Bias-Aware Inference in Fuzzy Regression Discontinuity Designs

Claudia Noack* Christoph Rothe†

Abstract

Fuzzy regression discontinuity (FRD) designs are widely used in applied economics. However, the confidence intervals based on nonparametric local linear regression that are commonly reported in empirical FRD studies can have poor finite sample coverage properties. This is particularly the case under weak identification, meaning that the jump in treatment probabilities at the threshold is small, and if the running variable is discrete. These issues are due to the general construction of these confidence intervals based on the delta method, and to how they account for smoothing bias. We therefore propose new confidence sets that are based on an Anderson-Rubin construction and bias-aware, in the sense that they explicitly take into account the exact local linear smoothing bias. These confidence sets are simple to compute, highly efficient, and have excellent coverage properties in finite samples. They are also valid under weak identification and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

First Version: June 11, 2019. This Version: December 19, 2019. We thank Tim Armstrong, Michal Kolesár and numerous seminar participants for helpful comments and suggestions. The authors gratefully acknowledge financial support by the European Research Council (ERC) through grant SH1-77202.

*Department of Economics, University of Mannheim, 68131 Mannheim, Germany. E-Mail: cnoack@mail.uni-mannheim.de.

†Department of Economics, University of Mannheim, 68131 Mannheim, Germany. E-Mail: rothe@vwl.uni-mannheim.de. Website: http://www.christophrothe.net.

1. INTRODUCTION

Regression discontinuity (RD) designs have become a popular empirical strategy for estimating causal treatment effects from observational data in economics. In sharp regression discontinuity (SRD) designs units receive a treatment if and only if a running variable falls above some known threshold value, whereas in fuzzy regression discontinuity (FRD) designs the probability of treatment jumps discontinuously at the threshold, but generally not from zero to one. Methods for estimation and inference based on local linear regression are widely used in empirical research with both types of designs, and their theoretical properties have been studied extensively (e.g. Hahn et al., 2001; Imbens and Kalyanaraman, 2012; Calonico et al., 2014; Armstrong and Kolesár, 2018).

With local linear SRD inference, the main issue for forming a valid confidence interval (CI) is the handling the estimator's smoothing bias. Due to the estimator's simple structure, this problem is by now well understood. In particular, Armstrong and Kolesár (2019) show that "bias aware" CIs that adjust the critical value for the exact "worst case" bias perform much better in this context than CIs based on traditional undersmoothing or robust bias correction. These techniques cannot be directly used to construct CIs in FRD designs, however, due to the additional nonlinearity of the usual FRD estimator, which is the ratio of two local linear estimates. This issue is typically addressed through a "delta method" approach in which the estimator is linearized, and conditions are imposed that make the resulting error negligible in large samples. The leading term in such an expansion then effectively behaves like an SRD estimator, and asymptotically valid CIs are constructed using one of aforementioned approaches to handling smoothing bias.

While delta method CIs are commonly used in applied economics, they have at least three main shortcomings First, by construction they do not account for the the actual bias of the FRD estimator, but only that of a first-order approximation. In finite samples, even bias aware versions of these CIs are therefore subject to distortions that are not present in the SRD case. Second, delta method CIs generally perform poorly if the jump in treatment probabilities at the threshold is small. This issue is analogous to weak identification in the instrumental variables literature (Staiger and Stock, 1997; Feir et al., 2016). Third, delta method CIs are not valid if the running variable is discrete because the error that arises from the linearization of the FRD estimator generally cannot be negligible in this case.[1] None

[1]Following Lee and Card (2008), it has become common practice in the applied literature to use standard errors that are clustered at the level of the running variable whenever the latter is discrete. Kolesár and Rothe (2018) show that this practice does not alleviate the issues caused by a discrete running variable for RD inference. Indeed, they show that clustering by the running variable tends to produce CIs with poor

of these issues seems to be widely recognized in the applied literature, and collectively they put into question the suitability of delta method CIs in many empirical settings.

The main contribution of this paper is to propose new confidence sets (CSs) for treatment effects in FRD designs based on local linear regression that are not subject to such shortcomings. We avoid issues related to "linearizing" the FRD estimator by basing our CSs on alternative auxiliary parameters that can be estimated via a single local linear regression step. The construction was considered (but implemented quite differently) by Feir et al. (2016) in an RD context, and is similar to that of Anderson-Rubin CSs in the instrumental variables literature. We combine this approach with methods for bias-aware inference developed in Armstrong and Kolesár (2018, 2019). Our resulting CSs are simple to compute, highly efficient, and have excellent coverage properties in finite samples because they explicitly take into account the exact smoothing bias from the local linear regression steps. They are also valid under weak identification and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

Two functions of the running variable play a key role in our analysis: the conditional expectation of the outcome and the conditional treatment probability. The derivation of our CSs assumes that both functions are smooth on either side of the cutoff, in the sense that their second derivatives exist and are bounded in absolute value by some constant that is specified explicitly. Our main theoretical result is that the resulting CSs are honest, in the sense that have correct asymptotic coverage uniformly over the class of candidates for these two key functions. Honesty, or uniform asymptotic validity, is an important property for a CS to have, as it implies good performance across the entire range of plausible data generating processes, and is thus necessary for good finite-sample coverage. A lack of honesty is why some widely used methods for RD inference often perform poorly in practice (Kamat, 2018; Armstrong and Kolesár, 2019).

The bounds on second derivatives are the main tuning parameters required for the construction of our bias-aware Anderson-Rubin-type CSs. Once they are specified, our CSs are valid for any bandwidth choice, and optimal bandwidths are determined automatically from the data. Choosing a bound close to zero effectively imposes the assumption that the respective function is close to linear, while choosing a larger bound allows for functions with increasingly higher curvature. In general, subject-specific knowledge is necessary to determine whether a particular derivative bound is plausible, and in practice we recommend reporting our CSs for a range of bound values in order to assess the robustness of empirical

coverage properties. Such standard errors, and corresponding CIs, should therefore not be used.

findings. Note that one cannot avoid specifying the derivative bounds in RD inference in general: alternative methods like undersmoothing and robust bias correction, that seem to avoid this choice, must make it implicitly in order to maintain correct coverage over the same class of functions (Armstrong and Kolesár, 2019).

Our paper contributes to an extensive methodological literature on RD inference; see Imbens and Lemieux (2008) or Lee and Lemieux (2010) for survey articles, and Cattaneo et al. (2019) for a textbook treatment. It is particularly related to Feir et al. (2016), who also consider Anderson-Rubin-type CSs in an FRD context. The main difference is that Feir et al. (2016) use undersmoothing instead of a bias-aware approach, which means their CSs are subject to a number of practical limitations common to all methods based on undersmoothing. See Section 6.3 for a more detailed discussion, and Section 7 for simulation results regarding the relative merits of our approach compared to that in Feir et al. (2016).[2]

The remainder of this paper is structured as follows. In the following section, we describe our setup and introduce some baseline notation. Section 3 describes our proposed CSs, and Section 4 establishes their theoretical properties. Section 5 discusses a number of possible extensions, and Section 6 compares our CSs to others that have been proposed in the literature. In Section 7, we present our simulation study, and Section 8 contains an empirical application. Finally, Section 9 concludes. Proofs are given in an online appendix.

2. SETUP

2.1. **Model and Parameter of Interest.** Let $Y_i \in \mathbb{R}$ be the outcome, $T_i \in \{0,1\}$ be the actual treatment status, $Z_i \in \{0,1\}$ be the assigned treatment, and $X_i \in \mathbb{R}$ be the running variable of the ith unit in a random sample of size n from a large population. Units are assigned to treatment if the running variable falls above a known threshold, which we normalize to zero, so that $Z_i = \mathbf{1}\{X_i \geq 0\}$. Due to limited compliance, we could have $Z_i \neq T_i$ for some units. We then write $\mu_Y(x) = \mathbb{E}(Y_i|X_i = x)$ and $\mu_T(x) = \mathbb{E}(T_i|X_i = x)$ for the conditional expectation functions of the outcome and the treatment status indicator, respectively, given the running variable; and $\mu_+ = \lim_{x\downarrow 0} \mu(x)$ and $\mu_- = \lim_{x\uparrow 0} \mu(x)$ for the right and left limit, respectively, of a generic function μ at zero. Our parameter of interest

[2]After circulating the first draft of this paper, we were made aware that Huang and Zhaoguo (2018) also discuss the idea of combining an Anderson-Rubin-type CS construction with the approach of Armstrong and Kolesár (2019) to handle smoothing bias. However, Huang and Zhaoguo (2018) misinterpret the results of Armstrong and Kolesár (2019), and their proposed CS is actually invalid under both uniform and pointwise asymptotics.

θ is the ratio of jumps in the functions μ_Y and μ_T at the cutoff:

$$\theta = \frac{\tau_Y}{\tau_T}, \quad \tau_Y = \mu_{Y+} - \mu_{Y-}, \quad \tau_T = \mu_{T+} - \mu_{T-}. \tag{2.1}$$

Under certain continuity and monotonicity conditions (e.g., Hahn et al., 2001; Dong, 2017), this parameter has a causal interpretation as the local average treatment effect among compliers at the cutoff, where compliers units whose treatment decision is affected by the assignment rule (Imbens and Angrist, 1994).

2.2. **Goal of the Paper.** Our goal is to construct confidence sets (CSs) that asymptotically cover the parameter θ with at least some pre-specified probability, uniformly in (μ_Y, μ_T) over some suitably chosen function class $\mathcal{F}$ that embodies shape restrictions that the analyst is willing to impose. That is, we want to construct data-dependent sets $\mathcal{C}^\alpha \subset \mathbb{R}$ that satisfy

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

for some $\alpha > 0$. Note that throughout the paper we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on the functions μ_Y and μ_T implicit in our notation. Following Li (1989), we refer to such CSs that satisfy (2.2) as *honest with respect to* $\mathcal{F}$, which is a much stronger requirement than correct pointwise asymptotic coverage

$$\lim_{n\to\infty} \mathbb{P}(\theta \in \mathcal{C}^\alpha) \geq 1 - \alpha \text{ for all } (\mu_Y, \mu_T) \in \mathcal{F}. \tag{2.3}$$

In particular, under (2.2) we can always find a sample size n such that the coverage probability of $\mathcal{C}^\alpha$ does not subceed $1 - \alpha$ by more than an arbitrarily small amount for every $(\mu_Y, \mu_T) \in \mathcal{F}$. Under (2.3) 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 (2.2) is necessary for good finite sample coverage of $\mathcal{C}^\alpha$ across data generating processes. Of course, we also want our CSs to be efficient, in the sense that they should be "small" while still maintaining honesty.

2.3. **Smoothness Conditions.** Following Armstrong and Kolesár (2018, 2019), we specify the class $\mathcal{F}$ of plausible candidates for (μ_Y, μ_T) 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 zero, are twice differentiable almost everywhere on either side of the threshold, and have second derivatives

uniformly bounded by some constant $B > 0$; and let

$$\mathcal{F}_H^\delta(B) = \{f \in \mathcal{F}_H(B) : |f_+ - f_-| > \delta\},$$

for some $\delta \geq 0$ be a similar Hölder-type class of functions whose discontinuity at zero exceeds δ in absolute magnitude. We then assume that

$$(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}_H^0(B_T) \equiv \mathcal{F}, \tag{2.4}$$

where B_Y and B_T are constants chosen by the researcher based on her subject knowledge about the respective application. In addition to imposing smoothness on μ_Y and μ_T, the condition (2.4) has two further implications. First, since $\mathcal{F}$ is a Cartesian product of two function classes, it rules out cross-restrictions between the shapes of μ_Y and μ_T. This seems reasonable for empirical applications in economics. Second, by imposing $\mu_T \in \mathcal{F}_H^0(B_T)$ instead of $\mu_T \in \mathcal{F}_H(B_T)$, we assume that $\tau_T \neq 0$. However, this is only a technicality whose role is to ensure that the parameter of interest $\theta = \tau_Y/\tau_T$ is well-defined. Our setup explicitly allows τ_T to be arbitrarily close to zero.

The interpretation of thwe smoothness constants B_Y and B_T warrants some further discussion. Intuitively, small values mean that the respective functions are assumed to be "close" to being linear on either side of the cutoff, whereas for larger values the functions can be increasingly "curved". Without explicit bounds on the smoothness of μ_Y and μ_T, 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; Kamat, 2018). This is because, roughly speaking, without such restrictions the true data generating process would always be arbitrarily close, in some appropriate sense, to one in which the parameter of interest takes an arbitrary value on the real line. We discuss the considerations for the choice of the smoothness constants in practice in the context of our empirical application in Section 8.

2.4. **Support of the Running Variable.** Conditional expectation functions are only well-defined over the support of the conditioning variable. The assumption (2.4) must thus be interpreted with some care if the running variable is discrete, or more generally such that there are gaps in its support. Following Kolesár and Rothe (2018) and Imbens and Wager (2019), in this paper assumption (2.4) is formally understood to mean that there exists a pair of functions $(\mu_Y, \mu_T) \in \mathcal{F}$ that is such that $(\mu_Y(X_i), \mu_T(X_i)) = (\mathbb{E}(Y_i|X_i), \mathbb{E}(T_i|X_i))$ with probability 1.

With this convention, the functions (μ_Y, μ_T) are conceptually well defined over the entire real line, even though they are identified from the distribution of observable quantities on the support of the running variable only. Outside the support of X_i, the values of (μ_Y, μ_T) are then only partially identified through their membership in the class $\mathcal{F}$. This means that the jumps τ_Y and τ_T, and thus their ratio θ, are well-defined in (2.1) irrespective of whether the running variable has full support. These quantities are point identified if the support of X_i contains an open neighborhood around the cutoff, and partially identified otherwise. In particular, assumption (2.4) implies that

$$\theta \in \Theta = \left\{ \frac{m_{Y+} - m_{Y-}}{m_{T+} - m_{T-}} : (m_Y, m_T) \in \mathcal{F} \text{ and } (m_Y(X_i), m_T(X_i)) = (\mu_Y(X_i), \mu_T(X_i)) \text{ with probability } 1 \right\},$$

where the identified set Θ is a singleton if the support of X_i contains an open neighborhood around the cutoff, and a more general subset of the real line otherwise.[3] While it is not possible to consistently estimate τ_Y and τ_T, and hence θ, under partial identification, it is possible to conduct valid inference in this case (Imbens and Manski, 2004). Indeed, the question whether θ is point or partially identified is immaterial, for our CSs described below.

2.5. **Local Linear Estimation.** Local linear regression (Fan and Gijbels, 1996) is arguably the most popular empirical strategy for estimation and inference in RD designs. Formally, for some generic dependent variable W_i, that could be equal to Y_i or T_i, for example, the local linear estimator of $\tau_W = \mu_{W+} - \mu_{W-}$, the jump in the generic dependent variable's conditional expectation given the running variable at zero, is

$$\widehat{\tau}_W(h) = e_1' \operatorname*{argmin}_{\beta \in \mathbb{R}^4} \sum_{i=1}^{n} K(X_i/h)(W_i - \beta'(Z_i, X_i, Z_i X_i, 1))^2, \tag{2.5}$$

where $K(\cdot)$ is a bounded kernel function that is zero outside $[-1, 1]$, $h > 0$ is a bandwidth,[4] and $e_1 = (1, 0, 0, 0)'$ is the first unit vector. With this notation, a natural point estimator

[3] More specifically, Θ can take one of three general forms in this case: a closed interval $[a_1, a_2]$; the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$, $a_1 < a_2$; or the entire real line. This is because the range of $(m_{Y+} - m_{Y-}, m_{T+} - m_{T-})$ over the functions $(m_Y, m_T) \in \mathcal{F}$ that are such that $P(m_Y(X_i), m_T(X_i)) = (\mu_Y(X_i), \mu_T(X_i)) = 1$ is the Carthesian product of two intervals. Thus Θ is a closed interval if the range of $m_{T+} - m_{T-}$ does not contain zero, equal to the union of two disjoint half-lines if the range of $m_{T+} - m_{T-}$ contains zero but the range of $m_{Y+} - m_{Y-}$ does not, and equal to the real line if both ranges contain zero.

[4] Note that we use the same bandwidth on each side of the cutoff to keep the notation simple. It would be conceptually straightforward to work with different bandwidths above and below the treatment threshold; see Section 5.3 for details.

of θ is given by $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$, for example. Estimators of the form in (2.5) are the building blocks of our honest CSs described below, and we refer to such estimators $\widehat{\tau}_W(h)$ as *SRD-type estimators of* τ_W in the following, as they are the usual estimators in the context of a hypothetical SRD in which W_i is the outcome variable.

2.6. **Delta Method Inference.** Inference in SRD designs based on SRD-type estimators is by now well understood. Due to the linear structure of SRD-type estimators, the main issues are the choice of bandwidth and the handling of the resulting smoothing bias. Armstrong and Kolesár (2019) compare the properties of the most widely used approaches, and we summarize their results in Section 6.1. None of these approaches can be directly applied in FRD designs as the point estimator $\widehat{\theta}(h) = \widehat{\tau}_Y(h)/\widehat{\tau}_T(h)$ is a nonlinear transformation of two SRD-type estimators. Instead, FRD inference typically relies on a linearization argument. Specifically, one can write $\widehat{\theta}(h) - \theta = \widetilde{\theta}^L(h) + \widetilde{\theta}^R(h)$ with

$$\widetilde{\theta}^L(h) = \frac{\widehat{\tau}_Y(h) - \tau_Y}{\tau_T} - \frac{\tau_Y(\widehat{\tau}_T(h) - \tau_T)}{\tau_T^2}, \quad \widetilde{\theta}^R(h) = \frac{\widehat{\tau}_Y(h)(\widehat{\tau}_T(h) - \tau_T)^2}{\widehat{\tau}_T^*(h)^3},$$

and $\widehat{\tau}_T^*(h)$ an intermediate value between τ_T and $\widehat{\tau}_T(h)$. Under certain regularity and bandwidth conditions, the term $\widetilde{\theta}^L(h)$ is the leading term in this expansion of $\widehat{\theta}(h) - \theta$, and $\widetilde{\theta}^R(h)$ is an asymptotically negligible remainder term. The first order asymptotic properties of $\widehat{\theta}(h)$ then coincide with those of $\widetilde{\theta}^L(h)$, which is again a SRD-type estimator:

$$\widetilde{\theta}^L(h) = \widehat{\tau}_U(h), \quad U_i = \frac{Y_i - \tau_Y}{\tau_T} - \frac{\tau_Y(T_i - \tau_T)}{\tau_T^2}.$$

Inference can then be carried out by applying one of the methods for handling smoothing bias in SRD inference to this leading term, and we refer to any CI based on such a construction as a *delta method CI*. Clearly, such an approach can at best control the bias of a first-order approximation of $\widehat{\theta}(h)$, and not the bias of $\widehat{\theta}(h)$ itself. Moreover, since U_i is unobserved, any such method must be made feasible by instead using an estimate $\widehat{U}_i$ in which τ_Y and τ_T are replaced by suitable estimators, which introduces additional uncertainty into the problem.

A more principal issue with delta method CIs is that the central condition for their validity, namely that $\widetilde{\theta}^R(h)$ is asymptotically negligible relative to $\widetilde{\theta}^L(h)$, is not innocuous. First, it generally rules out weakly identified settings in which τ_T is close to zero. To see this, note that in order for any delta method CI to be honest with respect to $\mathcal{F}$, the term $\widetilde{\theta}^R(h)$ must be of smaller order than $\widetilde{\theta}^L(h)$ 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 $\sup_{\mu_Y,\mu_T} |\widetilde{\theta}^R(h)| = \infty$, which means that delta method CIs break down.[5] A second issue is that requiring $\widetilde{\theta}^L(h)$ to be negligible implicitly rules out a discrete running variable. To see this, note that consistent estimation of τ_T (and τ_Y, for that matter) is generally not possible with a discrete running variable as pointed out in Section 2.4. The terms $\widetilde{\theta}^L(h)$ and $\widetilde{\theta}^R(h)$ therefore generally have non-zero probability limits, and $\widetilde{\theta}^R(h)$ cannot be ignored for the purpose of inference on θ. Since discrete running variables are ubiquitous in empirical applications, this is an important limitation.

Of course, delta method CIs, and particularly their bias aware versions, generally do perform well in the settings for which they have been developed, namely those with strong identification and a continuously distributed running variable (Armstrong and Kolesár, 2019). However, we show in Section 6.2 that our honest CSs described below perform just as well in such canonical setups.

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

3.1. **General Approach.** Our preferred approach to inference in FRD designs avoids the linearization errors of delta method CIs by directly considering an object that can be estimated by an SRD-type estimator. To describe the general idea, we introduce some notation. For any $c \in \mathbb{R}$, we define the "auxiliary" parameter $\tau_M(c)$ as

$$\tau_M(c) = \mu_{M+}(c) - \mu_{M-}(c), \quad \mu_M(x,c) = \mathbb{E}(M_i(c)|X_i = x), \quad M_i(c) = Y_i - cT_i.$$

In other words, $\tau_M(c)$ is the jump in the conditional expectation function $\mu_M(x,c)$ of the constructed outcome $M_i(c)$ given the running variable at $x = 0$. This jump can be estimated by the SRD-type estimator $\widehat{\tau}_M(h,c)$, which is as in (2.5) but with $M_i(c)$ replacing W_i, and we can use methods developed in Armstrong and Kolesár (2018, 2019) to form a bias-aware CI for $\tau_M(c)$ based on this estimator. To obtain a CS for our *actual* parameter of interest $\theta = \tau_Y/\tau_T$, we simply note that $\tau_M(c) = \tau_Y - c \cdot \tau_T$ by linearity of conditional expectations, and that therefore

$$\tau_M(c) = 0 \text{ if and only if } c = \theta.$$

[5]Feir et al. (2016) also point out the failure of delta method inference under weak identification, but do so using different technical arguments. Specifically, they show that delta method CIs based on "undersmoothing" bandwidths do not have correct asymptotic coverage under pointwise asymptotics when τ_T tends to zero with the sample size at an appropriate rate.

We can thus construct a CS for θ by collecting all values of $c \in \mathbb{R}$ for which the auxiliary CI for $\tau_M(c)$ contains the value zero. That is, our proposed CS is of the form

$$\mathcal{C}^{\alpha}_{\text{ar}} = \{c : \text{a } (1-\alpha) \text{ bias-aware CI for } \tau_M(c) \text{ contains } 0\}.$$

This construction is a version of Fieller's (1954) method for inference on ratios, which in turn is a special case of Anderson and Rubin's (1949) for inference in linear instrumental variable models. We refer to $\mathcal{C}^{\alpha}_{\text{ar}}$ as a *bias-aware Anderson-Rubin-type CS* for θ in the following.

3.2. **Formal Description.** To describe the precise implementation of the approach outlined in the previous subsection, note that the SRD-type estimator $\widehat{\tau}_M(h,c)$ can be written as

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

with $\tilde{X}_i = (1, X_i)'$. We also write

$$b_M(h,c) = \mathbb{E}(\widehat{\tau}_M(h,c)|\mathcal{X}_n) - \theta_M(c) \quad \text{and} \quad s_M(h,c) = \mathbb{V}(\widehat{\tau}_M(h,c)|\mathcal{X}_n)^{1/2}$$

for the conditional bias and standard deviation, respectively, of $\widehat{\tau}_M(h,c)$, given the realizations $\mathcal{X}_n = (X_1, \ldots, X_n)'$ of the running variable. Since the weights $w_i(h) = w_{i,+}(h) - w_{i-}(h)$ depend on the data through the realizations of the running variable only, and since $\mu_M(x,c) = \mu_Y(x) - c\mu_T(x)$ and $\sum_{i=1}^n w_{i,+}(h) = \sum_{i=1}^n w_{i-}(h) = 1$, these quantities can be written as

$$\begin{aligned}
b_M(h,c) &= \sum_{i=1}^{n} w_{i,+}(h)(\mu_Y(X_i) - c\mu_T(X_i) - (\mu_{Y+} - c\mu_{T+})) \\
&\quad - \sum_{i=1}^{n} w_{i-}(h)(\mu_Y(X_i) - c\mu_T(X_i) - (\mu_{Y-} - c\mu_{T-})), \\
s_M(h,c) &= \left(\sum_{i=1}^{n} w_i(h)^2 \sigma^2_{M,i}(c) \right)^{1/2},
\end{aligned}$$

where $\sigma^2_{M,i}(c) = \mathbb{V}(M_i(c)|X_i)$ is the conditional variance of $M_i(c)$ given X_i. The conditional bias and standard deviation of $\widehat{\tau}_M(h,c)$ are generally unknown in applications, but can be

bounded and estimated, respectively. First, a natural standard error, or estimate of $s_M(h,c)$, is of the form

$$\widehat{s}_M(h,c) = \left(\sum_{i=1}^{n} w_i(h)^2 \widehat{\sigma}^2_{M,i}(c) \right)^{1/2},$$

where $\widehat{\sigma}^2_{M,i}(c)$ is an appropriate estimate of $\sigma^2_{M,i}(c)$. We discuss in more detail how to construct such estimates below. Second, considering the bias, note that $b_M(h,c)$ depends on (μ_Y, μ_T) through the transformation $\mu_Y - c \cdot \mu_T$ only. Since $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_Y) \times \mathcal{F}^0_H(B_T)$, linearity of the second derivatives operator implies that

$$\mu_Y - c \cdot \mu_T \in \mathcal{F}_H(B_Y + |c| B_T).$$

It then follows from Armstrong and Kolesár (2019) that the "worst case" absolute bias over the functions contained in $\mathcal{F}$, for any value of the bandwidth h, can be bounded as follows:

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

with the supremum being achieved by a pair of piecewise quadratic functions with second derivatives equal to $(B_Y \cdot \text{sign}(x), B_T \cdot \text{sign}(x))$ over $x \in [-h,h]$.[6] We then write the t-statistic for $\widehat{\tau}_M(h,c)$ as

$$\frac{\widehat{\tau}_M(h,c) - \tau_M(c)}{\widehat{s}_M(h,c)} = \frac{\widehat{\tau}_M(h,c) - \tau_M(c) - b_M(h,c)}{\widehat{s}_M(h,c)} + \frac{b_M(h,c)}{\widehat{s}_M(h,c)}. \tag{3.1}$$

Under standard regularity conditions, a Central Limit Theorem (CLT) implies that the first term on the right hand side of the previous equation is approximately standard normal conditional on $\mathcal{X}_n$ in large samples. The second term, on the other hand, is bounded in absolute value by

$$\widehat{r}_M(h,c) = \frac{\bar{b}_M(h,c)}{\widehat{s}_M(h,c)},$$

[6]Note that the bound $\bar{b}_M(h,c)$ on the conditional bias of $\widehat{\tau}_M(h,c)$ may not be sharp if no such pair of piecewise quadratic functions is a feasible candidate for (μ_Y, μ_T). To give an example of a situation in which this might be the case, recall that T_i is binary, and that the range of any candidate for μ_T therefore must be a subset of the unit interval. Then there is no function μ_T with $\mu''_T(x) = B_T \cdot \text{sign}(x)$ and $\mu_T(x) \in [0,1]$ for all $x \in [-h,h]$ if $h > (2B_T)^{-1/2}$. A similar point applies if the support of Y_i is bounded. Still, the bias bound $\bar{b}_M(h,c)$ is valid in such cases.

the "worst case" bias to standard error ratio. For every $c \in \mathbb{R}$ and bandwidth $h > 0$, we can thus construct an auxiliary CI for the pseudo parameter $\tau_M(c)$ as

$$(\widehat{\tau}_M(h,c) \pm \mathrm{cv}_{1-\alpha}(\widehat{r}_M(h,c)) \cdot \widehat{s}_M(h,c)), \tag{3.2}$$

where the critical value $\mathrm{cv}_{1-\alpha}(r)$ is the $(1-\alpha)$-quantile of the $|N(r,1)|$ distribution, the distribution of the absolute value of a normal random variable with mean r and variance 1. In statistical software packages, this critical value can easily be computed as the square root of the $(1-\alpha)$-quantile of a non-central χ^2 distribution of one degree of freedom and non-centrality parameter r^2.

The construction (3.2) is analogous to that of the bias-aware CI of Armstrong and Kolesár (2019, 2018) for SRD designs. Since it is conditional on the realizations of the running variable, it is valid irrespective of whether the distribution of the latter is continuous or discrete; and since it takes into account the exact conditional bias, the CI is also valid for any choice of bandwidth, including fixed ones that do not depend on the sample size. The bandwidth

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

minimizes the length of the auxiliary CI, and thus maximizes the efficiency of inference. If the term on the right of the previous equation does not have a unique minimizer (which might be the case if the running variable is discrete, for example), we define $\hat{h}_M(c)$ as the smallest value of h for which the minimum is attained. The auxiliary CI can be shown to remain honest with this choice of bandwidth under standard conditions. Roughly, this is because, as long as the standard error satisfies a mild uniform consistency property, $\hat{h}_M(c)$ is a consistent estimate of the infeasible optimal bandwidth

$$h_M(c) = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha}(r_M(h,c)) \cdot s_M(h,c), \quad r_M(h,c) = \frac{\overline{b}_M(h,c)}{s_M(h,c)}$$

which depends on neither the outcomes $M_i(c)$ nor the functions μ_Y and μ_T. Our proposed CS for our actual parameter of interest θ is then given by the collection of all $c \in \mathbb{R}$ such that the auxiliary CI in (3.2) with the optimized bandwidth $\hat{h}_M(\mathrm{c})$ contains zero:

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

Note that $\mathcal{C}^{\alpha}_{\mathrm{ar}}$ is not necessarily an interval, although it will take this form in many applica-

tions. We provide details on its shape and computation in Section 5 below.

3.3. **Standard Errors.** There are a number of ways to construct the estimates $\widehat{\sigma}^2_{M,i}(c)$ of the conditional variances $\sigma^2_{M,i}(c) = \mathbb{V}(M_i(c)|X_i)$ that enter the standard error $\widehat{s}_M(h,c)$, with the nearest-neighbor approach of Abadie et al. (2014) being the most common recommendation in the RD literature (e.g. Calonico et al., 2014; Armstrong and Kolesár, 2018, 2019). This procedure defines $\widehat{\sigma}^2_{M,i}(c)$ as the squared difference between outcome $M_i(c)$ of unit i and the sample average of the outcomes of the units whose realizations of the running variable are the closest to that of unit i. While this approach works well in simulations and many empirical applications, it formally does not lead to a standard error that is consistent uniformly over $\mathcal{F}$. This is because the leading bias of this conventional nearest-neighbor estimator is proportional to the first derivative of $\mu_M(\cdot, c)$ at X_i, which is unbounded over $\mathcal{F}$.

In this paper, we therefore consider a variant of the approach of Abadie et al. (2014) in which the sample average among the nearest neighbors is replaced with a best linear predictor. We present this variant with a general notation that allows for ties among the realizations of the running variable. Specifically, let $R > 0$ be a small, fixed integer, denote the rank of $|X_j - X_i|$ among the elements of the set $\{|X_s - X_i| : s \in \{1, \ldots, n\} \setminus \{i\}, X_s X_i > 0\}$ 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 corresponding to at least two distinct realizations of the running variable, and let R_i be the resulting cardinality of $\mathcal{R}_i$. If every realization of X_i is unique, then $R = Q_i = R_i$, and $\mathcal{R}_i$ is simply the set of unit i's R nearest neighbors' indices. With ties in the data, multiple units could be equally far from unit i, and hence R_i could be greater than R.

If the realization of X_i is observed at least R times, we then simply put $\widehat{\sigma}^2_i(c)$ equal to the sample variance of the outcomes of the units with that realization; and otherwise we put $\widehat{\sigma}^2_{M,i}(c)$ equal to a scaled version of the squared difference between $M_i(c)$ and its best linear predictor given its R_i nearest neighbors. That is,

$$\widehat{\sigma}^2_{M,i}(c) = \begin{cases} \dfrac{1}{R_i - 1} \displaystyle\sum_{j: X_j = X_i} \left(M_j(c) - \frac{1}{R_i} \sum_{l: X_l = X_i} M_l(c) \right)^2 & \text{if } \#\{j : X_j = X_i\} \geq R, \\ \dfrac{1}{1 + H_i} \left(M_i(c) - \widehat{M}_i(c) \right)^2 & \text{else,} \end{cases}$$

with

$$\widehat{M}_i(c) = \tilde{X}_i \left(\sum_{j \in \mathcal{R}_i} \tilde{X}_j' \tilde{X}_j \right)^{-1} \sum_{j \in \mathcal{R}_i} \tilde{X}_j' M_j(c), \quad H_i = \tilde{X}_i \left(\sum_{j \in \mathcal{R}_i} \tilde{X}_i' \tilde{X}_j \right)^{-1} \tilde{X}_i{}',$$

and $X_i = (1, X_i)'$ as already defined above. The role of the adjustment term H_i is to ensure that $\widehat{\sigma}^2_{M,i}(c)$ is approximately unbiased in large samples. Its form follows from standard arguments for out-of-sample forecast error evaluation in linear regression models. In contrast to the conventional nearest-neighbor estimator, the bias of $\widehat{\sigma}^2_{M,i}(c)$ is proportional to the second derivative of $\mu_M(\cdot, c)$, which is bounded in absolute value over $\mathcal{F}$ by $B_Y + |c|B_T$. In the simulations and empirical applications in this paper, we implement the estimator $\widehat{\sigma}^2_{M,i}(c)$ with $R = 5$, and the results are not sensitive to this choice.

4. THEORETICAL PROPERTIES

Our main theoretical result in this paper is that $\mathcal{C}^{\alpha}_{\text{ar}}$ is an honest CS for θ with respect to $\mathcal{F}$, as defined in (2.2), under rather weak conditions. We first derive this result under general "high level" assumptions, and then verify these conditions for two specific setups.

Assumption 1. *(i) The data $\{(Y_i, T_i, X_i), i = 1, \ldots, n\}$ are an i.i.d. sample from a fixed population; (ii) $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^q|X_i = x)$ exists and is bounded uniformly over $x \in$ supp(X_i) and $(\mu_Y, \mu_T) \in \mathcal{F}$ for some $q > 2$ and every $c \in \mathbb{R}$; (iii) $\mathbb{V}(M_i(c)|X_i = x)$ is bounded and bounded away from zero uniformly over $x \in$ supp(X_i) and $(\mu_Y, \mu_T) \in \mathcal{F}$ for every $c \in \mathbb{R}$; (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 is standard in 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)–(iii) are standard moment conditions. Since $M_i(c) = Y_i - cT_i$ and T_i is binary, these conditions mainly restrict the conditional moments of the outcome variable. Part (iv) is satisfied by all 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) $(\widehat{\tau}_M(\widehat{h}_M(c), c) - \widehat{\tau}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$; (ii) $(\bar{b}_M(\widehat{h}_M(c), c) - \bar{b}_M(h_M(c), c))/s_M(h_M(c), c) = o_P(1)$ (iii) $\widehat{s}_M(\widehat{h}_M(c), c) = s_M(h_M(c), c)(1 + o_P(1))$; and (iv) $w_{\text{ratio}}(h_M(c)) = o_P(1)$, where $w_{\text{ratio}}(h) = \max_{i=1,\ldots,n} w_i(h)^2 / \sum_{i=1}^n w_i(h)^2$.*

Assumption 2 is a high-level condition that applies to a wide range of settings. We discuss more "low level" conditions for its validity below. Parts (i)–(ii) state that using an estimate of the optimal bandwidth instead of its population version has a minor impact, in some appropriate sense, on the quantities involved in the construction of our CS. Part (iii) states that the standard error $\widehat{s}_M(\widehat{h}_M(c), c)$ is consistent for the true standard deviation $s_M(h_M(c), c)$ at the infeasible optimal bandwidth. Finally, part (iv) implies that the magnitude of each of the weights $w_i(h_M(c))$ is arbitrarily small relative to the others' in large samples. Together with part (i) and the moment conditions in Assumptions 1, this ensures that a CLT applies to an appropriately standardized version of the estimator $\widehat{\tau}_M(\widehat{h}_M(c), c)$. We then have the following result.

Theorem 1. *Suppose that Assumptions 1–2 hold. Then*

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

Again, we leave the dependence of the probability measure $\mathbb{P}$ and the parameter θ on the functions μ_Y and μ_T implicit in our notation. The theorem shows that we can expect $\mathcal{C}^{\alpha}_{\text{ar}}$ to have accurate coverage in finite samples if a CLT applies to our estimates of $\tau_M(c)$, and we have a uniformly consistent standard error. For empirical practice, it is important to also give more low-level conditions for the validity of our approach. The two following ones cover the case of a discrete and a continuously distributed running variable, respectively.

Assumption LL1. The support of the running variable X_i is finite over some open neighborhood of the cutoff.

Assumption LL2. (i) The running variable X_i is continuously distributed with density f_X that is bounded and bounded away from zero over an open neighborhood of the cutoff; (ii) $\mathbb{V}(M_i(c)|X_i = x)$ is Lipschitz continuous uniformly over $x \in \mathbb{R}$ for every $c \in \mathbb{R}$; and (iii) $\mathbb{E}((M_i(c) - \mathbb{E}(M_i(c)|X_i))^4|X_i = x)$ exists and is uniformly bounded over $x \in \mathbb{R}$ for every $c \in \mathbb{R}$.

Theorem 2. *Suppose that Assumption 1 and either Assumption LL1 or Assumption LL2 are satisfied. Then Assumption 2 holds with the standard error described in Section 3.3.*

Some implications of the two low-level conditions for the behavior of $\mathcal{C}^{\alpha}_{\text{ar}}$ are as follows. Under Assumption LL1, it is easy to see $h_M(c)$ approaches a positive constant as the sample size tends to infinity, and thus $\mathcal{C}^{\alpha}_{\text{ar}}$ does not shrink to a singleton asymptotically, which is as

expected given the analysis in Section 2.5. Under Assumption LL2, on the other hand, it follows from results in Armstrong and Kolesár (2019) that

$$h_M(c) = \left(\frac{1}{n} \cdot \frac{\int K(u)^2 du}{(\int K(u)u^2 du)^2} \cdot \frac{\sigma^2_{M+}(c) + \sigma^2_{M-}(c)}{(B_Y + |c|B_T)^2 f_X(0)} \cdot r_\alpha^2 \right)^{1/5} \cdot (1 + o_P(1)),$$

where $r_\alpha = \text{argmin}_{r>0} r^{-1/5} \text{cv}_{1-\alpha}(r)$. Their results also imply that $\text{cv}_{1-\alpha}(r_M(h_M(c), c)) = \text{cv}_{1-\alpha}(r_\alpha) + o_P(1)$. For the common choice $\alpha = 0.05$, for example, $r_\alpha \approx 0.53$, and the corresponding critical value $\text{cv}_{1-\alpha}(r_\alpha) \approx 2.21$ is slightly larger than the usual critical value of 1.96 based on the standard normal distribution.

5. EXTENSIONS AND REMARKS

5.1. **Improving Finite-Sample Coverage Accuracy.** When constructing $\mathcal{C}^\alpha_{\text{ar}}$, using the bandwidth $\widehat{h}_M(c)$ that minimizes the length of the auxiliary CI in (3.2) is attractive, as it balances bias and standard error in way that is optimal for inference. In finite samples, however, this choice can potentially cause some distortions. Too see why, recall from the discussion after Assumption 2 that for any bandwidth h asymptotic normality of $\widehat{\tau}_M(h, c) = \sum_{i=1}^n w_i(h) M_i(c)$ follows from a CLT under the assumption that $w_{\text{ratio}}(h) = o_P(1)$. Normality should thus be a "good" approximation in finite samples if $w_{\text{ratio}}(h)$ is "close" to zero. For $B_Y + |c|B_T$ large, however, the bandwidth $\widehat{h}_M(c)$ can be rather small. This in turn leads to weights $w_i(\widehat{h}_M(c))$ concentrating on a few observations close to the cutoff, to $w_{\text{ratio}}(\widehat{h}_M(c))$ becoming large, and to $\widehat{\tau}_M(\widehat{h}_M(c), c)$ effectively behaving like a sample average of a small number of observations. CLT approximations could then be inaccurate in practice.

To address this issue, one can impose a lower bound on the bandwidth used in the construction of the auxiliary CI in (3.2), where the bound is chosen such that the resulting value of $w_{\text{ratio}}(\cdot)$ remains below some reasonable threshold. Specifically, one can consider replacing $\widehat{h}_M(c)$ in (3.2) with

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

where $\eta > 0$ is a small constant. Note that using $\widehat{h}^*_M(c)$ instead of $\widehat{h}_M(c)$ does not affect the validity of the auxiliary CI in (3.2), as the latter is valid for *any* choice of bandwidth. It is easy to see that in standard setups like those described by Assumptions LL1 or LL2 the lower bound on the bandwidth never binds asymptotically, but in simulations we found that it can potentially improve the finite-sample coverage of our CSs.

To give some intuition for what could be a plausible choice of η, suppose that $\mathcal{X}_n =$

$\{\pm.02, \pm.04, \ldots, \pm1\}$, that $K(t) = (1 - |t|)\mathbf{1}\{|t| < 1\}$ is the triangular kernel, and that $h = 1$. In this case $w_{\text{ratio}}(h) \approx .075$, and a CLT approximation to the distribution of $\widehat{\theta}(h, c)$ should be reasonably accurate in finite samples, as the estimator corresponds to a weighted linear regression with 50 observations on each side of cutoff. Choosing $\eta \in [0.05, 0.1]$ therefore seems reasonable in applications; and we actually use $\eta = .1$ in our simulations.

As $h_M^*(c) \geq h_M(c)$, in finite samples this bandwidth potentially over-smooths the data relative to the one that would be asymptotically optimal for inference. By using it, we accept the cost of a larger finite-sample bias in return for normality being a better finite-sample approximation. This idea could also be used in other settings where the finite sample accuracy of inference faces a similar "bias vs. normality" trade-off, such as inference on average treatment effects under unconfoundedness with limited overlap (e.g. Rothe, 2017).

5.2. **Shape and Computation of CS.** It is difficult to give formal results regarding the shape of our proposed CS $\mathcal{C}_{\text{ar}}^{\alpha}$ in finite samples. To see why, recall the definition from (3.3) that $c \in \mathcal{C}_{\text{ar}}^{\alpha}$ if and only if

$$|\widehat{\tau}_M(\widehat{h}_M(c), c)| - \text{cv}_{1-\alpha}(\widehat{r}_M(\widehat{h}_M(c), c)) \cdot \widehat{s}_M(\widehat{h}_M(c), c) < 0. \tag{5.1}$$

The left-hand-side of (5.1) depends on c both directly and indirectly through the bandwidth $\widehat{h}_M(c)$. It is therefore generally not possible to find the set of values of c that satisfy the above inequality analytically. In practice, we compute $\mathcal{C}_{\text{ar}}^{\alpha}$ as follows. For every $c \in \mathbb{R}$, let $p(c) = \sup\{\alpha : c \in \mathcal{C}_{\text{ar}}^{\alpha}\}$, so that $1 - p(c)$ is the smallest nominal level at which our CS contains c. We then calculate function $p(c)$ exactly over a grid $\{c_1, \ldots, c_m\}$, and approximate it at intermediate points through piecewise linear interpolation. Denoting the resulting approximation function by $\widetilde{p}(c)$, we then compute a numerical approximation to $\mathcal{C}_{\text{ar}}^{\alpha}$ as $\widetilde{\mathcal{C}}_{\text{ar}}^{\alpha} = \{c : \widetilde{p}(c) < \alpha\}$. In simulations and empirical applications, we find that $\mathcal{C}_{\text{ar}}^{\alpha}$ almost always takes one of three general forms: a closed interval $[a_1, a_2]$; the union of two disjoint half-lines, $(-\infty, a_1] \cup [a_2, \infty)$, $a_1 < a_2$; or the entire real line.

While we do not have a formal result regarding the shape of $\mathcal{C}_{\text{ar}}^{\alpha}$ in finite samples, one can prove such a result for a variant of our CS that uses a bandwidth that does not depend on the parameter value under consideration. The result suggests that as long as the dependence of the terms on the left-hand-side of (5.1) on the value of c dominates the indirect dependence through the bandwidth $\widehat{h}_M(c)$, our actual CS $\mathcal{C}_{\text{ar}}^{\alpha}$ should also take one of the three general shapes mentioned above.

Proposition 1. *Let* $\mathcal{C}_{\text{ar}}^{\alpha}(h) = \{c : |\widehat{\tau}_M(h, c)| < \text{cv}_{1-\alpha}(\widehat{r}_M(h, c)) \cdot \widehat{s}_M(h, c))\}$ *be a variant of*

$\mathcal{C}^{\alpha}_{\text{ar}}$, *where* $h > 0$ *is an arbitrary bandwidth that does not depend on* c. *Then* $\mathcal{C}^{\alpha}_{\text{ar}}(h) = [a_1, a_2]$, *or* $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, a_1] \cup [a_2, \infty)$, *or* $\mathcal{C}^{\alpha}_{\text{ar}}(h) = (-\infty, \infty)$, *for some constants* $a_1 < a_2$.

5.3. **Side-Specific Bandwidths and Smoothness Bounds.** So far, our local linear regressions use the same bandwidth on either side of the cutoff, and we have imposed that 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. Both conditions can easily be relaxed. Regarding the bandwidth, however, it follows from results in Armstrong and Kolesár (2019) that there is little to be gained from allowing for side-specific values unless there is a substantial shift in $\mathbb{V}(M_i(c)|X_i = x)$ at the cutoff.

Allowing for different smoothness constants on each side of the cutoff could be of more practical importance, at least for some applications. If the running variable measures time, for example, and the cutoff represents the introduction of a policy, researchers might know in advance that the shape of conditional expected outcomes and/or conditional treatment probabilities become much more "erratic" after the reform. For such scenarios, one could define a more general Hölder-type class 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 $\mathcal{F}^{\delta}_H(B_+, B_-)$ analogously, and then seek to obtain CSs that are honest uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}_H(B_{Y,+}, B_{Y-}) \times \mathcal{F}^0_H(B_{T,+}, B_{T-})$. It is easy to see that this would affect our analysis above by changing the explicit expression of the bound on the absolute value of the conditional bias of $\widehat{\tau}_m(h, c)$ to

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

but every other step of our derivation would remain the same. Of course, in this case it would also make sense to consider different bandwidths on either side of the cutoff.

6. COMPARISON WITH OTHER METHODS

6.1. **Accounting for Smoothing Bias.** Our construction of $\mathcal{C}^{\alpha}_{\text{ar}}$ involves creating a bias-aware CI for the auxiliary parameter $\tau_M(c)$ based on the estimator $\widehat{\tau}_M(h, c)$. Since the latter is an SRD-type estimator, such an auxiliary CI for $\tau_M(c)$ could in principle also be obtained through one of the several alternative approaches to handling smoothing bias from the literature on SRD inference. Armstrong and Kolesár (2019, Section 4) compare the

theoretical properties of such alternatives to bias-aware inference in a more general context. We now briefly review their main findings.

Adapting notation appropriately to our context, the approaches discussed in Armstrong and Kolesár (2019, Section 4) are: *(i)* a naive approach that simply ignores the presence of the bias term. In practice, with this approach the bandwidth is often chosen as an estimate $\hat{h}_{\text{mse}}(c)$ of the value $h_{\text{mse}}(c)$ that minimizes the pointwise asymptotic MSE of $\widehat{\tau}_M(h,c)$ at the "true" function $\mu_Y - c \cdot \mu_T$, see Imbens and Kalyanaraman (2012); *(ii)* Undersmoothing, or using a "small" bandwidth that makes the "bias to standard error" ratio asymptotically negligible. This type of approach was considered by Feir et al. (2016) in the context of FRD Anderson-Rubin inference; see Section 6.3 below for details. In practice, undersmoothing bandwidths are often chosen in an ad-hoc way as $\widehat{h}_{\text{mse}}(c) \cdot n^{-\epsilon}$ for some $\epsilon > 0$; *(iii)* Robust bias correction (Calonico et al., 2014), which involves constructing a new estimate of $\theta_M(c)$ as the difference between $\widehat{\tau}_M(\hat{h}_{\text{mse}}(c), c)$ and an estimate of its bias, obtained via local quadratic regression with another estimated "pilot bandwidth", and adjusting the standard error appropriately.

Armstrong and Kolesár (2019) argue that these three methods have substantial shortcomings relative to bias-aware inference. One is their reliance on the particular empirical bandwidth selector. The issue is that the bandwidth that minimizes the pointwise asymptotic MSE can be very large even if the underlying functions are highly nonlinear, which in turn leads to large smoothing biases in finite samples. Its estimator therefore involves a regularization step that is supposed to prevent extreme bandwidth values, but in practice the result is often unstable and depends critically on the values of tuning parameters that are difficult to pick. Another issue is that, even with reasonable infeasible bandwidth choices, none of the three methods lead to honest CIs. Armstrong and Kolesár (2019) report that naive and undersmoothing CIs generally undercover in finite samples, as they are not correctly centered. The undercoverage of robust bias correction CIs is typically less pronounced,[7] but they are inefficient and tend to be much longer than bias-aware CIs.

On the other hand, the bias-aware auxiliary CI for $\tau_M(c)$, is highly efficient, in the sense that no other approach can produce substantially shorter CIs and still maintain uniform coverage. It is also valid when the running variable is discrete, and comes with a straightforward way to select the bandwidth. The bias-aware approach thus seems to be the most appropriate one to construct the auxiliary CI for $\tau_M(c)$.

[7]Kamat (2018) shows that the robust bias correction CIs based on infeasible MSE-optimal bandwidths are honest with respect to a smaller function class that puts bounds on the absolute value of the third derivatives, instead of only the second.

6.2. **Bias-Aware Delta Method Inference.** In this subsection, we formally describe bias-aware delta method CIs for FRD designs, and compare them to ours based on the Anderson-Rubin principle. Bias-aware delta method CIs are obtained by applying the techniques of Armstrong and Kolesár (2018, 2019) to the term $\widetilde{\theta}^L(h)$, and imposing assumptions that ensure that this term is of smaller order than $\widetilde{\theta}^R(h)$. Armstrong and Kolesár (2019) describe such an approach, but it is instructive for our analysis to repeat the construction explicitly.

In order to derive formal results, we of course have to make assumptions that ensure that the delta method approach is valid in the first place. Specifically, we assume that Assumption LL2 holds, which implies, among other things, that the running variable is 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 $\delta > 0$ to rule out weak identification.

To keep the notation similar to that in Section 3, we write $\widehat{\tau}_U(h)$ instead of $\widetilde{\theta}^L(h)$ in the following, 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 its conditional bias and standard deviation, respectively. Exploiting linearity, we write these quantities as

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

where $\mu_U(x) \equiv \mathbb{E}(U_i|X_i = x) = (\mu_Y(x) - \tau_Y)/\tau_T - \tau_Y(\mu_T(x) - \tau_T)/\tau_T^2$ is a linear combination of 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. Since 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 $\mathcal{F}$, for any value of the bandwidth h, is given by

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

This bound on the bias involves the unknown population quantities τ_Y and τ_T, and thus needs to be estimated. An obvious candidate for such an estimate is

$$\widehat{\overline{b}}_U(h) = \frac{1}{2}\left(\frac{B_Y}{|\widehat{\tau}_T|} + \frac{|\widehat{\tau}_Y|B_T}{\widehat{\tau}_T^2}\right)\sum_{i=1}^n w_i(h)X_i^2,$$

where $\widehat{\tau}_Y = \widehat{\tau}_Y(g_Y)$ and $\widehat{\tau}_T = \widehat{\tau}_T(g_T)$ are local linear estimates based on some preliminary bandwidths g_Y and g_T, respectively. Under the regularity conditions we consider in this sub-

section, the preliminary bandwidths can be chosen such $\widehat{\tau}_Y$ and $\widehat{\tau}_T$ are uniformly consistent over $\mathcal{F}^\delta$, converging with the usual optimal rate of $n^{-2/5}$. One can also construct a feasible standard error of the form

$$\widehat{s}_U(h) = \left(\sum_{i=1}^n w_i(h)^2 \widehat{\sigma}^2_{\widehat{U},i} \right)^{1/2}$$

based on estimates $\widehat{U}_i = (Y_i - \widehat{\tau}_Y)/\widehat{\tau}_T - \widehat{\tau}_Y(T_i - \widehat{\tau}_T)/\widehat{\tau}_T^2$ of the realizations of the U_i. For every value of h, we then define the bias-aware delta method CI for θ with nominal level $1-\alpha$ as

$$\mathcal{C}^\alpha_\Delta(h) = \left(\widehat{\theta}(h) \pm \mathrm{cv}_{1-\alpha} \left(\frac{\widehat{\overline{b}}_U(h)}{\widehat{s}_U(h)} \right) \cdot \widehat{s}_U(h) \right),$$

The bandwidth value that minimizes the length of this CI is

$$\widehat{h}_U = \operatorname*{argmin}_h \mathrm{cv}_{1-\alpha} \left(\widehat{\overline{b}}_U(h)/\widehat{s}_U(h) \right) \cdot \widehat{s}_U(h),$$

and we write $\mathcal{C}^\alpha_\Delta = \mathcal{C}^\alpha_\Delta(\widehat{h}_U)$ for the CI that corresponding to this bandwidth choice. Results in Armstrong and Kolesár (2019) then imply that this CS is honest with respect to $\mathcal{F}^\delta$, and that it is near-optimal, in the sense that no other method can substantially improve upon its length asymptotically.

There are two main downsides to bias-aware delta method CIs relative to our bias-aware Anderson-Rubin CSs. First, as mentioned above, validity of any delta method approach to FRD inference requires strong identification and a continuously distributed running variable. Neither is required for the Anderson-Rubin approach. Second, when combined with the delta method, the bias-aware approach does not account for the actual bias of the estimator of interest, but only for the bias of the leading term in a stochastic approximation. Moreover, even the bound on the approximate bias needs to be estimated. This naturally affects the finite-sample coverage properties of the resulting CI. With the Anderson-Rubin approach, on the other hand, smoothing biases are controlled exactly even in finite samples.

The following theorem shows that bias-aware delta method CIs are also not more efficient than our bias-aware Anderson-Rubin-type CSs in settings where the former are valid. The following result shows that in this case both procedures have the same asymptotic local coverage probabilities for a drifting parameter that is within an $n^{-2/5}$ neighborhood of θ.[8]

[8]Note that $n^{-2/5}$ neighborhoods are the appropriate ones to consider here because the length of $\mathcal{C}^\alpha_\Delta$ is $O_P(n^{-2/5})$ uniformly over $\mathcal{F}^\delta$.

The robustness of bias-aware Anderson-Rubin CSs against weak identification and discrete running variables does thus not come with a loss of efficiency relative to the bias-aware delta method CI in more canonical settings.

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

$$\limsup_{n\to\infty} \sup_{(\mu_Y,\mu_T)\in\mathcal{F}^\delta} |\mathbb{P}(\theta^{(n)} \in \mathcal{C}^\alpha_{\mathrm{ar}}) - \mathbb{P}(\theta^{(n)} \in \mathcal{C}^\alpha_\Delta)| = 0.$$

The theorem is analogous to the result that there is no loss of efficiency when using the Anderson-Rubin approach for inference in exactly identified moment condition models relative to one based on a conventional t-test.

6.3. **Comparison with Feir et al. (2016).** In related work, Feir et al. (2016) also propose an Anderson-Rubin-type CSs for FRD inference. The main practical difference is that Feir et al. (2016) do not explicitly account for the bias from their local linear regression steps, and instead assume that the chosen bandwidth is sufficiently small for the bias to be negligible. Formally, in our notation the CS that they propose is

$$\mathcal{C}^\alpha_{\mathrm{ar,fml}}(h) = \left\{c : |\widehat{\tau}_M(h,c)| < q_{1-\alpha/2} \cdot \widehat{s}_M(h,c)\right\}.$$

where q_α is the usual α quantile of the standard normal distribution, and $\widehat{s}_M(h,c)$ is a standard error that is slightly different from ours. We note that Feir et al. (2016) use a bandwidth for estimating $\widehat{\tau}_M(h,c)$ that does not depend on the value of $c \in \mathbb{R}$. This is clearly not optimal because, as we pointed out above, both the bias and the variance of $\widehat{\tau}_M(h,c)$ depend on the value of $c \in \mathbb{R}$. Feir et al. (2016) also do not specify how this bandwidth should be chosen in practice. In their empirical application, they report $\mathcal{C}^\alpha_{\mathrm{ar,fml}}(h)$ for a range of bandwidth values. In our simulations below, we study a version of their procedure in which a different bandwidth is chosen for every $c \in \mathbb{R}$ as $\widehat{h}_{\mathrm{mse}}(c) \cdot n^{-1/20}$, where $\widehat{h}_{\mathrm{mse}}(c)$ is an estimate of the pointwise-MSE-optimal bandwidth of Imbens and Kalyanaraman (2012). With this common implementation of undersmoothing, the CS from Feir et al. (2016) exhibits moderate finite-sample coverage distortions, and is less efficient than our procedure.

6.4. **Optimized Linear Estimation.** We focus on methods based on local linear regression for inference in RD designs in this paper. An alternative approach, considered for example by Armstrong and Kolesár (2018) and Imbens and Wager (2019), is to directly compute the minimax linear estimator of the respective object of interest through numerical optimization,

and then use this estimator as a basis for inference. Existing results suggest that proceeding like this in the context of an FRD Anderson-Rubin-type CS construction should yield more efficient inference in settings with a multi-dimensional running variable, or in ones where the support of the running variable is rather coarse. It would, however, be expensive from a computational point of view, as an involved numerical optimization would have to be repeated for every $c \in \mathbb{R}$ under consideration; and little is to be gained in terms of efficiency for the most common setup of a univariate running variable with relatively rich support.

7. SIMULATIONS

7.1. **Setup.** In this section, we report the results of a Monte Carlo study of the performance of our bias-aware Anderson-Rubin-type CS, and that of competing procedures. We consider a number of data generating processes that vary with respect to the degrees of nonlinearities of the conditional expectation functions, the richness of the running variable's support, and the strength of identification. Specifically, we generate data as

$$\begin{aligned} Y_i &= (B_Y/2)\text{sign}(X_i) \cdot f(X_i) + \mathbf{1}\{X_i \geq 0\}\tau_Y + 0.1 \cdot \varepsilon_{1i}, \\ T_i &= \mathbf{1}\{-(B_T/2)\text{sign}(X_i) \cdot f(X_i) + \mathbf{1}\{X_i \geq 0\}\tau_T + 0.3 \geq \Phi(\varepsilon_{2i})\} \end{aligned}$$

where $(\varepsilon_{1i}, \varepsilon_{2i})$ are bivariate standard normal random variables with covariance 0.5; the running variable X_i either follows a continuous uniform distribution over $[-1, 1]$ or a discrete uniform distribution over $\{\pm 1/15, \pm 2/15, \ldots \pm 1\}$; and

$$f(x) = x^2 - 1.5 \cdot \max(0, |x| - 0.1)^2 + 1.25 \cdot \max(0, |x| - 0.6)^2.$$

The latter choice implies that the functions μ_Y and μ_T are second order splines whose maximal absolute second derivative over $[-1, 1]$ is B_Y and B_T, respectively. To illustrate the general shape of these functions, we plot μ_Y in Figure 1 for different values of B_Y and $\tau_Y = 1$. We then consider the parameter values $(\tau_Y, \tau_T) \in \{(1, 0.2), (0.5, 0.1)\}$, $B_T \in \{0.2, 1\}$, and $B_Y \in \{1, 10, 100\}$; and set the sample size to $n = 1,000$. Note that the values of (τ_Y, τ_T) are such that $\theta = 2$ in all of these settings. We refer to DGPs with $\tau_T = 0.1$ as weakly identified, and those with $\tau_T = 0.5$ as strongly identified.

We consider the performance of a number of different Anderson-Rubin type CSs in our simulations: (i) our bias-aware CSs, using the true values of the smoothness constants B_Y and B_T of the respective data generating process; (ii) our bias-aware CSs that uses estimates of the smoothness constants B_Y and B_T based on a rule-of-thumb, discussed in Armstrong

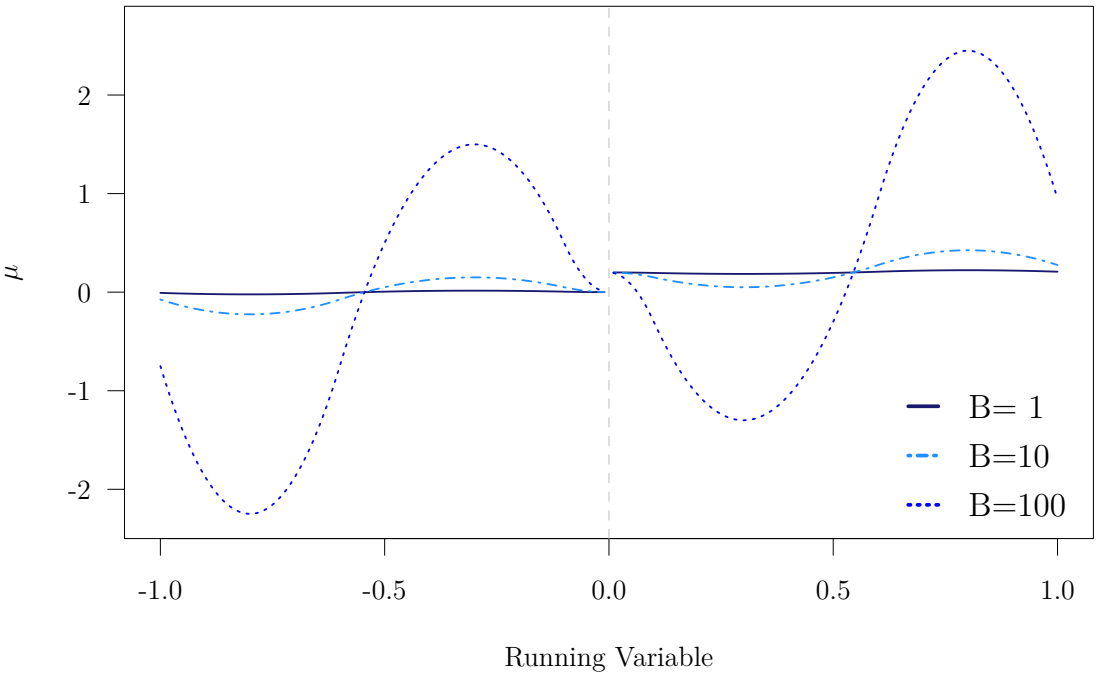

Figure 1: Shape of conditional expectation function μ_Y for different parameter values used in the simulation.

and Kolesár (2019), that fits a global fourth order polynomial on each side of the cutoff, and then calculates its maximal second derivatives; (iii) CSs based on robust bias correction, using a local quadratic specification to estimate the bias, and estimates of the Imbens and Kalyanaraman (2012) "pointwise-MSE-optimal" bandwidth, henceforth IK bandwidth; (iv) naive CSs that ignore the bias, and also use the IK bandwidth; and (v) undersmoothing CS that use the estimated IK bandwidth multiplied by $n^{-1/20}$. In addition, we also consider delta-method-type CIs for θ as described in Section 6 with all of the five just-mentioned methods to handle the bias.

Following standard practice, we use a triangular kernel to compute the local linear (and local quadratic, in the case of methods based on robust bias correction) estimators involved in the construction of the CSs we consider. We remark that these estimators are only well-defined with a discrete running variable if the bandwidth is such that positive kernel weights are assigned to at least two (or three, in the case of methods based on robust bias correction) support points on either side of the cutoff. In our simulations, the estimated IK bandwidth is

often very small and does not satisfy this criterion. Correspondingly, the standard software implementation of all methods for inference based on an estimated IK bandwidth (that is, all non-bias-aware methods) breaks down in such cases. In our results below, we report the rate at which the estimated IK bandwidth fails in this sense across the different DGPs. If such a failure occurs, we manually set the bandwidth equal to 4/15, the value of the fourth largest support point, to compute the respective CS. With a triangular kernel, this means that positive weights are given to three support points on each side of the cutoff.

7.2. **Results.** Table 1 shows the simulated frequencies at which the various CSs that we consider cover the true parameter $\theta = 2$ across data generating processes. We first discuss results for Anderson-Rubin-type CSs, shown in the left panel. As predicted by our theoretical results, our bias-aware CSs have simulated coverage rates close to or greater than the nominal level irrespective of the distribution of the running variable, the degree of nonlinearity of the unknown functions, and the degree of identification strength. Using the rule-of-thumb choice for the smoothness bounds leads to some over-coverage, especially for setups with a discrete running variable. This is because the global quadratic approximation tends to over-estimate the smoothness bounds in these setups. Combining a naive approach, undersmoothing, or robust bias correction with an Anderson-Rubin-type construction leads to CSs that undercover for all data generating processes we consider, with the distortions being more severe (up to about 20 percentage points) for larger values of the smoothness constants. When the running variable is discrete, the failure rate of the IK bandwidth is between 6 and 22 percent, depending on the data generating process.

Turning to result for delta method CIs in the right panel of Table 1, we see that combining a bias-aware approach with this construction does not necessarily lead to a CI with correct coverage even under strong identification. This is because bias-aware delta method CIs only control the bias of a first-order approximation of the estimator on which the CI is based, but not the bias itself. Such coverage distortions are further amplified by weak identification. Discreteness of the running variable, however, does not have a strong detrimental effect on bias-aware delta method CIs in our simulations. Using the rule-of-thumb choice for the smoothness bounds leads to further distortions in some setups. The coverage of delta method CIs that use the naive approach, undersmoothing, or robust bias correction is again severely distorted for most data generating processes, particularly those with weak identification.

While Table 1 clearly shows that bias-aware Anderson-Rubin-type CSs have better coverage properties than their competitors, this by itself does not mean that these CS are particularly powerful. To address this, we plot the simulated rates at which the various

Table 1: Simulated coverage rates (in %) of true parameter for various types of confidence sets

			Anderson-Rubin						Delta Method					
τ_T	B_Y	B_T	BA	BA-RT	Naive	US	RBC	IK Fail	BA	BA-RT	Naive	US	RBC	IK-Fail
Running Variable with Continuous Distribution														
0.5	1	0.2	96.0	96.3	90.9	91.1	91.4	-	94.8	93.1	88.7	88.0	89.5	-
0.5	1	1.0	95.3	95.8	90.8	90.7	91.1	-	94.1	93.2	88.6	88.0	89.4	-
0.5	10	0.2	94.7	96.0	90.1	90.2	90.4	-	95.7	94.8	90.2	89.7	91.8	-
0.5	10	1.0	94.7	96.2	90.1	90.0	90.2	-	95.4	94.6	89.9	89.0	91.5	-
0.5	100	0.2	94.4	97.6	78.7	85.3	75.4	-	92.5	97.1	97.4	94.6	97.8	-
0.5	100	1.0	94.4	97.7	77.8	84.4	74.6	-	94.0	97.1	97.3	94.5	97.8	-
0.1	1	0.2	96.3	96.9	91.5	91.6	92.0	-	89.3	81.1	73.1	70.8	76.4	-
0.1	1	1.0	96.2	96.9	91.5	91.4	91.8	-	87.0	80.5	72.7	70.6	75.7	-
0.1	10	0.2	95.8	96.9	91.3	91.3	91.5	-	88.9	85.6	76.4	73.9	80.6	-
0.1	10	1.0	95.9	97.1	91.4	91.5	91.7	-	89.3	85.4	75.9	73.5	80.0	-
0.1	100	0.2	95.9	98.3	83.3	89.1	79.7	-	90.8	93.7	91.2	85.3	92.7	-
0.1	100	1.0	95.9	98.1	83.5	89.0	80.2	-	91.9	93.9	91.5	85.5	92.9	-
Running Variable with Discrete Distribution														
0.5	1	0.2	96.8	99.0	93.0	92.7	93.7	7.8	95.7	95.1	88.6	87.3	90.4	7.8
0.5	1	1.0	96.6	99.1	92.9	92.9	93.3	7.7	94.6	94.9	87.8	86.9	89.9	7.7
0.5	10	0.2	96.7	99.2	92.0	91.4	92.4	7.8	97.6	97.9	90.5	89.8	93.4	7.7
0.5	10	1.0	96.6	99.2	92.6	92.3	92.7	8.1	97.6	97.8	90.3	89.4	93.0	7.9
0.5	100	0.2	96.5	100.0	71.6	65.3	63.9	21.0	90.5	96.4	99.0	99.2	98.3	15.6
0.5	100	1.0	96.4	100.0	70.4	64.4	63.1	22.2	92.7	96.7	99.2	99.3	98.3	15.4
0.1	1	0.2	97.1	99.3	93.8	94.2	94.0	7.8	89.7	78.8	66.4	62.7	72.0	8.0
0.1	1	1.0	96.9	99.2	93.6	93.4	93.9	7.9	87.4	78.1	66.0	62.2	71.3	7.9
0.1	10	0.2	97.7	99.5	93.6	93.5	93.7	8.0	93.2	88.7	72.0	68.7	78.5	7.7
0.1	10	1.0	97.9	99.5	93.3	93.2	93.6	7.8	93.6	88.4	71.3	68.3	77.6	8.0
0.1	100	0.2	97.1	100.0	77.9	72.1	70.0	15.1	95.0	96.4	92.1	93.4	94.9	13.7
0.1	100	1.0	97.2	100.0	77.2	71.4	69.3	15.5	95.8	96.6	92.0	93.6	95.1	13.2

Notes: Results based on 20,000 Monte Carlo draws for a nominal confidence level of 95%. BA: bias-aware approach with known smoothness bounds; BA-RT bias-aware approach with estimates smoothness bounds via rule of thumb; Naive: naive approach that ignores bias; US: undersmoothing approach; RBC: robust bias correction; IK-Fail: rate at which IK bandwidth selector fails to produce a bandwidth such that positive kernel weights are given to at least two support points on either side of the cutoff.

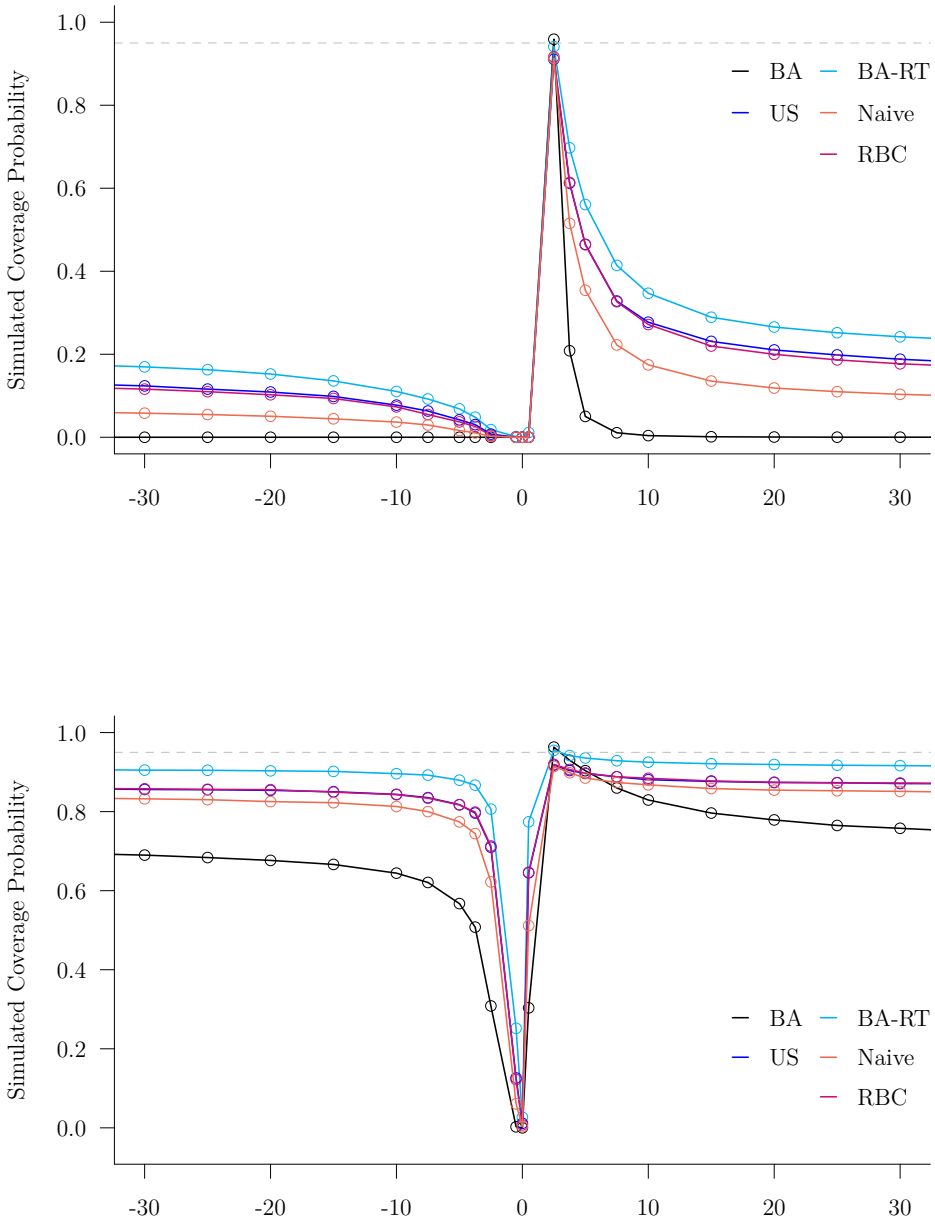

Figure 2: Simulated Rejection Probabilities. Results based on 20,000 Monte Carlo draws for a nominal confidence level of 95%. BA: bias-aware approach with known smoothness bounds; BA-RT bias-aware approach with estimates smoothness bounds via rule of thumb; Naive: naive approach that ignores bias; US: undersmoothing approach; RBC: robust bias correction

Anderson-Rubin-type CSs that we consider cover a range of parameter values other than the true one in Figure 2. We focus on the special case $\tau_T = .5$, $B_Y = 1$, and $B_T = 0.2$ in this plot. This is because, as one can see from the first line of Table 1, the coverage of the true parameter is reasonably close to the nominal level for all procedures, and thus comparison of coverage rates at "non-true" parameter values is meaningful across CSs.

Roughly speaking, a CS can be considered more "powerful" than a competing procedure if its coverage of non-true parameters is closer to zero over a wide range of the relevant parameter space. Figure 2 shows that the coverage rate of bias-aware Anderson-Rubin-type CSs drops very quickly to zero as we move away from the true parameter, and is clearly below that of all competing procedures over almost all of the parameter space. This confirms that the accurate coverage of our bias-aware Anderson-Rubin-type CSs does not come at the expense of statistical power.

8. EMPIRICAL APPLICATION

In this section, we illustrate the application of our bias-aware Anderson-Rubin-type CSs in empirical practice. We use data from Oreopoulos (2006), who studied the effects of a 1947 education reform in the United Kingdom that raised the minimum school-leaving age from 14 to 15 years. The data are a sample of UK workers who turned 14 between 1935 and 1965, obtained by combining the 1984-2006 waves of the UK General Household Survey; see Oreopoulos (2006) for details. We focus on a single parameter of interest, the effect of attending school beyond age 14 on annual earnings measured in 1998 UK pounds. The running variable is the year in which the worker turned 14, and the threshold is 1947. For simplicity, we refer to data on workers who turned 14 in year x as "data for x" below. Figure 3 shows the average of log annual earnings and the empirical proportions of students who attended school beyond age 14 as a function of the running variable. The RD design is clearly seen to be fuzzy.

For reasons explained below, we conduct every analysis in this section separately for both the entire data and for the subset that excludes the data for 1947. Oreopoulos (2006) used global specifications in which the respective dependent variable is regressed on a dummy for turning 14 in or after 1947 and a quadratic polynomial in age to estimate FRD parameters. Here this approach yields the point estimate $\widehat{\theta} = 0.111$ with a heteroscedasticity-robust standard error of 0.033 and a 95% delta method CI of $(0.046, 0.176)$ for the full data; and a point estimate $\widehat{\theta} = .088$ with a standard error of 0.060 and a 95% delta method CI of $(-0.029, 0.205)$ if we exclude data for 1947. However, these results do not account for the

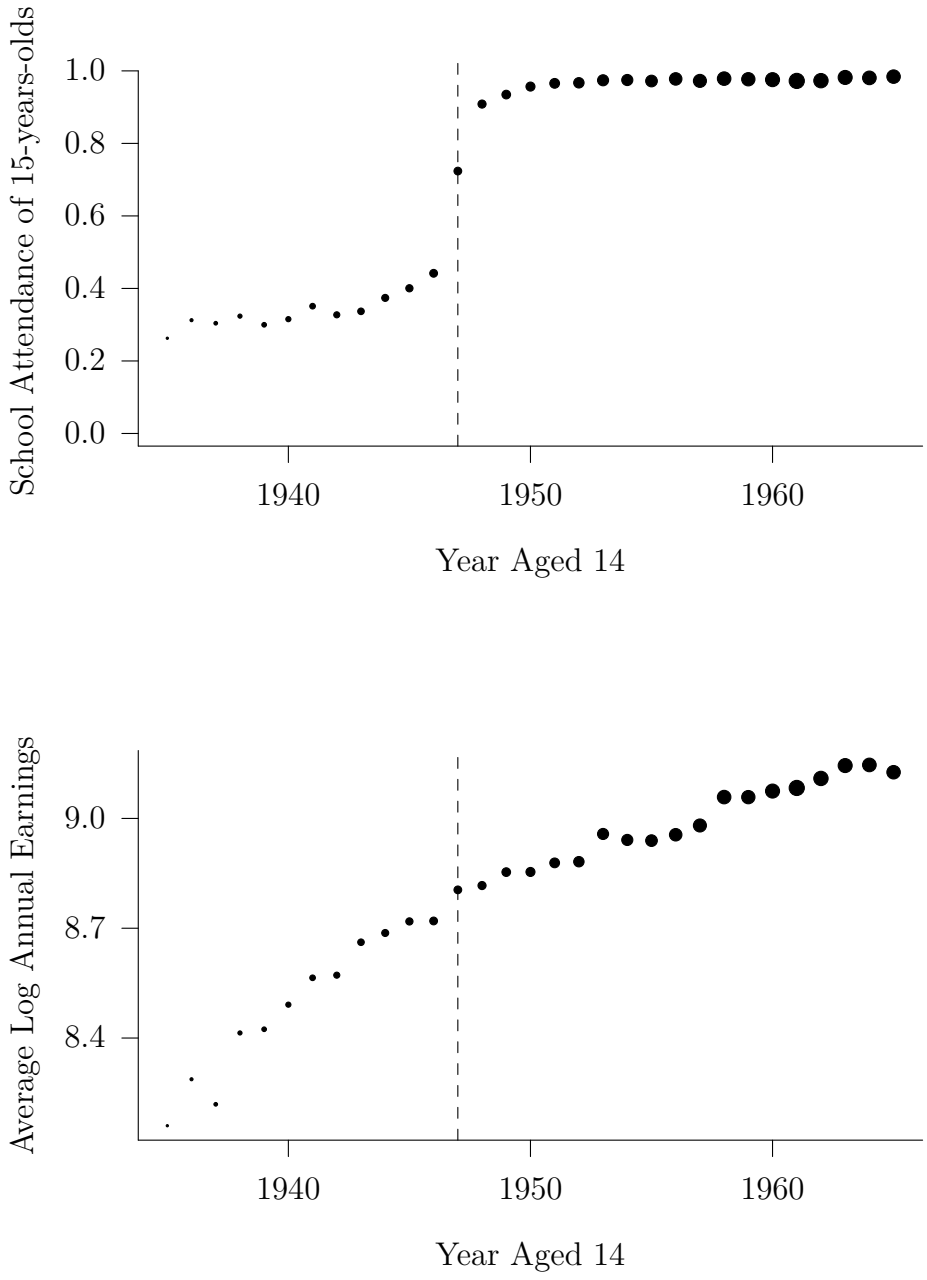

Figure 3: Fraction leaving full time education and average log annual earnings by cohort. The vertical lines indicate the year 1947, in which the minimum school leaving age changed from 14 to 15 years. Size of the dots is proportional to the cohort size. Only data from Great Britain are consider.

Table 2: Confidence sets for the effect of one additional year of compulsory schooling for various values of the smoothness bounds

Results for full data set

		B_T				
		0.0025	0.025	0.05	0.15	0.2
B_Y	0.0025	(-0.128, 0.304)	(-0.136, 0.665)	(-0.150, 0.942)	(-0.522, 3.010)	$(-\infty, \infty)$
	0.025	(-0.248, 0.753)	(-0.286, 0.931)	(-0.343, 1.277)	(-0.824, 3.440)	$(-\infty, \infty)$
	0.05	(-0.390, 0.964)	(-0.437, 1.242)	(-0.496, 1.424)	(-1.185, 3.937)	$(-\infty, \infty)$
	0.15	(-0.840, 1.575)	(-0.936, 1.761)	(-1.074, 2.032)	(-2.906, 6.076)	$(-\infty, \infty)$
	0.2	(-1.070, 1.810)	(-1.195, 2.027)	(-1.375, 2.345)	(-3.908, 7.203)	$(-\infty, \infty)$

Results excluding data for 1947

		B_T				
		0.0025	0.025	0.05	0.15	0.2
B_Y	0.0025	(-0.132, 0.261)	(-0.136, 0.513)	(-0.177, 0.684)	(-1.094, 2.597)	$(-\infty, \infty)$
	0.025	(-0.294, 0.611)	(-0.356, 0.728)	(-0.454, 0.871)	(-1.707, 3.316)	$(-\infty, \infty)$
	0.05	(-0.478, 0.785)	(-0.555, 0.890)	(-0.653, 1.046)	(-2.461, 4.175)	$(-\infty, \infty)$
	0.15	(-1.020, 1.313)	(-1.160, 1.493)	(-1.371, 1.764)	(-6.053, 7.997)	$(-\infty, \infty)$
	0.2	(-1.289, 1.581)	(-1.468, 1.800)	(-1.739, 2.132)	(-8.037, 10.025)	$(-\infty, \infty)$

Notes: Estimates use data for Great Britain only.

potential misspecification of the global linear regression model.

In Table 2, we report our bias-aware Anderson-Rubin-type CSs with nominal level 95% for various values of the smoothness bounds, namely $(B_Y, B_T) \in \{0.0025, 0.025, 0.05, 0.15, 0.2\}^2$, separately for the entire data (top panel) and for the subsample that excludes data for 1947 (bottom panel). All CSs shown are formally valid given the respective choice of B_Y and B_T. To decide which particular CSs should be considered a reasonable description of sampling uncertainty, however, one needs to consider the empirical content of these smoothness bounds (one can always compute a CSs for any combination of B_Y and B_T, irrespective of whether these values make sense in a particular context).

The smallest value of B_Y that we consider corresponds to the assumption that μ_Y is very close to linear on either side of the cutoff, while larger values allow for increasing degrees of curvature of μ_Y. An analogous statement applies to B_T and μ_T. Following Kolesár and Rothe (2018), we can also interpret the value of these smoothness bounds through the heuristic that a function $f \in \mathcal{F}_H(B)$ cannot deviate by more than $B/8$ from a straight line between two points that are one unit apart. When it comes to the choice of B_Y, we can use such reasoning

together with the fact that a typical increase in log earnings per extra year in age is about 0.02 in the data to deduce that $B_Y = 0.025$ and $B_Y = 0.05$ should reasonable choices.

Regarding the choice of B_T, one has to be a bit more careful. From the top panel of Figure 3, we see that the empirical share of "treated" students increases very slowly after 1948, but jumps sharply from 0.724 for 1947 to 0.909 for 1948. If we consider the latter change as a natural variation in treatment probabilities that was not directly associated with the reform, a minimum value for B_T of about 0.15 is needed to capture the "hockey stick" shape of observed treatment probabilities on the right of the threshold.[9] Using such a B_T implies that a similar increase in treatment probabilities between 1946 and 1947 would have been plausible in the absence of the reform. For $B_T \geq 0.15$, the data are thus consistent with a value of τ_T that is very close to zero, which means that our parameter of interest is rather weakly identified. Indeed, for $B_Y \in \{0.025, 0.05\}$ and $B_T = 0.15$ the CSs in the top panel of Table 2 are extremely wide (in the sense that they contain values that are implausible candidates for the returns an additional year of compulsory schooling), and for $B_T = 0.2$ the CSs are in fact all equal to the entire real line.

If we take the arguably more realistic position that the change in treatment probabilities between 1947 and 1948 was still mostly caused by the introduction of the reform through delayed implementation, a more natural approach is to exclude the 1947 data for the analysis. With this sample selection the typical increase in treatment probability per year is about 0.015 on the left of the threshold and 0.005 right, which again suggests that $B_T = 0.025$ or $B_T = 0.0025$ should be reasonable choices. The parameter of interest is then rather strongly identified. The resulting CSs for $B_Y \in \{0.025, 0.05\}$ and $B_T \in \{0.0025, 0.025\}$ in the bottom panel of Table 2 are substantially more narrow than their counterparts discussed above. However, they still quite wide from an empirical point of view, which shows that the data are not very informative about the returns to schooling.

9. CONCLUSIONS

Fuzzy regression discontinuity designs occur frequently in many areas of applied economics. Motivated by the various shortcomings of existing methods of inference, we propose new confidence sets for the causal effect in such designs, which are based on a bias-aware Anderson-Rubin-type construction. Our CSs are simple to compute, highly efficient, and have excel-

[9]Kolesár and Rothe (2018) propose a method to estimate a lower bound on the value of B_T. Their procedure gives a point estimate of 0.158, with 95% CI of $[0.126, \infty)$, for the data to the right of the threshold. For the data on the left the point estimate is zero, with a 95% CI of $[0, \infty)$. Recall that it is not possible to deduce an upper bound on the value of B_T from the data.

lent coverage properties in finite samples because they explicitly take into account the exact smoothing bias from the local linear regression steps. They are also valid under weak identification and irrespective of whether the distribution of the running variable is continuous, discrete, or of some intermediate form.

A. APPENDIX

To simplify the notation, we write $A_n(\mu) = o_{P,\mathcal{F}}(1)$ if $\sup_{\mu\in\mathcal{F}} |A_n(\mu)| = o_P(1)$ for a generic sequence $A_n(\mu)$ of random variables indexed by $\mu \in \mathcal{F}$. We also drop the subscript "opt" from the symbol for the optimal bandwidth, and omit its dependency on c in the following. For example, we write h_M instead of $h_M(c)$.

A.1. **Proof of Theorem 1.** Since $\theta \in \mathcal{C}^\alpha_{\text{ar}}$ if and only if $\tau_M(\theta) \in \mathcal{C}^\alpha(\theta)$, it suffices to show that for any $c \in \mathbb{R}$

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

To see this, note that it follows from Assumption 2, uniform continuity of the function $\text{cv}_{1-\alpha}$, and our basic notation, that

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

We now apply Lyapunov's CLT to show that $(\hat\tau_M(h_M, c) - \mathbb{E}\left[\hat\tau_M(h_M, c)|\mathcal{X}_n\right])/s_M(h_M, c)$ converges in distribution to a standard normally distributed random variable, uniformly over $(\mu_Y, \mu_T) \in \mathcal{F}$. Specifically, let C be a positive constant, let $\delta > 2$, and recall that $\hat\tau_M(h_M, c) = \sum_{i=1}^n w_i(h_M) M_i(c)$. Then Lyapunov's CLT can be applied conditional on $\mathcal{X}_n$ since

$$\begin{aligned}
\lim_{n\to\infty} \frac{\sum_{i=1}^n \mathbb{E}\left[|w_i(h_M)(M_i(c) - \mathbb{E}[M_i(c)|\mathcal{X}_n])|^\delta |\mathcal{X}_n\right]}{\left(\sqrt{\sum_{i=1}^n w_i(h_M)^2 \sigma^2_{M,i}}\right)^\delta} &\leq \lim_{n\to\infty} C \sum_{i=1}^n \frac{|w_i(h_M)|^\delta}{\left(\sqrt{\sum_{i=1}^n w_i(h_M)^2}\right)^\delta} \\
&\leq \lim_{n\to\infty} C \max_{i=1,\dots n} \left(\frac{|w_i(h_M)|}{\sqrt{\sum_{i=1}^n w_i(h_M)^2}} \right)^{\delta-2} = o_{P,\mathcal{F}}(1)
\end{aligned}$$

by Assumption 1(i)–(iii) and Assumption 2(iv). Standard arguments then yield that

$$\liminf_{n\to\infty}\left(\inf_{(\mu_Y,\mu_T)\in\mathcal{F}}\mathbb{P}(\tau_M(c)\in\mathcal{C}^\alpha(c))-\inf_{(\mu_Y,\mu_T)\in\mathcal{F}}\mathbb{P}\left(\left|Z+\frac{b_M(h_M,c)}{s_M(h_M,c)}\right|\leq \mathrm{cv}_{1-\alpha}(r_M(h_M,c))\right)\right)=0,$$

with Z a generic standard normal random variable. The statement of the theorem now follows from the definition of the critical value function $\mathrm{cv}_{1-\alpha}$ 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 (2019, Theorem B.1) show that the last statement would actually hold with equality if μ_Y and μ_T had unbounded domain. In our setup, we then only have a weak inequality because 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.2. **Proof of Theorem 2.** We split the proof of Theorem 2 into four parts. Lemma A.1 shows that the term $w_{\mathrm{ratio}}(\cdot)$ defined in Assumption 2(iv) converges to zero in probability uniformly over the unknown functions and the bandwidth. This then immediately implies that Assumption 2(iv) holds. Lemma A.2 establishes consistency of our standard error, again uniformly over the unknown functions and the bandwidth. Lemma A.3 establishes consistency of our estimate of the optimal bandwidth, uniformly over the unknown functions. Lemma A.4 finally uses the previous results to show that Assumption 2(i)–(iii) hold.

In order to show the uniformity of convergence over the bandwidth, we introduce the set $\mathcal{H}_n$ whose precise definition depends on whether Assumption LL1 or Assumption LL2 are supposed to hold. Under Assumption LL1 the optimal bandwidth approaches a positive constant $\underline{h}$ in large samples, and we put $\mathcal{H}_n$ equal to $[\underline{h}+n^{-\epsilon},\bar{h}]$ for some sufficiently large $\epsilon>0$ and some fixed sufficiently large constant $\bar{h}$. Under Assumption LL2, we put $\mathcal{H}_n$ equal to the interval $\left[n^{-1/5-\delta},\bar{h}\right]$, where $\bar{h}$ is again some fixed, sufficiently large constant and $\delta\in(0,4/5)$. Since the optimal bandwidth is proportional to $n^{-1/5}$ in this case, it is guaranteed to fall into $\mathcal{H}_n$ for n sufficiently large.

To further simplify the notation, we write $A_n(\mu,h)=o_{P,\mathcal{F},\mathcal{H}_n}(1)$ if $\sup_{h\in\mathcal{H}_n}\sup_{\mu\in\mathcal{F}}|A_n(\mu)|=o_P(1)$ for a generic sequence $A_n(\mu,h)$ of random variables indexed by $\mu\in\mathcal{F}$ and $h\in\mathcal{H}_n$.

Lemma A.1. *Let either Assumption LL1 or LL2 hold. Then*

$$\max_{i\in\{1,\ldots,n\}}\frac{w_i(h)^2}{\sum_{j=1}^n w_j(h)^2}=o_{P,\mathcal{F},\mathcal{H}_n}(1).$$

Proof. Suppose that Assumption LL1 is satisfied. With probability approaching 1, we have

$$\max_{i\in\{1,\dots,n\}} \frac{w_i(h)^2}{\sum_{j=1}^n w_j(h)^2} \leq \max_{i\in\{1,\dots,n\}} \frac{w_i(h)^2}{\sum_{j:X_j=X_i} w_j(h)^2} = \max_{i\in\{1,\dots,n\}} \frac{1}{\sum_{j:X_j=X_i} \mathbb{1}[X_i = X_j]}$$

for any $h \in \mathcal{H}_n$. As the sample size increases, the number of units whose realization of the running variable is equal to any particular value in its support tends to infinity, and we obtain the statement of the lemma.

Now suppose that Assumption LL2 is satisfied. Clearly, it holds for any $h \in \mathcal{H}_n$ that

$$\max_{i\in\{1,\dots,n\}} \frac{w_i(h)^2}{\sum_{j=1}^n w_j(h)^2} \leq \max_{i:Z_i=1} \frac{w_i(h)^2}{\sum_{j:Z_j=1} w_j(h)^2} + \max_{i:Z_i=0} \frac{w_i(h)^2}{\sum_{j:Z_j=0} w_j(h)^2}.$$

It then suffices to show that the first term on the right-hand side of the last inequality tends to zero in probability uniformly over $\mathcal{F}$ and $\mathcal{H}_n$, as the same arguments can be used to prove an analogous result for the second term. Note that

$$\begin{aligned}
&\max_{i:Z_i=1} \frac{w_i(h)^2}{\sum_{j:Z_j=1} w_j(h)^2} \\
&\quad = \max_{i:Z_i=1} \frac{K(X_i/h)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2 - X_i \sum_{l:Z_l=1} X_l K(X_l/h)]^2}{\sum_{j:Z_j=1} K(X_j/h)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h)]^2}.
\end{aligned}$$

Treating the numerator of the right-hand side of the second line as a function of X_i, it follows from the fact that the kernel is bounded from above by Assumption 1 that this function is convex in $X_i \in [0, h]$. The maximum of this function is thus bounded by its value at the boundaries, which are equal to $[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2]^2$ for $X_i = 0$, and from above bounded by $[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2]^2 + h^2[\sum_{l:Z_l=1} X_l K(X_l/h)]^2$ for $X_i = h$. We also have that $h < \bar{h}$ for all $h \in \mathcal{H}_n$ for n sufficiently large. Taken together, this means that

$$\begin{aligned}
&\max_{i:Z_i=1} \frac{w_i(h)^2}{\sum_{j:Z_j=1} w_j(h)^2} \\
&\quad \leq C \max_{i:Z_i=1, X_i\leq h} \frac{[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2 - X_i \sum_{l:Z_l=1} X_l K(X_l/h)]^2}{\sum_{j:Z_j=1} K(X_j/h)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h)]^2} \\
&\quad \leq C \frac{(\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2)^2 + \bar{h}^2(\sum_{l:Z_l=1} X_l K(X_l/h))^2}{\sum_{j:Z_j=1} K(X_j/h)^2[\sum_{l:Z_l=1} X_l^2 K(X_l/h)^2 - X_j \sum_{l:Z_l=1} X_l K(X_l/h)]^2}.
\end{aligned}$$

for some finite constant C, for all $h \in \mathcal{H}_n$ and for n sufficiently large. Standard kernel calculations than yield that the numerator on the right-hand side of the last inequality is of the order $O_P(n^2h^2)$, while the denominator is of the order $O_P(n^3h^3)$. Since $\inf_{h\in\mathcal{H}_n} nh \to \infty$

as $n \to \infty$ by assumption, this completes our proof. □

Lemma A.2. *Let Assumption 1 and either Assumption LL1 or LL2 hold. Then*

$$\hat{s}^2_M(h,c) = s^2_M(h,c)(1 + o_{P,\mathcal{F},\mathcal{H}_n}(1))$$

Proof. Our proof is similar in structure to that of Abadie and Imbens (2006, Theorem 6). To simplify the presentation, we supress the dependence on c of various quantities that apprear in this proof, and thus write $\hat{s}^2_M(h)$ instead of $\hat{s}^2_M(h,c)$, etc. We also write

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

so that $\sum_{i=1}^n q_i(h)\hat{\sigma}^2_{M,i} = \hat{s}^2_M(h)/s^2_M(h)$. We note that $\max_{i=1,\dots,n} q_i(h) = o_{P,\mathcal{F},\mathcal{H}_n}(1)$ and $\sum_{i=1}^n q_i(h) = O_{P,\mathcal{F},\mathcal{H}_n}(1)$ by Lemma A.1 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 we are considering a kernel with compact support by Assumption 1 and the bandwidth is bounded, the number of support points for which $q_i(h) > 0$ is finite. It follows that $\sum_{i=1}^n \mathbb{1}[X_i = g]$ tends to infinity uniformly over all support points g with $q_i(h) > 0$ if $X_i = g$. Moreover, it holds that

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

Since $\sum_{i=1}^n q_i(h) = O_{P,\mathcal{F},\mathcal{H}_n}(1)$ and $q_i(h)$ is positive, the statement of the lemma follows because

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

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

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

All three statements follow from basic algebra. Next, note that the variance estimators $\hat{\sigma}^2_{M,i}$,

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

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

Our proof then proceeds by showing that

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

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

which together imply the statement of the lemma. We begin by noting that (A.1) follows from the triangle inequality and the fact that $\sum_{i=1}^n q_i(h) = O_{P,\mathcal{F},\mathcal{H}_n}(1)$ if

$$\max_{i=1,\ldots,n} |\sigma^2_{M,i} - \mathbb{E}[\hat{\sigma}^2_{M,i}|\mathcal{X}_n]| = o_{P,\mathcal{F},\mathcal{H}_n}(1). \tag{A.3}$$

To show (A.3), note that

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

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

$$x_{\max} \equiv \max_{i=1,\ldots,n} \max_{r \in \mathcal{R}_i} |X_i - X_r| = o_{P,\mathcal{F},\mathcal{H}_n}(1). \tag{A.4}$$

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

then have that

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

To show (A.3), it thus only remains to be shown that

$$\max_i \frac{1}{1+H_i}\,\check{\mu}_M(X_i)^2 = o_{P,\mathcal{F},\mathcal{H}_n}(1). \tag{A.5}$$

To do so, note that

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

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

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

Here first inequality follows from Cauchy-Schwartz 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.1).

To show that (A.2) holds, write $\tilde{q}_i(h)=q_i(h)(1+H_i)^{-1}$. Note that since $|\tilde{q}_i(h)|\leq|q_i(h)|$,

it follows from Lemma A.1 that $\max_{i=1,\dots,n} \tilde{q}_i(h) = o_{P,\mathcal{F},\mathcal{H}_n}(1)$ and $\sum_{i=1}^n \tilde{q}_i(h) = O_{P,\mathcal{F},\mathcal{H}_n}(1)$. We now write our quantity of interest as the sum of five terms:

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

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 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

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

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

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

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

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

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

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

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

We now consider the third term, which satisfies

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

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

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

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

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

This completes the proof of the statement (A.2); and thus the proof of the lemma. □

Lemma A.3. *Suppose that Assumption 1 holds. If additionally Assumption LL1 holds, then* $\hat{h}_M = h_M(1+o_{P,\mathcal{F}}(n^{-1/2}))$. *If additionally Assumption LL2 holds, then* $\hat{h}_M = h_M(1+o_{P,\mathcal{F}}(1))$.

Proof. Suppose that Assumption LL1 holds. By definition the optimal bandwidth h_M minimizes the criterion function $h \mapsto \mathrm{cv}_{1-\alpha}(r(h))s_M(h)$, and its empirical version $\hat{h}_M$ minimizes the criterion function's sample counterpart $h \mapsto \mathrm{cv}_{1-\alpha}(\hat{r}(h))\hat{s}_M(h)$. It follows from Lemma A.2 and from uniform continuity of the function $\mathrm{cv}_{1-\alpha}(\cdot)$ that the criterion function

is uniformly close to its sample counterpart, in the sense that

$$\mathrm{cv}_{1-\alpha}(\hat{r}(h))\hat{s}_M(h) = \mathrm{cv}_{1-\alpha}(r(h))s_M(h)(1 + o_{P,\mathcal{F},\mathcal{H}_n}(1)). \tag{A.6}$$

Simple algebra also shows that the criterion function is asymptotically strictly convex, and hence has a well-separated minimum. The claim then follows since $h_M \in \mathcal{H}_n$ with probability approaching one by construction.

For the case that Assumption LL1 holds, the result follows through similar arguments from the fact that h_M converges in probability to a strictly positive constant, and the fact that the standard error tends to zero with rate $O_P(n^{-1/2})$. □

Lemma A.4. *Let Assumption 1 and either Assumption LL1 or LL2 hold. Then*

$$\frac{\bar{b}_M(\hat{h}_M) - \bar{b}_M(h_M)}{s_M(h_M)} = o_{P,\mathcal{F}}(1), \tag{A.7}$$

$$\frac{\hat{\tau}_M(\hat{h}_M) - \hat{\tau}_M(h_M)}{s_M(h_M)} = o_{P,\mathcal{F}}(1), \tag{A.8}$$

$$\frac{\hat{s}^2_M(\hat{h}_M) - s^2_M(h_M)}{s^2_M(h_M)} = o_{P,\mathcal{F}}(1). \tag{A.9}$$

Proof. Suppose Assumption LL1 holds. We show that with probability approaching one the absolute values of the terms on the respective left-hand side of (A.7)–(A.9) can be bounded by the product of constant and a term of the form

$$\frac{\sum_{i=1}^n |w_i(\hat{h}_M)^j - w_i(h_M)^j|}{(\sum_{i=1}^n w_i(h)^2)^{j/2}}. \tag{A.10}$$

with either $j = 1$ or $j = 2$. The statement of the lemma then follows by showing that the term in (A.10) converges uniformly to zero in probability.

We first show that the left-hand side of (A.7) is of the required form. As $\hat{h}_M/h_M = 1 + o_{P,\mathcal{F}}(1)$, the probability of the event $\hat{h}^2_M \leq 2h^2_M$ approaches one. On this event, it holds that

$$\frac{\left|\bar{b}_M(\hat{h}_M) - \bar{b}_M(h_M)\right|}{s_M(h_M)} = \frac{B_M}{2}\frac{|\sum_{i=1}^n (w_i(\hat{h}_M) - w_i(h_M))X_i^2|}{s_M(h_M)} \leq \frac{2B_M h^2}{\underline{\sigma}}\frac{\sum_{i=1}^n |w_i(\hat{h}_M) - w_i(h_M)|}{(\sum_{i=1}^n w_i(h)^2)^{1/2}}.$$

The inequality follows as the variance is uniformly bounded from below, the support of the kernel is bounded by $[-1, 1]$ and therefore the weights are zero if $|x| \geq h$.

Second, we show that the left-hand side of (A.8) is of the required form. Writing $\tilde{w}_g =$

$\sum_{i:X_i=X_g} w_i$, we find that by the weak law of large numbers

$$\begin{aligned}\hat{\tau}_M(\hat{h}_M) - \hat{\tau}_M(h_M) &= \sum_{g=1}^{G} (\tilde{w}_g(\hat{h}_M) - \tilde{w}_g(h_M)) \left(\mu_M(x_g) + \sum_{i:X_i=x_g} \varepsilon_i \right) \\ &= \sum_{g=1}^{G} (\tilde{w}_g(\hat{h}_M) - \tilde{w}_g(h_M))(\mu_M(x_g) + o_{P,\mathcal{F}}(1)).\end{aligned}$$

From a second order Taylor expansion of μ_M, noting that $\sum_{g=1}^{G} \tilde{w}_g(h) = 0$ and $\sum_{g=1}^{G} \tilde{w}_g(h) x_g = 0$, and using similar arguments as for the first term, it follows that with probability approaching one

$$\begin{aligned}|\sum_{g=1}^{G} (\tilde{w}_g(\hat{h}_M) - \tilde{w}_g(h_M))\mu_M(x_g)| &= \frac{1}{2} |\sum_{g=1}^{G} (\tilde{w}_g(\hat{h}_M) - \tilde{w}_g(h_M))\mu''(\tilde{x}_g) x_g^2| \\ \leq \frac{1}{2} B_M h^2 \sum_{g=1}^{G} |\tilde{w}_g(\hat{h}_M) - \tilde{w}_g(h_M)| &= \frac{1}{2} B_M h^2 \sum_{i=1}^{n} |w_i(\hat{h}_M) - w_i(h_M)|.\end{aligned}$$

where $\tilde{X}_k$ is some value between zero and x_g.

Third, we show that the left-hand side of (A.9) is of the required form. This follows by Lemma A.2 if $\hat{s}_M(\hat{h}_M) - \hat{s}_M(h_M) = o_{P,\mathcal{F}}(s_M(h_M))$. As the running variable has countable many support points, it follows from the weak law of large numbers that $\hat{\sigma}^2_{M,i} = \sigma(X_i)(1 + o_{P,\mathcal{F}}(1))$ uniformly over $\{i : \max\{w_i(h), w_i(\hat{h}_M)\} > 0\}$. As the $\sigma^2_{M,i}$ are bounded from below and from above it holds that

$$\frac{|\hat{s}^2_M(\hat{h}_M) - \hat{s}^2_M(h_M)|}{s^2_M(h_M)} \leq \frac{\overline{\sigma}^2 + o_{P,\mathcal{F}}(1)}{\underline{\sigma}^2} \frac{\sum_{i=1}^{n} |w_i^2(\hat{h}_M) - w_i^2(h_M)|}{\sum_{i=1}^{n} w_i^2(h_M)}.$$

It thus remains to show that the term in (A.10) uniformly converges to zero. We only prove the case $j = 1$ as the case $j = 2$ follows from the same kinds of arguments. First, note that the numerator of (A.10) satisfies

$$\sum_{i=1}^{n} |w_i(\hat{h}) - w_i(h)| \leq n \sum_{g=1}^{G} |\tilde{w}_g(\hat{h}) - \tilde{w}_g(h)| \leq nG \max_{j=1,\ldots,G} |\tilde{w}_j(\hat{h}_M) - \tilde{w}_j(h)| = o_{P,\mathcal{F}}(\sqrt{n})$$

because $\tilde{w}_j(h_M)$ is a continuous function in h and by Lemma A.3 $\sqrt{n}|\hat{h}_M - h_M| = o_{P,\mathcal{F}}(1)$.

Second, we have that n times the square of the denominator of (A.10) satisfies

$$n\sum_{i=1}^{n} w_i(h)^2 = n\sum_{g=1}^{G} \frac{\tilde{w}_g(h)^2}{\sum_{i=1}^{n} \mathbb{1}[X_i = x_g]} \geq \sum_{g=1}^{G} \tilde{w}_g(h)^2 \geq \max_{j=1,\ldots,K} \tilde{w}_j(h)^2,$$

where the last term is bounded away from zero by construction. The term in (A.10) is thus the ratio of a term of order $o_{P,\mathcal{F}}(\sqrt{n})$, and a term that is bounded away from zero if pre-multiplied by $\sqrt{n}$. Hence the term in (A.10) is of order $o_{P,\mathcal{F}}(1)$ as claimed. This completes the proof of the lemma for the case that Assumption LL1 holds.

Now suppose that Assumption LL2 holds. The statements (A.7) and (A.8) then follow from arguments analogous to those used in the proof of Theorem E.1 in Armstrong and Kolesár (2019). A similar line of reasoning can also be used to show (A.9). We now describe the latter part of the proof in detail. Since $s_M^2(h_M) = O_P((nh)^{-1})$, it suffices to show that

$$nh(\hat{s}_M^2(\hat{h}_M) - \hat{s}_M^2(h_M)) = o_{P,\mathcal{F}}(1).$$

To do so, write $\widehat{\eta}_i = \hat{\sigma}^2(X_i) - \mathbb{E}[\hat{\sigma}_{M,i}^2|\mathcal{X}_n]$, and note that

$$\hat{s}_M^2(\hat{h}_M) - \hat{s}_M^2(h_M) = \sum_{i=1}^{n}(w_i(\hat{h}_M)^2 - w_i(h_M)^2)\widehat{\eta}_i + \sum_{i=1}^{n}(w_i^2(\hat{h}_M) - w_i^2(h_M))\mathbb{E}[\hat{\sigma}_{M,i}^2|\mathcal{X}_n].$$

In the proof of Lemma A.2, we have shown in (A.3) that $\max_{i=1,\ldots,n} |\sigma_{M,i}^2 - \mathbb{E}[\hat{\sigma}_{M,i}^2|\mathcal{X}_n]| = o_{P,\mathcal{F}}(1)$, and by assumption the conditional variance terms $\sigma_{M,i}^2$ are bounded. We thus only need to show that to show that

$$nh\sum_{i=1}^{n} |w_i(\hat{h}_M)^2 - w_i(h_M)^2| = o_{P,\mathcal{F}}(1), \tag{A.11}$$

$$nh\left|\sum_{i=1}^{n}(w_i(\hat{h}_M)^2 - w_i(h_M)^2)\widehat{\eta}_i\right| = o_{P,\mathcal{F}}(1). \tag{A.12}$$

Armstrong and Kolesár (2019) prove a version of these two results in the proof of their of Theorem E.1. In their setup, the scaling is $\sqrt{nh}$ instead of nh, the original weights are used instead of their squared values, and their analogue of $\widehat{\eta}_i$ is i.i.d. whereas in our case these terms are generally not independent. Still, as mentioned above, one can proceed using arguments similar to theirs.

By the triangular inequality it suffices that both (A.11) and (A.12) hold with $w_{+,i}(\cdot)^2$

replacing $w_i(\cdot)^2$, as the same arguments apply to $w_{-,i}(\cdot)^2$. These weights can be written as

$$w_{+,i}^2(h) = \frac{1}{nh}\varphi(h)'\psi_i(h)\varphi(h)$$

where

$$\varphi(h) = \left(\frac{1}{nh}\sum_{i:X_i>0} K(X_i/h)\tilde{X}_i'\tilde{X}_i\right)^{-1} e_1, \quad \psi_i(h_M) = K(X_i/h)^2\tilde{X}_i'\tilde{X}_i/(nh).$$

Let $\|\cdot\|$ be the L_1-norm of a vector or a matrix. By the triangular inequality, it follows that the left-hand side of (A.11) is bounded by

$$\|\varphi(\hat{h}_M)-\varphi(h_M)\|\,(2\,\|\varphi(h_M)|+\|\varphi(\hat{h}_M)-\varphi(h_M)\|)\sum_{i:Z_i=1}\|\psi_i(\hat{h}_M)\|+\|\varphi(h_M)\|^2\sum_{i:Z_i=1}\|\psi_i(\hat{h}_M)-\psi_i(h_M)\|.$$

As shown in the proof of Lemma E.1 in Armstrong and Kolesár (2019), this term is of the order $o_{P,\mathcal{F}}(1)$. Moreover, equation (A.12) is bounded by

$$\|\varphi(\hat{h}_M)-\varphi(h_M)\|\,(2\,\|\varphi(h_M)|+\|\varphi(\hat{h}_M)-\varphi(h_M)\|)\|\sum_{i:Z_i=1}\psi_i(\hat{h}_M)\eta_i\|+\|\varphi(h_M)\|^2\,\|\sum_{i:Z_i=1}(\psi_i(\hat{h}_M)-\psi_i(h_M))\widehat{\eta}_i\|.$$

By Lemma E.1 Armstrong and Kolesár (2019) it follows that $\|\varphi(h_M)\|^2 = O_{P,\mathcal{F}}(1)$ and $\|\varphi(\hat{h}_M)-\varphi(h_M)\| = o_{P,\mathcal{F}}(1)$. It therefore suffices to show that $\|\sum_{i:Z_i=1}(\psi_i(\hat{h}_M)-\psi_i(h_M))\widehat{\eta}_i\| = o_{P,\mathcal{F}}(1)$. The elements of $\psi_i(h)$ are given by the function $g(z) = z^v K(z)^w$ for $v,\ w\ \in\{0,1,2\}$. We therefore show that

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\left(\sup_{s\in[1-\delta,1+\delta]}\left|\sqrt{nh}\sum_{i:Z_i=1}(g(X_i/(sh_M))-g(X_i/h_M))\widehat{\eta}_i\right|>\varepsilon\right)=0. \tag{A.13}$$

For δ small enough, it holds that for s and $\tilde{s}$ in a neighborhood of 1, and C a positive constant that can take different values in different equations, that

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{i:Z_i=1}(g(X_i/sh_M)-g(X_i/\tilde{s}h_M))\widehat{\eta}_i\right)^2\right] &\le \frac{C}{nh}\sum_{i:Z_i=1}(g(X_i/(sh_M))-g(X_i/(\tilde{s}h_M)))^2\\ &\le |1/s-1/\tilde{s}|^2\frac{C}{nh}\sum_{i:Z_i=1}\mathbb{1}[X_i/h\le C]. \end{aligned}\tag{A.14}$$

Here the first inequality holds because $\widehat{\eta}_i$ has a finite second moment, Cauchy-Schwarz, and the fact that for all i the cardinality of the set of indicies j such that $\widehat{\eta}_i$ containes data points that are also used in $\widehat{\eta}_j$ is bounded by a finite constant (this was already shown in

our proof of Lemma A.2). The second inequality then holds because the function $g(\cdot)$ is Lipschitz continuous and the kernel is bounded from above and has compact support. For n large enough, the term in (A.14) is bounded by $|1/s - 1/\tilde{s}|^2$ times a constant that does not depend on the sample size. The result (A.13) then follows from Example 2.2.12 in van der Vaart and Wellner (1996). □

A.3. **Proof of Theorem 3.** For the proof of this result, we make the dependence of quantities like $h_M(c)$ on c again explicit in our notation. We begin by noting that the events $\theta^{(n)} \in \mathcal{C}_\Delta^\alpha$ and $\theta^{(n)} \in \mathcal{C}_{ar}^\alpha$ occur if and only if

$$\frac{|\widehat{\theta}(\widehat{h}_U) - \theta^{(n)}|}{\widehat{s}_U(\widehat{h}_U)} - \mathrm{cv}_{1-\alpha}\left(\frac{\widehat{\overline{b}}_U(\widehat{h}_U)}{\widehat{s}_U(\widehat{h}_U)}\right) < 0 \tag{A.15}$$

$$\text{and} \quad \frac{|\widehat{\tau}_M(\widehat{h}_M(\theta^{(n)}), \theta^{(n)})|}{s_M(\widehat{h}_M(\theta^{(n)}))} - \mathrm{cv}_{1-\alpha}\left(\frac{\overline{b}_M(\widehat{h}_M(\theta^{(n)}), \theta^{(n)})}{\widehat{s}_M(\widehat{h}_M(\theta^{(n)}), \theta^{(n)})}\right) < 0, \tag{A.16}$$

respectively. To prove the theorem, it thus suffices to show that the difference between the respective left-hand sides of the last two displays converges to zero in probability, uniformly over $\mathcal{F}^\delta$. Using arguments analogous to those in the proofs above, one can show that the left-hand side of (A.15) is equal to

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

where $h_U = \mathrm{argmin}_h \mathrm{cv}_{1-\alpha}(\overline{b}_U(h)/s_U(h))s_U(h)$ is the population version of the empirical bandwidth $\widehat{h}_U$. Now recall the definition that $U_i = (Y_i - \tau_Y)/\tau_T - \tau_Y(T_i - \tau_T)/\tau_T^2$, and note that $U_i = M_i(\theta)/\tau_T$. For any bandwidth $h > 0$, we thus have that

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

Substituting these identities into the formula for h_U, we also find that

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

We thus have that the left-hand side of (A.15) is equal to

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

Now consider the term on the left-hand side of (A.16). Using again similar arguments as before, and the fact that $\widehat{\tau}_M(h,c) = \widehat{\tau}_Y(h) - c\widehat{\tau}_M(h)$, one can show that this term is equal to

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

This completes our 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.

Armstrong, T. and M. Kolesár (2019): "Simple and honest confidence intervals in nonparametric regression," *Working Paper*.

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

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

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

Cattaneo, M. D., N. Idrobo, and R. Titiunik (2019): *A Practical Introduction to Regression Discontinuity Designs: Foundations*, Elements in Quantitative and Computational Methods for the Social Sciences, Cambridge University Press.

Dong, Y. (2017): "Alternative Assumptions to Identify LATE in Fuzzy Regression Discontinuity Designs," *Working Paper*.

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.

Fieller, E. C. (1954): "Some problems in interval estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, 16, 175–185.

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.

Huang, X. and Z. Zhaoguo (2018): "Reliable and Efficient Inference in Regression Discontinuity," *Working Paper*.

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.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

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.

Lee, D. S. and D. Card (2008): "Regression discontinuity inference with specification error," *Journal of Econometrics*, 142, 655–674.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

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.

Oreopoulos, P. (2006): "Estimating average and local average treatment effects of education when compulsory schooling laws really matter," *American Economic Review*, 96, 152–175.

Rothe, C. (2017): "Robust confidence intervals for average treatment effects under limited overlap," *Econometrica*, 85, 645–660.

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

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