

# Noise-Induced Randomization in Regression Discontinuity Designs\*

Dean Eckles  
MIT

Nikolaos Ignatiadis  
Stanford University

Stefan Wager  
Stanford University

Han Wu  
Stanford University

Draft version September 2020

## Abstract

Regression discontinuity designs are used to estimate causal effects in settings where treatment is determined by whether an observed running variable crosses a pre-specified threshold. While the resulting sampling design is sometimes described as akin to a locally randomized experiment in a neighborhood of the threshold, standard formal analyses do not make reference to probabilistic treatment assignment and instead identify treatment effects via continuity arguments. Here we propose a new approach to identification, estimation, and inference in regression discontinuity designs that exploits measurement error in the running variable. Under an assumption that the measurement error is exogenous, we show how to consistently estimate causal effects using a class of linear estimators that weight treated and control units so as to balance a latent variable of which the running variable is a noisy measure. We find this approach to facilitate identification of both familiar estimands from the literature, as well as policy-relevant estimands that correspond to the effects of realistic changes to the existing treatment assignment rule. We demonstrate the method with a study of retention of HIV patients and evaluate its performance using simulated data and a regression discontinuity design artificially constructed from test scores in early childhood.

## 1 Introduction

Regression discontinuity designs are a popular approach to causal inference that rely on known, discontinuous treatment assignment mechanisms to identify causal effects [Hahn, Todd, and van der Klaauw, 2001, Imbens and Lemieux, 2008, Thistlethwaite and Campbell, 1960]. More specifically, we assume existence of a running variable  $Z_i \in \mathbb{R}$  such that unit  $i$  gets assigned treatment  $W_i \in \{0, 1\}$  whenever the running variable exceeds a cutoff  $c \in \mathbb{R}$ , i.e.,  $W_i = 1 (\{Z_i \geq c\})$ . For example, in an educational setting where admission to a program hinges on a test score exceeding some cutoff, we could evaluate the effect of the program on marginal admits by comparing outcomes for students whose test scores fell right above and below the cutoff.

---

\*Authors are listed in alphabetical order. We thank Alex D’Amour, Jan Gleixner, Michal Kolesár, David Hirshberg, Guido Imbens, Fabrizia Mealli, Johan Ugander, and José Zubizarreta for helpful discussions.

Recent explanations and qualitative justifications of identification in regression discontinuity designs typically appeal to implicit, local randomization: There are many factors outside of the control of decision-makers that determine the running variable  $Z_i$  such that if some unit barely clears the eligibility cutoff for the intervention then the same unit could also plausibly have failed to clear the cutoff with a different realization of these chance factors [Lee and Lemieux, 2010]. This is sometimes illustrated by reference to sampling error or other errors in measurement that cause units to have a measured running variable just above or just below the threshold. For example, again in our educational setting, there may be a group of marginal students who might barely pass or fail pass the test due to unpredictable variation in their test score, thus resulting in an effectively exogenous treatment assignment rule. Likewise, medical assays frequently involve a degree of random measurement error, whether because of sampling techniques or other sources of random variation [Bor et al., 2014].

Most formal and practical approaches to identification, estimation, and inference for treatment effects in regression discontinuity designs, however, do not use exogenous noise in the running variable to drive inference. Instead, following Hahn, Todd, and van der Klaauw [2001], the dominant approach relies on a continuity argument. As in Imbens and Lemieux [2008], assume potential outcomes  $\{Y_i(0), Y_i(1)\}$  such that  $Y_i = Y_i(W_i)$ . Then, we can identify a weighted causal effect  $\tau_c = \mathbb{E}[Y_i(1) - Y_i(0) \mid Z_i = c]$  via

$$\tau_c = \lim_{z \downarrow c} \mathbb{E}[Y \mid Z = z] - \lim_{z \uparrow c} \mathbb{E}[Y \mid Z = z], \quad (1)$$

provided that the conditional response functions  $\mu_{(w)}(z) = \mathbb{E}[Y(w) \mid Z = z]$  are continuous. Furthermore, if we are willing to posit quantitative smoothness bounds on  $\mu_{(w)}(z)$ , e.g., we could assume  $\mu_{(w)}(z)$  to have a uniformly bounded second derivative, we can use this continuity-based argument to derive confidence intervals for  $\tau_c$  with well understood asymptotics [Armstrong and Kolesár, 2018, Calonico, Cattaneo, and Farrell, 2018, Calonico, Cattaneo, and Titiunik, 2014, Cheng, Fan, and Marron, 1997, Imbens and Kalyanaraman, 2012, Imbens and Wager, 2019, Kolesár and Rothe, 2018].

Despite its simplicity and interpretability, the continuity-based approach to regression discontinuity inference does not satisfy the criteria for rigorous design-based causal inference as outlined by Rubin [2008]. According to the design-based paradigm, even in observational studies, a treatment effect estimator should be justifiable based on randomness in the treatment assignment mechanism alone; the leading example of this paradigm is the analysis of randomized controlled trials following Neyman [1923] and Rubin [1974]. In contrast, the formal guarantees provided by the continuity-based regression discontinuity analysis often take smoothness of  $\mu_{(w)}(z)$  as a primitive. While continuous measurement error in (or “imprecise control” of) the running variable by units implies continuity of the conditional expectation function [Lee, 2008], this result is not used in estimation and inference and, as we show, only makes limited use of the identifying power of measurement error, perhaps most notably for discrete running variables.

Here we propose a new approach to regression discontinuity inference—one that goes back to the qualitative argument above used to justify regression discontinuity designs and directly exploits noise in the running variable  $Z_i$  for inference. Formally, we assume the existence of a latent variable  $U_i$  such that  $\mathbb{E}[Z_i \mid U_i] = U_i$ , and that any variation in the running variable  $Z_i$  around  $U_i$  is exogenous. For example, again revisiting our educational setting, we can take  $U_i$  to be a measure of the student’s true ability; then the test score  $Z_i$  is a noisy measurement of  $U_i$  with well-documented psychometric properties. Likewise,

in a medical setting, the running variable  $Z_i$  may be a measurement of an underlying condition  $U_i$  (e.g., CD4 counts); such diagnostic measurements often have well-studied test-retest reliability. In both cases, it is plausible that the measurements  $Z_i$  are independent of relevant potential outcomes conditional on the underlying quantity  $U_i$ .

Our main result is that, if the measurement error in  $Z_i$  has a known distribution and the measurement error is conditionally independent of potential outcomes, then we can estimate a weighted marginal treatment effect by solving an integral equation. We then propose a practical approach to estimation and inference in regression discontinuity designs that builds on this result. Unlike in the classical regression discontinuity design, our inference is design-based because it can be—at least in the case of bounded outcomes—purely driven by random treatment assignment induced by noise in  $Z_i$ .

## 1.1 Motivating application: Treatment Eligibility and Retention

In this section, we motivate the applicability of our approach by considering a medical application. Bor et al. [2017] study 11,306 patients in South Africa (in 2011-2012) who were diagnosed with HIV, and seek to understand whether immediately initiating antiretroviral therapy (ART) helps retain patients in the medical system. According to health guidelines used in South Africa at the time, an HIV-positive patient would receive immediate ART if their measured CD4 count<sup>1</sup> was below 350 cells/ $\mu$ L. This setting can naturally be analyzed as a regression discontinuity design, with running variable  $Z_i$  corresponding to the log of the CD4 count (in cells/ $\mu$ L) and a treatment cutoff  $c = \log(350)$ .<sup>2</sup> Figure 1(a) shows a histogram of  $Z_i$  from Bor et al. [2017], with treatment cutoff  $c$  denoted by a dashed line.

Given this setting, Bor et al. [2017] proceed to estimate the effect of ART on retention via local linear regression. We implement local linear regression in our setting as follows:<sup>3</sup> We start by choosing a bandwidth  $h$ , and then estimate the treatment effect parameter by regressing  $Y_i$  on a fully interacted linear model<sup>4</sup> in terms of  $W_i$  and  $Z_i - c$  on all observations  $i$  for which  $|Z_i - c| \leq h$ . Here,  $W_i = 1(\{Z_i < c\})$  denotes treatment assignment and  $Y_i$  is a binary indicator of retention of the  $i$ -th patient at 12 months measured by the presence of a clinic visit, lab test, or ART initiation 6 to 18 months after the initial HIV diagnosis. We choose the bandwidth  $h$  using the approach of Armstrong and Kolesár [2020]<sup>5</sup> and, also following that paper, use an inflated critical value when building confidence intervals for the

<sup>1</sup>CD4 cells are specialized immune system cells, and low CD4 count is indicative of poor immune function.

<sup>2</sup>We discard the patients with zero CD4 count.

<sup>3</sup>The local linear regression approach used by Bor et al. [2017] differs from ours in two respects: First, they did not use an approach to inference that formally accounts for misspecification of the local linear regression due to curvature effects near the boundary; and, second, they used CD4 count itself (as opposed to log CD4 count) as the running variable. Regarding the first point, we here prefer to discuss the local linear regression approach of Armstrong and Kolesár [2020] as a baseline, as this approach formally accounts for curvature effects and so gives an honest assessment of the power of local linear regression (in contrast, confidence intervals that do not account for curvature will be over-optimistically short). We note that Bor et al. [2017] do address this question qualitatively: They start with standard (as opposed to bias-adjusted) inference for the treatment effect parameter using the bandwidth choice recommended in Imbens and Kalyanaraman [2012], but then conduct a sensitivity analysis to potential bias by varying the bandwidth. Meanwhile, the use of CD4 versus log CD4 count as the running variable appeared to have little qualitative impact on the performance of local linear regression so here, for consistency, we chose to use log CD4 count as the running variable in all our analyses.

<sup>4</sup>Like in Bor et al. [2017], we use a uniform kernel, i.e., use  $K(x) = 1(\{|x| \leq 1\})$  in (2).

<sup>5</sup>This approach assumes that the conditional response functions  $\mu_{(w)}(x)$  have a bounded second derivative. One then proceeds to estimate the worst-case curvature via polynomial regression, and picking a bandwidth  $h$  that minimizes the worst-case mean-squared error of the local linear regression treatment effect estimator given this curvature bound.

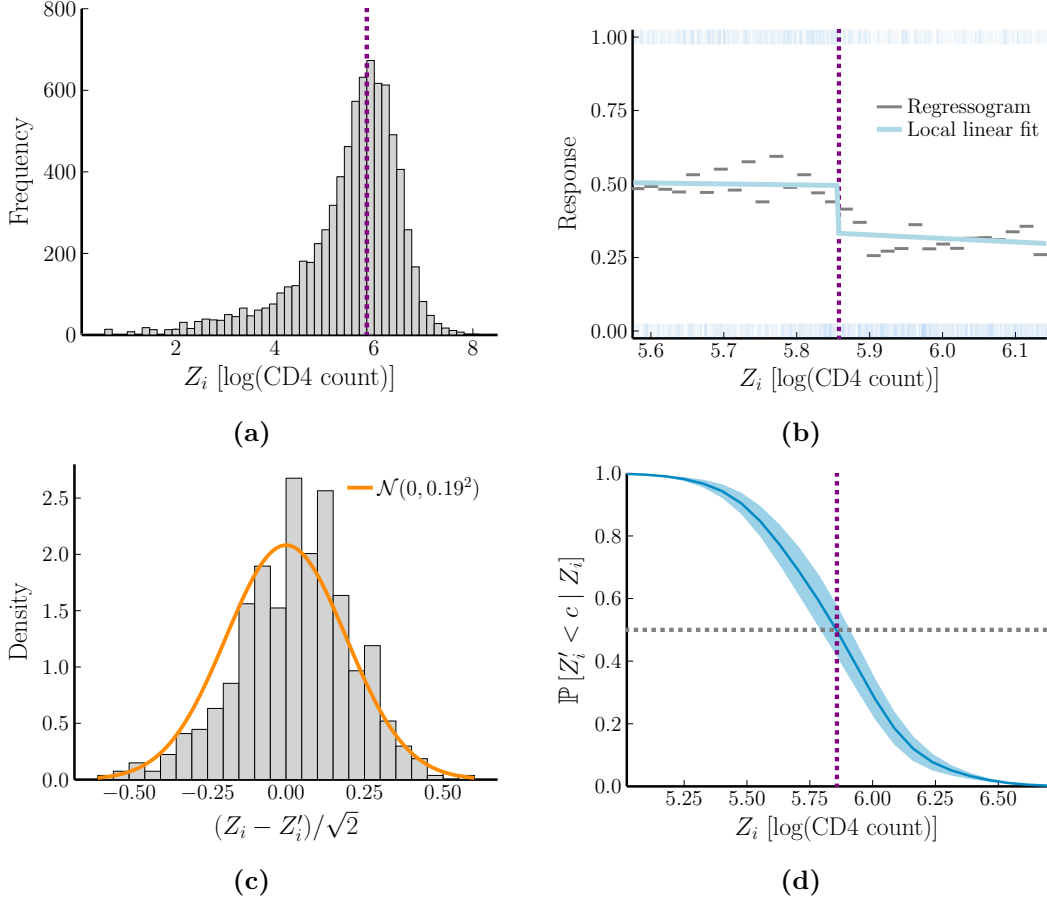


Figure 1: Illustration of a regression discontinuity analysis for estimating the effect of ART on patient retention. **(a)** Histogram of the running variable  $Z_i$  in the dataset of [Bor et al. \[2017\]](#). **(b)** Illustration of a local linear regression for treatment effect estimation. **(c)** Differences  $(Z_i - Z'_i)/\sqrt{2}$  between repeated measurements in the dataset of [Venter et al. \[2018\]](#), overlaid with a Gaussian probability density function. **(d)** Empirical Bayes estimates of  $\mathbb{P}[Z'_i \leq c \mid Z_i = z]$  on the dataset of [Bor et al. \[2017\]](#), with bias-aware 95% confidence intervals obtained using the method of [Ignatiadis and Wager \[2019\]](#).

treatment effect parameter that accounts for potential curvature effects.<sup>6</sup> Figure 1(b) shows the raw data for the subset of analyzed individuals, as well as the fitted regression. We get a point estimate of 0.16 and a 95% confidence interval of (0.08, 0.25).

A potential criticism of this analysis is that it hinges on approximate well-specification of the fitted linear model for  $Y_i$  within the selected bandwidth around  $c$  and this condition is not satisfied “by design” in the sense of [Rubin \[2008\]](#).<sup>7</sup> Now, [Bor et al. \[2017\]](#) do emphasize

<sup>6</sup>Specifically, for 95% confidence intervals, we use a critical value of 2.18 as opposed to 1.96. [Armstrong and Kolesár \[2020\]](#) show that this approach provides honest inference provided we can accurately upper-bound the second derivative of  $\mu_{(w)}(x)$ .

<sup>7</sup>Noise in the running variable does imply certain continuity properties for the functions  $\mu_{(w)}(x)$ , which

that CD4 count measurements are noisy; causes of this noise include instrument imprecision and variability in the blood sample taken [see, e.g., [Glencross et al., 2008](#), [Hughes et al., 1994](#), [Wade et al., 2014](#)]. They then use the existence of such noise to qualitatively argue that treatment  $W_i = 1 (\{Z_i < c\})$  is effectively random close to the cutoff  $c$ , thus strengthening the credibility of the regression discontinuity analysis. However, the local linear regression strategy above doesn't directly rely on this noise to justify the resulting inference.

Here, in contrast, we seek an explicitly design-based approach to estimating the effect of ART on retention that is purely driven by measurement error in  $Z_i$ . To this end, we need to start by modeling this measurement error. [Venter et al. \[2018\]](#) provide pairs of repeated measurements  $Z_i, Z'_i$  of the log CD4 count on 553 individuals (with measurements taken in the same laboratory). Figure 1(c) compares a histogram of the normalized differences  $(Z_i - Z'_i)/\sqrt{2}$  on the data of [Venter et al. \[2018\]](#) to a fitted Gaussian probability density function with noise  $\nu = 0.19$ .<sup>8</sup> Based on this observation, we revisit the analysis of [Bor et al. \[2017\]](#) under the assumption the measurement error in the log CD4 counts can be modeled as  $Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$ , where  $U_i$  is the true underlying log CD4 count of patient  $i$ . Importantly, we find that the noise in  $Z_i$  is large enough, so that treatment is essentially random for patients close to the 350 cells/ $\mu L$  cutoff.

Given these preliminaries, we can now proceed with inference. The approach developed in this paper relies on both on an estimate of the noise level  $\nu$  of  $Z_i$  and a bound  $M$  on the range of the expectation of  $Y_i$  given the true CD4 count  $U_i$ . Here, however, we know that  $Y_i \in \{0, 1\}$  and so can without loss of generality use  $M = 1$ , resulting in inference that only relies on measurement error. As discussed further in Section 5, an application of our method to the dataset of [Bor et al. \[2017\]](#) assuming a noise model  $Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$  with  $\nu = 0.19$  results in a point estimate of 0.11, along with a 95% confidence interval of (0.01, 0.21). While this confidence interval is wider than that obtained by local linear regression it is still statistically significant and—unlike the latter—is directly justified by the sampling design.<sup>9</sup> In Section 5, we also consider how our intervals can be tightened by making further assumptions on the data-generating distribution, and compare our approach to several continuity-based alternatives.

*Remark 1.* To illustrate the noise-induced randomization present in this design, one can consider the following hypothetical question. Suppose patient  $i$  has measured log CD4 count  $Z_i$ . If we were to re-measure their CD4 counts, what is the probability that they would be assigned to treatment, i.e., what is  $\pi(z) = \mathbb{P}[Z'_i < c | Z_i = z]$  where  $Z'_i$  is an independent measurement of log CD4 counts? [Ignatiadis and Wager \[2019\]](#) develop methods for both

---

can then be used to argue for approximate well-specification of the local linear regression model. Existing approaches to local linear regression, however, do not use bounds on the curvature implied by the noise in  $Z_i$  to build confidence intervals for the treatment effect; rather, they rely on curvature estimates obtained by fitting the data via polynomial regression, and the success of this approach cannot be guaranteed by design. As discussed further in Sections 2.3 and 5, it is also possible to build on our results to derive confidence intervals for local linear regression that are justified by design; however, these will be much longer than the ones discussed here (and will also be longer than the ones obtained via our preferred approach).

<sup>8</sup>We estimated the noise level  $\nu = 0.19$  using a robust method that ignores outliers by Winsorizing the smallest and largest 5% of the normalized differences  $(Z_i - Z'_i)/\sqrt{2}$  and rescaling so as to be unbiased under Gaussian noise. In practice, this conservative approach may cause us to underestimate the noise level. As discussed in Remark 2, however, underestimating the noise level in  $Z_i$  will not in general compromise the validity of our inference.

<sup>9</sup>As observed in the simulation study from Section 7, depending on the design, our approach may yield shorter or longer confidence intervals than continuity-based approaches like local linear regression. The fact that the local linear regression confidence intervals are shorter than ours here reflects the fact that the conditional response functions  $\mu_{(w)}(x)$  are estimated to be smoother here than can be justified via the noise process alone.

point estimation and inference in problems of this type, and we display resulting estimates of  $\pi(z)$  in Figure 1(d). As expected, treatment is essentially a coin-flip close to the cutoff.<sup>10</sup>

## 1.2 Related Work

As discussed above, the dominant approach to inference in regression discontinuity designs is via continuity-based arguments that build on (1). Perhaps the most popular continuity-based approach is to use local linear regression, and to estimate the treatment effect at  $Z_i = c$  via [Hahn, Todd, and van der Klaauw, 2001, Imbens and Lemieux, 2008]

$$\hat{\tau}_c = \underset{\tau}{\operatorname{argmin}} \left\{ \sum_{i=1}^n K \left( \frac{|Z_i - c|}{h_n} \right) (Y_i - a - \tau W_i - \beta_- (Z_i - c)_- - \beta_+ (Z_i - c)_+)^2 \right\}, \quad (2)$$

where  $K(\cdot)$  is a weighting function,  $h_n \rightarrow 0$  is a bandwidth, and  $a$  and  $\beta_{\pm}$  are nuisance parameters. In general, this approach can be used for valid estimation and inference of  $\tau_c$  provided the function  $\mu_{(w)}(z)$  is smooth and that  $h_n$  decays at an appropriate rate; the rate of convergence of  $\hat{\tau}_c$  and appropriate choice of  $h_n$  depend on the degree of smoothness assumed. Notable results in this line of work, including robust confidence intervals and data-adaptive choices for  $h_n$ , include Armstrong and Kolesár [2020], Calonico, Cattaneo, and Farrell [2018], Calonico, Cattaneo, and Titiunik [2014], Cheng, Fan, and Marron [1997], Imbens and Kalyanaraman [2012] and Kolesár and Rothe [2018].

More recently, extensions have been considered to the continuity-based approaches to regression discontinuity inference that improve over local linear regression (2) by directly exploiting the assumed smoothness properties of  $\mu_{(w)}(z)$ . Under the assumption that  $\mu_{(w)}(z)$  belongs to a convex class, e.g.,  $|\mu''_{(w)}(z)| \leq B$  for all  $z \in \mathbb{R}$ , Armstrong and Kolesár [2018] and Imbens and Wager [2019] use numerical convex optimization to derive minimax linear estimators of  $\hat{\tau}_c$ . This optimization-based approach also directly extends to more complex regression discontinuity designs, e.g., where  $Z_i$  is multivariate and the treatment assignment is determined by a set  $\mathcal{A}$ , i.e.,  $W_i = 1(\{Z_i \in \mathcal{A}\})$ .

One alternative approach to inference in regression discontinuity designs, which Cattaneo, Frandsen, and Titiunik [2015], Li, Mattei, and Mealli [2015] and Mattei and Mealli [2016] refer to as randomization inference, starts by positing a non-trivial interval  $\mathcal{I}$  with  $c \in \mathcal{I}$ , such that

$$[Z_i \perp \{Y_i(0), Y_i(1)\}] \mid 1(\{Z_i \in \mathcal{I}\}). \quad (3)$$

They then focus on the subset of units with  $Z_i \in \mathcal{I}$ , and perform classical randomized study inference on this subset. Unlike the continuity-based analysis, this approach is design-based in the sense of Rubin [2008].

In practice, however, the assumption (3) is often unrealistic and limits the applicability of methods relying on it. One testable implication of (3) is that  $\mu_{(w)}(z)$  should be constant over  $\mathcal{I}$  for both  $w = 0$  and 1, but this structure rarely plays out in the data.<sup>11</sup> Furthermore,

<sup>10</sup>We emphasize that, although  $\pi(z)$  qualitatively captures the extent to which the treatment assignment is randomized here, it is not a propensity score in the sense of Rosenbaum and Rubin [1983], and in particular cannot be used for inverse-propensity weighting. The relevant propensity score here would be  $e(u) = \mathbb{P}[Z_i < c \mid U_i = u]$ , but this depends on the unobserved  $U_i$  and is thus inaccessible. Our approach to inference will not involve weighting by a transformation of  $\pi(z)$ ; rather, we need to implicitly solve an integral equation in order to account for confounding.

<sup>11</sup>Some authors, e.g. Sales and Hansen [2020], have argued that one can fix this issue by first de-trending outcomes, and then assuming (3) on the residuals. Any such approach, however, relies on well specification of the trend removal, and is thus no longer justified by randomization.

it is not clear how to choose the interval  $\mathcal{I}$  used in (3) via the types of methods typically used for regression discontinuity inference. There’s no data-driven way of discovering an interval  $\mathcal{I}$  over which (3) holds that is itself justified by randomization; conversely, if the interval  $\mathcal{I}$  is known a-priori, then the problem collapses to a basic randomized controlled trial where the regression discontinuity structure ends up not being used for inference.<sup>12</sup>

Some work has generalized this randomization inference approach to regression discontinuity designs, where the independence assumption (3) is only made after conditioning on a set of observed covariates  $X_i$  [Angrist and Rokkanen, 2015, Diaz and Zubizarreta, 2020]. Considering the role of such additional covariates is beyond the scope of the present manuscript; however, it is plausible that our identification strategy powered by noise in the running variable could also be extended to allow for additional, observed confounding variables.

While knowledge of the presence of measurement error (or other noise) in running variables is often mentioned [Bor et al., 2014, 2017, Harlow et al., 2020, Lee, 2008], we do not know of any work that exploits side information about measurement error for inference.<sup>13</sup> Perhaps closest to our work, Rokkanen [2015] posits a latent-factor model for the running variable and uses this for identification and estimation. Rather than using, as we do, e.g., biomedical knowledge, test–retest data, or prior modeling of item-level responses to tests, Rokkanen [2015] uses a particular factor model for scores on at least three, somewhat similar tests, which are assumed observed for the same units observed in the regression discontinuity design. Moreover, while his identification argument is nonparametric, estimation and inference relies on a parametric factor model.

We also contrast our setup with another design-based approach in which the cutoff, rather than the running variable, has an exogenous random component. Ganong and Jäger [2018] posit that the cutoff is randomly drawn according to a known distribution. This may be plausible when the cutoff is set based on, e.g., aggregate statistics for a past year’s data when there are random year-to-year fluctuations. In contrast to our approach, this hypothetical experiment involves highly correlated treatment assignments for units with similar values of the running variable, which should typically substantially decrease precision, as has been observed in the context of spatial boundaries [Kelly, 2019]. In cases where there is both known measurement error in the running variable (as we study) and the cutoff is plausibly random, we can think of our approach as simply conditioning on the observed cutoff, as is also common in other approaches to regression discontinuity designs.

Finally, we also note a related, but distinct line of work that considers regression discontinuity design with latent variables  $U_i$  and noisy measurements  $Z_i$  thereof: Bartalotti, Brummet, and Dieterle [2020], Davezies and Le Barbanchon [2017], Pei and Shen [2016], Yanagi [2014], Yu [2012] assume that treatment is assigned according to  $U_i$ , i.e.,  $W_i = 1 (\{U_i \geq c\})$ ; however only  $Z_i$  is observed. Identification becomes subtle and estimation difficult because of the perils of nonparametric estimation with measurement error [Meister, 2009]. Instead, we use measurement error as our identifying assumption.

<sup>12</sup>Campbell and Stanley [1963] considered an analogy to an imagined “tie-breaking” experiment, where treatment assignment is explicitly randomized for units whose running variable  $Z_i$  is close to a cutoff, and argued that such tie-breaking experiments should be conducted (see also Owen and Varian [2018]). A regression discontinuity design then “attempts to substitute” for such an experiment [Campbell and Stanley, 1963]. However, such tie-breaking designs do not fall under (3), because the running variable is not taken to be randomized—rather the treatment assignment rule is randomized after seeing  $Z_i$  for a subset of units.

<sup>13</sup>Some early work has studied such measurement error under an assumed (e.g., linear) outcome model, simply showing that its presence does not induce bias [Trochim, Cappelleri, and Reichardt, 1991].



## 2 Identification via Noisy Running Variables

Our goal is to develop an approach to identification and inference of causal effects in regression discontinuity designs that exploits noise in the running variable. We start with the classical regression discontinuity design with potential outcomes as described below, and then add an assumption about how the running variable is generated.

**Assumption 1** (Regression discontinuity design). There are  $i = 1, \dots, n$  independent and identically distributed samples  $\{Y_i(0), Y_i(1), Z_i\} \in \mathbb{R}^3$  and a cutoff  $c \in \mathbb{R}$  such that units are assigned treatment according to  $W_i = 1(\{Z_i \geq c\})$ . For each sample, we observe pairs  $\{Y_i, Z_i\}$  with  $Y_i = Y_i(W_i)$ .

**Assumption 2** (Noisy running variable). There is a latent variable  $U_i$  with (unknown) distribution  $G$  such that  $Z_i | U_i \sim p(\cdot | U_i)$  for a known conditional density  $p(\cdot | \cdot)$  with respect to a measure  $\lambda$ , such that  $\mathbb{E}[Z_i | U_i = u] = u$ .<sup>14</sup> We denote the implied marginal distribution of  $Z$  by  $F$  and its  $d\lambda$ -density by  $f$ , i.e.,  $f(z) = dF(z)/d\lambda = \int p(z | u)dG(u)$ .

Qualitatively, we interpret the latent variable  $U_i$  in Assumption 2 as a true measure of the property we want to use for treatment assignment, e.g.,  $U_i$  could capture ability in an educational setting or health in a medical one. The actual observed running variable  $Z_i$  is then a noisy realization of  $U_i$ . One common example of measurement error we consider in this paper is Gaussian measurement error, i.e.,

$$Z_i | U_i \sim \mathcal{N}(U_i, \nu^2), \quad \nu > 0; \quad (4)$$

however the assumption also accommodates discrete running variables, such as  $Z_i | U_i \sim \text{Binomial}(N, U_i)/N$  for some  $N \in \mathbb{N}$ .

Finally, in order to use the noise in  $Z_i$  to identify treatment effects, we need for this noise to be exogenous. The assumption below formalizes this requirement in terms of an unconfoundedness condition following Rosenbaum and Rubin [1983].

**Assumption 3** (Exogeneity). The noise in  $Z_i$  is exogenous, i.e.,  $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp Z_i | U_i$ .

An implication of Assumption 3 is that

$$\mathbb{E}[Y_i | U_i, Z_i] = \alpha_{(W_i)}(u), \quad \alpha_{(w)}(u) = \mathbb{E}[Y_i(w) | U_i = u], \quad (5)$$

where the  $\alpha_{(w)}(u)$  are the response function for the potential outcomes conditionally on the latent variable  $u$ . Then, following Frangakis and Rubin [2002] we can think of  $u$  as indexing over unobserved principal strata, such that

$$\tau(u) = \mathbb{E}[Y_i(1) - Y_i(0) | U_i = u] \quad (6)$$

is the conditional average treatment effect of the stratum with  $U_i = u$ ; see also Heckman and Vytlacil [2005]. Given Assumptions 1–3, we know that the treatment assignment  $W_i = 1(\{Z_i \geq c\})$  is exogenous conditionally on  $U_i$ . The remaining difficulty is that we do not know  $U_i$ , and so we cannot directly estimate the (heterogeneous) treatment assignment probabilities  $\mathbb{P}[W_i = 1 | U_i]$ . However, as shown below, we can get around this difficulty by

<sup>14</sup>The mean-parametrization condition  $\mathbb{E}[Z_i | U_i = u] = u$  is not strictly necessary, but we use it throughout for interpretability.



solving an integral equation that lets us balance out confounding due to the latent variable  $U_i$ .

Our main result is that, given our 3 assumptions, we can use noise in  $Z_i$  to identify weighted average of causal effects over principal strata (6). To this end, consider any averaging treatment effect estimator of the form

$$\tilde{\tau}_\gamma = \frac{1}{n} \left( \sum_{Z_i \geq c} \gamma_+(Z_i) Y_i - \sum_{Z_i < c} \gamma_-(Z_i) Y_i \right), \quad (7)$$

with a weighting function satisfying  $\mathbb{E}[\gamma_+(Z_i); Z_i \geq c] = 1$  and  $\mathbb{E}[\gamma_-(Z_i); Z_i < c] = 1$ ; we will later also consider ratio-form estimators that do not require this moment assumption on the weights. Our first result characterizes the expectation of this estimator under our sampling model.

**Theorem 1.** *Under Assumptions 1, 2 and 3, the estimator (7) has expectation*

$$\mathbb{E}[\tilde{\tau}_\gamma] = \underbrace{\int h_+(u) \tau(u) dG(u)}_{\text{weighted treatment effect}} - \underbrace{\int (h_+(u) - h_-(u)) \alpha_{(0)}(u) dG(u)}_{\text{confounding bias}}, \quad (8)$$

where  $h_-(u)$  and  $h_+(u)$  are given by

$$h_+(u) = \int_{[c, \infty)} \gamma_+(z) p(z | u) d\lambda(z), \quad h_-(u) = \int_{(-\infty, c)} \gamma_-(z) p(z | u) d\lambda(z). \quad (9)$$

*Proof:* Conditioning on the latent variable  $U_i$ , we find that

$$\begin{aligned} \mathbb{E}[\gamma_+(Z_i) \cdot Y_i \cdot 1(\{Z_i \geq c\}) | U_i] &\stackrel{(i)}{=} \mathbb{E}[\gamma_+(Z_i) \cdot Y_i(1) \cdot 1(\{Z_i \geq c\}) | U_i] \\ &\stackrel{(ii)}{=} \underbrace{\mathbb{E}[Y_i(1) | U_i]}_{\alpha_{(1)}(U_i)} \underbrace{\mathbb{E}[\gamma_+(Z_i) 1(\{Z_i \geq c\}) | U_i]}_{h_+(U_i) = \int_{[c, \infty)} \gamma_+(z) p(z | U_i) d\lambda(z)} \end{aligned}$$

In (i) we used the fact that  $Y_i = Y_i(1)$  for  $Z_i \geq c$  by Assumption 1 and in (ii) we used exogeneity of the noise (Assumption 3). Finally, the expression for  $\mathbb{E}[\gamma_+(Z_i) 1(\{Z_i \geq c\}) | U_i]$  follows from Assumption 2. Similarly, we find that

$$\mathbb{E}[\gamma_-(Z_i) Y_i 1(\{Z_i < c\}) | U_i] = \underbrace{\mathbb{E}[Y_i(0) | U_i]}_{\alpha_{(0)}(U_i)} \underbrace{\mathbb{E}[\gamma_-(Z_i) 1(\{Z_i < c\}) | U_i]}_{h_-(U_i) = \int_{(-\infty, c)} \gamma_-(z) p(z | U_i) d\lambda(z)}$$

Thus, unconditionally,

$$\mathbb{E}[\tilde{\tau}_\gamma] = \int h_+(u) \alpha_{(1)}(u) dG(u) - \int h_-(u) \alpha_{(0)}(u) dG(u).$$

We conclude with (8) by noting that  $\tau(u) = \alpha_{(1)}(u) - \alpha_{(0)}(u)$  and rearranging.  $\square$

We emphasize that Theorem 1 made no use of any properties of the  $Z$ -conditional response function  $\mu_{(w)}(z)$  other than those induced by random noise in the running variable  $Z_i$ . We can apply this result to estimate several different causal parameters of interest.

*Remark 2.* The assumption that we know the noise distribution  $p(z|u)$  exactly may appear restrictive in some applications. We note, however, that all our results remain valid if we work with a noise distribution  $\hat{p}(z|u)$  that under-estimates the true noise level, in the sense that  $p(z|u) = \int \hat{p}(z|u')\lambda(u'|u) du'$  for some distribution function  $\lambda(u'|u)$  that captures the noise left out by  $\hat{p}(\cdot)$ .<sup>15</sup> For example, if the true noise process involves heteroskedastic Gaussian measurement errors  $Z_i|U_i \sim \mathcal{N}(U_i, \nu_i^2)$ , where  $U_i$  and  $\nu_i$  may be correlated, then our approach would remain valid if we posit a homoskedastic noise model  $Z_i|U_i \sim \mathcal{N}(U_i, \hat{\nu}^2)$  so long as  $\nu_i \geq \hat{\nu}$  almost surely. This fact is helpful when choosing which noise model to use in practice: For example, again with Gaussian errors, one can estimate the noise scale  $\hat{\nu}^2$  by considering a conservative lower bound on measurement accuracy obtained via repeated measurement, e.g., in education via repeatedly administering similar tests or in medicine by repeatedly administering the same diagnostic.

## 2.1 Constant treatment effects

A first interesting special case of the setting considered above is where the stratum-wise conditional treatment effect function is taken to be constant  $\tau(u) = \tau$ , and our goal is to estimate this constant treatment effect. In this case, Theorem 1 simplifies as follows, implying that when we estimate a constant treatment effect, we can guarantee low bias by using a weighted estimator of the form (19) whose induced  $h$ -function from (9) makes  $\sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\}$  small.

**Corollary 2.** *In the setting of Theorem 1, suppose furthermore that  $\tau(u) = \tau$  for all  $u$ , and that there are constants  $M$  and  $\alpha_{(0)}$  for which  $|\alpha_{(0)}(u) - \alpha_{(0)}| \leq M$ .<sup>16</sup> Then, (8) from Theorem 1 implies*

$$|\mathbb{E}[\tilde{\tau}_\gamma] - \tau| \leq M \sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\}. \quad (10)$$

*Proof.* The stated result follows directly from (8) by noting that

$$\int h_+(u)dG(u) = \int_{[c,\infty)} \gamma_+(z)dF(z) = 1, \quad \int h_-(u)dG(u) = \int_{(-\infty,c)} \gamma_-(z)dF(z) = 1.$$

Thus, the constant treatment effect  $\tau$  is captured exactly and the contribution of  $\alpha_{(0)}$  gets canceled out, and so the only remaining leading-order bias term is

$$\int (h_+(u) - h_-(u)) (\alpha_{(0)}(u) - \alpha_{(0)}) dG(u) \leq M \sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\},$$

and we recover the bound (10).

*Remark 3.* Even when the stratum-wise treatment effect function  $\tau(u)$  is not constant, using an estimator of the above type that makes  $\sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\}$  small may be desirable. By direct analogy to (10) we immediately see that

$$|\mathbb{E}[\tilde{\tau}_\gamma] - \tau_{h,+}| \leq M \sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\}, \quad \tau_{h,+} = \int h_+(u)\tau(u) dG(u), \quad (11)$$

<sup>15</sup>To check this fact, note that we can generate  $Z_i|U_i$  by first drawing  $U'_i|U_i$  with distribution  $\lambda(u'|u)$ , and then drawing  $Z_i|U'_i$  with distribution  $\hat{p}(z|u')$ . Our analysis then goes through with  $U_i$  replaced by  $U'_i$  (provided Assumption 3 still holds with  $U'_i$ ). In general, under estimating the measurement error will result in a loss of power (since it reduces the number of units that may plausibly both get treated or not treated depending on the realization of  $Z_i$ ), but does not cause any conceptual problems (since our results will hold regardless of the distribution of  $U_i$ , and in particular also hold for latent states distributed as  $U'_i$ ).

<sup>16</sup>The result is identical if instead there exists  $\alpha_{(1)}$  such that  $|\alpha_{(1)}(u) - \alpha_{(1)}| \leq M$ .

meaning that  $\hat{\tau}_\gamma$  always approximates some weighted treatment effect functional—and this may be of interest if we are not directly interested in treatment heterogeneity [Crump, Hotz, Imbens, and Mitnik, 2009, Li, Morgan, and Zaslavsky, 2017, Imbens and Wager, 2019, Kallus, 2020].

*Remark 4.* Formally, Corollary 2 is a partial identification result, as (10) only provides an upper bound on the bias of our estimator. In general, we will consider sequences of weight functions in (7) whose induced  $h_\pm(u)$  functions make  $\sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\}$ , and thus the bias bound in (10), progressively smaller. As discussed further in Section 3, the choice of weighting functions is governed by a bias-variance tradeoff, whereby reducing the bias bound in (10) entails increasing the variance of the estimator (7). In some settings, e.g., when  $Z_i|U_i$  has a Gaussian distribution, we will be able to push the imbalance term to zero, and so Corollary 2 can also be used for point identification of  $\tau$ . However, in other settings, e.g., when  $Z_i|U_i$  has a binomial distribution, it is not possible to get zero bias via (10) and so we can only achieve partial identification—even asymptotically. For a further discussion of point versus partial identification in regression discontinuity designs, see Section II.A of Imbens and Wager [2019].

## 2.2 Targeted treatment effects

Theorem 1 demonstrates that noisy running variables enable identification of the weighted treatment effects  $\tau_{h,+}$  (11). Depending on the application, an analyst may want to target a weighted treatment effect  $\tau_w$  of their choice,

$$\tau_w = \int w(u)\tau(u) dG(u), \quad w(u) \geq 0, \quad \int w(u)dG(u) = 1 \quad (12)$$

This target may be identified as shown in Corollary 3 below. The remainder of this section provides examples of statistical targets that may be expressed as in (12).

**Corollary 3.** *In the setting of Theorem 1, suppose furthermore that there are constants  $M, M', \alpha_{(0)}$  and  $\tau$  for which  $|\alpha_{(0)}(u) - \alpha_{(0)}| \leq M$  and  $|\tau(u) - \tau| \leq M'$ . Then,*

$$\begin{aligned} |\mathbb{E}[\tilde{\tau}_\gamma] - \tau_w| &\leq M \sup\{|h_+(u) - h_-(u)| : u \in \mathbb{R}\} \\ &\quad + M' \sup\{|h_+(u) - w(u)| : u \in \mathbb{R}\}, \end{aligned} \quad (13)$$

where  $\tau_w$  and  $w(\cdot)$  are defined in (12).

*Proof.* This result follows by the same argument as used for Corollary 2 along with the decomposition

$$\int \tau(u)w(u)dG(u) = \int \tau(u)h_+(u)dG(u) + \int (\tau(u) - \tau)(w(u) - h_+(u)) dG(u)$$

**Regression discontinuity parameter** One statistical target that may be of interest is the standard regression discontinuity parameter  $\tau_c$  as defined in (1). Interest in this parameter may not arise directly from first principles; however, it has traditionally been a key focus of the continuity-based inference literature, and obtaining estimates of this quantity that rely only on implicit randomization via noise in  $Z_i$  may be helpful in comparing our approach to traditional approaches. To write  $\tau_c$  as in (12), note that by Bayes' rule,

$$\tau_c = \mathbb{E}[Y_i(1) - Y_i(0) | Z_i = c] = \mathbb{E}[\tau(U_i) | Z_i = c] = \int \tau(u)p(c|u)dG(u)/f(c) \quad (14)$$

Recall here that  $f(c)$  is the density of the running variable  $Z_i$  at the cutoff  $c$ , defined in Assumption 2. Thus, the representation from (12) holds with  $w(u) = p(c|u)/f(c)$ . The result from Corollary 3 thus implies that we can estimate  $\tau_c$  using weighted estimators that make the right-hand side bound of (13) small.

Another closely related target is  $\tau_{c'}$  as defined in (14), but for some other value  $c' \neq c$  of the running variable. Formally, this approach again fits withing our setting, with  $w(u) = p(c'|u)/f(c')$ . Conceptually, estimating  $\tau_{c'}$  away from  $c$  involves extrapolating treatment effects away from cutoff [Angrist and Rokkanen, 2015, Rokkanen, 2015]. Estimating  $\tau_{c'}$  away from the cutoff is also possible using continuity-based approaches, though the inference can quickly become uninformative.

**Changing the cutoff** As argued in Heckman and Vytlacil [2005], in many settings we may be most interested in evaluating the effect of a policy intervention. One simple case of a policy intervention involves changing the eligibility threshold, i.e., that standard practice involves prescribing treatment to subjects whose running variable crosses  $c$ , but we are now considering changing this cutoff to a new value  $c' < c$ .<sup>17</sup> For example, in a medical setting, we may consider lowering the severity threshold at which we intervene on a patient. In this case, we need to estimate the average treatment effect of patients affected by the treatment which, in this case, amounts to  $\tau_\pi = \mathbb{E}[Y_i(1) - Y_i(0) | c' \leq Z_i < c]$ . The following result shows how to identify this quantity via a weighted estimator. Corollary 3 is applicable by noting that effect of this policy change is

$$\tau_\pi = \mathbb{E}[Y_i(1) - Y_i(0) | c' \leq Z_i < c] = \int_{[c',c)} \int \tau(u)p(z|u)dG(u)d\lambda(z) / \int_{[c',c)} dF(z), \quad (15)$$

and so by Fubini's theorem,  $\tau_\pi$  can be written in the form (12) with weight function  $w(u) = \int_{[c',c)} p(z|u)d\lambda(z) / \int_{[c',c)} dF(z)$ .

**Reducing measurement error** Another policy intervention of potential interest could involve switching to a more (or less) accurate device for measuring  $Z_i$ , thus changing the noise level  $\nu$  in the running variable. For example, one could imagine that a policy maker has the option to reduce measurement error by using a new (potentially more expensive) measurement device, and wants to know whether improved outcomes from more reproducible targeting are worth the cost. Specifically, suppose that we currently assign treatment as  $W_i = 1(\{Z_i \geq c\})$  for  $Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$ , and are considering a switch to a new treatment rule  $W'_i = 1(\{Z'_i \geq c\})$  based on a measurement  $Z'_i | U_i \sim \mathcal{N}(U_i, \nu'^2)$  with a different noise level  $\nu'$ . Writing  $\Phi_\nu(\cdot)$  for the standard normal cumulative distribution function with variance  $\nu^2$ , we see that the effect of this policy change is

$$\tau_\pi = \mathbb{E}[(Y_i(1) - Y_i(0))(W'_i - W_i)] = \int \tau(u)(\Phi_{\nu'}(c - u) - \Phi_\nu(c - u))dG(u), \quad (16)$$

which again is covered by (12) and Corollary 3.

### 2.3 Noise-induced or continuity-based identification?

As mentioned in the introduction, the traditional approach to inference in regression discontinuity designs relies on smoothness assumptions for the conditional response functions

<sup>17</sup>We also note that the hypothetical experiment that Thistlethwaite and Campbell [1960] offer as analogous to a regression discontinuity is equivalent to randomizing some units to a different threshold  $c'$ .

$\mu_{(w)}(z)$ . Now, one can verify that noise in the running variable as in Assumption 2 implies smoothness properties on the conditional response function: Under Assumptions 1–3,

$$\mu_{(w)}(z) = \mathbb{E} [Y_i(w) | Z_i = z] = \int \alpha_{(w)}(u)p(z | u) dG(u) / \int p(z | u) dG(u), \quad (17)$$

so if  $\alpha_{(w)}(u)$  is bounded and  $z \mapsto p(z | u)$  is continuous, then by the dominated convergence theorem we can show that  $\mu_{(w)}(z)$  is also continuous (see also Proposition 2 of Lee [2008]).

Given this observation, it is natural to ask whether we can usefully exploit smoothness induced by measurement error to drive inference using classical continuity-based methods like local linear regression. Recall that many continuity-based methods, including Armstrong and Kolesár [2020], Imbens and Kalyanaraman [2012], Imbens and Wager [2019] and Kolesár and Rothe [2018], rely on  $\mu_{(w)}(z)$  having a bounded second derivative to drive inference. Thus, to build a formal connection between our setting and this line of work, we need to derive upper bounds on  $|\mu_{(w)}''(z)|$  that are justified by (17).

The following result provides such bounds in the case of Gaussian measurement error, i.e., with  $Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$ . The lower bound for  $\mu_{(w)}'(z)$  below is obtained by considering a distribution  $G(u)$  with two point masses symmetrically positioned around the cutoff, while the lower bound for  $\mu_{(w)}''(z)$  adds a third point mass to  $G(u)$  at the cutoff. Meanwhile, the upper bounds below build on a lemma of Jiang and Zhang [2009].

**Proposition 4.** *Suppose that Assumptions 1–3 hold with noise model  $Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$ , and that  $|\alpha_{(w)}(u)|$  is uniformly bounded. Then,  $\mu_{(w)}(z)$  is infinitely differentiable. Furthermore, for any point  $z \in \mathbb{R}$ , given a lower bound  $\rho$  on  $f(z) = \int \varphi_\nu(z - u)dG(u)$ , the density of  $Z_i$  at  $z$ , and a uniform upper bound  $M$  on the variation of  $\alpha_{(w)}(u)$ , the worst-case first and second derivatives of  $\mu_{(w)}(z)$  at  $z$  can be bounded as follows:*

$$\begin{aligned} & \frac{M}{\nu} \cdot \sqrt{-\log(2\pi\nu^2\rho^2)} \\ & \leq \sup \left\{ \left| \frac{d\mu_{(w)}(z)}{dz} \right| : f(z) \geq \rho, |\alpha_{(w)}(u) - \alpha_{(w)}| \leq M \text{ for all } u \in \mathbb{R} \right\} \\ & \leq \frac{5M}{\nu} \cdot \sqrt{-\log(2\pi\nu^2\rho^2/25)} \\ & \frac{M}{5\nu^2} \cdot (-\log(2\pi\nu^2\rho^2)) \\ & \leq \sup \left\{ \left| \frac{d^2\mu_{(w)}(z)}{dz^2} \right| : f(z) \geq \rho, |\alpha_{(w)}(u) - \alpha_{(w)}| \leq M \text{ for all } u \in \mathbb{R} \right\} \\ & \leq \frac{13M}{\nu^2} \cdot (-\log(2\pi\nu^2\rho^2/25)), \end{aligned} \quad (18)$$

where  $\varphi_\nu(\cdot)$  is the standard Gaussian probability density function with variance  $\nu^2$ .

In other words, given a Gaussian noise model and a lower bound  $\rho$  on the density of the running variable, one can use (18) to obtain quantitative upper bounds on the second derivative of  $\mu_{(w)}(z)$ <sup>18</sup> that can then be used in conjunction with, e.g., the estimators of Imbens and Wager [2019] and Armstrong and Kolesár [2020] that provide uniform inference for the regression discontinuity parameter given a curvature bound on the response function.

<sup>18</sup>Instead of (18), we use the sharper (yet more complicated) lower and upper bounds that are derived in the proof of Proposition 4.

This result may be of conceptual interest, as adaptively discovering the curvature of  $\mu_{(w)}(z)$  is not possible in general [Armstrong and Kolesár, 2018].

In practice, however, bounds based on (18) appear to be quite wide. For example, revisiting our example from Section 1.1, one could adapt the approach of Armstrong and Kolesár [2020] to build local linear based confidence intervals for the regression discontinuity parameter that only rely on the curvature upper bound provided by (18); however, doing so would result in a 95% confidence interval of  $(-0.21, 0.19)$ , which is not particularly useful here.<sup>19</sup> Furthermore, in our simulation experiments, we again find such intervals to be considerably wider than our proposed ones that directly exploit the noise model on the running variable. Thus, although measurement error does imply some smoothness in the running variable  $\mu_{(w)}(z)$ , this connection does not reduce the problem of accurate regression discontinuity inference with measurement error to one of accurate continuity-based inference.

### 3 From Identification to Inference

In the previous section, we discussed how weighted estimators of the form (7) can identify causal effects in regression discontinuity designs using only variation in the running variable. In order to make use of such an estimator in practice, however, it's not enough to just bound its bias; we also need to understand its sampling distribution. In this section, instead of (7), we study ratio estimators of the form

$$\hat{\tau}_\gamma = \hat{\mu}_{\gamma,+} - \hat{\mu}_{\gamma,-}, \quad \hat{\mu}_{\gamma,+} = \frac{\sum_{Z_i \geq c} \gamma_+(Z_i) Y_i}{\sum_{Z_i \geq c} \gamma_+(Z_i)}, \quad \hat{\mu}_{\gamma,-} = \frac{\sum_{Z_i < c} \gamma_-(Z_i) Y_i}{\sum_{Z_i < c} \gamma_-(Z_i)}. \quad (19)$$

The self-normalization accounts for the fact that the constraints,  $\mathbb{E}[\gamma_+(Z_i); Z_i \geq c] = 1$  and  $\mathbb{E}[\gamma_-(Z_i); Z_i < c] = 1$ , which we assumed to hold in Section 2, cannot be enforced exactly unless the distribution of the running variable  $Z_i$  were known a-priori. The ratio estimator is more stable in finite samples and invariant to translations of the response  $Y_i$ .

We first provide a general central limit theorem for (19), then we show how our identifying assumptions enable the construction of asymptotic confidence intervals for various treatment effect parameters. In Section 4, we propose a concrete approach to constructing weighting functions  $\gamma$  via quadratic programming that appropriately trades off the bias and variance of the resulting estimator.

We start by studying the asymptotic distribution of the weighted ratio estimator (19). We treat the weighting kernels  $\gamma_+, \gamma_-$  as deterministic but allow them to vary with  $n$ , i.e.,  $\gamma_+ = \gamma_+^{(n)}$  and  $\gamma_- = \gamma_-^{(n)}$ . Our results allow considerable flexibility in choosing  $\gamma_+, \gamma_-$ . In Section 4 we provide concrete guidance for constructing  $\gamma_+, \gamma_-$ ; however, the abstract limiting results given here would hold for other choices of weighting functions also (for example, they would hold for weights derived from local linear regression as in (2) with decreasing bandwidth).

Our first formal result is the following central limit theorem. We note that the conditions on the response noise are mild and similar to commonly made assumptions. The assumption

<sup>19</sup>For completeness, we note that an application of the method of Armstrong and Kolesár [2020] to this problem with curvature controlled by the *lower* bound from (18) would result in a 95% confidence of  $(-0.06, 0.20)$ , which is still larger than ours. This is of course not a rigorous confidence interval (since we need to use an upper bound on the curvature to justify inference); however, what this argument shows is that, even though the upper bounds in (18) may be somewhat loose, it is impossible to sharpen Proposition 4 so much as to make local linear regression with design-based curvature bounds competitive with our approach.

on regular weighting kernels is also easy to satisfy, and in particular the weights proposed in Section 4 will satisfy this property.

**Theorem 5** (Asymptotic normality of Ratio estimators). *Suppose that the pairs  $(Z_i, Y_i)$  are independent and identically distributed, and that:*

1. *The sequences of weighting kernels  $\gamma_+^{(n)}$  and  $\gamma_-^{(n)}$  are deterministic,<sup>20</sup> and there exist  $\beta \in (0, 1/2)$ ,  $C > 0$  such that*

$$\mathbb{P} \left[ 0 < \max_z \left| \gamma_\diamond^{(n)}(z) \right| \leq n^\beta \cdot C \cdot \mathbb{E} \left[ \gamma_\diamond^{(n)}(Z_i) \right], \diamond \in \{+, -\} \right] \rightarrow 1 \text{ as } n \rightarrow \infty.$$

2. *There exist  $q > 0$  and  $\bar{\sigma}, \underline{\sigma} > 0$  such that the response noise satisfies  $(\diamond \in \{+, -\})$*

$$\text{Var} [Y_i | Z_i = z] \geq \underline{\sigma}^2, \quad \mathbb{E} \left[ |Y_i - \mu_{\gamma, \diamond}|^{2+q} | Z_i = z \right] \leq \bar{\sigma}^{2+q} \text{ for all } z \text{ with } \gamma_\diamond(z) \neq 0.$$

Then,  $\hat{\tau}_\gamma$  is asymptotically normal, i.e.,

$$\sqrt{n} (\hat{\tau}_\gamma - \tau_\gamma) / \sqrt{V_\gamma} \Rightarrow \mathcal{N}(0, 1), \quad (20)$$

where

$$\tau_\gamma = \mu_{\gamma,+} - \mu_{\gamma,-}, \quad \mu_{\gamma,+} = \frac{\mathbb{E} [\gamma_+(Z_i) Y_i; Z_i \geq c]}{\mathbb{E} [\gamma_+(Z_i); Z_i \geq c]}, \quad \mu_{\gamma,-} = \frac{\mathbb{E} [\gamma_-(Z_i) Y_i; Z_i < c]}{\mathbb{E} [\gamma_-(Z_i); Z_i < c]} \quad (21)$$

$$V_\gamma = \frac{\mathbb{E} [\gamma_+^2(Z_i) (Y_i - \mu_{\gamma,+})^2; Z_i \geq c]}{\mathbb{E} [\gamma_+(Z_i); Z_i \geq c]^2} + \frac{\mathbb{E} [\gamma_-^2(Z_i) (Y_i - \mu_{\gamma,-})^2; Z_i < c]}{\mathbb{E} [\gamma_-(Z_i); Z_i < c]^2}. \quad (22)$$

### 3.1 Feasible confidence intervals

Given our result from Theorem 5, we can design confidence intervals for various targets “ $\tau$ ” discussed in Section 2, including a constant treatment parameter  $\tau$  as in Corollary 2 or  $\tau := \tau_w$  from (12). In doing so, we need to account for both the variance term  $V_\gamma$  as in (22), and potential bias  $|b_\gamma| = |\tau_\gamma - \tau|$ . Here, we will not assume that the bias is negligible (i.e., we do not assume “undersmoothing”). Rather, we will derive an upper bound  $\hat{B}_\gamma$  for the bias  $|b_\gamma|$ , and then build confidence intervals for  $\tau$  that are robust to estimation bias up to  $\hat{B}_\gamma$ : Following Imbens and Manski [2004] and Imbens and Wager [2019], we set

$$\tau \in \hat{\tau}_\gamma \pm \ell_\alpha, \quad \ell_\alpha = \min \left\{ \ell : \mathbb{P} \left[ |b + n^{-1/2} \hat{V}_\gamma^{1/2} \tilde{Z}| \leq \ell \right] \geq 1 - \alpha \text{ for all } |b| \leq \hat{B}_\gamma \right\}, \quad (23)$$

where  $\tilde{Z}$  is a standard Gaussian random variable,  $\alpha$  is the significance level, and  $\hat{V}_\gamma$  is an estimate of the sampling variance  $V_\gamma$ .

**Corollary 6** (Valid confidence intervals). *Assume the conditions from (5) are satisfied. Furthermore, let  $b_\gamma = \tau_\gamma - \tau$  be the (asymptotic) bias for estimating  $\tau$  and let  $\hat{B}_\gamma$  be a (potentially data-driven) upper bound on  $|b_\gamma|$  and  $\hat{V}_\gamma$  an estimate of  $V_\gamma$  such that*

$$\sqrt{n} \left( \hat{B}_\gamma - |b_\gamma| \right) / \sqrt{V_\gamma} \geq 0 + o_p(1), \quad \frac{\hat{V}_\gamma}{V_\gamma} = 1 + o_p(1). \quad (24)$$

Then, the confidence intervals from (23) satisfy  $\liminf_{n \rightarrow \infty} \mathbb{P} [\tau \in \hat{\tau}_\gamma \pm \ell_\alpha] \geq 1 - \alpha$ .

<sup>20</sup>It suffices for  $\gamma_+, \gamma_-$  to be independent of  $(U_i, Z_i, Y_i(0), Y_i(1))$ ,  $1 \leq i \leq n$ .



In order to make use of this result, it remains to design data-driven choices for  $\widehat{B}_\gamma$  and  $\widehat{V}_\gamma$ . We start with the latter, which admits a simple plug-in estimator.

**Proposition 7.** *Under the assumptions of Theorem 5,  $V_\gamma$  can be consistently estimated with the following plug-in estimator:  $\widehat{V}_\gamma / V_\gamma = 1 + o_P(1)$  for*

$$\widehat{V}_\gamma = \frac{\frac{1}{n} \sum_{i:Z_i \geq c} \gamma_+(Z_i)^2 (Y_i - \hat{\mu}_{\gamma,+})^2}{\left(\frac{1}{n} \sum_{i:Z_i \geq c} \gamma_+(Z_i)\right)^2} + \frac{\frac{1}{n} \sum_{i:Z_i < c} \gamma_-(Z_i)^2 (Y_i - \hat{\mu}_{\gamma,-})^2}{\left(\frac{1}{n} \sum_{i:Z_i < c} \gamma_-(Z_i)\right)^2}, \quad (25)$$

where  $\hat{\mu}_{\gamma,+}, \hat{\mu}_{\gamma,-}$  are defined in (19).

We next turn to bounding bias. This task is more involved, and our proposed solution is built around fractional linear programming. Here, we consider bounding the bias for estimators of a constant treatment effect  $\tau$  as in Corollary 2,<sup>21</sup> and defer a discussion of bias for weighted targets  $\tau_w$  as in Corollary 3 to Section 4.2. To this end, we start by stating a variant of Corollary 3 that holds for the ratio-form estimator (19).

**Corollary 8.** *Under the conditions of Corollary 2, suppose furthermore that  $|\alpha_{(0)}(u) - \mu_{\gamma,-}| \leq M$ . Then the limit  $\tau_\gamma$  as defined in (21) satisfies*

$$|\tau_\gamma - \tau| \leq \frac{M \mathbb{E}[|h_+(U_i) - h_-(U_i)|]}{\mathbb{E}[h_+(U_i)]} = \frac{\int M|h_+(u) - h_-(u)|dG(u)}{\int h_+(u)dG(u)}. \quad (26)$$

A challenge in using this result is that we do not know the expectations precisely since they involve integrals over the latent variable  $U_i$ . To get around this issue, we instead seek to bound the worst-case bias over any data-generating distribution that appears consistent with the observed data. Specifically, let  $\mathcal{G}_n$  be the class of latent variable distributions that lead to marginal distributions that lie within the Dvoretzky—Kiefer—Wolfowitz band of the empirical measure  $\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(Z_i \leq t)$ , i.e.,<sup>22</sup>

$$\mathcal{G}_n = \left\{ G \text{ distrib. : } \sup_{t \in \mathbb{R}} \left| \int_{(-\infty, t]} p(z | u) dG(u) d\lambda(z) - \widehat{F}_n(t) \right| \leq \sqrt{\frac{\log(2/\alpha_n)}{2n}} \right\}, \quad (27)$$

where  $\alpha_n = \min\{0.05, n^{-\frac{1}{4}}\}$ . Then we solve the following optimization problem to get  $\widehat{B}_\gamma$ :

$$\widehat{B}_\gamma = \sup_{G \in \mathcal{G}_n} \frac{\int M|h_+(u) - h_-(u)|dG(u)}{\int h_+(u)dG(u)}. \quad (28)$$

The above optimization problem is tractable: Both the denominator and numerator of the objective are linear in  $G$  and furthermore  $\mathcal{G}_n$  is a convex class of densities that may be represented by linear inequality constraints. Thus (28) is a fractional program that may be solved through linear programming via the Charnes and Cooper [1962] transformation.

**Proposition 9.** *Under the conditions of Corollary 8,  $\widehat{B}_\gamma$  from (28) satisfies  $\mathbb{P}[|b_\gamma| \leq \widehat{B}_\gamma] \rightarrow 1$  almost surely as  $n \rightarrow \infty$ , and the condition (24) from Corollary 6 is satisfied.*

<sup>21</sup>Following Remark 3, all the following discussion about inference for  $\tau$  also goes through as a discussion for inference about the weighted target  $\tau_{h,+}$  under potential heterogeneity.

<sup>22</sup>More generally, any class of distributions  $\mathcal{G}_n$  such that  $\mathbb{P}[G \in \mathcal{G}_n] \rightarrow 1$  could be used instead.

Given any choice of weighting functions  $\gamma$  for (19), Propositions 7 and 9 provide a complete recipe for building valid confidence intervals via Corollary 6 (at least, so far, in the case of constant treatment effects). As discussed above, at this point, one could already take weighting functions implied by various regression discontinuity estimators (e.g., local linear regression), and use these results to build valid confidence intervals for  $\tau$  that are directly justified by noise-induced randomization. Existing weighting functions  $\gamma$ , however, were not designed for this purpose, and so may not yield particularly short confidence intervals. In the following section, we discuss how to design estimators  $\hat{\tau}_\gamma$  of the type (19) with an eye towards making confidence intervals obtained via Corollary 6 short.

*Remark 5.* The confidence intervals (23) are not pointwise exact, i.e., there may exist data-generating distributions for which  $\liminf_{n \rightarrow \infty} \mathbb{P}[\tau \in \hat{\tau}_\gamma \pm \ell_\alpha] > 1 - \alpha$ . However, if our upper bound on the bias is sharp, i.e.,  $\lim_{n \rightarrow \infty} \sqrt{n}(\hat{B}_\gamma - |b_\gamma|) / \sqrt{V_\gamma} = 0$  in probability, then our intervals are exact in a minimax sense, i.e., there exists some data-generating distribution for which  $\liminf_{n \rightarrow \infty} \mathbb{P}[\tau \in \hat{\tau}_\gamma \pm \ell_\alpha] = 1 - \alpha$ . Here, our goal is to provide practical and valid confidence intervals for  $\tau$ , and a discussion of optimal inference is beyond the scope of this paper. We note, however, that in many statistical applications confidence intervals of the type (23), i.e., ones that seek robustness to worst-case bias, have strong optimality properties: Intervals of this type are effectively the shortest possible intervals that achieve coverage uniformly over a class of data-generating distributions [Armstrong and Kolesár, 2018, Donoho, 1994]. Whether a phenomenon of this type also plays out here would be an interesting topic for further investigation.

## 4 Designing estimators via quadratic programming

We now turn to the problem of deriving weighting functions  $\gamma_+$ ,  $\gamma_-$  that make the estimator discussed in the previous section perform well. We proceed according to the following roadmap. If we knew the distribution  $F(\cdot)$  of the running variable exactly, then we could use quadratic programming to derive weights that control the error of the unnormalized estimator (7), as captured by its variance and the worst-case bias bounds given in Corollaries 2 and 3. In practice, of course, we do not know  $F(\cdot)$ , so what we do is we obtain a “guess”  $\bar{F}(\cdot)$  for  $F(\cdot)$ ,<sup>23</sup> and then solve for the optimal unnormalized weighted estimator given this guess  $\bar{F}(\cdot)$ .

We emphasize that this approach is heuristic, and may not recover the optimal weights, i.e., weights that would make the confidence intervals from Corollary 6 as short as possible. This is because our weights are optimized for the unnormalized estimator of the form (7) rather than the ratio-form estimator (19) we use in practice,<sup>24</sup> and because we rely on potentially inaccurate guesses  $\bar{F}(\cdot)$  for  $F(\cdot)$ . However, as evidenced by our numerical experiments, this heuristic appears to yield weighting functions  $\gamma_+$ ,  $\gamma_-$  that yield powerful inference in practice; and, as discussed above, we leave a topic of minimax-optimal inference under noise-induced randomization to further work.

<sup>23</sup>This  $\bar{F}(\cdot)$  may represent an actual guess, or may be a pilot estimator derived from either held out or unsupervised data (i.e., data for which we do not observe the response  $Y_i$ ). In general, using a poor choice of  $\bar{F}(\cdot)$  would not impact the validity of our inference from Corollary 6, but could result in less accurate point estimates and longer confidence intervals. We discuss possible constructions for  $\bar{F}(\cdot)$  further in Section 4.3.

<sup>24</sup>Recall that the construction of the unnormalized estimator (7) requires enforcing constraints of the type  $\mathbb{E}[\gamma_+(Z_i); Z_i \geq c] = \int \gamma_+(z)\mathbf{1}(z \geq c)dF(z) = 1$  on the weights, and this is only possible if we know  $F(\cdot)$ . When  $F(\cdot)$  is not known a-priori, we need to use the ratio-form estimator (19) instead.

## 4.1 Targeting constant treatment effects

We start our discussion in the same setting as in Section 3.1, i.e., where we target constant treatment effect parameters. We consider weighted treatment effects in Section 4.2. To this end, we start by assuming that we have access to a constant  $\sigma^2$  such that  $\text{Var}[Y_i | Z_i = z] \leq \sigma^2$  for all  $z$ , which gives us the following bound for the variance of the estimator (7):

$$\text{Var}[\tilde{\tau}_\gamma] \leq \sigma^2 \left( \int_{(-\infty, c)} \gamma_-^2(z) dF(z) + \int_{[c, \infty)} \gamma_+^2(z) dF(z) \right). \quad (29)$$

We also assume that we have a guess  $\bar{F}(\cdot)$  for  $F(\cdot)$ . Given these ingredients, we choose  $\gamma_\pm(\cdot)$  by solving the following optimization problem:

$$\gamma_\pm(\cdot) \in \text{argmin} \left\{ \frac{\sigma^2}{n} \left( \int_{(-\infty, c)} \gamma_-^2(z) d\bar{F}(z) dz + \int_{[c, \infty)} \gamma_+^2(z) d\bar{F}(z) dz \right) + t^2 \right\} \quad (30)$$

$$\text{s.t. } M \cdot \left| \int_{(-\infty, c)} \gamma_-(z) p(z | u) d\lambda(z) - \int_{[c, \infty)} \gamma_+(z) p(z | u) d\lambda(z) \right| \leq t \text{ for all } u \quad (31)$$

$$\int_{(-\infty, c)} \gamma_-(z) d\bar{F}(z) = 1, \quad \int_{[c, \infty)} \gamma_+(z) d\bar{F}(z) = 1 \quad (32)$$

$$|\gamma_-(z)|, |\gamma_+(z)| \leq C \cdot n^\beta \quad (33)$$

$$\gamma_-(z) = \gamma_+(z) = 0 \text{ for } z \notin [\ell, u]. \quad (34)$$

Here (31) is the bound for the bias of the estimator given in Corollary 2, while the objective (30) is the sum of the variance bound (29) and this worst case bias. Thus, the above optimization problem is trying to minimize a bound on the worst-case mean-squared error of the estimator (7), but with  $F(\cdot)$  replaced with the guess  $\bar{F}(\cdot)$  in the optimization problem. Finally, (32) enforces the moment constraint required by weights used in (7), (33) ensures that no single observation is given excessive influence, and (34) forces the weights  $\gamma_\pm$  to be zero outside of  $[\ell, u]$  for numerical stability reasons. The above optimization problem is a quadratic program, and an appropriately discretized version of it can be solved using standard convex optimization software.

The following results shows that the resulting weights satisfy the conditions of Theorem 5 and thus enable valid inference.

**Proposition 10** (Sufficient condition for regular weighting kernels). *Assume we derive  $\gamma_\pm$  by solving optimization problem (30) with potentially random choices for tuning parameters, e.g.,  $\bar{F}(\cdot)$ . Furthermore, assume the expectation of  $\gamma_+, \gamma_-$  is lower bounded by a strictly positive number, i.e., there exists  $\delta > 0$  such that*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left[ \int_{(-\infty, c)} \gamma_-(z) dF(z), \int_{[c, \infty)} \gamma_+(z) dF(z) \geq \delta \right] = 1. \quad (35)$$

*Then the weights derived from optimization problem (30) are regular, i.e., satisfy the first condition from Theorem 5.*

## 4.2 Targeting weighted treatment effects

Our inference strategy can be naturally extended to the setting of estimating weighted treatment effect of the type discussed in (12), i.e., where we target  $\tau_w = \int \tau(u) w(u) dG(u)$ .

To do so, we need to derive a feasible bias bound analogous to that in Corollary 8, and then incorporate it into the optimization problem used to generate the weights  $\gamma_{\pm}(\cdot)$ . One difficulty, however, is that in many of the applications considered in Section 2.2, we did not know the weighting function  $w(u)$  exactly; rather, we only knew it up to a multiplicative constant. For example, in order to identify the classical regression discontinuity parameter at the cutoff  $c$ , we need to use  $w(u) = p(c|u)/f(c)$ . Here  $p(c|u)$  is known given our noise model (Assumption 2), but  $f(c)$ , i.e., the density of the running variable at  $c$ , is not known a priori. To address this difficulty, we start below by providing a bias bound that does not need  $w(u)$  as input, and instead can be applied with any choice  $\bar{w}(u) \propto w(u)$ .

**Corollary 11.** *Under the conditions of Corollary 3, suppose also that  $|\alpha_{(0)}(u) - \mu_{\gamma,-}| \leq M$ . Furthermore let  $\bar{w}(u)$  any function that satisfies  $\bar{w}(u) \propto w(u)$ . Then the limit  $\tau_{\gamma}$  as defined in (21) satisfies*

$$|\tau_{\gamma} - \tau_w| \leq M \frac{\mathbb{E}[|h_+(U) - h_-(U)|]}{\mathbb{E}[h_+(U)]} + M' \frac{\mathbb{E}[|h_+(U) - \bar{w}(U)|] + \mathbb{E}[|h_+(U) - \bar{w}(U)|]}{\mathbb{E}[h_+(U)]} \quad (36)$$

Given a choice of  $\bar{w}(u) \propto w(u)$ , we can turn Corollary 11 into a practical bias bound as before in (28): Writing  $\mathcal{G}_n$  for the Dvoretzky–Kiefer–Wolfowitz confidence set for  $G$  as in (27), we set the bias bound  $\hat{B}_{\gamma}$  to<sup>25</sup>

$$\sup_{G \in \mathcal{G}_n} \frac{\int (M|h_+(u) - h_-(u)| + M'|h_+(u) - \bar{w}(u)|) dG(u) + M' \int |h_+(u) - \bar{w}(u)| dG(u)}{\int h_+(u) dG(u)}, \quad (37)$$

and note that the conclusion of Proposition 9 remains valid in this setting.

Finally, we can solve for  $\gamma_{\pm}(\cdot)$  via an optimization problem almost identical to (30), the only change being that we replace the worst-case bias inequality (31) with

$$\begin{aligned} t_1 + t_2 &\leq t \\ M \cdot \left| \int_{(-\infty, c)} \gamma_-(z) p(z|u) d\lambda(z) - \int_{[c, \infty)} \gamma_+(z) p(z|u) d\lambda(z) \right| &\leq t_1 \text{ for all } u \\ M' \cdot \left| \int_{[c, \infty)} \gamma_+(z) p(z|u) d\lambda(z) - \bar{w}(u) \right| &\leq t_2 \text{ for all } u. \end{aligned} \quad (38)$$

The only remaining ambiguity is in how we choose the weighting function  $\bar{w}(u) \propto w(u)$ . In practice, we seek to make  $\bar{w}(u)$  closely match the true weighting function  $w(u)$ , e.g., for the case of the regression discontinuity parameter discussed above, we use  $\bar{w}(u) = p(c|u)/\hat{f}(c)$  where  $\hat{f}(c)$  is an estimate of  $f(c)$ . If the constant of proportionality between  $\bar{w}(u)$  and  $w(u)$  is far from 1, the whole derivation above remains valid; however, it may be difficult to make the bias bound  $t_2$  in (38) small while enforcing the constraint (32).

### 4.3 Practical considerations

We summarize our approach to inference in Algorithm 1. As emphasized above, valid inference hinges on correctly modeling the measurement error as in Assumption 2 (or its

<sup>25</sup>The solution of this optimization problem is still tractable: It suffices to solve two fractional programs, one taking plus sign for the last term in numerator of (37) and the other one taking a minus sign. Each individual program can then be solved by linear programming through the Charnes and Cooper [1962] transformation. Then we take  $\hat{B}_{\gamma}$  as the maximum of the two objective values.

**Algorithm 1:** Confidence intervals for treatment effects in regression discontinuity designs identified via noise-induced randomization (NIR). This algorithm can either be used for constant treatment effects  $\tau$  as specified below, or for targeted treatment effect estimands  $\tau_w$  using [modified steps in brackets].

**Input:** Samples  $Z_i, Y_i, W_i, i = 1, \dots, n$  and RD cutoff  $c$

Bounds  $M$  [resp.  $M, M'$ ]

Nominal significance level  $\alpha$

Hyperparameters  $\sigma^2, \ell, u, C, \beta$  for (30) [resp. (38)]

- 1 Form a guess or estimate  $\hat{f} = \hat{f}_m$  of the marginal  $Z$ -density.
- 2 Solve the minimax problem (30) [resp. (38)] to get  $\gamma_+, \gamma_-$  and the induced  $h_+, h_-$ .
- 3 Form the point estimate  $\hat{\tau}_\gamma$  as in (19).
- 4 Estimate the variance of  $\hat{\tau}_\gamma$  by  $\hat{V}_\gamma$  as in (25).
- 5 Estimate the worst case bias  $\hat{B}_\gamma$  by (28) [resp. (37)]
- 6 Form bias-aware confidence intervals at level  $\alpha$  as in (23).

relaxation in Remark 2), and having conservative bounds  $M$  and  $M'$  on the range of the functions  $\alpha_{(0)}(u)$  and  $\tau(u)$  respectively. If the measurement error model and  $M$  and  $M'$  are available a-priori, then our inference is valid by design.

In practice, having domain-specific knowledge about the distribution of the running variable (e.g., from test-retest data or a physical model for the measurement device) is essentially a pre-requisite for applying our approach. On the other hand, for  $M$  and  $M'$ , there may be more flexibility. In some cases, e.g., if we have a binary outcome  $Y_i \in \{0, 1\}$ , then we can use  $M = M' = 1$  as a purely uninformative choice. In other cases, however, this may not be desirable: Perhaps our outcomes are not uniformly bounded, or perhaps using this uniform bound (e.g.,  $M = M' = 1$  for binary data) may seem needlessly conservative. In such cases, the following heuristic may be helpful: Fit a cubic smoothing spline of  $Y_i \sim Z_i$  in the control group, with smoothing parameter chosen by generalized cross-validation (GCV) and finally take  $M$  to be one half of the range of fitted values. As a sensitivity analysis, we also recommend running our procedure with  $2M$  and  $4M$  when using heuristic choices of  $M$ . For  $M'$ , we try  $M, \frac{M}{2}$  and  $\frac{M}{4}$ . The larger the value of  $M$  and  $M'$  we use, the more conservative the resulting inference.

We now discuss the remaining parameters/steps required by our algorithm. These are not as critical, in the sense that poor parameter choices will not compromise valid inference; however, making good choices here may enable shorter confidence intervals. Our first task here is to estimate the noise density  $f(\cdot)$ . In doing so, we make use of the structure provided by Assumption 2, and fit  $f(\cdot)$  via non-parametric maximum likelihood [Kiefer and Wolfowitz, 1956] as implemented in the R package REBayes [Koenker and Gu, 2017]. We take  $\sigma^2$  as the residual error from running a linear regression of  $Y_i$  on  $Z_i, W_i$  and their interaction (recall that, here,  $\sigma^2$  only scales the bias-variance tradeoff for learning  $\gamma_\pm(\cdot)$ , but is not used for inference). Our final parameters are chosen for practical convenience and numerical stability. For  $\ell, u$ , in the case of Gaussian running variables, we set  $\ell, u = c \pm 3\nu$  where  $c$  is the cutoff and  $\nu^2 = \text{Var}[Z_i | U_i]$  is the noise variance, while for binomial running variables, we set  $\ell, u = c \pm 3 \cdot 0.5/\sqrt{N}$ .<sup>26</sup> Finally, we set  $C = +\infty, \beta = 1$ , i.e., we do not use the constraint (33) in our experiments, as we found this constraint to hardly ever be active for

<sup>26</sup>These choices are recommended for either estimating a constant effect  $\tau$ , or the regression discontinuity parameter at the cutoff  $c$ . To estimate targets away from the cutoff, wider intervals may be helpful.

reasonable choices of  $C$  and  $\beta$ .

## 5 Revisiting the HIV Example

As a first application of our method, we revisit our motivating example on estimating the effect of ART on retention among HIV positive patients. We discussed the statistical setting and existing analyses in Section 1.1. Here, our goal is to flesh-out the application of our method to this dataset, and also compare our approach to state-of-the-art continuity-based approaches to regression discontinuity estimation.

To this end, we consider the following broad strategies to the problem. Our first baseline builds on the identifying assumption used for local linear regression in Section 1.1, i.e., that there is a constant  $B$  such that  $|\mu''_{(w)}(z)| \leq B$  for all  $w \in \{0, 1\}$  and  $z \in \mathbb{R}$ . Many different approaches of this type have been recently discussed in the literature, including by [Armstrong and Kolesár \[2018, 2020\]](#), [Imbens and Wager \[2019\]](#) and [Kolesár and Rothe \[2018\]](#). Here, we consider the optimized regression discontinuity (optrdd) method of [Imbens and Wager \[2019\]](#), which uses convex optimization to derive the minimax linear estimator of  $\tau$  under the assumed curvature bound. Relative to the local linear approach from Section 1.1, optrdd does not require centering the confidence interval at a point estimate derived from local linear regression, thus potentially allowing us to shorten the interval.

The main difficulty in using optrdd is in choosing the curvature bound  $B$ . Being able to choose a good  $B$  fundamentally requires further assumptions, because if all we can assume is that  $|\mu''_{(w)}(z)| \leq B$  for some unknown  $B$ , then estimating  $B$  in a way that enables valid yet adaptive inference is impossible [[Armstrong and Kolesár, 2018](#)]. Here, we consider two approaches to choosing  $B$ . First, as recommended in [Armstrong and Kolesár \[2020\]](#), we fit fourth-degree polynomials to  $\mu_{(0)}(z)$  and  $\mu_{(1)}(z)$ , and take the largest estimated curvature obtained anywhere. This approach is heuristic and not justified by design, but appears to yield reasonable results. Second, we consider a design-based approach: Given values for  $M$  as in Section 4.3, we use Proposition 4 to obtain an upper bound  $B$  on curvature. This bound is rigorously justified given our noise model, but results in much wider confidence intervals.

Our next baseline relies on higher-order smoothness for inference. This approach, which has recently become popular in applications, involves first fitting the regression discontinuity parameter via local linear regression as in (2), and then estimating and correcting for its bias in a way that's asymptotically justified under higher-order smoothness assumptions [[Calonico, Cattaneo, and Titiunik, 2014](#)]. We implement this approach via the R package `rdrobust` of [Calonico, Cattaneo, and Titiunik \[2015\]](#). Relative to our first baseline, `rdrobust` essentially uses higher-order smoothness assumptions to automate discovery of the curvature of  $\mu_{(w)}(z)$ ; see [Calonico, Cattaneo, and Farrell \[2018\]](#) for further discussion. A major advantage of `rdrobust` is that it does not require any tuning parameters; a potential downside, however, is that it may be more heavily reliant on asymptotics (in contrast, for optrdd and associated methods, choosing  $B$  is difficult but once  $B$  is given inference is essentially exact in finite samples).

Finally, we consider our proposed method identified via noise-induced randomization, assuming Gaussian noise with standard deviation  $\nu = 0.19$  as discussed in Section 1.1. In this application, we focus on estimating a constant treatment effect parameter. For the bound  $M$  from Corollary 2, we consider several choices. First, we estimate  $M$  as discussed in Section 4.3. Next, following our discussion in Section 1.1, we note that our outcome is

<i>Method</i>	<i>Bound</i>			
	M = 0.08	M = 0.16	M = 0.32	M = 1
NIR	<b>0.136 ± 0.087</b>	0.122 ± 0.092	0.116 ± 0.097	<b>0.112 ± 0.102</b>
optrdd	0.066 ± 0.133	0.063 ± 0.135	0.063 ± 0.136	0.062 ± 0.136
	B = 1.46	B = 2.92	B = 5.84	
optrdd	<b>0.153 ± 0.080</b>	0.136 ± 0.091	0.110 ± 0.102	
rdrobust	<b>0.170 ± 0.076</b>			

Table 1: Estimates and nominally 95% confidence intervals for the effect of ART on retention rate of HIV patients, as given by our noise-induced randomization (NIR) method, optimized regression discontinuity design (optrdd), robust nonparametric confidence interval (rdrobust). For our method, we consider  $M = 0.08$  as estimated following Section 4.3, 2 and 4 times this value, and the uninformative choice  $M = 1$ . For optrdd, we consider  $B = 1.46$  estimated following [Armstrong and Kolesár \[2020\]](#), as well as 2 and 4 times this value. We also consider values for  $B$  implied by our  $M$  bounds following Proposition 4; quantitatively, this yields  $B = 51.31, 102.63, 205.26$  and  $641.43$ . The function rdrobust is run using the default specification of [Calonico, Cattaneo, and Titiunik \[2015\]](#). As discussed further in the text, estimates most likely to be used in practice are highlighted in bold-face; other estimates serve as sensitivity analysis.

binary so we can also use an uninformative bound of  $M = 1$ . We run our method following the recommendations in Section 4.3.

We present the results in Table 1. We show results for several different choices of  $B$  and  $M$ ; moreover, for each case where we use data-driven choices of  $B$  or  $M$ , we also conduct a sensitivity analysis where we multiply these bounds by 2 or 4. For clarity, we display in bold-face estimates that are most likely to be used in practice, i.e., using the estimate  $M$  or the uninformative  $M = 1$  for our method, taking the worst-case curvature as in [Armstrong and Kolesár \[2020\]](#), or using rdrobust our of the box.

Overall, at least in this example, we find that when we run our method with the estimated  $M = 0.08$ , we get confidence intervals whose width is roughly in line with those provided by continuity-based methods that estimate the curvature of  $\mu_{(w)}(z)$ . But unlike the continuity-based approaches, our method also allows for a purely uninformative choice of  $M = 1$  for which inference becomes purely design based and only relies on the noise in the running variable  $Z_i$ ; and yet the treatment effect estimate remains significant at the 5% level. Noting the difficulty of accurately estimating curvature (especially in finite samples), we believe the ability of our method to deliver reasonably short confidence intervals that are purely justified by randomization to be potentially useful in practice.

## 6 Test Scores in Early Childhood

We next consider the behavior of our method in a semi-synthetic experiment built using data from the Early Childhood Longitudinal Study [[Tourangeau et al., 2015](#)]. This dataset



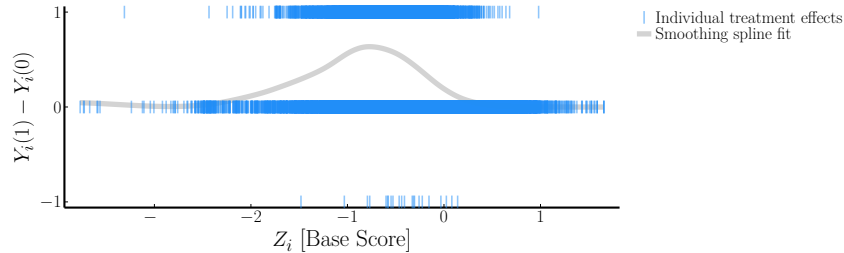


Figure 2: Scatterplot of the individual treatment effects  $Y_i(1) - Y_i(0)$  against the running variable  $Z_i$ , with data derived from the Early Childhood Longitudinal Study [Tourangeau et al., 2015] as discussed in Section 6. We fit the ground truth treatment effect function  $\mathbb{E}[Y_i(1) - Y_i(0) | Z_i = z]$ , shown as a line, using a smoothing spline.

has scaled test scores for  $n = 18,174$  children from kindergarten to fifth grade. Furthermore, each test score is accompanied by a noise estimate obtained via item response theory; see Tourangeau et al. [2015] for further details.

We build a semi-synthetic regression discontinuity experiment using this dataset as follows, where each sample  $i = 1, \dots, n$  is built using the sequence of test scores from a single child. We set the running variable  $Z_i$  to be child’s kindergarten spring semester test score, and set treatment as  $W_i = 1(\{Z_i \geq c\})$  for a cutoff  $c = -0.2$ . We then set control potential outcomes  $Y_i(0) \in \{0, 1\}$  to indicate whether the child’s test scores were above  $a = 0.5$  in spring semester of their first grade, while  $Y_i(1) \in \{0, 1\}$  measures the same quantity in spring semester of their second grade; these are analogous to typically studied outcomes such as passing subsequent examinations. Thus, the “treatment effect”  $Y_i(1) - Y_i(0)$  measures the child’s improvement in passing the test (i.e., clearing the cutoff  $a = 0.5$ ) between first and second grades.

As shown in Figure 2, there is considerable heterogeneity in the regression discontinuity parameter  $\tau_{c'} = \mathbb{E}[Y_i(1) - Y_i(0) | Z_i = c']$  as we vary  $c'$  away from the cutoff: For children with either very good or very bad values of  $Z_i$  the treatment effect is essentially 0 (since they will pass or, respectively, fail to pass the cutoff  $a$  in both first and second grade with high probability), while for students with intermediate values of  $Z_i$  there is a large treatment effect (we chose the parameters  $a$  and  $c$  in our experiment specification to accentuate this type of heterogeneity).

Our main question here is whether our procedure is able to estimate this heterogeneity, i.e., whether it can accurately recover variation in treatment effects away from the cutoff. To this end, we consider two statistical targets: First, we consider estimation of the regression discontinuity parameter (14) at  $c'$  away from the cutoff, and second we consider the policy-relevant parameter (15) quantifying the effect of changing the cutoff from  $c$  to  $c'$ . When applying our method, we assume Gaussian errors in the running variable as in Assumption 2 and, following Remark 2, set  $\nu = 0.2043$  to match the lowest noise estimate provided in the Early Childhood Longitudinal Study dataset [Tourangeau et al., 2015]. We run our method as discussed in Section 4.3 with a data-driven choice of bounds  $M = M' = 0.31$ ; however, to accommodate values of  $c'$  further from the cutoff, we consider non-zero weights on the slightly wider interval  $[\ell, u] = c \pm 4\nu$ .

Results for both targets are shown in Figure 3. We see that our method is able to recover

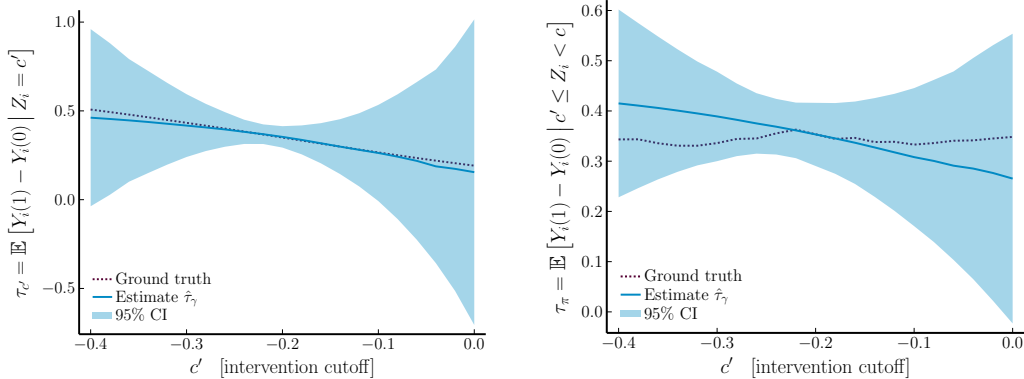


Figure 3: Regression discontinuity inference using our method (NIR) with data generated as in Figure 2. The left panel shows estimates of the regression discontinuity parameter (14), while the right panel shows estimates of the policy relevant parameter (15), both with 95% confidence intervals.

heterogeneity in the regression discontinuity parameter (14) with remarkable accuracy, while we have some more difficulties in fitting the policy-relevant parameter (15). In both cases, the confidence intervals provided by our approach cover the ground truth. Furthermore, as expected, they are narrowest near the cutoff  $c = -0.2$ , and get wider as we move  $c'$  away from the cutoff.

## 7 Simulation Study

In order to complete the picture given by our applications, we consider a simulation study to more precisely assess the performance of our method in terms of both its accuracy and coverage. To this end, we compare the following five methods closely related to those considered in Section 5: An implementation of `optrdd` with  $B$  chosen to match the true worst-case curvature in the simulation specification (oracle `optrdd`);<sup>27</sup> an implementation of `optrdd` with  $B$  estimated by polynomial regression as used in Section 5; `rdrobust` with default specification (taking the debiased estimate as the point estimate); our method with  $M$  set to the true range of  $\alpha_{(w)}(u)$  in the simulation specification (oracle NIR); and our method with the uninformative choice  $M = 1$  (all simulations have binary outcomes). We run NIR such as to target a constant treatment parameter and following recommendations from Section 4.3. The two new baselines, oracle `optrdd` and oracle NIR, are provided to illustrate any loss from feasible tuning parameter choice.

We first consider two simulation specifications where the running variable  $Z_i$  has continuous support and Gaussian measurement error. In our first setting, we generate independent

<sup>27</sup>In the case of discrete running variables, oracle `optrdd` sets  $B$  to the worst-case curvature of a cubic spline interpolator to  $\mu_{(w)}(z)$ ; see Imbens and Wager [2019] and Kolesár and Rothe [2018] for further discussion.

		$\nu^2$	setup 1			setup 2		
			0.25	0.5	1	0.25	0.5	1
$n = 1,000$	coverage	oracle oprtdd	97.2%	94.2%	95.8%	96.6%	97.4%	96.6%
		optrdd	97.4%	94.4%	96.8%	96.6%	97.0%	97.8%
		rdrobust	95.4%	91.8%	94.0%	94.0%	95.0%	94.6%
		oracle NIR	98.6%	93.6%	96.0%	100.0%	96.6%	95.2%
		NIR ( $M = 1$ )	97.4%	93.0%	96.2%	99.0%	97.6%	95.4%
	length	oracle oprtdd	0.202	0.192	0.174	0.273	0.238	0.201
		optrdd	0.289	0.280	0.262	0.313	0.306	0.278
		rdrobust	0.248	0.243	0.229	0.279	0.273	0.249
		oracle NIR	0.338	0.274	0.218	0.361	0.251	0.209
		NIR ( $M = 1$ )	0.354	0.289	0.230	0.339	0.288	0.229
	MAE	oracle oprtdd	0.075	0.076	0.070	0.102	0.084	0.074
		optrdd	0.102	0.110	0.101	0.118	0.112	0.103
		rdrobust	0.098	0.108	0.099	0.118	0.114	0.104
		oracle NIR	0.115	0.109	0.087	0.088	0.093	0.082
		NIR ( $M = 1$ )	0.125	0.116	0.091	0.110	0.104	0.090
$n = 5,000$	coverage	oracle oprtdd	96.4%	96.4%	97.0%	97.4%	96.6%	96.4%
		optrdd	96.6%	97.6%	97.4%	97.8%	96.8%	97.0%
		rdrobust	94.4%	94.4%	96.2%	95.2%	95.8%	93.2%
		oracle NIR	96.2%	95.0%	97.2%	97.4%	96.2%	95.6%
		NIR ( $M = 1$ )	96.4%	95.0%	97.4%	96.6%	96.0%	95.8%
	length	oracle oprtdd	0.107	0.101	0.092	0.141	0.125	0.106
		optrdd	0.139	0.131	0.121	0.157	0.147	0.129
		rdrobust	0.112	0.108	0.102	0.134	0.127	0.110
		oracle NIR	0.148	0.122	0.101	0.119	0.118	0.098
		NIR ( $M = 1$ )	0.153	0.127	0.105	0.133	0.129	0.105
	MAE	oracle oprtdd	0.040	0.038	0.033	0.052	0.045	0.037
		optrdd	0.052	0.049	0.044	0.058	0.053	0.045
		rdrobust	0.046	0.045	0.041	0.055	0.053	0.044
		oracle NIR	0.058	0.048	0.038	0.044	0.045	0.038
		NIR ( $M = 1$ )	0.061	0.050	0.040	0.051	0.049	0.040

Table 2: Simulation results in the Gaussian noise settings (39) (setup 1) and (40) (setup 2) for different choices of sample size  $n$  and noise variance  $\nu^2$ . We report the actual coverage, mean length of confidence intervals and mean absolute error (MAE) of the nominal 95% confidence intervals in simulations by oracle oprtdd (with true  $B$ ), optrdd (adaptively chosen  $B$ ), rdrobust, oracle NIR (with true  $M$ ) and NIR with uninformative  $M = 1$ . All numbers are averaged over 500 simulation replications.

		$K$	setup 3			setup 4		
			50	100	200	50	100	200
$n = 1,000$	coverage	oracle optrdd	96.6%	94.4%	95.2%	97.2%	98.4%	97.4%
		optrdd	98.2%	96.4%	96.8%	98.6%	98.2%	97.6%
		rdrubust	93.8%	91.8%	94.4%	94.2%	92.2%	92.8%
		oracle NIR	97.2%	96.2%	97.0%	97.0%	97.2%	97.2%
		NIR ( $M = 1$ )	96.6%	96.4%	96.6%	97.4%	97.2%	97.2%
	length	oracle optrdd	0.203	0.221	0.249	0.15	0.144	0.142
		optrdd	0.434	0.373	0.368	0.303	0.253	0.248
		rdrubust	0.378	0.357	0.341	0.267	0.243	0.237
		oracle NIR	0.252	0.290	0.346	0.204	0.226	0.264
		NIR ( $M = 1$ )	0.268	0.304	0.361	0.212	0.230	0.270
	MAE	oracle optrdd	0.079	0.089	0.101	0.054	0.051	0.048
		optrdd	0.140	0.144	0.141	0.091	0.090	0.090
		rdrubust	0.169	0.169	0.155	0.115	0.113	0.107
		oracle NIR	0.097	0.111	0.132	0.071	0.080	0.097
		NIR ( $M = 1$ )	0.103	0.118	0.139	0.074	0.084	0.100
$n = 5,000$	coverage	oracle optrdd	95.4%	96.6%	96.0%	96.6%	98.2%	96.4%
		optrdd	98.0%	98.0%	95.8%	99.4%	97.2%	95.6%
		rdrubust	92.2%	95.6%	93.8%	94.8%	93.8%	92.4%
		oracle NIR	95.2%	95.4%	94.8%	96.4%	95.8%	94.0%
		NIR ( $M = 1$ )	95.6%	95.4%	94.2%	96.2%	95.6%	94.4%
	length	oracle optrdd	0.108	0.117	0.132	0.081	0.076	0.075
		optrdd	0.200	0.169	0.171	0.141	0.117	0.114
		rdrubust	0.162	0.150	0.151	0.117	0.106	0.102
		NIR	0.113	0.128	0.152	0.091	0.098	0.112
		NIR ( $M = 1$ )	0.119	0.134	0.158	0.094	0.100	0.114
	MAE	oracle optrdd	0.040	0.043	0.050	0.029	0.028	0.028
		optrdd	0.063	0.055	0.063	0.043	0.042	0.043
		rdrubust	0.070	0.060	0.064	0.051	0.048	0.046
		oracle NIR	0.042	0.047	0.059	0.034	0.038	0.045
		NIR ( $M = 1$ )	0.045	0.049	0.062	0.035	0.039	0.046

Table 3: Simulation results in the binomial noise setting (41), with baseline conditional response function (42) (setup 3) and (43) (setup 4), for different choices of sample size  $n$  and binomial parameter  $K$ . We report the actual coverage, mean length of confidence intervals and mean absolute error (MAE) of the nominal 95% confidence intervals in simulations by oracle optrdd (with true  $B$ ), optrdd (adaptively chosen  $B$ ), rdrubust, oracle NIR (with true  $M$ ) and NIR with uninformative  $M = 1$ . All numbers are averaged over 500 simulation replications.

samples as below

$$\begin{aligned} U_i &\sim \text{Uniform}([-3, 3]), & Z_i | U_i &\sim \mathcal{N}(U_i, \nu^2), \\ W_i &= 1(\{Z_i \geq 0\}), & Y_i(w) &\sim \text{Bernoulli}\left(\frac{\sin(U_i)}{4} + 0.3 + 0.25w\right), \end{aligned} \quad (39)$$

while in our second setting, we use

$$\begin{aligned} U_i &\sim p\delta_0 + \frac{1-p}{2}\delta_k + \frac{1-p}{2}\delta_{-k}, & Z_i | U_i &\sim \mathcal{N}(U_i, \nu^2), \\ W_i &= 1(\{Z_i \geq 0\}), & Y_i(w) &\sim \text{Bernoulli}(0.3 \cdot 1(\{U_i = 0\}) + 0.2 + 0.25w), \end{aligned} \quad (40)$$

with  $k$  chosen such that  $\varphi(k) = 0.1$  and let  $p = \varphi(k)/\varphi(0)$  where  $\varphi$  is the standard normal density. In both settings, the treatment effect is constant  $\tau = 0.25$ , so that both our noise-based approach and the continuity-based approach unambiguously target the same estimand.

Second, we consider a pair of data-generating distributions wherein  $Z_i$  has discrete support, and has a binomial distribution conditionally on the latent  $U_i$ . In both settings, we generate

$$\begin{aligned} U_i &\sim \text{Uniform}([0.5, 0.9]), & Z_i | U_i &\sim \text{Bin}(K, U_i), \\ W_i &= 1(\{Z_i \geq 0.6K\}), & Y_i(w) &\sim \text{Bernoulli}(\mathbb{E}[Y_i(w)|U_i]), \end{aligned} \quad (41)$$

but we consider two different choices for  $\mathbb{E}[Y_i(w)|U_i]$ ,

$$\mathbb{E}[Y_i(w)|U_i = u] = 0.25 \cdot 1(\{u < 0.6\}) + 0.75 \cdot 1(\{u \geq 0.6\}), \quad (42)$$

$$\mathbb{E}[Y_i(w)|U_i = u] = \sin(9u)/3 + 0.4. \quad (43)$$

In both cases, we use a null treatment effect  $\tau = 0$ , and consider various choices of  $K$ .

Table 2 shows the results for the simulation settings with a continuous  $Z_i$ , for different values of sample size  $n$  and measurement error  $\nu^2$ . Here, by far the best performing method is “oracle optrdd”, i.e., the continuity-based approach that gets a-priori access to the exact worst-case bound on  $\mu_{(w)}(z)$ . In practice, however, oracle optrdd is of course not a feasible baseline; and the feasible alternatives of optrdd and rdrobust get wider intervals. In this comparison, our approach is roughly competitive with feasible optrdd and rdrobust, with a closer comparison depending on the noise level  $\nu^2$ . As one might expect, when  $\nu^2$  is large, our method performs well because there is a lot of measurement error to exploit; meanwhile, when  $\nu^2$  gets small, our approach has less information to exploit and our intervals get wider. In contrast, the continuity-based baselines are fairly insensitive to the noise level  $\nu^2$  here. All methods get reasonable coverage here.

Next, we consider specifications with a binomial running variable in Table 3. Here, the results are more clearly in favor of NIR. Our method achieves good coverage throughout, and yields shorter confidence intervals than feasible optrdd (including with the uninformative choice  $M = 1$ ). Our method yields confidence intervals of comparable length to those from rdrobust; however, rdrobust does not achieve nominal coverage here.<sup>28</sup>  $K$  provides a natural analogue to the noise level  $\nu^2$ : The smaller  $K$  is the more effective noise there is in the running variable, and so the better our method does.

<sup>28</sup>Rdrobust is justified via asymptotic arguments that do not hold for discrete running variables, so its failure to achieve coverage here is not in contradiction with formal results backing the method.

At a high level, this simulation experiment shows that our method, NIR, can flexibly turn assumptions about exogenous noise in the running variable  $Z_i$  into a practical, design-based procedure for regression discontinuity inference in regression discontinuity designs. We achieve nominal coverage across a wide variety of simulation settings, and are overall competitive with continuity-based alternatives in terms of power. Thus, these results suggest that the pursuit of design-based inference in regression discontinuity may be practical in applications; in other words, concerns about power need not necessarily get in the way of a statistician who would prefer to rely on design-based inference for conceptual reasons.

Our results also cautiously point to the possibility that NIR may in fact result in improved power in settings where running variables are discrete with binomial noise. This would not be unreasonable, as continuity-based approaches (especially `rdrubust`) were not necessarily designed for this setting,<sup>29</sup> whereas NIR can directly exploit structure of the binomial distribution. However, a detailed study of the power (as opposed to feasibility) of designed based regression inference across settings of practical interest is beyond the scope of this paper.

## 8 Discussion

Informal descriptions of regression discontinuity designs often appeal to an analogy to a local randomized experiment, whereby units near the cutoff are as if randomly assigned to treatment. In perhaps the most common version of this analogy, one posits that units near the cutoff have had their running variable randomized [Cattaneo, Frandsen, and Titiunik, 2015]. However, this analogy is typically undermined by the clear relevance of the running variable to the outcome—even within a region near the cutoff. Here, we proposed a new approach to inference in regression discontinuity designs that formalizes measurement error or other exogenous noise in the running variable  $Z_i$  to capture the stochastic nature of the assignment mechanism in regression discontinuity designs. In the presence of measurement error, units are indeed randomly assigned to treatment—but with unknown, heterogeneous probabilities determined by a latent variable of which  $Z_i$  is a noisy measure.

Regression discontinuity designs with known or estimable measurement error in the running variable arise in many settings. We have already considered applications to educational and biomedical tests. Public policies that target interventions based on, e.g., proxy means testing [e.g., Alatas, Banerjee, Hanna, Olken, and Tobias, 2012] may also readily admit analysis with the noise-induced randomization approach. Furthermore, this approach is applicable to settings where thresholds for statistical significance are used to make numerous decisions.

Finally, while this noise-induced randomization approach applies to many settings of interest, we emphasize that it does not apply to all regression discontinuity designs, as some running variables are not readily interpretable as having measurement error. For example, numerous studies have used geographic boundaries as discontinuities [Keele and Titiunik, 2014, Rischard, Branson, Miratrix, and Bornn, 2018], but it would be questionable to model the location of a household in space as having meaningful measurement error (rather, it may be more plausible to argue that the location of the boundary itself is random [Ganong and Jäger, 2018]). Likewise, analyses of close elections—a central example of regression discontinuity designs in political science and economics [Caughey and Sekhon, 2011, Lee,

<sup>29</sup>Although, as discussed in Kolesár and Rothe [2018] they can rigorously be used in this setting given appropriate interpretation.

2008]—may not allow for a natural noise model for  $Z_i$  that would arise from, e.g., noisy counting of the number of ballots cast for each candidate. These considerations call attention to the limits of the proposed approach, but also highlights a difference in the foundational assumptions required for identification, estimation, and inference in regression discontinuity designs with a noisy running variable versus the assumptions required when the running variable is noiseless.

## References

- Vivi Alatas, Abhijit Banerjee, Rema Hanna, Benjamin A Olken, and Julia Tobias. Targeting the poor: Evidence from a field experiment in Indonesia. *American Economic Review*, 102(4):1206–40, 2012.
- Joshua D Angrist and Miikka Rokkanen. Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344, 2015.
- Timothy B Armstrong and Michal Kolesár. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.
- Timothy B Armstrong and Michal Kolesár. Simple and honest confidence intervals in non-parametric regression. *Quantitative Economics*, 11(1):1–39, 2020.
- Otávio Bartalotti, Quentin Brummet, and Steven Dieterle. A correction for regression discontinuity designs with group-specific mismeasurement of the running variable. *Journal of Business & Economic Statistics*, pages 1–16, 2020.
- Jacob Bor, Ellen Moscoe, Portia Mutevedzi, Marie-Louise Newell, and Till Bärnighausen. Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology (Cambridge, Mass.)*, 25(5):729, 2014.
- Jacob Bor, Matthew P Fox, Sydney Rosen, Atheendar Venkataramani, Frank Tanser, Deenan Pillay, and Till Bärnighausen. Treatment eligibility and retention in clinical HIV care: A regression discontinuity study in South Africa. *PLoS medicine*, 14(11), 2017.
- Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.
- Sebastian Calonico, Matias D. Cattaneo, and Rocio Titiunik. rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. *The R Journal*, 7(1):38–51, 2015.
- Sebastian Calonico, Matias D Cattaneo, and Max H Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.
- Donald T Campbell and Julian C Stanley. Experimental and quasi-experimental designs for research on teaching. In *Handbook of Research on Teaching*, chapter 5. 1963.
- Matias D Cattaneo, Brigham R Frandsen, and Rocio Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference*, 3(1):1–24, 2015.



- Devin Caughey and Jasjeet S Sekhon. Elections and the regression discontinuity design: Lessons from close US house races, 1942-2008. *Political Analysis*, 19(4):385–408, 2011.
- A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, 1962.
- Ming-Yen Cheng, Jianqing Fan, and James S Marron. On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.
- Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.
- Laurent Davezies and Thomas Le Barbanchon. Regression discontinuity design with continuous measurement error in the running variable. *Journal of Econometrics*, 200(2):260–281, 2017.
- Juan D Diaz and Jose R Zubizarreta. Complex discontinuity designs using covariates. *arXiv preprint arXiv:2004.05641*, 2020.
- David L Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270, 1994.
- Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.
- Peter Ganong and Simon Jäger. A permutation test for the regression kink design. *Journal of the American Statistical Association*, 113(522):494–504, 2018.
- Deborah K Glencross, George Janossy, Lindi M Coetzee, Denise Lawrie, Hazel M Aggett, Lesley E Scott, Ian Sanne, James A McIntyre, and Wendy Stevens. Large-scale affordable PanLeucogated CD4+ testing with proactive internal and external quality assessment: In support of the South African national comprehensive care, treatment and management programme for HIV and AIDS. *Cytometry Part B: Clinical Cytometry: The Journal of the International Society for Analytical Cytology*, 74(S1):S40–S51, 2008.
- Jinyong Hahn, Petra Todd, and Wilbert van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.
- Alyssa F Harlow, Jacob Bor, Alana T Brennan, Mhairi Maskew, William MacLeod, Sergio Carmona, Koleka Mlisana, and Matthew P Fox. Impact of viral load monitoring on retention and viral suppression: A regression discontinuity analysis of south africa’s national laboratory cohort. *American Journal of Epidemiology*, 2020.
- James J Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005.
- Michael D Hughes, Daniel S Stein, Holly M Gundacker, Fred T Valentine, John P Phair, and Paul A Volberding. Within-subject variation in CD4 lymphocyte count in asymptomatic human immunodeficiency virus infection: implications for patient monitoring. *Journal of Infectious Diseases*, 169(1):28–36, 1994.

- Nikolaos Ignatiadis and Stefan Wager. Bias-aware confidence intervals for empirical Bayes analysis. *arXiv preprint arXiv:1902.02774*, 2019.
- Guido Imbens and Stefan Wager. Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278, 2019.
- Guido W Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959, 2012.
- Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.
- Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.
- W. Jiang and C.H. Zhang. General maximum likelihood empirical Bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.
- Nathan Kallus. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020.
- Luke J Keele and Rocio Titiunik. Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155, 2014.
- Morgan Kelly. The standard errors of persistence. 2019.
- Jack Kiefer and Jacob Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *The Annals of Mathematical Statistics*, pages 887–906, 1956.
- Arlene KH Kim. Minimax bounds for estimation of normal mixtures. *Bernoulli*, 20(4):1802–1818, 2014.
- Roger Koenker and Jiaying Gu. REBayes: Empirical Bayes mixture methods in R. *Journal of Statistical Software*, 82(8):1–26, 2017.
- Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. *American Economic Review*, 108(8):2277–2304, 2018.
- David S Lee. Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, 142(2):675–697, 2008.
- David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.
- Fan Li, Alessandra Mattei, and Fabrizia Mealli. Evaluating the causal effect of university grants on student dropout: Evidence from a regression discontinuity design using principal stratification. *The Annals of Applied Statistics*, 9(4):1906–1931, 2015.
- Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, pages 1–11, 2017.
- Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

- Alessandra Mattei and Fabrizia Mealli. Regression discontinuity designs as local randomized experiments. *Observational Studies*, 2:156–173, 2016.
- Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*. Lecture Notes in Statistics. Springer Berlin Heidelberg, 2009.
- Jersey Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.
- Art B Owen and Hal Varian. Optimizing the tie-breaker regression discontinuity design. *arXiv preprint arXiv:1808.07563*, 2018.
- Zhuan Pei and Yi Shen. The devil is in the tails: Regression discontinuity design with measurement error in the assignment variable. *arXiv preprint arXiv:1609.01396*, 2016.
- Maxime Rischard, Zach Branson, Luke Miratrix, and Luke Bornn. A Bayesian nonparametric approach to geographic regression discontinuity designs: Do school districts affect NYC house prices? *arXiv preprint arXiv:1807.04516*, 2018.
- Miikka Rokkanen. Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design. Technical report, 2015.
- Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.
- Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.
- Donald B Rubin. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840, 2008.
- Adam C Sales and Ben B Hansen. Limitless regression discontinuity. *Journal of Educational and Behavioral Statistics*, 45(2):143–174, 2020.
- Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6):309–317, 1960.
- K. Tourangeau, C. Nord, T. Le, A.G. Sorongon, M.C. Hagedorn, P. Daly, and M. Najarian. Early childhood longitudinal study, kindergarten class of 2010-11 (ecls-k:2011), user’s manual for the ECLS-K:2011 kindergarten data file and electronic codebook, public version (nces 2015-074). Technical report, U.S. Department of Education. Washington, DC: National Center for Education Statistics., 2015.
- William MK Trochim, Joseph C Cappelleri, and Charles S Reichardt. Random measurement error does not bias the treatment effect estimate in the regression-discontinuity design: II. When an interaction effect is present. *Evaluation Review*, 15(5):571–604, 1991.
- Willem Daniel Francois Venter, Matthew F Chersich, Mohammed Majam, Godspower Akpomemie, Natasha Arulappan, Michelle Moorhouse, Nonkululeko Mashabane, and Deborah K Glencross. CD4 cell count variability with repeat testing in South Africa: Should reporting include both absolute counts and ranges of plausible values? *International Journal of STD & AIDS*, 29(11):1048–1056, 2018. doi: 10.1177/0956462418771768. URL <https://doi.org/10.1177/0956462418771768>. PMID: 29749876.

- Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the  $r$ th absolute moment of a sum of random variables,  $1 \leq r \leq 2$ . *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.
- Djibril Wade, Géraldine Daneau, Said Aboud, Gaby H Vercauteren, Willy SK Urassa, and Luc Kestens. WHO multicenter evaluation of FACSCount CD4 and Pima CD4 T-cell count systems: instrument performance and misclassification of HIV-infected patients. *Journal of acquired immune deficiency syndromes*, 66(5):e98, 2014.
- Takahide Yanagi. The effect of measurement error in the sharp regression discontinuity design. KIER working papers, Kyoto University, Institute of Economic Research, 2014.
- Ping Yu. Identification of treatment effects in regression discontinuity designs with measurement error. *Working paper*, 2012.

## A Proofs

### A.1 Proof of Proposition 4

*Proof.* For the first result, note that  $\mu_{(w)}(z)$  may in fact be extended to an analytic function across all of  $\mathbb{C}$ , cf. Kim [2014]. We proceed with the quantitative claims and first note that it suffices to consider the Standard Normal case, i.e.,  $\nu = 1$ . To see this, take  $Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$ . Then  $\tilde{Z}_i = Z_i/\nu_i \sim \mathcal{N}(U_i/\nu_i, 1)$  and we may apply the results to  $\tilde{Z}_i$ . Concretely, let  $\tilde{m} : \mathbb{R} \mapsto \mathbb{R}$  be an arbitrary function and  $m : z \mapsto \tilde{m}(z/\nu) = \tilde{m}(\tilde{z})$ . This defines a bijection between functions that enables us to translate results for  $\tilde{Z}_i$  into results for  $Z_i$  and vice versa (by applying the chain rule). It only remains to express the density  $\tilde{f}(\tilde{z})$  of  $\tilde{Z}_i$  at  $\tilde{z} = z/\nu$  in terms of the density  $f$  of  $Z_i$ ; by transformation we have  $\tilde{f}(\tilde{z}) = \nu \cdot f(z)$ . Furthermore, we derive all of our results for the control arm  $\mu_{(0)}(z)$ ; the arguments for  $\mu_{(1)}(z)$  are identical.

**Upper bound: First derivative.** Fix  $c > 0$ . Since the Lipschitz constant remains the same after shifting the response, we may assume (without loss of generality) that  $\alpha_{(0)}(u) \in [c, 2M + c]$ . Let  $H \ll G$  be the probability measure with  $\frac{dH}{dG}(u) = \frac{\alpha_{(0)}(u)}{\int \alpha_{(0)}(u)dG(u)}$  and write  $h(z) = \int \varphi(z - u)dH(u)$ . Then we can write:

$$\mu_{(0)}(z) = \mathbb{E} [\alpha_{(0)}(U_i) | Z_i = z] = \frac{h(z) \cdot \int \alpha_{(0)}(u)dG(u)}{f(z)}$$

$$\frac{d}{dz}\mu_{(0)}(z) = \int \alpha_{(0)}(u)dG(u) \cdot \left( \frac{h'(z)}{f(z)} - \frac{h(z)}{f(z)} \cdot \frac{f'(z)}{f(z)} \right) = \int \alpha_{(0)}(u)dG(u) \cdot \frac{h(z)}{f(z)} \cdot \left( \frac{h'(z)}{h(z)} - \frac{f'(z)}{f(z)} \right)$$

We next bound the three terms appearing in the expression above. First, we already saw that  $\mu_{(0)}(z) = \int \alpha_{(0)}(u)dG(u) \cdot \frac{h(z)}{f(z)}$  and so this term is upper bounded in absolute value by  $2M + c$ . Next, by Lemma A.1. in Jiang and Zhang [2009] it holds that:

$$\left| \frac{f'(z)}{f(z)} \right| \leq \sqrt{-\log(2\pi f^2(z))}, \quad \left| \frac{h'(z)}{h(z)} \right| \leq \sqrt{-\log(2\pi h^2(z))}$$

It remains to lower bound  $h(z)/f(z)$ :

$$h(z) = \frac{\int \alpha_{(0)}(u)\varphi(z - u)dG(u)}{\int \alpha_{(0)}(u)dG(u)} \geq \frac{c}{2M + c} \cdot \int \varphi(z - u)dG(u) = \frac{c}{2M + c} \cdot f(z)$$

So putting everything together:

$$\left| \frac{d}{dz}\mu_{(0)}(z) \right| \leq \inf_{c>0} \left\{ (2M + c) \cdot \left( \sqrt{-\log(2\pi f^2(z))} + \sqrt{-\log\left(\frac{2\pi c^2}{(2M + c)^2} f^2(z)\right)} \right) \right\} \quad (44)$$

Taking  $c = 0.5M$  leads to the stated bound. We can computationally sharpen this bound by optimizing over  $c$ .

**Upper bound: Second derivative.** Continuing with the notation from the part above, we can verify by taking the second derivative of  $\mu_{(0)}(z)$ , that:

$$\mu''_{(0)}(z) = \mu_{(0)}(z) \cdot \left\{ \left( \frac{h''(z)}{h(z)} + 1 \right) - \left( \frac{f''(z)}{f(z)} + 1 \right) \right\} - 2\mu'_{(0)}(z) \frac{f'(z)}{f(z)}$$

Applying Lemma A.1. in [Jiang and Zhang \[2009\]](#) again we find that:

$$0 \leq \frac{f''(z)}{f(z)} + 1 \leq -\log(2\pi f^2(z)), \quad 0 \leq \frac{h''(z)}{h(z)} + 1 \leq -\log(2\pi h^2(z))$$

Using the fact that  $|\mu_{(0)}(z)| \leq 2M + c$ , that we already bounded  $|\mu'_{(0)}(z)|$ ,  $f'(z)/f(z)$  above and the triangle inequality we conclude (by using  $c = 0.5M$  as we did for the first derivative; better bounds can be obtained computationally by optimizing over  $c$ ).

**Lower bound: First derivative.** Take  $G = \frac{1}{2}(\delta_{-c} + \delta_c)$ , where  $\delta_u$  is a point mass at  $u$ ,  $\alpha_{(0)}(u) = M \cdot (\mathbf{1}(u = c) - \mathbf{1}(u = -c))$ .

$$f(z) = \frac{1}{2}(\varphi(z-c) + \varphi(z+c)), \quad f'(z) = \frac{1}{2}((c-z)\varphi(z-c) - (z+c)\varphi(z+c))$$

In particular,  $f(0) = \varphi(c)$ ,  $f'(0) = 0$  and

$$\mu_{(0)}(z) = \frac{M(\varphi(z-c) - \varphi(z+c))}{\varphi(z-c) - \varphi(z+c)}$$

Hence  $\mu'_{(0)}(z) = Mc$ . Now take  $c$  such that  $\varphi(c) = f(z)$ , i.e.,  $c = \sqrt{-\log(2\pi f^2(z))}$  to get:

$$\mu'_{(0)}(0) = M \cdot \sqrt{-\log(2\pi f^2(z))}$$

**Lower bound: Second derivative.** In this case we take  $G = \frac{1-w}{2} \cdot (\delta_{-c} + \delta_c) + w \cdot \delta_0$  for parameters  $w \in [0, 1]$ ,  $c > 0$  which we will specify later and  $\alpha_{(0)}(u) = 2M \cdot \mathbf{1}(u = 0)$ . Then:

$$f(z) = \frac{1-w}{2} \cdot (\varphi(z-c) + \varphi(z+c)) + w \cdot \varphi(z), \quad \mu_{(0)}(z) = 2 \cdot M \cdot w \cdot \frac{\varphi(z)}{f(z)}$$

By direct calculation we can verify that

$$\mu''(0) = -2 \cdot M \cdot w \frac{\varphi(0)f(0) + \varphi(0)f''(0)}{f^2(0)}, \quad f''(0) = (1-w)(c^2 - 1)\varphi(c) - w\varphi(0)$$

To get a lower bound computationally, we conduct a grid search over  $c$  and  $w$  and find the parameter values that maximize  $|\mu''(0)|$  subject to the constraint that  $f(0) \geq \rho$ .

For the analytic lower bound, we proceed as follows: We choose  $w = \varphi(c)$ , so that  $f(0) = (1 + \varphi(0) - \varphi(c))\varphi(c)$  and

$$\mu''(0) = -2 \cdot M \cdot \varphi(0) \frac{(1 - \varphi(c)) \cdot c^2}{(1 + \varphi(0) - \varphi(c))^2}$$

We now pick  $c$  so that  $\varphi(c) = \rho$ . It then holds in particular that  $f(0) \geq \rho$  and also:

$$|\mu''(0)| \geq \frac{M}{5} c^2 = \frac{M}{5} (-\log(2\pi \rho^2))$$

□

## A.2 Proof of Theorem 5

*Proof.* Notation: We use  $\mathbb{E}_n[\cdot]$  to denote empirical averages. We omit dependence on  $n$  of the weighting kernels. Finally, we write  $\gamma_+(z) = \gamma_+(z)\mathbf{1}(z \geq c)$ . We only prove a central limit theorem for  $\hat{\mu}_{\gamma,+}$ . The CLT for  $\hat{\mu}_{\gamma,-}$  and  $\hat{\tau}_\gamma = \hat{\mu}_{\gamma,+} - \hat{\mu}_{\gamma,-}$  follow similarly.

**CLT for  $\sum_i \gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})$ :** We seek to prove the following central limit Theorem:

$$\frac{\sum_{i=1}^n \gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})}{\sqrt{n\mathbb{E}[\gamma_+(Z_i)^2(Y_i(1) - \mu_{\gamma,+})^2]}} \Rightarrow \mathcal{N}(0, 1)$$

To this end, first note that the numerator has expectation 0, since:

$$\mathbb{E}[\gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})] = \mathbb{E}[\gamma_+(Z_i)Y_i(1)] - \mathbb{E}[\gamma_+(Z_i)] \frac{\mathbb{E}[\gamma_+(Z_i)Y_i(1)]}{\mathbb{E}[\gamma_+(Z_i)]} = 0.$$

Next we will check the condition of Lyapunov's central limit theorem.

$$\begin{aligned} \text{Var}[\gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})] &\geq \mathbb{E}[\text{Var}[\gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+}) \mid Z_i]] \\ &= \mathbb{E}[\gamma_+(Z_i)^2 \text{Var}[Y_i(1) - \mu_{\gamma,+} \mid Z_i]] \\ &= \mathbb{E}[\gamma_+(Z_i)^2 \text{Var}[Y_i(1) \mid Z_i]] \\ &\geq \sigma^2 \mathbb{E}[\gamma_+(Z_i)^2] \\ &\geq \bar{\sigma}^2 \mathbb{E}[\gamma_+(Z_i)]^2 \end{aligned} \tag{45}$$

$$\begin{aligned} \mathbb{E}[|\gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})|^{2+q}] &= \mathbb{E}[|\gamma_+(Z_i)|^{2+q} \mathbb{E}[|(Y_i(1) - \mu_{\gamma,+})|^{2+q} \mid Z_i]] \\ &\leq \bar{\sigma}^{2+q} \mathbb{E}[|\gamma_+(Z_i)|^{2+q}] \\ &\leq \max_z |\gamma_+(z)|^q \cdot \bar{\sigma}^{2+q} \cdot \mathbb{E}[\gamma_+(Z_i)^2] \end{aligned}$$

So:

$$\begin{aligned} \frac{n\mathbb{E}[|\gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})|^{2+q}]}{(n\text{Var}[\gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})])^{(2+q)/2}} &\leq \frac{\max_z |\gamma_+(z)|^q \cdot \bar{\sigma}^{2+q} \cdot \mathbb{E}[\gamma_+(Z_i)^2]}{n^{q/2} \bar{\sigma}^{2+q} \mathbb{E}[\gamma_+(Z_i)^2]^{(2+q)/2}} \\ &\leq \left(\frac{\bar{\sigma}}{\sigma}\right)^{2+q} \cdot \frac{\max_z |\gamma_+(z)|^q}{n^{q/2} \mathbb{E}[\gamma_+(Z_i)^2]^{q/2}} \\ &\leq \left(\frac{\bar{\sigma}}{\sigma}\right)^{2+q} \cdot \frac{\max_z |\gamma_+(z)|^q}{n^{q/2} \mathbb{E}[\gamma_+(Z_i)]^q} \\ &\leq \left(\frac{\bar{\sigma}}{\sigma}\right)^{2+q} \cdot (Cn^{\beta-1/2})^q \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

This proves the central limit theorem.



**Estimation of normalization factor:** Here we prove that  $\mathbb{E}_n [\gamma_+(Z_i)] / \mathbb{E} [\gamma_+(Z_i)] = 1 + o_P(1)$ . For any  $\varepsilon > 0$ , by Chebyshev's inequality:

$$\begin{aligned} \mathbb{P} [|\mathbb{E}_n [\gamma_+(Z_i)] - \mathbb{E} [\gamma_+(Z_i)]| \geq \varepsilon \mathbb{E} [\gamma_+(Z_i)]] &\leq \frac{\text{Var} [\gamma_+(Z_i)]}{n\varepsilon^2 \mathbb{E} [\gamma_+(Z_i)]^2} \\ &\leq \frac{\max_z \gamma_+(z)^2}{n\varepsilon^2 \mathbb{E} [\gamma_+(Z_i)]^2} \\ &\leq \left( \frac{C}{\varepsilon} \cdot n^{\beta-1/2} \right)^2 \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

**CLT for  $\hat{\mu}_{\gamma,+}$ :** Note that

$$\hat{\mu}_{\gamma,+} - \mu_{\gamma,+} = \frac{\sum_{i=1}^n \gamma_+(Z_i)(Y_i(1) - \mu_{\gamma,+})}{\sum_{i=1}^n \gamma_+(Z_i)}$$

The above display, along with our preceding results and Slutsky yield the CLT:

$$\frac{\sqrt{n}(\hat{\mu}_{\gamma,+} - \mu_{\gamma,+})}{\sqrt{\mathbb{E} [\gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2] / \mathbb{E} [\gamma_+(Z_i)]^2}} \Rightarrow \mathcal{N}(0, 1)$$

□

### A.3 Proof of Corollary 6

First note that by the Assumption of this Corollary on the bias, there exists  $\eta_n = o_P(n^{-1/2}V_\gamma^{1/2})$  such that:

$$b_\gamma = \tilde{b}_\gamma + \eta_n, \quad |\tilde{b}_\gamma| \leq \hat{B}_\gamma$$

Then, Theorem 5, along with Slutsky (and recalling that  $b_\gamma = \tau_\gamma - \tau$ ) imply that:

$$\frac{\sqrt{n}(\hat{\tau}_\gamma - \tau - \tilde{b}_\gamma)}{\hat{V}_\gamma^{1/2}} \Rightarrow \mathcal{N}(0, 1)$$

So, letting  $\tilde{Z} \sim \mathcal{N}(0, 1)$  independent of everything else:

$$\begin{aligned} \mathbb{P} [\tau \in \hat{\tau}_\gamma \pm \ell_\alpha] &= \mathbb{P} [-\ell_\alpha - \tilde{b}_\gamma \leq \hat{\tau}_\gamma - \tau - \tilde{b}_\gamma \leq \ell_\alpha - \tilde{b}_\gamma] \\ &= \mathbb{P} [-\sqrt{n}\hat{V}_\gamma^{-1/2}(\ell_\alpha + \tilde{b}_\gamma) \leq \sqrt{n}\hat{V}_\gamma^{-1/2}(\hat{\tau}_\gamma - \tau - \tilde{b}_\gamma) \leq \sqrt{n}\hat{V}_\gamma^{-1/2}(\ell_\alpha - \tilde{b}_\gamma)] \\ &\stackrel{(i)}{=} \mathbb{E} \left[ \mathbb{P} [-\sqrt{n}\hat{V}_\gamma^{-1/2}(\ell_\alpha + \tilde{b}_\gamma) \leq \tilde{Z} \leq \sqrt{n}\hat{V}_\gamma^{-1/2}(\ell_\alpha - \tilde{b}_\gamma) \mid \hat{V}_\gamma, \hat{B}_\gamma, \hat{\tau}_\gamma] \right] + o(1) \\ &= \mathbb{E} \left[ \mathbb{P} [-\ell_\alpha \leq \tilde{b}_\gamma + n^{-1/2}\hat{V}_\gamma^{1/2}\tilde{Z} \leq \ell_\alpha \mid \hat{V}_\gamma, \hat{B}_\gamma, \hat{\tau}_\gamma] \right] + o(1) \\ &\stackrel{(ii)}{\geq} \mathbb{E} [1 - \alpha] + o(1) \\ &= 1 - \alpha + o(1) \end{aligned}$$

In (i) we used the fact that the central limit theorem implies that the distribution function of the (asymptotic) pivot converges to the standard Normal distribution function  $\Phi(\cdot)$  uniformly. In (ii) we used the definition of  $\ell_\alpha$  in (23) and the fact that  $|\tilde{b}_\gamma| \leq \hat{B}_\gamma$ .

## A.4 Proof of Proposition 7

*Proof.* The proof here continues from the argument used for the proof of Theorem 5. As we did there, we only prove the result for the variance of  $\hat{\mu}_{\gamma,+}$ , the result for  $\hat{\tau}_\gamma$  follows in the same way. Let us start by arguing that:

$$\mathbb{E}_n \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right] / \mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right] = 1 + o_P(1).$$

By the inequality of von Bahr and Esseen [1965], there exists a constant  $C_q < \infty$ , such that:

$$\begin{aligned} & \mathbb{E} \left[ \left| \sum_{i=1}^n \left\{ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 - \mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right] \right\} \right|^{1+q/2} \right] \\ & \leq C_q \sum_{i=1}^n \mathbb{E} \left[ \left| \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 - \mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right] \right|^{1+q/2} \right] \\ & \leq n \tilde{C}_q \mathbb{E} \left[ |\gamma_+(Z_i) (Y_i(1) - \mu_{\gamma,+})|^{2+q} \right] \end{aligned}$$

In the last step we used Jensen's inequality and  $\tilde{C}_q$  is a finite positive constant. The above display then is equivalent to:

$$\frac{\mathbb{E} \left[ \left| (\mathbb{E}_n - \mathbb{E}) \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right] \right|^{1+q/2} \right]}{\mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right]^{(2+q)/2}} \leq \frac{\tilde{C}_q \mathbb{E} \left[ |\gamma_+(Z_i) (Y_i(1) - \mu_{\gamma,+})|^{2+q} \right]}{n^{q/2} \mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right]^{(2+q)/2}}$$

This is precisely the expression we already showed converges uniformly to 0 during the verification of Lyapunov's condition in the proof of Theorem 5. It remains to show that the feasible estimator is also asymptotically equivalent. To this end note the decomposition.

$$\begin{aligned} & \gamma_+(Z_i)^2 (Y_i(1) - \hat{\mu}_{\gamma,+})^2 - \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \\ & = \gamma_+(Z_i)^2 (\hat{\mu}_{\gamma,+} - \mu_{\gamma,+})^2 + 2\gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+}) (\mu_{\gamma,+} - \hat{\mu}_{\gamma,+}) \end{aligned}$$

From the CLT of Theorem 5, we know that:

$$(\hat{\mu}_{\gamma,+} - \mu_{\gamma,+})^2 = O_P \left( n^{-1} \mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right] / \mathbb{E} \left[ \gamma_+(Z_i)^2 \right] \right) = O_P \left( n^{-1+2\beta} \right) = o_P(1)$$

And so:

$$\frac{\frac{1}{n} \sum_{i=1}^n \gamma_+(Z_i)^2}{\mathbb{E} \left[ \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2 \right]} \cdot (\hat{\mu}_{\gamma,+} - \mu_{\gamma,+})^2 = O_P(1) \cdot o_P(1) = o_P(1)$$

The fact that the first term is  $O_P(1)$  follows by arguing with Chebyshev's inequality. First note that from (45), we know that  $\mathbb{E}[\gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2] \geq \sigma^2 \mathbb{E}[\gamma_+(Z_i)^2]$  and so it suffices to show that  $\mathbb{E}_n [\gamma_+(Z_i)^2] / \mathbb{E} [\gamma_+(Z_i)^2]$  is  $O_P(1)$ . Indeed this term is also  $1 + o_P(1)$ ,

since for any  $\varepsilon > 0$ :

$$\begin{aligned} \mathbb{P} \left[ \left| \mathbb{E}_n [\gamma_+(Z_i)^2] - \mathbb{E} [\gamma_+(Z_i)^2] \right| \geq \varepsilon \mathbb{E} [\gamma_+(Z_i)^2] \right] &\leq \frac{\text{Var} [\gamma_+(Z_i)^2]}{n\varepsilon^2 \mathbb{E} [\gamma_+(Z_i)^2]^2} \\ &\leq \frac{\max_z \gamma_+(z)^2}{n\varepsilon^2 \mathbb{E} [\gamma_+(Z_i)^2]^2} \cdot \frac{\mathbb{E} [\gamma_+(Z_i)^2]}{\mathbb{E} [\gamma_+(Z_i)^2]} \\ &\leq \left( \frac{C}{\varepsilon} \cdot n^{\beta-1/2} \right)^2 \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned}$$

This proves the first term is negligible. To show that the second term is negligible, our basic argument is that

$$\frac{\frac{1}{n} \sum_{i=1}^n \gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})}{\mathbb{E} [\gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2]} \cdot (\hat{\mu}_{\gamma,+} - \mu_{\gamma,+}) = O_P(1) \cdot o_P(1) = o_P(1),$$

and it remains to prove that the first term is indeed  $O_P(1)$ . Now by Cauchy-Schwarz

$$\begin{aligned} \left| \mathbb{E}_n [\gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})] \right| &= \left| \mathbb{E}_n [\gamma_+(Z_i) \cdot \gamma_+(Z_i) (Y_i(1) - \mu_{\gamma,+})] \right| \\ &\leq \left( \mathbb{E}_n [\gamma_+(Z_i)^2] \right)^{1/2} \left( \mathbb{E}_n [\gamma_+(Z_i)^2 (Y_i(1) - \mu_{\gamma,+})^2] \right)^{1/2} \end{aligned}$$

But the above is the product of two  $O_P(\mathbb{E}[\gamma_+(Z_i)^2(Y_i(1) - \mu_{\gamma,+})^2]^{1/2})$  terms (as we showed above), so we conclude upon dividing by  $\mathbb{E}[\gamma_+(Z_i)^2(Y_i(1) - \mu_{\gamma,+})^2]$ .  $\square$

## A.5 Proof of Corollary 8 and Corollary 11

*Proof.* For generality, we here consider the case where  $\tau(u)$  may not be constant and so, as discussed in Remark 3, our estimator should be interpreted as targeting  $\tau_{h,+}$ . The asymptotic expectation of the estimator is equal to (letting formally  $\gamma_+(z)$  to be 0 for  $z < c$  and  $\gamma_-(z) = 0$  for  $z \geq c$ ).

$$\tau_\gamma = \mu_{\gamma,+} - \mu_{\gamma,-} = \frac{\mathbb{E} [\gamma_+(Z_i)Y_i(1)]}{\mathbb{E} [\gamma_+(Z_i)]} - \frac{\mathbb{E} [\gamma_-(Z_i)Y_i(0)]}{\mathbb{E} [\gamma_-(Z_i)]}$$

$\tau_\gamma$  remains invariant upon translating  $Y_i(w) \mapsto Y_i(w) - \mu_{\gamma,-}$ . Thus let  $\tilde{Y}_i(w)$  be the shifted  $Y_i(w)$ 's, i.e.,  $\tilde{Y}_i(w) = Y_i(w) - \mu_{\gamma,-}$ . Similarly let  $\tilde{\alpha}_{(w)}(u) = \mathbb{E}[\tilde{Y}_i(w) | U_i = u]$  be the shifted conditional response surfaces. Then:

$$\tau_\gamma = \frac{\mathbb{E} [\gamma_+(Z_i)\tilde{Y}_i(1)]}{\mathbb{E} [\gamma_+(Z_i)]} - \frac{\mathbb{E} [\gamma_-(Z_i)\tilde{Y}_i(0)]}{\mathbb{E} [\gamma_-(Z_i)]} = \frac{\mathbb{E} [\gamma_+(Z_i)\tilde{Y}_i(1)]}{\mathbb{E} [\gamma_+(Z_i)]} - 0 = \frac{\mathbb{E} [\gamma_+(Z_i)\tilde{Y}_i(1) - \gamma_-(Z_i)\tilde{Y}_i(0)]}{\mathbb{E} [\gamma_+(Z_i)]}$$

It is helpful to note that:

$$\begin{aligned} \mathbb{E} [h_+(U_i)] &= \int \int_{[c,\infty)} \gamma_+(z)p(z|u)d\lambda(z)dG(u) = \int_{[c,\infty)} \gamma_+(z) \int p(z|u)dG(u)d\lambda(z) \\ &= \int_{[c,\infty)} \gamma_+(z)dF(Z) = \mathbb{E} [\gamma_+(Z_i)] \end{aligned}$$

Now, following the Proof of Theorem 1 verbatim and noting that  $\tau(u) = \tilde{\alpha}_{(1)}(u) - \tilde{\alpha}_{(0)}(u)$ , we get:

$$\begin{aligned}
& \mathbb{E} \left[ \gamma_+(Z_i) \tilde{Y}_i(1) - \gamma_-(Z_i) \tilde{Y}_i(0) \right] \\
&= \mathbb{E} [h_+(U_i) \tilde{\alpha}_{(1)}(U_i)] - \mathbb{E} [h_-(U_i) \tilde{\alpha}_{(0)}(U_i)] \\
&= \mathbb{E} [h_+(U_i) \tau(U_i)] + \mathbb{E} [\tilde{\alpha}_{(0)}(U_i) (h_+(U_i) - h_-(U_i))] \\
&= \mathbb{E} [h_+(U_i) \tau(U_i)] + \mathbb{E} [(\alpha_{(0)}(U_i) - \mu_{\gamma,-}) (h_+(U_i) - h_-(U_i))]
\end{aligned}$$

So we have

$$\tau_\gamma = \frac{\mathbb{E} [h_+(U_i) \tau(U_i)] + \mathbb{E} [(\alpha_{(0)}(U_i) - \mu_{\gamma,-}) (h_+(U_i) - h_-(U_i))]}{\mathbb{E} [h_+(U_i)]}$$

and

$$\begin{aligned}
|\tau_\gamma - \tau_{h,+}| &\leq \frac{\mathbb{E} [ |(\alpha_{(0)}(U_i) - \mu_{\gamma,-}) (h_+(U_i) - h_-(U_i))| ]}{\mathbb{E} [h_+(U_i)]} \\
&\leq \frac{M \mathbb{E} [|h_+(U_i) - h_-(U_i)|]}{\mathbb{E} [h_+(U_i)]}.
\end{aligned}$$

This proves (26). Further, we have

$$\begin{aligned}
& \left| \frac{\mathbb{E} [h_+(U_i) \tau(U_i)]}{\mathbb{E} [h_+(U_i)]} - \frac{\mathbb{E} [w(U_i) \tau(U_i)]}{\mathbb{E} [w(U_i)]} \right| \\
&= \left| \frac{\mathbb{E} [h_+(U_i) \tau(U_i)]}{\mathbb{E} [h_+(U_i)]} - \frac{\mathbb{E} [\bar{w}(U_i) \tau(U_i)]}{\mathbb{E} [\bar{w}(U_i)]} \right| \\
&= \left| \frac{\mathbb{E} [h_+(U_i) \tau(U_i)] - \mathbb{E} [\bar{w}(U_i) \tau(U_i)]}{\mathbb{E} [h_+(U_i)]} + \mathbb{E} [\bar{w}(U_i) \tau(U_i)] \left( \frac{1}{\mathbb{E} [h_+(U_i)]} - \frac{1}{\mathbb{E} [\bar{w}(U_i)]} \right) \right| \\
&\leq M' \frac{\mathbb{E} [|h_+(U_i) - \bar{w}(U_i)|]}{\mathbb{E} [h_+(U_i)]} + M' \mathbb{E} [\bar{w}(U_i)] \left| \frac{1}{\mathbb{E} [h_+(U_i)]} - \frac{1}{\mathbb{E} [\bar{w}(U_i)]} \right| \\
&= \frac{M' \mathbb{E} [|h_+(U_i) - \bar{w}(U_i)|] + M' \mathbb{E} [h_+(U_i) - \bar{w}(U_i)]}{\mathbb{E} [h_+(U_i)]}
\end{aligned}$$

Along with the triangle inequality, this proves (36). □

## A.6 Proof of Proposition 9

*Proof.* Consider the event  $\{G \in \mathcal{G}_n\}$ . On this event, in view of Corollary 8, it holds for  $b_\gamma = \tau_\gamma - \tau$ , that

$$|b_\gamma| \leq \frac{\int M |h_+(u) - h_-(u)| dG(u)}{\int h_+(u) dG(u)} \leq \sup_{\tilde{G} \in \mathcal{G}_n} \frac{\int M |h_+(u) - h_-(u)| d\tilde{G}(u)}{\int h_+(u) d\tilde{G}(u)} = \widehat{B}_\gamma.$$

This implies that  $\{G \in \mathcal{G}_n\} \subset \{|b_\gamma| \leq \widehat{B}_\gamma\}$  and so  $\mathbb{P} [ |b_\gamma| \leq \widehat{B}_\gamma ] \geq \mathbb{P} [G \in \mathcal{G}_n]$ . It thus suffices to show that the RHS converges to 1 as  $n \rightarrow \infty$ . By construction of  $\mathcal{G}_n$  in (27) and Massart's tight constant for the DKW inequality [Massart, 1990], it holds that

$$\mathbb{P} [G \in \mathcal{G}_n] \geq \mathbb{P} \left[ \sup_{t \in \mathbb{R}} |F(t) - \widehat{F}_n(t)| \leq \sqrt{\log(2/\alpha_n)/(2n)} \right] \geq 1 - \alpha_n$$

Since  $\alpha_n \rightarrow 0$ , we conclude with the proof of the first statement of the proposition. The second statement also follows, since for any  $\varepsilon > 0$ :

$$\mathbb{P} \left[ \sqrt{n} \left( \widehat{B}_\gamma - |b_\gamma| \right) / \sqrt{V_\gamma} \leq -\varepsilon \right] \leq \mathbb{P} \left[ \widehat{B}_\gamma < |b_\gamma| \right] \rightarrow 0 \text{ as } n \rightarrow \infty$$

□

## A.7 Proof of Proposition 10

*Proof.* Let us look only at  $\gamma_+$ , the results for  $\gamma_-$  are analogous. We need to check that there exists a constant  $\tilde{C} > 0$  and  $\tilde{\beta} \in (0, 1/2)$  such that the event  $A_n$  has asymptotic probability 1, where:

$$A_n = \left\{ 0 < \max_z \left| \gamma_+^{(n)}(z) \right| \leq n^{\tilde{\beta}} \cdot \tilde{C} \cdot \mathbb{E} \left[ \gamma_+^{(n)}(Z_i) \right] \right\}$$

We will show that we can use  $\tilde{\beta} = \beta$  and  $\tilde{C} = C/\delta$ , where  $C, \beta$  are specified in constraint (33) of the optimization problem and  $\delta$  is defined in (35).

To see this, first note that  $\max_z \left| \gamma_+^{(n)}(z) \right| > 0$  must hold, otherwise constraint (32) of the optimization problem would not be satisfied. Second, note that on the event

$$B_n = \left\{ \int_{[c, \infty)} \gamma_+(z) dF(z) \geq \delta \right\},$$

from (35), indeed it holds that:

$$\max_z \left| \gamma_+^{(n)}(z) \right| \leq C n^\beta \leq \tilde{C} n^\beta \cdot \delta \leq \tilde{C} n^\beta \int_{[c, \infty)} \gamma_+(z) dF(z)$$

Thus  $\mathbb{P} [A_n] \geq \mathbb{P} [B_n] \rightarrow 1$  as  $n \rightarrow \infty$  and the weights are regular. □