete case.

In Lemma A.1 in the Appendix, we show that our assumptions have two main implications: (i) using an estimate of the optimal bandwidth instead of its population version has a small impact, in some appropriate sense, on the quantities involved in the construction of our CS; (ii) the magnitude of each weight $w_i(h_M(c))$ is small relative to the others' in large samples, in the sense that $w_{\text{ratio}}(h) \equiv \max_{j=1,\dots,n} w_j(h)^2 / \sum_{i=1}^n w_i(h)^2 = o_P(1)$, so that a CLT applies to an appropriately standardized version of the estimator of $\tau_M(c)$. This yields the following formal result.

Theorem 1. *Suppose that Assumptions 1–2 and either LL1 or LL2 hold. Then $\mathcal{C}_{\text{ar}}^{\alpha}$ is honest with respect to $\mathcal{F}$ in the sense of* (2.1)*.*

4.2. **Shape.** Because our CS $\mathcal{C}_{\text{ar}}^{\alpha}$ is defined through an inversion argument, it is interesting to study its shape. A simple sufficient condition for $\mathcal{C}_{\text{ar}}^{\alpha}$ to be non-empty is that the bandwidth $\widehat{h}_M(c)$ is continuous in c, but beyond that it is difficult to make general statements. To see why, recall that $c \in \mathcal{C}_{\text{ar}}^{\alpha}$ if and only if

$$|\widehat{\tau}_M(\widehat{h}_M(c), c)| \leq \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c), c))\widehat{s}_M(\widehat{h}_M(c), c).$$

The above quantities depend on c directly, but also indirectly through $\widehat{h}_M(c)$. While the former dependence is rather simple in structure, the latter introduces complicated nonlinearities that make it impossible to give a simple analytical result regarding the shape of our CS. Such a result is possible, however, for a version that uses a fixed bandwidth.

Theorem 2. *Let $\mathcal{C}_{\text{ar}}^{\alpha}(h)$ be a version of $\mathcal{C}_{\text{ar}}^{\alpha}$ that uses a bandwidth h that does not depend on c. Then either $\mathcal{C}_{\text{ar}}^{\alpha}(h) = [a_1, a_2]$, or $\mathcal{C}_{\text{ar}}^{\alpha}(h) = (-\infty, a_1] \cup [a_2, \infty)$, or $\mathcal{C}_{\text{ar}}^{\alpha}(h) = (-\infty, \infty)$, or $\mathcal{C}_{\text{ar}}^{\alpha}(h) = [a_1, \infty)$ or $\mathcal{C}_{\text{ar}}^{\alpha}(h) = (-\infty, a_1]$, for some constants $a_1 < a_2$.*

The result mirrors the discussion at the end of Section 2. It suggests that our actual CS should take one of these general shapes as long as $\widehat{h}_M(c)$ does not vary "too much" with c. We found this to be the case in every simulation run and every empirical analysis that we conducted in the context of this paper. The second case in Theorem 2, in which the CS is not an interval, is a natural description of uncertainty in settings in which the data clearly suggest that τ_Y is non-zero, but cannot rule out the possibility that τ_T is not. The last two cases in Theorem 2, in which $\mathcal{C}_{\text{ar}}^{\alpha}(h)$ is a half-line, are also "knife-edge" cases: they only occur if one of the boundaries of a bias-aware CI for τ_T is exactly equal to zero, and are thus largely irrelevant for empirical practice.

5. COMPARISON WITH DELTA METHOD INFERENCE

5.1. **Method and Limitations.** The CIs commonly reported in empirical FRD studies are based on a linearization or "delta method" (DM) argument. It starts by noting that, after centering, the FRD point estimator $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$ can be written as the sum of an SRD-type estimator $\widehat{\tau}_U(h)$ with unobserved dependent variable U_i, and a remainder $\widehat{\rho}(h)$:

$$\widehat{\theta}(h) - \theta = \widehat{\tau}_U(h) + \widehat{\rho}(h), \quad \widehat{\tau}_U(h) = \sum_{i=1}^{n} w_i(h) U_i, \quad U_i = \frac{Y_i - \tau_Y}{\tau_T} - \frac{\tau_Y(T_i - \tau_T)}{\tau_T^2},$$
$$\widehat{\rho}(h) = \frac{\widehat{\tau}_Y(h)(\widehat{\tau}_T(h) - \tau_T)^2}{2\widehat{\tau}_T^*(h)^3} - \frac{(\widehat{\tau}_Y(h) - \tau_Y)(\widehat{\tau}_T(h) - \tau_T)}{\tau_T^2},$$

with $\widehat{\tau}_T^*(h)$ an intermediate value between τ_T and $\widehat{\tau}_T(h)$. One then imposes conditions under which $\widehat{\rho}(h)$ is asymptotically negligible relative to $\widehat{\tau}_U(h)$, and forms a CI for θ by applying some method for SRD inference to $\widehat{\tau}_U(h)$, which differ mainly in how they handle potential bias. Such DM CIs are proposed, for example, by Calonico et al. (2014) and Armstrong and Kolesár (2020) in combination with robust bias correction and a bias-aware approach, respectively.[7] As U_i is unobserved, any such method must also be made feasible by using an estimate $\widehat{U}_i$ in which τ_Y and τ_T are replaced by suitable preliminary estimators.

Obvious downsides of such constructions include that they only control the bias of a first-order approximation of $\widehat{\theta}(h)$, and not the bias of $\widehat{\theta}(h)$ itself; and that replacing U_i with an estimate $\widehat{U}_i$ introduces additional uncertainty, which is asymptotically second-order and hence generally unaccounted for in practice. In FRD designs such CIs are therefore generally subject to additional finite-sample distortions, relative to SRD designs.

A more principal issue with DM CIs is that a central condition for their validity, namely that $\widehat{\rho}(h)$ is asymptotically negligible relative to $\widehat{\tau}_U(h)$, is not innocuous. In particular, this condition is not compatible with a discrete running variable, or more generally one with support gaps around the cutoff. This is because τ_T and τ_Y are generally only partially identified in this case, and hence cannot be consistently estimated; see Section 2. The term $\widehat{\rho}(h)$ then generally has a non-zero probability limit, and cannot be ignored for the purpose of inference on θ. This issue occurs irrespective of the method chosen to control the bias of $\widehat{\tau}_U(h)$, including bias-aware inference. Because running variables with discrete or irregular support are ubiquitous in practice, this is an important limitation.

[7] In empirical papers, FRD estimates are sometimes obtained through the two-stage least squares regression $Y_i = \theta T_i + \beta_+ X_i Z_i + \beta_- X_i(1 - Z_i) + \varepsilon_i$ with Z_i as an instrument for T_i, using only data in some window around the cutoff. This is numerically equivalent to a ratio of local linear regressions with a uniform kernel, and the resulting CI is thus of the DM type (Hahn et al., 2001; Imbens and Lemieux, 2008).

Another issue with DM CIs is that the conditions for their uniform validity rule out weakly identified settings with τ_T close to zero. This problem occurs even if the running variable is continuously distributed, and with any method chosen to control the bias of $\widehat{\tau}_U(h)$, including bias-aware inference. This is because for any DM CI to be honest with respect to $\mathcal{F}$, the term $\widehat{\rho}(h)$ must be of smaller order than $\widehat{\tau}_U(h)$ not only at the "true" function pair (μ_Y, μ_T), but uniformly over all $(\mu_Y, \mu_T) \in \mathcal{F}$. But since τ_T can be arbitrarily close to zero over $(\mu_Y, \mu_T) \in \mathcal{F}$, we have that $\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |\widehat{\rho}(h)| = \infty$, which means that DM CIs can be unreliable in such settings.[8]

5.2. **An Equivalence Result.** Armstrong and Kolesár (2020) study bias-aware DM CIs under conditions for which such DM CIs are asymptotically valid. These include Assumption LL2, which implies that is X_i continuously distributed, and that $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^\delta(B_T) \equiv \mathcal{F}^\delta$ for some $\delta > 0$, which means that τ_T is well-separated from zero. They show that bias-aware DM CIs are honest with respect to $\mathcal{F}^\delta$ in this case, and also near-optimal, in the sense that no other method can substantially improve upon their length in large samples. The next theorem shows that our bias-aware AR CSs are as efficient as their DM counterparts in such settings for which DM CIs are specifically designed.

To avoid introducing additional high-level assumptions about the implementation details of bias-aware DM CIs we consider an infeasible version $\mathcal{C}_\Delta^\alpha$, formally defined in (A.1), and compare it to its infeasible counterpart $\mathcal{C}_*^\alpha$ in our setup. Equal efficiency is established in the sense that both CSs have the same local asymptotic coverage for a drifting parameter within a $O(n^{-2/5})$ neighborhood of θ. Such neighborhoods are appropriate to consider as the length of $\mathcal{C}_\Delta^\alpha$ is $O_P(n^{-2/5})$ uniformly over $\mathcal{F}^\delta$.

Theorem 3. *Suppose that Assumptions 1–2 and LL2 hold, and put $\theta^{(n)} = \theta + \kappa n^{-2/5}$ for some constant κ. Then*

$$\limsup_{n\to\infty} \sup_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} \left|\mathbb{P}\left(\theta^{(n)} \in \mathcal{C}_*^\alpha\right) - \mathbb{P}\left(\theta^{(n)} \in \mathcal{C}_\Delta^\alpha\right)\right| = 0.$$

This result parallels the well-known finding that there is no loss of efficiency when using the AR approach in exactly identified IV models relative to one based on a conventional t-test (e.g. Andrews et al., 2019). It is not an obvious corollary, however, as there are, for

[8]Feir et al. (2016) also point out coverage issues of DM CIs under weak identification, but use different types of arguments. Specifically, they show that DM CIs based on infeasible "undersmoothing" bandwidths do not have correct asymptotic coverage under pointwise (with respect to the involved conditional expectation functions) asymptotics if τ_T tends to zero at an appropriate rate related to that of the bandwidth. They also show that undersmoothing AR CSs can have correct poinwise asymptotic coverage in this case.

example, no analogues to the bandwidth and the smoothing bias in such IV models.

6. IMPLEMENTATION DETAILS AND EXTENSIONS

6.1. **Standard Errors.** Natural standard errors for $\widehat{\tau}_M(h,c)$ are of the form $\widehat{s}_M(h,c) = (\sum_{i=1}^n w_i(h)^2 \widehat{\sigma}^2_{M,i}(c))^{1/2}$, with $\widehat{\sigma}^2_{M,i}(c)$ some estimate of $\sigma^2_{M,i}(c)$. Setting $\widehat{\sigma}^2_{M,i}(c)$ to the squared difference between the outcome of unit i and the average outcome among its nearest neighbors in terms of the running variable (Abadie and Imbens, 2006; Abadie et al., 2014) is commonly recommended in the RD literature (e.g. Calonico et al., 2014). However, this nearest-neighbor standard error is actually not uniformly consistent over $\mathcal{F}$ because the leading bias of $\widehat{\sigma}^2_{M,i}(c)$ is proportional to the first derivative of $\mu_M(\cdot,c)$ at X_i, which is unbounded over $\mathcal{F}$. We therefore propose a novel procedure that replaces the local sample average with a local best linear predictor. This modification makes the bias of $\widehat{\sigma}^2_{M,i}(c)$ proportional to the second derivative of $\mu_M(\cdot,c)$ at X_i, which is bounded in absolute value over $\mathcal{F}$ by $B_Y + |c|B_T$.

We propose a version that explicitly allows for ties among the realizations of the running variable. For R a small integer, denote the rank of $|X_j - X_i|$ among the elements of the set $\{|X_s - X_i| : s \in \{1,\dots,n\} \setminus \{i\}, X_s X_i > 0\}$ by $r(j,i)$, let $\mathcal{R}_i$ be the set of indices such that $r(j,i) \leq Q_i$, where Q_i is the smallest integer such that $\mathcal{R}_i$ contains at least R elements, and let R_i be the resulting cardinality of $\mathcal{R}_i$.[9] The estimator $\widehat{\sigma}^2_{M,i}(c)$ is then defined as the scaled squared difference between $M_i(c)$ and its best linear predictor given its R_i nearest neighbors:

$$\widehat{\sigma}^2_{M,i}(c) = \frac{1}{1+H_i}\left(M_i(c) - \widehat{M}_i(c)\right)^2, \text{ with}$$
$$\widehat{M}_i(c) = \widetilde{X}_i \left(\sum_{j\in\mathcal{R}_i} \widetilde{X}_j^\top \widetilde{X}_j\right)^{-1} \sum_{j\in\mathcal{R}_i} \widetilde{X}_j^\top M_j(c), \quad H_i = \widetilde{X}_i \left(\sum_{j\in\mathcal{R}_i} \widetilde{X}_j^\top \widetilde{X}_j\right)^{-1} \widetilde{X_i}^\top.$$

Here $\widetilde{X}_i = (1, X_i)^\top$ if the running variable takes at least two distinct values among the R_i nearest neighbors of unit i, and $\widetilde{X}_i = 1$ otherwise. The scaling term H_i ensures that $\widehat{\sigma}^2_{M,i}(c)$ is approximately unbiased in large samples. The next result, which we prove in Online Appendix A, shows that our new standard error is indeed uniformly consistent under general conditions. We recommend its use not just for our CS, but more generally for bias-aware inference methods that work with bounds on second derivatives.

Theorem 4. *Suppose that Assumption 1, Assumption 2(i), and either Assumption LL1 or Assumption LL2 are satisfied; that* $\mathbb{V}(Y_i|X_i = x)$, $\mathbb{V}(T_i|X_i = x)$ *and* $Cov(Y_i, T_i|X_i = x)$ *are*

[9]Note that if every realization of X_i is unique, then $R = Q_i = R_i$, and $\mathcal{R}_i$ is the set of unit i's R nearest neighbors' indices; but with ties in the data R_i could be greater than R. We use $R = 5$ in our simulations and the empirical application.

Lipschitz continuous on each side of the cutoff uniformly over $x \in supp(X_i)$ and $(\mu_Y, \mu_T) \in \mathcal{F}$; and that $\mathbb{E}((Y_i - \mathbb{E}(Y_i|X_i))^4|X_i = x)$ is uniformly bounded over $x \in \mathbb{R}$ and $(\mu_Y, \mu_T) \in \mathcal{F}$. Then Assumption 2(ii) holds for the standard error described in this subsection.

6.2. **Bandwidth Choice.** An obvious candidate for a feasible bandwidth is the empirical analogue of $h_M(c)$, which minimizes the length of the auxiliary CI in Section 4:

$$\widehat{h}^*_M(c) = \operatorname*{argmin}_h \operatorname{cv}_{1-\alpha}(\widehat{r}_M(h,c))\widehat{s}_M(h,c).$$

While this choice is generally attractive, it could lead to coverage distortions if $B_Y + |c|B_T$ is very large relative to sampling uncertainty. To see why, recall from the discussion at the end of Section 4.1 that $\widehat{\tau}_M(h,c) = \sum_{i=1}^n w_i(h)M_i(c)$ is asymptotically normal if $w_{\text{ratio}}(h) \equiv \max_{j=1,\dots,n} w_j(h)^2 / \sum_{i=1}^n w_i(h)^2 = o_P(1)$. Normality should thus be a "good" finite-sample approximation if $w_{\text{ratio}}(h)$ is "close" to zero (this reasoning also follows from a Berry-Esseen-type result). However, if $B_Y + |c|B_T$ (and thus the worst-case bias) is large, then $\widehat{h}^*_M(c)$ is typically small. The weights $w_i(\widehat{h}^*_M(c))$ then concentrate on few observations close to the cutoff, $w_{\text{ratio}}(\widehat{h}^*_M(c))$ is large, and CLT approximations can be inaccurate as $\widehat{\tau}_M(\widehat{h}^*_M(c), c)$ then effectively behaves like a sample average of a small number of observations.

To address this issue, we propose imposing a lower bound on the bandwidth, chosen such that the value of $w_{\text{ratio}}(h)$ remains below some reasonable threshold constant $\eta > 0$, which we set to $\eta = .075$ in our simulations and empirical application.[10]

$$\widehat{h}_M(c) = \max\left\{\widehat{h}^*_M(c), h_{\min}(\eta)\right\}, \quad h_{\min}(\eta) = \min\left\{h : w_{\text{ratio}}(h) < \eta\right\}.$$

Under standard conditions like Assumption LL1 or LL2 the lower bound on the bandwidth clearly never binds asymptotically, but imposing it can improve the finite-sample coverage of our CSs: as $\widehat{h}_M(c) \geq \widehat{h}^*_M(c)$, our construction trades off a possible increase in finite-sample bias against normality being a better finite-sample approximation. This improves coverage because our CSs explicitly account for the exact bias, but cannot capture deviations from normality. This idea can also be used for SRD inference, and more generally in all settings where finite-sample accuracy of inference faces a similar "bias vs. normality" trade-off. For example, Armstrong and Kolesár (2021) use our approach in the context of inference on average treatment effects under unconfoundedness with limited overlap.

[10]To motivate this choice, suppose that $\mathcal{X}_n = \{\pm.02, \pm.04, \dots, \pm 1\}$, that $K(t) = (1-|t|)\mathbf{1}\{|t| < 1\}$ is the triangular kernel, and that $h = 1$. Then $\widehat{\tau}_M(h,c)$ is a weighted least squares estimator that gives positive weight to 50 observations on each side of cutoff, and $w_{\text{ratio}}(h) \approx .075$.

6.3. **Choosing Smoothness Bounds.** In order to compute $\mathcal{C}^{\alpha}_{\text{ar}}$, researchers needs to choose the smoothness bounds B_Y and B_T. Such bounds cannot be estimated consistently without imposing strong additional assumptions; and without choosing such bounds it is generally not possible to conduct inference on θ that is both valid and informative, even in large samples (Low, 1997; Armstrong and Kolesár, 2018; Bertanha and Moreira, 2018). Methods that seem to such choices still require restrictions on smoothness implicitly to attain approximately correct CI coverage in practice (Armstrong and Kolesár, 2020).[11] Explicitly specifying B_Y and B_T makes it transparent on which assumptions the inferential statements are based.

Roughly speaking, "small" values of B_Y and B_T amount to the assumption that the respective functions are "close" to linear on either side of the cutoff, whereas larger values allow the functions to be increasingly "curved". The choice should be guided by subject knowledge but is arguably difficult in empirical practice, where there will be no single objectively correct value. In line with the previous literature, we hence recommend considering a range of plausible values as a form of sensitivity analysis. We also recommend estimating lowers bound $\widehat{B}_{Y,\text{low}}$ and $\widehat{B}_{T,\text{low}}$, and to compute one-sided CIs for B_Y and B_T, respectively, via the methods proposed in Armstrong and Kolesár (2018) and Kolesár and Rothe (2018), to guard against overly optimistic choices.

Two heuristic "rules of thumb" (ROT) for determining plausible values in practice have also been considered in the literature. Both are based on fitting global polynomial specifications $\widetilde{\mu}_{Y,k}$ and $\widetilde{\mu}_{T,k}$ of order k on either side of the cutoff by conventional least squares. Armstrong and Kolesár (2020) use fourth-order polynomials, and propose the ROT $\widehat{B}_{Y,\text{ROT1}} = \sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{Y,4}(x)|$ and $\widehat{B}_{T,\text{ROT1}} = \sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{T,4}(x)|$, where $\mathcal{X}$ denotes the support of the running variable. Imbens and Wager (2019) consider a ROT in which the maximal curvature implied by a quadratic fit is multiplied by some moderate factor, say 2, yielding $\widehat{B}_{Y,\text{ROT2}} = 2\sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{Y,2}(x)|$ and $\widehat{B}_{T,\text{ROT2}} = 2\sup_{x\in\mathcal{X}} |\widetilde{\mu}''_{T,2}(x)|$.

Such rules of thumb can provide useful first guidance, but should be complemented with other approaches in a sensitivity analysis. We strongly recommend to always check the fit of the respective polynomial specification, and to dismiss the ROT value if the fit is obviously poor. In Online Appendix C, we argue that in "roughly quadratic" settings the fourth-order polynomial specification that underlies ROT1 tends to produce quite erratic over-fits of the

[11] For example, an undersmoothing SRD CI can only be expected to have approximately correct coverage if the bias of the local linear estimator is "small" relative to its standard error. This can only be expected if the underlying function is "close" to linear, which is equivalent to its maximum second derivative being "close" to zero. A researcher that considers an undersmoothing SRD CI to be reliable has thus implicitly imposed a smoothness bound. An analogous argument applies to robust bias correction (Kamat, 2018).

data that can lead to vast over-estimates of the true smoothness bounds, and corresponding CSs with poor statistical power. ROT2, on the other hand, tends to produce more reasonable values in many such setups. This pattern is also visible in our simulations.

6.4. **Regression Kink Designs.** In Online Appendix D, we present an extension of our approach to CSs for the ratio $\theta^{(v)} \equiv \tau_Y^{(v)}/\tau_T^{(v)} \equiv (\mu_{Y+}^{(v)} - \mu_{Y-}^{(v)})/(\mu_{T+}^{(v)} - \mu_{T-}^{(v)})$ of the jumps in the vth-order derivatives of two generic conditional expectation functions (μ_Y, μ_T) at some threshold value. This setup covers the important Fuzzy Regression Kink Design (Card et al., 2015), where the parameter of interest is the ratio $\theta^{(1)}$ of two jumps in first derivatives.

We again form bias-aware CIs for an auxiliary parameter $\tau_M^{(v)}(c) = \tau_M^{(v)} - c\tau_T^{(v)}$, now based on pth-order local polynomial regression (where $p \geq v$ and typically $p = v+1$), and collect all values of c for which such CIs contain zero to create an AR CS for $\theta^{(v)}$. The construction is largely analogous to that described in Section 3, with the main difference concerning the bias bound. Specifically, we derive the apparently novel result that if μ_Y and μ_T are both $(p+1)$ times continuously differentiable on either side of the threshold, with derivatives of order $(p+1)$ uniformly bounded by B_Y and B_T, respectively, and $\widehat{\tau}_{M,p}^{(v)}(h,c) = \sum_{i=1}^n w_{vp,i}(h) M_i(c)$ is the local pth order polynomial estimator of $\tau_M^{(v)}(c)$ with bandwidth h, the conditional bias $\mathbb{E}(\widehat{\tau}_{M,p}^{(v)}(h,c)|\mathcal{X}_n) - \tau_{M,p}^{(v)}(c)$ is absolutely bounded by

$$\bar{b}_{M,vp}(h,c) \equiv (-1)^{p-v} \frac{B_Y + |c| B_T}{(p+1)!} \sum_{i=1}^n w_{vp,i}(h) X_i^{p+1} \text{sign}(X_i)^{v+1},$$

uniformly over the (μ_Y, μ_T); see Online Appendix D for details.

7. NUMERICAL ILLUSTRATIONS

7.1. **Empirical Application.** In this subsection, we illustrate our methods revisiting data from Battistin et al. (2009), who study the effects of retirement on consumption in Italy. The data are a sample of $n = 30,006$ individuals, obtained by combining several waves of the Bank of Italy Survey on Household Income and Wealth (SHIW) for the period 1993-2004. We take the natural logarithm of total household spending as the outcome, retirement as the treatment, and years of age from the formal retirement eligibility threshold, which is normalized to zero, as the running variable. The running variable is thus discrete, but still has somewhat rich support. Figure 1 shows the average of log consumption and the empirical proportion of retired individuals in the data as a function of the running variable.

We then compute our bias-aware AR CSs with both rules of thumb ROT1 and ROT2 to choose the smoothness bounds. The resulting CSs turn out to be intervals. We compare

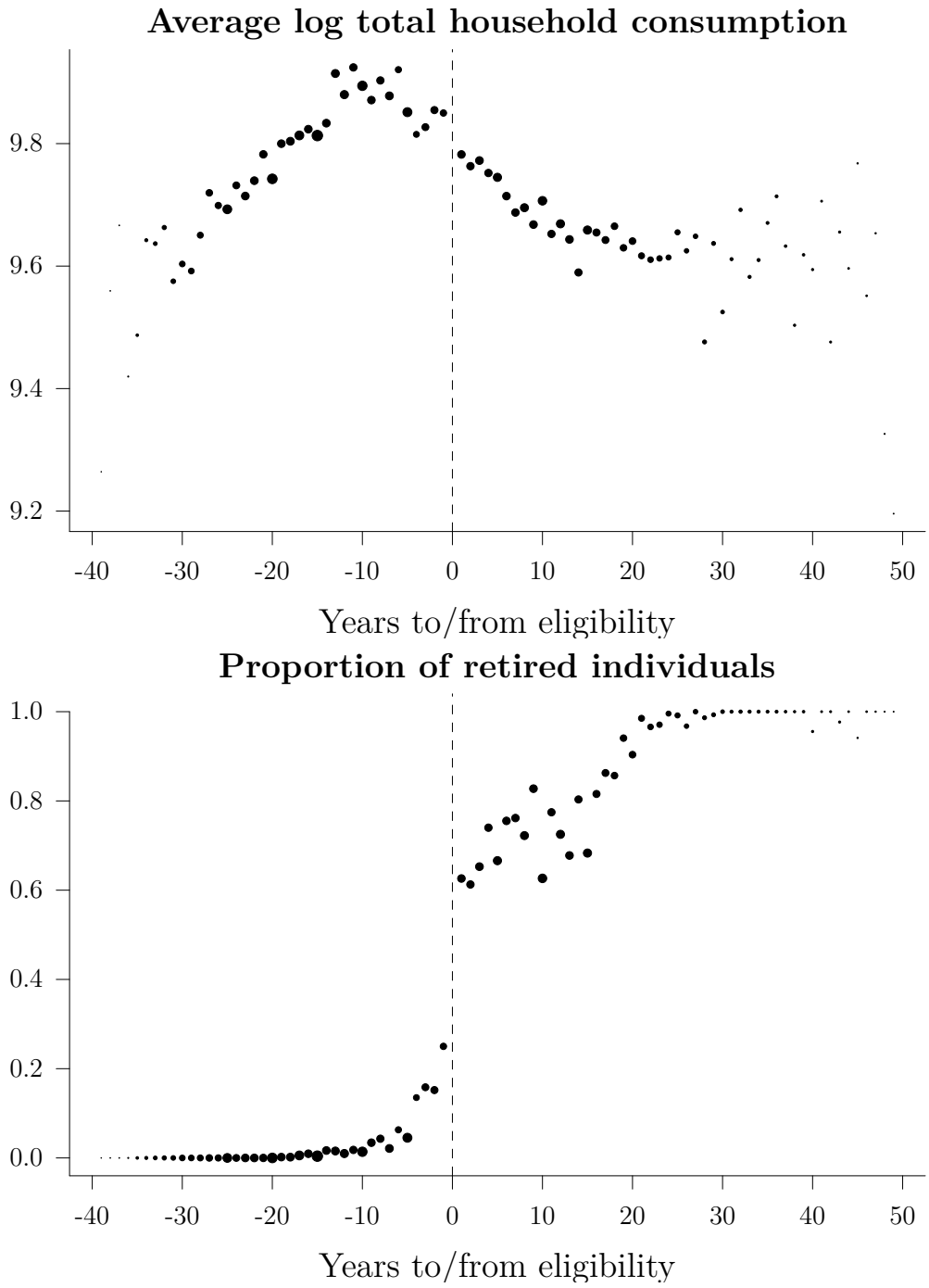

Figure 1: Average log consumption (top panel) and empirical proportion of retired individuals by years of age to/from the retirement eligibility threshold. Dashed vertical lines indicate indicates the cutoff, which is normalized to zero. Size of dots is proportional to the respective number of individuals in the data.

Table 1: Confidence sets for the effect of retiring on log consumption for various methods

Smoothness Bound	Method	Confidence Set
ROT1 ($B_Y = 0.004, B_T = 0.008$)	Bias-aware AR CS	-0.268 ± 0.356
	Bias-aware DM CS	-0.216 ± 0.304
ROT2 ($B_Y = 0.002, B_T = 0.002$)	Bias-aware AR CS	-0.150 ± 0.260
	Bias-aware DM CS	-0.136 ± 0.234
–	Robust bias correction DM CI	-0.269 ± 0.305
	Global Linear with DM CI	-0.252 ± 0.065
	Global Polynomial with DM CI	-0.376 ± 0.042

Notes: 30,006 data points, CSs with 95% nominal coverage.

them to bias-aware DM CIs that use the same smoothness bounds, to robust bias correction DM CIs, to DM CIs based on global linear regression with separate intercepts and slopes on each side of the cutoff, and to DM CIs based on a global 4th order polynomial regression with a dummy for retirement eligibility.[12] We report the results in Table 1 in the form "midpoint $\pm$ half-length" to make comparing differences in the CIs' location and length easier.

We see that bias-aware AR CSs can differ meaningfully from their bias-aware DM CI counterparts in terms of both length and location, even if the same smoothness bounds are used. Our preferred rule ROT2 produces markedly smaller smoothness bounds than ROT1, which is reflected in the shorter CSs. Robust bias correction yields DM CIs that are qualitatively closer to those obtained under ROT1 by bias-aware methods. The two global parametric methods are the only ones that yield CIs that do not cover zero, but of course this does not account for the model misspecification bias apparent from Figure 1.

7.2. **Simulations.** In this subsection, we compare the practical performance of our bias-aware AR CS to that of alternative procedures through simulations. We consider a number of data generating processes calibrated to the data from Battistin et al. (2009) with varying curvature of the conditional expectation functions, richness of the running variables support, and strength of identification.

Data Generating Processes. We first create three versions of each of the two CEFs of outcomes and treatment, shown in Figure 2. Specifically, for W_i equal to either T_i or Y_i, we create the sth CEF version $\mu_{W,s}(x)$ by fitting a second order spline with four knots on each side of the

[12] In Online Appendix F, we also report results for variants of these CSs that use the optimized RD estimator of Imbens and Wager (2019) instead of local linear regression.

cutoff, that is,

$$\mu_{W,s}(x) = \mathbf{1}\{x \geq 0\} \left(\beta_{0,W+} + \beta_{1,W+} x + \sum_{j=1}^{4} \beta_{j+1,W+} \lfloor x - v_j \rfloor^2 \right) + \mathbf{1}\{x < 0\} \left(\beta_{0,W-} + \beta_{1,W-} x + \sum_{j=1}^{4} \beta_{j+1,W-} \lfloor x - v_j \rfloor^2 \right)$$

via least squares to the data from Battistin et al. (2009), with $\lfloor \cdot \rfloor = \max(0, \cdot)$. Here our first version uses the knot points $v_j \in \{0, 10, 20, 30\}$; the second version uses the knot points $v_j \in \{0, 5, 15, 30\}$, creating greater curvature near the cutoff; while the third version also uses the knot points $v_j \in \{0, 5, 15, 30\}$ and additionally fixes the intercept parameters $\beta_{0,W+}$ and $\beta_{0,W-}$ to generate very high curvature near the cutoff.[13] We refer to these settings as having either low, moderate or high CEF curvature. Note that the magnitude of the jump in the conditional treatment probability, and hence the strength of identification, also decreases across versions. We then consider all combinations of these CEFs, but for brevity only report results for the two combinations $(\mu_{T,1}, \mu_{Y,1})$ and $(\mu_{T,2}, \mu_{Y,2})$. See Online Appendix E for the remaining results.

For each combination of CEFs, we also consider four different distributions for the running variable X_i: the "mildly discrete" empirical distribution of the data from Battistin et al. (2009); a continuous distribution obtained by adding uniformly distributed $U(0, 1)$ noise to a draw from the empirical distribution; and two more "coarsely discrete" distributions obtained by rounding draws from the empirical distribution to the integers $\{\pm 1, \pm(1 + d), \pm(1 + 2d), \dots\}$ for $d \in \{3, 6\}$. Finally, for any draw of X_i we draw treatment status T_i from a Bernoulli distribution with mean $\mu_{T,j}(X_i)$ and outcome Y_i from a normal distribution with mean $\mu_{T,k}(X_i)$ and variance $\sigma^2(X_i)$, for $(j, k) \in \{1, 2, 3\} \times \{1, 2, 3\}$ and $\sigma^2(x)$ the sampling variance of Y_i among units with running variable equal to the original draw of the running variable from the empirical distribution in the data.

Methods. We study the performance of eight different implementations of AR CSs in our simulations: (i) our bias-aware CS, using the respective true smoothness bounds B_Y and B_T; (ii) our bias-aware CS, using twice the true B_Y and B_T; (iii) our bias-aware CS, using half the true B_Y and B_T; (iv) our bias-aware CS, using ROT1 estimates of B_Y and B_T; (v) our bias-aware CS, using ROT2 estimates of B_Y and B_T; (vi) a naive CS that ignores bias, using

[13] For the conditional treatment probability, our least squares fits also impose the constraint that the function is increasing, and that it is equal to zero and one below and above the lowest and highest support point, respectively. Note that the setting with very high curvature setting might be unrealistic in practice.

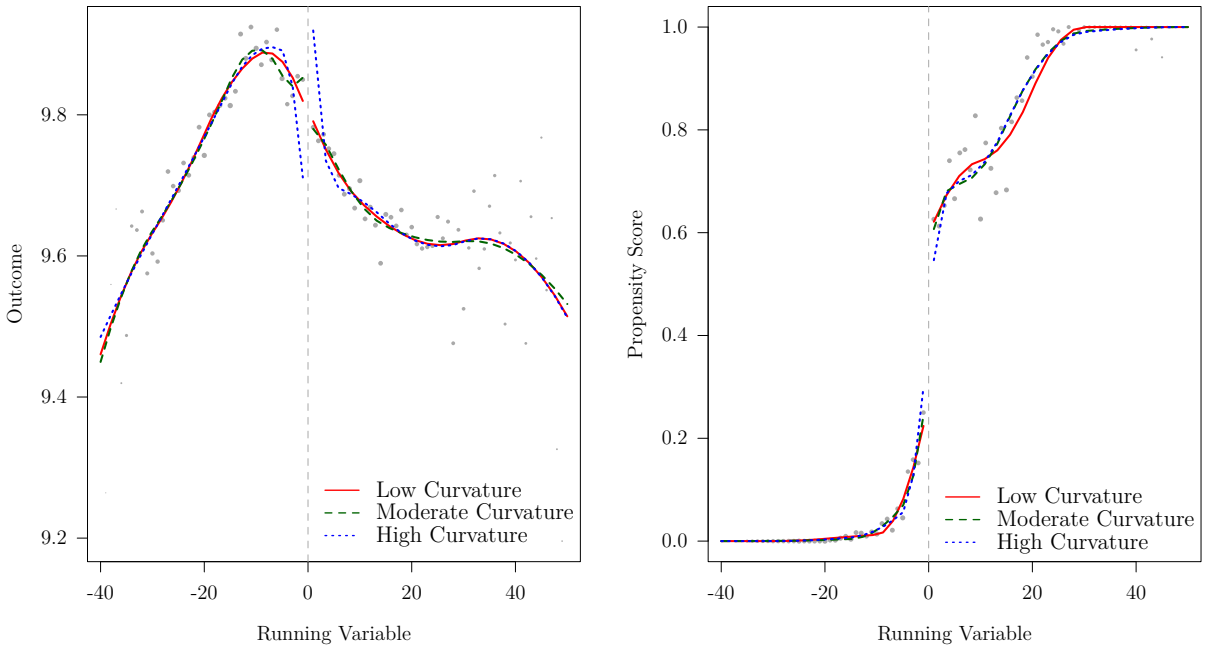

Figure 2: CEFs of outcome (left panel) and of the treatment (right panel) CEFs used in the simulations. Dashed vertical lines indicate indicates the cutoff, which is normalized to zero. Data from Figure 1 (grey dots) shown for reference.

an estimate of the "pointwise-MSE optimal" bandwidth (Imbens and Kalyanaraman, 2012, henceforth IK); (vii) an undersmoothing CS, using $n^{-1/20}$ times the estimated IK bandwidth; and (viii) a robust bias correction CS, using local quadratic regression to estimate the bias, and estimated IK bandwidths. In addition, we also consider the performance of eight different DM CIs using the just-mentioned approaches to handling bias.[14]

Results. Table 2 shows the simulated coverage rates of the various CSs under the sixteen different DGPs we consider in our simulations (four combinations of CEFs times four running variable distributions). We first discuss results for AR CSs, shown in the left panel. With the true smoothness bounds, the coverage rates of our bias-aware CSs are close to and mostly

[14]Computations are carried out with the statistical software `R`. All bias-aware CSs are computed using our own software, which builds on the package `RDHonest`. All other CSs are computed using functions from the package `rdrobust`. A triangular kernel is used in all cases. We note that the IK bandwidth estimates computed by `rdrobust` are sometimes too small for the respective CSs to be well-defined if the running variable is discrete. In those cases, we manually set the main bandwidth such that positive weights are given to three support points on each side of the cutoff (for the bias correction bandwidth we use four support points). In Online Appendix F, we also report results for variants of our CSs in which local linear regression is replaced with the optimized RD estimator of Imbens and Wager (2019).

slightly above the nominal level, irrespective of running variable distribution, curvature of the unknown functions, and identification strength. The slight overcoverage occurs because the function $\mu_Y(x) - \theta\mu_T(x)$ is not exactly quadratic in either setting, and thus the bias does not achieve its worst-case value. Using twice the true bounds increases simulated coverage as expected, while half the true value can result in meaningful undercoverage.

Using one of the ROTs for the smoothness bounds leads to potentially severe distortions in some settings, which highlights the need to investigate the fit of the respective underlying global polynomial approximation in practice (cf. Online Appendix C). Combining a naive approach, undersmoothing, or robust bias correction with an AR construction leads to CSs with undercoverage that is modest in some DGPs we consider, but can be substantial especially for those with more coarse running variable support, stronger curvature of the CEFs, and weak identification (cf. Online Appendix F).

Turning to results for DM CIs in the right panel of Table 2, we see that combining a bias-aware approach with this construction does not lead to CIs with correct coverage in all settings even when using the true smoothness bound. This is because bias-aware DM CIs only control the bias of a first-order approximation of the estimator on which they are based. Coverage distortions are particularly severe in settings of strong curvature, and they further amplify in settings with weak identification. Using the ROT choices for the smoothness bounds leads to further distortions in some cases. The coverage of DM CIs that use the naive approach, undersmoothing, or robust bias correction is distorted in most settings, and the distortions generally become more severe with a more coarse support, in settings with a higher curvature and it further amplifies in settings of weak identification (cf. Online Appendix F).

To show that our bias-aware AR CSs not only have good coverage properties but also yield comparatively powerful inference, we simulate the rates at which the various CSs we consider cover parameter values other than the true one. We report the results for one exemplary setting (low curvature of outcome and treatment CEF, continuously distributed running variable) in Figure 3.[15] To avoid having all 16 coverage curves in one plot, we split the results into four panels: the five bias-aware AR CSs in (a), the three other AR CSs in (b), the five bias-aware DM CIs in (c), and the three other DM CIs in (d). Panels (b)–(d) also show the curve for our bias-aware AR CS with the true constants to have a common point of reference.

[15]We focus on this setting because the coverage of the true parameter is reasonably close to the nominal level for all procedures, and thus a comparison of coverage rates at "non-true" parameter values is meaningful across CSs. Analogous plots for other DGPs are available from the authors.

Table 2: Simulated CS coverage (%)

	Anderson-Rubin								Delta Method							
	Bias-Aware								Bias-Aware							
Support	TC	TC×.5	TC×2	ROT1	ROT2	Naive	US	RBC	TC	TC×.5	TC×2	ROT1	ROT2	Naive	US	RBC
Setting 1 - *Low outcome CEF curvature, low treatment CEF curvature*																
Baseline	96.7	91.0	98.4	97.9	92.4	88.7	93.7	94.6	97.9	93.7	99.3	98.9	93.9	94.2	96.0	94.2
Continuous	96.6	92.7	97.4	97.1	94.3	89.6	94.0	94.7	97.2	94.4	98.0	97.8	95.1	94.7	95.4	94.7
$\{\pm 1, \pm 4, \dots\}$	95.9	92.6	98.4	97.8	94.4	88.1	93.8	94.6	96.7	93.8	99.3	98.7	95.0	94.0	89.3	94.6
$\{\pm 1, \pm 7, \dots\}$	97.8	91.6	100.0	99.6	95.5	78.0	84.8	93.2	98.6	92.3	100.0	99.8	96.0	85.1	77.5	93.1
Setting 2 - *Moderate outcome CEF curvature, moderate treatment CEF curvature*																
Baseline	97.2	88.7	99.2	91.9	36.4	47.6	80.9	58.6	91.1	74.0	98.3	81.9	24.7	73.3	93.0	72.9
Continuous	96.4	92.2	97.3	94.3	59.8	68.6	87.6	78.2	93.7	84.7	96.5	89.3	46.6	84.9	94.5	86.3
$\{\pm 1, \pm 4, \dots\}$	98.3	88.8	100	92.1	63.5	27.5	71.1	45.6	92.7	81.8	99.0	86.0	46.6	70.7	65.9	71.3
$\{\pm 1, \pm 7, \dots\}$	100	79.9	100	92.1	26.2	1.3	6.5	0.3	95.4	56.0	100	78.2	19.8	6.0	5.5	4.7

Notes: Results based on 50,000 Monte Carlo draws for a nominal confidence level of 95%. Columns show results for bias aware approach with true constants (TC), two times true constants (TC×2), half true constants (TC×.5), and with rule of thumb estimates (ROT1) and (ROT2); naive approach that ignores bias (Naive); undersmoothing (US); and robust bias correction (RBC).

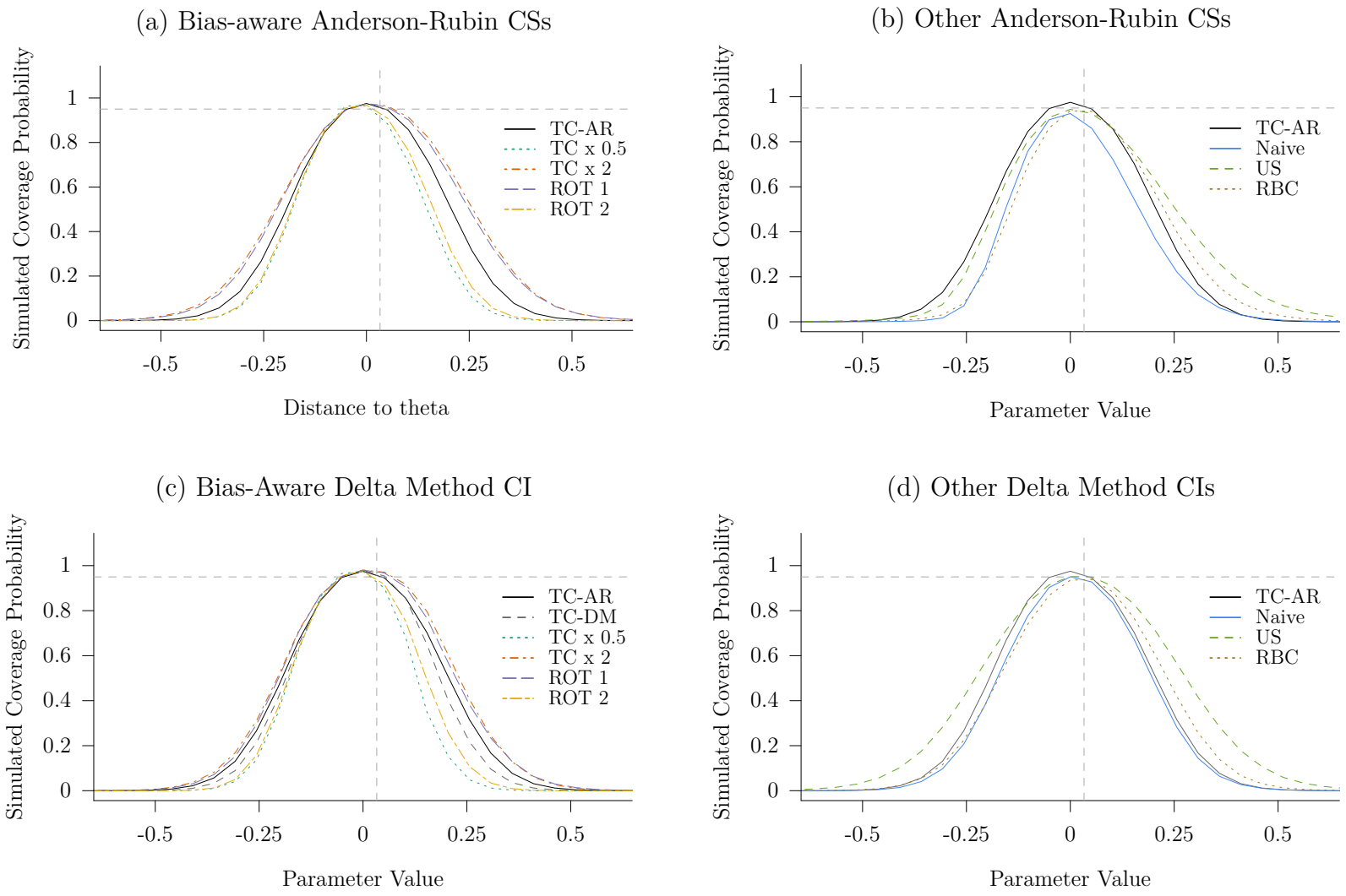

Figure 3: Simulated coverage rates of various values of parameter values and for different types of confidence sets. Based on the Setting 1 and a continuous running variable as described in the main text. Bias aware approach with true constants (TC (ref); as reference function in all graphs), two times true constants (TC×2), 0.5 times true constants (TC×.5), and with rule of thumb smoothness bounds (ROT1) and (ROT2); robust bias correction (RBC); naive approach that ignores bias (Naive); and undersmoothing (US).

Panel (a) then shows that the coverage rate of bias-aware AR CSs drops very quickly to zero away from the true parameter. Panels (b)–(d) show that the coverage of bias-aware AR CSs away from the true parameter is also below that of most competing procedures over almost all the parameter space, with the exception of those which exhibit a meaningful distortion at the true parameter value and are therefore not suitable for a direct comparison. This confirms that the accurate coverage of our CSs in settings with discrete running variables and weak identification does not come at the expense of statistical power in a canonical setup, for which most competing CSs are specifically constructed.

A. PROOFS OF THEOREMS 1–3

In this Appendix, we prove Theorems 1–3. See Online Appendix A for a proof of Theorem 4. We note that, by basic least squares algebra, the statistic $\widehat{\tau}_M(h,c)$ can be written as

$$\begin{aligned}
&\widehat{\tau}_M(h,c) = \sum_{i=1}^{n} w_i(h) M_i(c), \quad w_i(h) = w_{i,+}(h) - w_{i,-}(h), \\
&w_{i,+}(h) = e_1^\top Q_+^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i \geq 0\}, \quad Q_+ = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_i \widetilde{X}_i^\top \mathbf{1}\{X_i \geq 0\} \\
&w_{i,-}(h) = e_1^\top Q_-^{-1} \widetilde{X}_i K(X_i/h) \mathbf{1}\{X_i < 0\}, \quad Q_- = \sum_{i=1}^{n} K(X_i/h) \widetilde{X}_i \widetilde{X}_i^\top \mathbf{1}\{X_i < 0\},
\end{aligned}$$

where $\widetilde{X}_i = (1, X_i)^\top$. We write $A_n(\mu) = o_{P,\mathcal{F}}(1)$ if $\sup_{\mu \in \mathcal{F}} P(|A_n(\mu)| > \epsilon) = o(1)$ for all $\epsilon > 0$ and a generic sequence $A_n(\mu)$ of random variables indexed by $\mu \in \mathcal{F}$. We also drop the dependency on c from the notation for the optimal bandwidth in most instances, and thus write h_M instead of $h_M(c)$.

A.1. **Proof of Theorem 1.** We begin by establishing the following lemma.

Lemma A.1. *Suppose that Assumption 1–2 and either Assumption LL1 or Assumption LL2 are satisfied. Then the following holds uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}$: (i) $w_{\text{ratio}}(h_M(c)) = o_P(1)$; (ii) $(\widehat{\tau}_M(\widehat{h}_M(c), c) - \widehat{\tau}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$; and (iii) $(\bar{b}_M(\widehat{h}_M(c), c) - \bar{b}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$.*

Proof. First suppose that Assumption LL1 is satisfied, and note that it is clear with a discrete running variable that the optimal bandwidth h_M shrinks with the sample size, but that it cannot tend to zero as it has to be greater than the support point second closest to the cutoff in order for the local linear regression estimator to be well-defined. To show

part (i), note that $w_{\text{ratio}}(h_M)$ is well-defined with probability approaching 1, and that

$$w_{\text{ratio}}(h_M) \leq \max_{i\in\{1,\dots,n\}} \frac{w_i(h_M)^2}{\sum_{j:X_j=X_i} w_j(h_M)^2} = \max_{i\in\{1,\dots,n\}} \frac{1}{\sum_{j:X_j=X_i} \mathbf{1}\{X_i = X_j\}} = o_p(1)$$

as the number of units whose realization of the running variable is equal to any particular value in its support tends to infinity. To show parts (ii) and (iii), recall that h_M approaches some value between second and third support point closest to the cutoff, and that indeed any bandwidth h between the second and third support point closest to the cutoff implies the same local linear regression weights $w_i(h)$ for all i. Part (ii)–(iii) then follow trivially, as each expression under consideration depends on h only through $w_i(h)$.

Now suppose that Assumption LL2 holds. Note that the minimizer of the function $h \mapsto \text{cv}_{1-\alpha}(r_M(h,c))s_M(h,c)$ must balance the asymptotic bias and standard deviation of the the local linear regression estimator, and thus $h_M \propto n^{-1/5}(1+o(1))$. Statement (i) then follows from classical arguments for this bandwidth. On the other hand, statements (ii)–(iii) of Lemma A.1 follow by applying the arguments starting in the second paragraph of the proof of Theorem E.1 in Armstrong and Kolesár (2020) conditional on $\mathcal{X}_n$, and the assumption that the running variable density is continuous around the cutoff. □

We now proceed with the proof of the core statement of Theorem 1. As $\theta \in \mathcal{C}^\alpha_{\text{ar}}$ if and only if $\tau_M(\theta) \in \mathcal{C}^\alpha(\theta)$, it suffices to show that for any $c \in \mathbb{R}$

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

Note that it follows from Lemma A.1 (ii)–(iii) and uniform continuity of $\text{cv}_{1-\alpha}(\cdot)$ that

$$\begin{aligned}
&\frac{|\widehat{\tau}_M(\widehat{h}_M,c) - \tau_M(c)|}{\widehat{s}_M(\widehat{h}_M,c)} - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M,c)) \\
&\quad = \left|\frac{\widehat{\tau}_M(h_M,c) - \mathbb{E}\left[\widehat{\tau}_M(h_M,c)|\mathcal{X}_n\right]}{s_M(h_M,c)} + \frac{b_M(h_M,c)}{s_M(h_M,c)}\right| - \text{cv}_{1-\alpha}(r_M(h_M,c)) + o_{P,\mathcal{F}}(1).
\end{aligned}$$

By Lemma A.1 (i) and Lyapunov's CLT $(\widehat{\tau}_M(h_M,c) - \mathbb{E}\left[\widehat{\tau}_M(h_M,c)|\mathcal{X}_n\right])/s_M(h_M,c)$ converges in distribution to a standard normally distributed random variable, uniformly over $(\mu_Y,\mu_T) \in \mathcal{F}$. It then follows again from uniform continuity of $\text{cv}_{1-\alpha}(\cdot)$ that

$$\mathbb{P}(\tau_M(c) \in \mathcal{C}^\alpha(c)) = \mathbb{P}\left(\left|S + \frac{b_M(h_M,c)}{s_M(h_M,c)}\right| \leq \text{cv}_{1-\alpha}(r_M(h_M,c))\right) + o_{P,\mathcal{F}}(1),$$

with S a generic standard normal random variable. Honesty now follows from the definition

of the critical value function $\text{cv}_{1-\alpha}(\cdot)$ if

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}} |b_M(h_M, c)/s_M(h_M, c)| \leq r_M(h_M, c).$$

Armstrong and Kolesár (2020, Theorem B.3) implies that the last statement would hold with equality if μ_Y and μ_T had unbounded domain. We obtain a weak inequality because in our setup μ_T is naturally constrained to take values in $[0, 1]$, and the supremum is thus taken over a smaller set of functions. This completes our proof. □

A.1.1. *Alternative assumptions for discrete case.* The above asymptotic framework for the case of a discrete running variable implies that in large samples the optimal bandwidth h_M is the corner solution of the corresponding optimization problem, and is thus such that only the two closest support points on each side of the cutoff receive positive kernel weights. This in turn implies that the corresponding bias asymptotically dominates the corresponding standard deviation. Here we show that our CSs also remain honest, in the sense of (2.1), under an alternative asymptotic sequence under which the variance of $M_i(c)$ increases with the sample size at an appropriate rate. This rate is chosen such that bias and standard error are of the same stochastic order in large samples under the resulting optimal bandwidth.

Proposition 1. *Suppose that for each n, the data $\{(Y_i, T_i, X_i,), i = 1, \ldots, n\}$ are i.i.d., distributed according to a law P_n. Under each P_n, the support of X_i consists of the $J = J_+ + J_-$ fixed points $x_1 < \ldots < x_{J_-} < 0 \leq x_{J_-+1} < \ldots < x_J$; and the conditional expectation functions (μ_Y, μ_T) do not vary with n as well. Moreover, $\mathbb{V}(M_i(c)|X = x_j) = n\bar{\sigma}^2_{M,j}(c)$ for constants $\bar{\sigma}^2_{M,j}(c) > 0$ for every $c \in \mathbb{R}$, $j = 1, \ldots, J$, and $(\mu_Y, \mu_T) \in \mathcal{F}$; and $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^q|X_i = x_j) = O(n^{q/2})$ for some $q > 2$. Also, the kernel K is a continuous, unimodal, symmetric density function that is equal to zero outside some compact set, say $[-1, 1]$. Then $\mathcal{C}^\alpha_{\text{ar}}$ is honest with respect to $\mathcal{F}$ in the sense of* (2.1).

Proof. We show that the statements (i)–(iii) of Lemma A.1 also hold under the conditions of the Proposition. Honesty of our CS then follows as in the proof of Theorem 1 by noting that Lyapunov's CLT still applies under the conditions of the Proposition.

First note that for any fixed bandwidth $h > \max\{|x_{J_--1}|, x_{J_-+2}\}$, we have that both $\bar{b}_M(h, c)$ and $s_M(h, c)$ converge with rate $n^{-1/2}$ to strictly positive constants in probability, and thus $h_M = \bar{h}_M + o_P(1)$, with $\bar{h}_M > \max\{|x_{J_--1}|, x_{J_-+2}\}$. Statement (i) then follows from the same reasoning as in the proof of Lemma A.1(i). To show statement (ii), note that because $s_M(h_M, c) = O_P(1)$, it suffices to show that $\widehat{\tau}_M(\widehat{h}_M, c) - \widehat{\tau}_M(h_M, c)$ converges to zero in probability. With a discrete running variable, we can write the estimator $\widehat{\tau}_M(h, c)$

as follows:

$$\widehat{\tau}_M(h,c) = \sum_{j=1}^{J} w_{(j)}(h)\bar{M}_j(c), \quad \bar{M}_j(c) = \frac{1}{n\widehat{p}_j}\sum_{i:X_i=x_j} M_i(c), \quad w_{(j)}(h) = w_{(j)+}(h) - w_{(j)-}(h),$$

$$w_{(j)+}(h) = e_1^\top \tilde{Q}_+^{-1}\widehat{p}_j K(x_j/h)\widetilde{x}_j \mathbf{1}\{x_j \geq 0\}, \quad \tilde{Q}_+ = \sum_{j=1}^{J} \widehat{p}_j K(x_j/h)\widetilde{x}_j\widetilde{x}_j^\top \mathbf{1}\{x_j \geq 0\}$$

$$w_{(j)-}(h) = e_1^\top \tilde{Q}_-^{-1}\widehat{p}_j K(x_j/h)\widetilde{x}_j \mathbf{1}\{x_j < 0\}, \quad \tilde{Q}_- = \sum_{j=1}^{J} \widehat{p}_j K(x_j/h)\widetilde{x}_j\widetilde{x}_j^\top \mathbf{1}\{x_j < 0\},$$

with $\widetilde{x}_j = (1, x_j)^\top$ and $\widehat{p}_j = \sum_{i=1}^n \mathbf{1}\{X_i = x_j\}/n$ the relative frequency of the jth support point in the data. We then have

$$\begin{aligned}\widehat{\tau}_M(\widehat{h}_M, c) - \widehat{\tau}_M(h_M, c) = &\sum_{j=1}^{J}(w_{(j)}(\widehat{h}_M) - w_{(j)}(h_M))\mu_M(x_j, c) \\ &+ \sum_{j=1}^{J}(w_{(j)}(\widehat{h}_M) - w_{(j)}(h_M))(\bar{M}_j(c) - \mu_M(x_j, c)).\end{aligned}$$

From continuity of the kernel function, it follows that the mapping $h \mapsto w_{(j)}(h)$ is continuous for all $j = 1, \ldots, J$, and hence $w_{(j)}(\widehat{h}_M) - w_{(j)}(h_M) = o_P(1)$ for all $j = 1, \ldots, J$. The first term on the right-hand side of the previous equation then converges to zero in probability follows because $\mu_M(\cdot, c)$ has uniformly bounded second derivatives, and from the algebraic properties of the local linear regression weights. For the second term, convergence in probability to zero follows because $\bar{M}_j(c) - \mu_M(x_j, c) = O_P(1)$. Taken together, this proves statement (ii). Statement (iii) can then be shown analogously. □

A.2. **Proof of Theorem 2.** To simplify the exposition, we emphasize the dependence of various estimators on c in our notation, but suppress their dependency on the bandwidth h (which does not depend on c under the conditions of this theorem). The CS $\mathcal{C}_{\text{ar}}^{\alpha}(h)$ is given by the set of all values of c satisfying

$$\vartheta(c) \leq 0, \quad \text{where} \quad \vartheta(c) \equiv |\widehat{\tau}_Y - c\widehat{\tau}_T| - \text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c).$$

The function $\vartheta(c)$ is continuous because $\text{cv}_{1-\alpha}$ is uniformly continuous, and both the standard error $\widehat{s}_M(c) = (\widehat{s}_Y^2 - 2c\widehat{s}_{TY} + c^2\widehat{s}_T^2)^{1/2}$ and the worst case bias $\overline{b}_M(h,c) = -(B_Y + |c|B_T)/2 \cdot \sum_{i=1}^n w_i(h)X_i^2\text{sign}(X_i)$ are continuous in c. The term $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is also strictly convex in c, because both the standard error and the worst-case bias are convex in c and

$\text{cv}_{1-\alpha}(\cdot)$ is strictly convex and increasing. The shape of $\mathcal{C}^{\alpha}_{\text{ar}}(h)$ is then determined by the roots of $\vartheta(c)$. To prove the theorem, it suffices to show that the function $\vartheta(c)$ always fits into one of the following four categories: (i) $\vartheta(c) \leq 0$ for all c; (ii) $\vartheta(c)$ has two roots, and there exists $c^* > 0$ such that $\vartheta(c) < 0$ for all $|c| > c^*$; (iii) $\vartheta(c)$ has two roots, and there exists $c^* > 0$ such that $\vartheta(c) > 0$ for all $|c| > c^*$, and (iv) $\vartheta(c)$ has one root. Then $\mathcal{C}^{\alpha}_{\text{ar}}(h) = \mathbb{R}$ in case (i), $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1] \cup [a_2, \infty)$ for some $a_1 < a_2$ in case (ii); and by $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, a_2]$ for some $a_1 < a_2$ in case (iii), and $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_2]$ or $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, \infty)$ in case (iv). We now go through a number of case distinctions.

If $\widehat{\tau}_T = 0$, then $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ is a constant function in c. As $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is strictly convex in c and unbounded, $\vartheta(c)$ must be either of form (i) or (ii). We threfore suppose that $\widehat{\tau}_T \neq 0$ from now on, and write $\widehat{\theta} = \widehat{\tau}_Y/\widehat{\tau}_T$. As $\vartheta(\widehat{\theta}) < 0$ by construction, the function $\vartheta(c)$ cannot be strictly positive. As $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ is a piecewise linear function and $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ is strictly convex, the function $\vartheta(c)$ can also have at most two roots for $c \leq \widehat{\theta}$, and at most two roots for $c > \widehat{\theta}$. If it does not have any root, $\vartheta(c)$ is of the form (i).

Let us first assume that $\lim_{c\to\pm\infty} \vartheta(c) \neq 0$. It follows from basic algebra that there exists some c^* sufficiently large such that $\text{sign}(\vartheta(c)) = \text{sign}(\vartheta(-c)) = 1$ or $\text{sign}(\vartheta(c)) = \text{sign}(\vartheta(-c)) = -1$ and $\vartheta(c) \neq 0$ for all $c > c^*$. The function $\vartheta(c)$ therefore cannot have one or three roots; so it must have either four roots or two roots or none. If $\text{sign}(\vartheta(c)) = -1$ for all $|c| > c^*$, which means that $|\widehat{\tau}_Y - c\widehat{\tau}_T| > \text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$. The function $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ intersects once with the function $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ for $c < \widehat{\theta}$, and once for $c > \widehat{\theta}$. Therefore $\vartheta(c)$ must be of form (iii) in this case. If $\text{sign}(\vartheta(c)) = -1$ for all $|c| > c^*$, the above reasoning only yields that $\vartheta(c)$ has at most four roots. However, note that for $|c| \to \infty$ the absolute value of the first derivative of $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ with respect to c converges to some constant ϖ, and that for any value of $\varsigma \in \mathbb{R}$ the expression $\text{sign}(c)\cdot(\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M - |\varsigma + \varpi \cdot c|)$ converges to a constant. Choose ς such that the latter constant is zero, and set $\varrho(c) = |\varsigma + \varpi c|$. By construction, $\varrho(c)$ intersects with $|\widehat{\tau}_Y - c\widehat{\tau}_T|$ twice either for $c \leq \widehat{\theta}$ or $c \geq \widehat{\theta}$. It also holds that $\varrho(c) \leq \text{cv}_{1-\alpha}(\widehat{r}(c))\widehat{s}_M(c)$ for all c by strict convexity of $\text{cv}_{1-\alpha}(\widehat{r}(c))\widehat{s}_M(c)$. This reasoning implies that $\vartheta(c)$ can have at most two roots, and must be of form (ii) in this case.

Now suppose that $\lim_{c\to\pm\infty} \vartheta(c) = 0$, which only occurs if $\widehat{\tau}_T = \pm\text{cv}_{1-\alpha}(\widehat{r}_T(c)) \cdot \widehat{s}_T(c)$. It then follows from strict convexity of $\text{cv}_{1-\alpha}(\widehat{r}_M(c))\widehat{s}_M(c)$ that $\vartheta(c)$ cannot have three roots. $\vartheta(c)$ is therefore of form (i) if it does not have any root, and otherwise of form (iv). This completes the proof. □

A.3. **Proof of Theorem 3.** We begin by giving a formal description of a bias-aware DM CI. Recall the definition of U_i from Section 5, and let $b_U(h) = \mathbb{E}(\widehat{\tau}_U(h)|\mathcal{X}_n)$ and $s_U(h) =$

$\mathbb{V}(\widehat{\tau}_U(h)|\mathcal{X}_n)^{1/2}$ denote conditional bias and standard deviation, respectively, of the SRD-type estimator $\widehat{\tau}_U(h)$. Exploiting linearity, one can write

$$b_U(h) = \sum_{i=1}^{n} w_i(h)(\mu_U(X_i) - \tau_U) \text{ and } s_U(h) = \left(\sum_{i=1}^{n} w_i(h)^2 \sigma_{U,i}^2\right)^{1/2},$$

where $\mu_U(x) = (\mu_Y(x) - \tau_Y)/\tau_T - \tau_Y(\mu_T(x) - \tau_T)/\tau_T^2$ depends on the functions μ_Y and μ_T, and $\sigma_{U,i}^2 = \mathbb{V}(U_i|X_i)$ is the conditional variance of U_i given X_i. Because the bias depends on (μ_Y, μ_T) through the function $\mu_U \in \mathcal{F}_H(B_Y/|\tau_T| + |\tau_Y|B_T/\tau_T^2)$ only, its "worst case" magnitude over the functions contained in $\mathcal{F}^\delta$ is

$$\sup_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} |b_U(h)| = \overline{b}_U(h) \equiv -\frac{1}{2}\left(\frac{B_Y}{|\tau_T|} + \frac{|\tau_Y|B_T}{\tau_T^2}\right)\sum_{i=1}^{n} w_i(h)X_i^2 \text{sign}(X_i).$$

An infeasible bias-aware DM CI is then given by

$$\mathcal{C}_\Delta^\alpha = \left[\widehat{\theta}(h_U) \pm \text{cv}_{1-\alpha}\left(\overline{b}_U(h_U)/s_U(h_U)\right) s_U(h_U)\right], \tag{A.1}$$

where $h_U = \text{argmin}_h \, \text{cv}_{1-\alpha}\left(\overline{b}_U(h)/s_U(h)\right) s_U(h)$ is the bandwidth that minimizes its length. It is easy to see that this optimal bandwidth must such that neither bias nor variance dominate asymptotically, which means that $\lim n^{-1/5} h_U = c > 0$.

Making this CI feasible would require three main modifications: (i) replacing the unknown bias bound with an estimate $\widehat{\overline{b}}_U(h)$ which replaces τ_Y and τ_T with feasible estimates, such as local linear estimates $\widehat{\tau}_Y = \widehat{\tau}_Y(g_Y)$ and $\widehat{\tau}_T = \widehat{\tau}_T(g_T)$ based on preliminary bandwidths g_Y and g_T; (ii) replacing the standard deviation $s_U(h)$ with a valid standard error, which could be achieved as in Section 6.1 using estimates of the form $\widehat{U}_i = (Y_i - \widehat{\tau}_Y)/\widehat{\tau}_T - \widehat{\tau}_Y(T_i - \widehat{\tau}_T)/\widehat{\tau}_T^2$ of the U_i); (iii) replacing the bandwidth h_U with an empirical analogue, like an adaptation of the procedure in Section 6.2. As such modifications can be shown not to affect the first-order asymptotic coverage properties of the CI under standard regularity conditions, we base our result on a comparison of $\mathcal{C}_*^\alpha$ and $\mathcal{C}_\Delta^\alpha$.

To prove Theorem 3, we make the dependence of quantities like $h_M(c)$ on c again explicit. We begin by noting that the events $\theta^{(n)} \in \mathcal{C}_\Delta^\alpha$ and $\theta^{(n)} \in \mathcal{C}_*^\alpha$ occur if and only if

$$\frac{|\widehat{\theta}(h_U) - \theta^{(n)}|}{s_U(h_U)} - \text{cv}_{1-\alpha}\left(\frac{\overline{b}_U(h_U)}{s_U(h_U)}\right) \leq 0 \tag{A.2}$$

$$\text{and} \quad \frac{|\widehat{\tau}_M(h_M(\theta^{(n)}), \theta^{(n)})|}{s_M(h_M(\theta^{(n)}), \theta^{(n)})} - \text{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h_M(\theta^{(n)}), \theta^{(n)})}{s_M(h_M(\theta^{(n)}), \theta^{(n)})}\right) \leq 0, \tag{A.3}$$

respectively. Because the left-hand sides of the last two displays both approximately behave like a constant plus the absolute value of a normal random variable with variance 1 in large samples, it suffices to show that the difference between the respective left-hand sides of the last two displays converges to zero in probability, uniformly over $\mathcal{F}^\delta$. To show this, note first that standard delta method arguments yield that the left-hand side of (A.2) is equal to

$$\frac{|\widehat{\tau}_U(h_U) - \kappa n^{-2/5}|}{s_U(h_U)} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_U(h_U)}{s_U(h_U)}\right) + o_{P,\mathcal{F}^\delta}(1).$$

Next, note that $U_i = M_i(\theta)/\tau_T$, and that we thus have that

$$\widehat{\tau}_U(h) = \frac{\widehat{\tau}_M(h,\theta)}{\tau_T}, \qquad s_U(h) = \frac{s_M(h,\theta)}{|\tau_T|}, \qquad \overline{b}_U(h) = \frac{\overline{b}_M(h,\theta)}{|\tau_T|},$$

for any $h > 0$. Substituting these identities into the definition of h_U, we also find that

$$h_U = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h,\theta)}{s_M(h,\theta)}\right) \cdot \frac{s_M(h,\theta)}{|\tau_T|} = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h,\theta)}{s_M(h,\theta)}\right) s_M(h,\theta) = h_M(\theta).$$

The left-hand side of (A.2) is thus equal to

$$\frac{|\widehat{\tau}_M(h_M(\theta),\theta) - \tau_T\kappa n^{-2/5}|}{s_M(h_M(\theta),\theta)} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h_M(\theta),\theta)}{s_M(h_M(\theta),\theta)}\right) + o_{P,\mathcal{F}^\delta}(1).$$

Now consider the term on the left-hand side of (A.3). By simple algebra, we have that for n sufficiently large

$$\begin{aligned}
\overline{b}_M(h,\theta^{(n)}) &= \overline{b}_M(h,\theta) + (|\theta + n^{-2/5}\kappa| - |\theta|)\overline{b}_T(h) = \overline{b}_M(h,\theta) + |n^{-2/5}\kappa|\overline{b}_T(h), \\
s_M(h,\theta^{(n)})^2 &= s_M^2(h,\theta) + n^{-2/5}\kappa\left((2\theta + n^{-2/5}\kappa)s_T^2(h) - 2\tilde{s}_{M(\theta),T}(h)\right),
\end{aligned}$$

with $\tilde{s}_{M(\theta),T}(h) = (\sum_{i=1}^n w_i(h)^2\sigma_{M(\theta),T,i})^{1/2}$ a conditional covariance term of the same order as $s_T(h)$. These identities imply that $\widehat{\tau}_M(h,\theta)$ and $\widehat{\tau}_M(h,\theta^{(n)})$ have the same first-order bias and standard deviation along any bandwidth sequence h of order $n^{-1/5}$; and from Theorem 2.1(i) in Armstrong and Kolesár (2020), we know that to first order the optimal bandwidth that minimizes the length of a bias-aware CI depends on the first-order bias and standard deviation only. This yields that $h_M(\theta^{(n)}) = h_M(\theta)(1 + o_{P,\mathcal{F}^\delta}(1))$. Arguing as in the proof of Lemma A.1, the left-hand side of (A.3) is thus equal to

$$\frac{|\widehat{\tau}_M(h_M(\theta),\theta) - \tau_T\kappa n^{-2/5}|}{s_M(h_M(\theta),\theta)} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(h_M(\theta),\theta)}{s_M(h_M(\theta),\theta)}\right) + o_{P,\mathcal{F}^\delta}(1),$$

which completes the proof. □

REFERENCES

Abadie, A. and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

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

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

Andrews, I., J. H. Stock, and L. Sun (2019): "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753.

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

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics*.

——— (2021): "Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness," *Econometrica*, 89, 1141–1177.

Barreca, A. I., M. Guldi, J. M. Lindo, and G. R. Waddell (2011): "Saving babies? Revisiting the effect of very low birth weight classification," *Quarterly Journal of Economics*, 126, 2117–2123.

Battistin, E., A. Brugiavini, E. Rettore, and G. Weber (2009): "The retirement consumption puzzle: evidence from a regression discontinuity approach," *American Economic Review*, 99, 2209–26.

Bertanha, M. and M. J. Moreira (2018): "Impossible Inference in Econometrics: Theory and Applications," *Journal of Econometrics*.

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

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on causal effects in a generalized regression kink design," *Econometrica*, 83, 2453–2483.

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

Feir, D., T. Lemieux, and V. Marmer (2016): "Weak identification in fuzzy regression discontinuity designs," *Journal of Business & Economic Statistics*, 34, 185–196.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation

of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69, 201–209.

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

Imbens, G. and C. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.

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

Imbens, G. W. and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

Kamat, V. (2018): "On nonparametric inference in the regression discontinuity design," *Econometric Theory*, 34, 694–703.

Kolesár, M. and C. Rothe (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 108, 2277–2304.

Li, K.-C. (1989): "Honest confidence regions for nonparametric regression," *Annals of Statistics*, 17, 1001–1008.

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

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 557–586.

Thistlethwaite, D. L. and D. T. Campbell (1960): "Regression-discontinuity analysis: An alternative to the ex post facto experiment," *Journal of Educational Psychology*, 51, 309.

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.