Noise-Induced Randomization in Regression Discontinuity Designs*

Dean Eckles
MIT

Nikolaos Ignatiadis†
Stanford University

Stefan Wager
Stanford University

Han Wu
Stanford University

Draft version April 2022

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, or other noise, in the running variable. Under an assumption that the measurement error is exogenous, we show how to 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 inference for 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 both 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

*We thank Alex D'Amour, Timothy Armstrong, Matias Cattaneo, Jan Gleixner, David Hirshberg, Guido Imbens, Michal Kolesár, Fan Li, Charles Manski, Fabrizia Mealli, Jamie Robins, Jasjeet Sekhon, Johan Ugander, and José Zubizarreta for helpful suggestions and discussions. Some of the computing for this project was performed on the Sherlock cluster with support from the Stanford Research Computing Center.

†Authors are listed in alphabetical order. Corresponding author: ignat@stanford.edu

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.

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], we 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 $\tau_c = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, Z_i = c\right]$ via

$$\tau_c = \lim_{z \downarrow c} \mathbb{E}\left[Y \,\middle|\, Z = z\right] - \lim_{z \uparrow c} \mathbb{E}\left[Y \,\middle|\, Z = z\right], \tag{1}$$

provided that the conditional response functions $\mu_{(w)}(z) = \mathbb{E}\left[Y(w) \,\middle|\, Z = z\right]$ 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 τ_c with well understood asymptotics [Armstrong and Kolesár, 2018, 2020, 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 appeal and rigorous justification, 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, 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 weighted treatment effects. 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—at least in the case of bounded outcomes—driven entirely by random treatment assignment induced by noise in Z_i.

1.1 A latent variable model for regression discontinuity designs

Throughout this paper, we consider the classical sharp regression discontinuity design with potential outcomes as described below:

Assumption 1 (Sharp regression discontinuity design)**.** There are $i = 1, \ldots, n$ 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 $W_i = 1\left(\{Z_i \geq c\}\right)$. For each sample, we observe pairs $\{Y_i,\, Z_i\}$ with $Y_i = Y_i(W_i)$.

The pre-requisite for applying our approach is the existence of domain-specific knowledge about the distribution of the running variable Z_i, as formalized in the following:[1]

Assumption 2 (Noisy running variable)**.** There is a latent variable U_i with (unknown) distribution G such that $Z_i \,\big|\, U_i \sim p(\cdot \,\big|\, U_i)$ for a known conditional density $p(\cdot \,\big|\, \cdot)$ with respect to a measure λ.

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.,

$$Z_i \,\big|\, U_i \sim \mathcal{N}\left(U_i,\, \nu^2\right), \qquad \nu > 0; \tag{2}$$

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

We also require for the additional 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., $\left[\{Y_i(0),\, Y_i(1)\} \perp\!\!\!\perp Z_i\right] \,\big|\, U_i$.

[1] The idea that explicit structural modeling is valuable for causal inference has a long tradition in economics, going back to Roy [1951] and Heckman [1979], with recent developments by e.g., Heckman and Vytlacil [2005], Brinch, Mogstad, and Wiswall [2017] and Mogstad, Santos, and Torgovitsky [2018]. At a high level, our work can be seen as connecting this tradition to the regression discontinuity design, and demonstrating how structural assumptions enable inference of policy-relevant causal estimands.

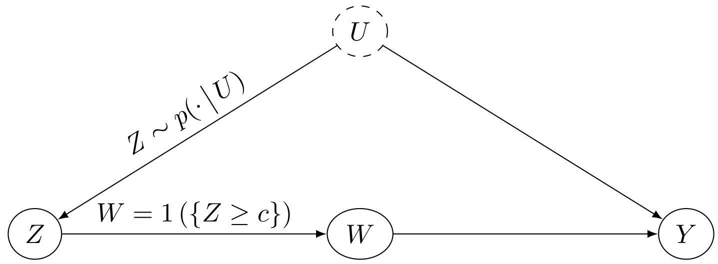

Figure 1: Graphical illustration of the sharp regression discontinuity design with a noisy running variable. U is an unobserved latent variable with unknown distribution G and Z is the running variable with known density $p(\cdot \mid U)$ conditionally on U. Treatment is assigned deterministically as $W = \mathbf{1}(\{Z \geq c\})$ for a known cutoff c and $Y = Y(W)$ is the observed response.

An implication of Assumption 3 is that

$$\mathbb{E}\left[Y_i \,\middle|\, U_i,\, Z_i\right] = \alpha_{(W_i)}(u), \qquad \alpha_{(w)}(u) = \mathbb{E}\left[Y_i(w) \,\middle|\, U_i = u\right], \tag{3}$$

where the $\alpha_{(w)}(u)$ are the response functions for the potential outcomes conditionally on the latent variable u. Following Frangakis and Rubin [2002] we can think of u as indexing over unobserved principal strata; see also Heckman and Vytlacil [2005].

A graphical illustration of our assumptions is presented in Figure 1. In view of Assumptions 2 and 3, the key argument for our identification, estimation and inference strategy is captured by the following proposition.

Proposition 1. *Suppose that $\mathbb{E}\left[Y^2\right] < \infty$ and let $\gamma_+(\cdot), \gamma_-(\cdot)$ be measurable functions of Z with $\mathbb{E}\left[\gamma_-(Z)^2\right], \mathbb{E}\left[\gamma_+(Z)^2\right] < \infty$ such that $\gamma_+(z) = 0$ for $z < c$, $\gamma_-(z) = 0$ for $z \geq c$. Then, under Assumptions 1, 2, and 3:*

$$\mathbb{E}\left[\gamma_+(Z)Y\right] = \mathbb{E}\left[\alpha_{(1)}(U)h(U,\, \gamma_+)\right], \;\; \mathbb{E}\left[\gamma_-(Z)Y\right] = \mathbb{E}\left[\alpha_{(0)}(U)h(U,\, \gamma_-)\right], \tag{4}$$

where

$$h(u,\, \gamma) := \int \gamma(z)p(z \mid u)d\lambda(z). \tag{5}$$

We will apply this result by choosing functions γ_+, γ_- and then averaging the response $Y_i(1)$ of treated units with weights $\gamma_+(Z_i)$ and the response $Y_i(0)$ of control units with weights $\gamma_-(Z_i)$. While there is no overlap between treated and control units in a regression discontinuity design, Proposition 1 establishes that by weighting treated units by γ_+ and control units by γ_- we may achieve balance in the latent variable, as long as $h(\cdot,\, \gamma_+) \approx h(\cdot,\, \gamma_-)$.

Our assumptions do not impose any restriction on G, which is a property of the population studied in the regression discontinuity design; however, we require some precise knowledge about the noise distribution $p(z \mid u)$. Such knowledge is a prerequisite for applying our approach and it may be available for example from test–retest data, prior modeling of item-level responses to tests, a physical model for the measurement device, or biomedical knowledge. All our results also remain valid if we work with a noise distribution $\hat{p}(z \mid u)$ that underestimates the true noise level, in the sense that $p(z \mid u) = \int \hat{p}(z \mid u')\lambda(u' \mid u)\, du'$

for some distribution function $\lambda(u' \,|\, u)$ that captures the noise left out by $\hat{p}(\cdot)$.[2] For example, if the true noise process involves heteroskedastic Gaussian measurement errors $Z_i \,|\, U_i \sim \mathcal{N}\left(U_i,\, \nu_i^2\right)$, where U_i and ν_i may be correlated, then our approach would remain valid if we posit a homoskedastic noise model $Z_i \,|\, U_i \sim \mathcal{N}\left(U_i,\, \hat{\nu}^2\right)$ 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, with Gaussian errors, one can estimate the noise scale $\hat{\nu}^2$ by considering a conservative lower bound on measurement reliability obtained via repeated measurement, e.g., in education via repeatedly administering similar tests or in medicine by repeatedly administering the same diagnostic.

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]

$$\hat{\tau}_c = \underset{\tau}{\operatorname{argmin}} \left\{ \sum_{i=1}^n K\left(\frac{|Z_i - c|}{h_n}\right) \left(Y_i - a - \tau W_i - \beta_- \left(Z_i - c\right)_- - \beta_+ \left(Z_i - c\right)_+\right)^2 \right\}, \tag{6}$$

where $K(\cdot)$ is a weighting function, $h_n \to 0$ is a bandwidth, and a and $\beta_\pm$ are nuisance parameters. In general, this approach can be used for valid estimation and inference of τ_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 (6) 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., $W_i = 1\left(\{Z_i \in \mathcal{A}\}\right)$.

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

$$\left[Z_i \perp\!\!\!\perp \{Y_i(0),\, Y_i(1)\}\right] \;\Big|\; \{Z_i \in \mathcal{I}\}. \tag{7}$$

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

[2] 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, underestimating 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').

in the sense of Rubin [2008]. In practice, however, the assumption (7) is often unrealistic and limits the applicability of methods relying on it [Sekhon and Titiunik, 2017]. One testable implication of (7) 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.[3] Furthermore, it is not clear how to choose the interval $\mathcal{I}$ used in (7) 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 (7) 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.

Knowledge of the presence of measurement error (or other noise) in running variables is often mentioned [Bor et al., 2014, 2017, Fraga and Merseth, 2016, Harlow et al., 2020, Lee, 2008], yet this side-information is typically not directly used for inference. In a rare quantitative use of information about measurement error, Fraga and Merseth [2016] make explicit use of margin of error statistics provided by the Census Bureau for the fraction or size of a voting-aged population that has limited English proficiency; they report some analyses using only units that are within a 90% margin of error of the cutoff. Trochim, Cappelleri, and Reichardt [1991] studied measurement error under an assumed (e.g., linear) outcome model, and showed that its presence does not induce bias.

Closer to our approach, Rokkanen [2015] considers the regression discontinuity design under Assumptions 2 and 3. Instead of assuming prior knowledge of the noise distribution $p(\cdot \mid u)$, Rokkanen [2015] assumes that for each unit in the design we observe at least two noisy measurements Z_i', Z_i'' of the underlying latent variable U_i in addition to the running variable Z_i. While Rokkanen [2015] provides conditions for the nonparametric identification of $\alpha_{(w)}(\cdot)$ in (3) and consequently of treatment effects, the estimation and inference strategy posits strong parametric assumptions, namely joint normality of (U_i, Z_i, Z_i', Z_i'') and linearity of $\alpha_{(w)}(u)$ as a function of u.[4] In contrast, in our work we assume knowledge of the noise distribution through e.g., biomedical knowledge or test–retest data, however we impose no parametric restrictions on G and $\alpha_{(w)}(u)$. Furthermore, we develop a practical and intuitive method for estimation and inference, that provides valid coverage even when treatment effects are only partially identified (e.g., when $p(\cdot \mid u)$ is finitely supported).

We also note a connection between our result and a line of research on treatment effect estimation under "biased allocation" or the "risk-based allocation design" [Bilodeau, 1997, Finkelstein, Levin, and Robbins, 1996a,b, Robbins and Zhang, 1988, 1989, 1991, Robbins, 1993]. As discussed further by Cook [2008], these authors appear to have effectively reinvented the regression discontinuity design without being aware of the work of Thistlethwaite and Campbell [1960] and subsequent developments. They focus on settings where sequential measurements of the same quantity function as both the running variable and the outcome; for example, Finkelstein et al. [1996b] discuss an application where patients with high blood cholesterol are given a drug whose purpose is to lower cholesterol, and we are interested in measuring the extent to which the drug succeeded in lowering the patients' blood cholesterol as measured at future visits. Then, in order to estimate treatment effects in this class of problems, they posit a noise model similar to the one we use,[5] together with a parametric

[3]Some authors, e.g., Sales and Hansen [2020], have argued that one can fix this issue by first de-trending outcomes, and then assuming (7) on the residuals. Any such approach, however, relies on well specification of the trend removal, and is thus no longer justified by randomization.

[4]Morell [2020] and Morell, Yang, and Liu [2020] consider fully parametric specifications for regression discontinuity designs with latent variables and demonstrate their utility in education research.

[5]The line of work on "biased allocation" was motivated from an empirical Bayes [Robbins, 1956] interpretation of the noise model in Assumption 2.

model linking the unobserved types U_i with expected outcomes. Robbins and Zhang [1989] study treatment effect estimation under what effectively amount to our Assumptions 2 and 3 as well as a requirement that noise is Gaussian and control potential outcomes are linked to U_i via an additive shift:

$$Z_i \,\big|\, U_i \sim \mathcal{N}\left(U_i,\, \nu^2\right), \qquad \alpha_{(0)}(u) = \mathbb{E}\left[Y_i(0) \,\big|\, U_i = u\right] = u + c, \qquad c \in \mathbb{R}. \tag{8}$$

Meanwhile Robbins and Zhang [1991] consider a Poisson noise model for the running variable together with a linear baseline model, $\alpha_{(0)}(u) = cu$ for some $c > 0$. The strong parametric assumptions on $\alpha_{(0)}(u)$ play a central role in their approach and—while potentially plausible in some applications involving sequential measurements of the same quantity—these parametric assumptions are not appropriate in examples considered in this paper. Thus, the methods developed in Robbins and Zhang [1988, 1989, 1991] and Finkelstein et al. [1996a,b] do not provide a methodological baseline for our approach. However, from a conceptual point of view, these papers present a notable yet largely overlooked chapter in the history of regression discontinuity designs.

Li, Mercatanti, Mäkinen, and Silvestrini [2021] study the regression discontinuity design with an ordinal running variable that, similar to our setting, is a noisy measurement of a latent variable U_i. However, Li et al. [2021] assume that U_i is a linear function of observed pre-treatment variables, and so inference can proceed by inverse-propensity weighting [Rosenbaum and Rubin, 1983] with estimated propensities $e(u) = \mathbb{P}\left[Z_i \geq c \mid U_i = u\right]$. In our setting, U_i is unobservable, and so, the propensities $e(U_i)$ are inaccessible. Our approach to inference does not involve inverse-propensity weighting; rather, we need to solve an integral equation in order to account for confounding.

A related, but distinct line of work studies the regression discontinuity design when the running variable is unobserved, and instead a noisy measurement thereof is observed [Bartalotti, Brummet, and Dieterle, 2020, Davezies and Le Barbanchon, 2017, Dong and Kolesár, 2021, Pei and Shen, 2016, Yanagi, 2014, Yu, 2012]. Identification becomes subtle and estimation can be difficult because of the perils of nonparametric estimation with measurement error [Meister, 2009]. Instead, we use measurement error as our identifying assumption; that is, the noise in our setup is beneficial for our estimation strategy rather than a barrier (and we observe the running variable). We also note a wider literature dealing with measurement error in causal inference beyond the regression discontinuity design, e.g., Pearl [2010], Kuroki and Pearl [2014], Jiang and Ding [2020].

Finally, we 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.

2 Ratio-form estimators and weighted treatment effects

In our approach to estimation and inference, motivated by Proposition 1, we consider ratio-form estimators,

$$\hat{\tau}_\gamma = \hat{\mu}_{\gamma,+} - \hat{\mu}_{\gamma,-}, \;\; \hat{\mu}_{\gamma,+} = \frac{\sum_i \gamma_+(Z_i) Y_i}{\sum_i \gamma_+(Z_i)}, \;\; \hat{\mu}_{\gamma,-} = \frac{\sum_i \gamma_-(Z_i) Y_i}{\sum_i \gamma_-(Z_i)}, \tag{9}$$

where γ_+, γ_- are pre-specified weighting functions such that $\gamma_+(z) = 0$ for $z < c$, $\gamma_-(z) = 0$ for $z \geq c$. (9) is a broad and intuitive class of estimators that includes, for example, the difference-in-means of units that are close to the cutoff (with the choice $\gamma_+(z) = \mathbf{1}(z \in [c, c+h])$ and $\gamma_-(z) = \mathbf{1}(z \in [c-h, c))$ for $h > 0$).

We seek to conduct inference for weighted treatment effects,

$$\tau_w = \int \frac{w(u)}{\mathbb{E}_G\left[w(U)\right]} \tau(u)\, dG(u), \quad w(\cdot) \geq 0, \tag{10}$$

where $\tau(u)$ is the conditional average treatment effect (CATE) of the stratum with $U_i = u$,

$$\tau(u) = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, U_i = u\right] = \alpha_{(1)}(u) - \alpha_{(0)}(u). \tag{11}$$

In the next two sections we take the choice of γ_+, γ_- as pre-specified by the researcher and seek to understand how to use the point estimate $\hat{\tau}_\gamma$ from (9) to form valid confidence intervals for τ_w (10) by also accounting for potential bias. In Section 4, we make a concrete recommendation for choosing γ_+, γ_-.

2.1 An asymptotic bias decomposition

We first derive the asymptotic limit of $\hat{\tau}_\gamma$ with fixed $\gamma_+(\cdot), \gamma_-(\cdot)$ given n i.i.d. copies of $(U_i, Z_i, Y_i(0), Y_i(1))$ satisfying Assumptions 1-3.

Theorem 2. *Suppose that Assumptions 1-3 hold and that $\mathbb{E}[\gamma_+(Z_i)^2]$, $\mathbb{E}[\gamma_-(Z_i)^2]$, $\mathbb{E}[Y_i^2]$ are finite. Then as $n \to \infty$, $\hat{\tau}_\gamma - \theta_\gamma \xrightarrow{\mathbb{P}} 0$, where:*

$$\theta_\gamma = \mu_{\gamma,+} - \mu_{\gamma,-}, \; \mu_{\gamma,+} = \frac{\mathbb{E}\left[\alpha_{(1)}(U) h(U,\, \gamma_+)\right]}{\mathbb{E}\left[h(U,\, \gamma_+)\right]}, \; \mu_{\gamma,-} = \frac{\mathbb{E}\left[\alpha_{(0)}(U) h(U,\, \gamma_-)\right]}{\mathbb{E}\left[h(U,\, \gamma_-)\right]}. \tag{12}$$

Proof. Apply the law of large numbers noting the result of Proposition 1 and that $\mathbb{E}\left[\gamma_+(Z)\right] = \mathbb{E}\left[h(U,\, \gamma_+)\right]$, $\mathbb{E}\left[\gamma_-(Z)\right] = \mathbb{E}\left[h(U,\, \gamma_-)\right]$. □

In view of Theorem 2 and the definition of τ_w in (10), we derive an asymptotic decomposition of the bias in estimating τ_w through $\hat{\tau}_\gamma$:

Corollary 3. *Under the conditions of Theorem 2, the asymptotic bias $\theta_\gamma - \tau_w$ can be decomposed as:*

$$\begin{aligned} \mathrm{Bias}\left[\gamma_\pm,\, \tau_w;\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right] = &\underbrace{\int \left(\frac{h(u,\, \gamma_+)}{\mathbb{E}_G\left[h(U,\, \gamma_+)\right]} - \frac{h(u,\, \gamma_-)}{\mathbb{E}_G\left[h(U,\, \gamma_-)\right]} \right) \alpha_{(0)}(u)\, dG(u)}_{\textit{Confounding bias}} \\ &+ \underbrace{\int \left(\frac{h(u,\, \gamma_+)}{\mathbb{E}_G\left[h(U,\, \gamma_+)\right]} - \frac{w(u)}{\mathbb{E}_G\left[w(U)\right]} \right) \tau(u) dG(u)}_{\textit{CATE heterogeneity bias}}. \end{aligned}$$

The bias decomposes into two terms. The first term, which we call 'Confounding bias', corresponds to how well we are balancing units through their latent variable u and will be small if $h(\cdot,\ \gamma_+) \approx h(\cdot,\ \gamma_-)$. The second term, which we call 'CATE heterogeneity bias' is equal to zero in the following two cases: when the CATE $\tau(u)$ is constant as a function of u, or when $h(u,\ \gamma_+) = w(u)$ for all u.

2.2 Examples of weighted treatment effects

The remainder of this section provides examples of statistical targets that may be expressed as in (10).

Regression discontinuity parameter: One statistical target that may be of interest is the standard regression discontinuity parameter τ_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 τ_c as in (10), note that by Bayes' rule,

$$\tau_c = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, Z_i = c\right] = \mathbb{E}\left[\tau(U_i) \,\middle|\, Z_i = c\right] = \int \tau(u) p(c \,|\, u) dG(u) / f(c), \tag{13}$$

where $f(c) = f_G(c) = \int p(c \mid u) dG(u)$ is the density of the running variable Z_i at the cutoff c. Thus, the representation from (10) holds with $w(u) = p(c \,|\, u)$ and $\mathbb{E}\left[w(U)\right] = f(c)$.

Another closely related target is $\tau_{c'}$ as defined in (13), but for some other value $c' \neq c$ of the running variable. Formally, this approach again fits within our setting, with $w(u) = p(c' \,|\, u)$ and $\mathbb{E}\left[w(U)\right] = f(c')$. 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, for example by noting that $\tau_{c'} \approx \tau_c + (d\tau_c/dc) \cdot (c' - c)$ [Dong and Lewbel, 2015].

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 $c' < c$.[6] 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 τ_π of patients affected by the treatment which, in this case, amounts to:

$$\tau_\pi = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, c' \leq Z_i < c\right] = \int_{[c',c)} \int \tau(u) p(z \,|\, u) dG(u) d\lambda(z) \Big/ \int_{[c',c)} dF(z), \tag{14}$$

where $F = F_G$ is the marginal Z-distribution (i.e., the distribution with $d\lambda$-density $f = f_G$). By Fubini's theorem, τ_π can be written in the form (10) with weight function $w(u) = \int_{[c',c)} p(z \,|\, u) d\lambda(z)$, $\mathbb{E}\left[w(U)\right] = \int_{[c',c)} dF(z)$.

[6] 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'.

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 ν 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 $W_i = 1\left(\{Z_i \geq c\}\right)$ for $Z_i \,\big|\, U_i \sim \mathcal{N}(U_i,\, \nu^2)$, and are considering a switch to a new treatment rule $W_i' = 1\left(\{Z_i' \geq c\}\right)$ based on a measurement $Z_i' \,\big|\, U_i \sim \mathcal{N}(U_i,\, {\nu'}^2)$ with a different noise level ν'. Writing $\Phi_\nu(\cdot)$ for the standard normal cumulative distribution function with variance ν^2, we see that the average treatment effect of patients who would be treated only with implementation of the policy change,[7] is equal to

$$\mathbb{E}\left[Y_i(1) - Y_i(0) \,\big|\, W_i' > W_i\right] = \frac{\int \tau(u)\left(1 - \Phi_{\nu'}\left(c-u\right)\right)\Phi_\nu\left(c-u\right)dG(u)}{\int \left(1 - \Phi_{\nu'}\left(c-u\right)\right)\Phi_\nu\left(c-u\right)dG(u)}, \tag{15}$$

which again is covered by (10).

3 Bias-aware confidence intervals

In the previous section, we discussed the asymptotic limit of the ratio-form estimator from (9) and the bias in estimating causal effects in regression discontinuity designs. In order to make use of such an estimator in practice, however, we also need to understand its sampling distribution and to control the bias. In this section, we describe our approach to inference.

We start by making the following additional assumption:[8]

Assumption 4 (Bounded response)**.** The response Y_i is bounded, $Y_i \in [0, 1]$.

We start by studying the asymptotic distribution of the weighted ratio estimator (9). We treat the weighting kernels γ_+, γ_- as deterministic but allow them to vary with n, i.e., $\gamma_+ = \gamma_+^{(n)}$ and $\gamma_- = \gamma_-^{(n)}$. Our first formal result is the following central limit theorem.

Theorem 4 (Asymptotic normality of ratio-form estimators)**.** *Suppose that the pairs $(Z_i,\, Y_i)$, $i = 1, \ldots, n$ are independent and identically distributed with $Y_i \in [0,1]$ (Assumption 4) and such that $\inf_z \operatorname{Var}\left[Y_i \,\big|\, Z_i = z\right] > 0$. Further suppose that the sequence of weighting kernels $\gamma_+^{(n)}$ and $\gamma_-^{(n)}$ is deterministic,[9] and that there exist $\beta \in (0, 1/2)$, $C, C' > 0$ such that for all n large enough:*

$$\sup_z \left|\gamma_\diamond^{(n)}(z)\right| < Cn^\beta \mathbb{E}\left[\gamma_\diamond^{(n)}(Z_i)\right], \quad \sup_u \left|h(u,\, \gamma_\diamond^{(n)})\right| < C'\mathbb{E}\left[\gamma_\diamond^{(n)}(Z_i)\right], \tag{16}$$

where $\diamond \in \{+, -\}$. Then, $\hat\tau_\gamma = \hat\tau_{\gamma^{(n)}}$ is asymptotically normal, i.e.,

$$\sqrt{n}\left(\hat\tau_\gamma - \theta_\gamma\right) \,/\, \sqrt{V_\gamma} \Rightarrow \mathcal{N}\left(0,\, 1\right),$$

where θ_γ is defined in (12) and

$$V_\gamma = \mathbb{E}\left[\gamma_+^2(Z_i)\left(Y_i - \mu_{\gamma,+}\right)^2\right] / \mathbb{E}\left[\gamma_+(Z_i)\right]^2 + \mathbb{E}\left[\gamma_-^2(Z_i)\left(Y_i - \mu_{\gamma,-}\right)^2\right] / \mathbb{E}\left[\gamma_-(Z_i)\right]^2. \tag{17}$$

[7] Here we assume that Z_i, Z_i' are independent conditionally on U_i.

[8] This assumption describes the most common use-case of our approach and streamlines exposition. It may be relaxed as follows: there exist known functions LB, UB such that $\mathrm{LB}(u) \leq \alpha_{(0)}(u),\, \alpha_{(1)}(u) \leq \mathrm{UB}(u)$ for all u.

[9] It suffices for $\gamma_+,\, \gamma_-$ to be independent of $\left(U_i, Z_i, Y_i(0), Y_i(1)\right), 1 \leq i \leq n$.

We note that the condition on the response noise is mild. The assumption on γ_+, γ_- is also easy to satisfy, and in particular the weights proposed in Section 4 will satisfy this property, as well as other choices of weighting functions.[10]

Given our result from Theorem 4, we can design confidence intervals for τ_w from (10). In doing so, we need to first account for the variance term V_γ as in (17):

Proposition 5. *Under the assumptions of Theorem 4, V_γ can be consistently estimated with the following plug-in estimator: $\widehat{V}_\gamma \,/\, V_\gamma = 1 + o_{\mathbb{P}}(1)$ for*

$$\widehat{V}_\gamma = \frac{\sum_i \gamma_+(Z_i)^2 \left(Y_i - \hat{\mu}_{\gamma,+}\right)^2}{n\left(\frac{1}{n}\sum_i \gamma_+(Z_i)\right)^2} + \frac{\sum_i \gamma_-(Z_i)^2 \left(Y_i - \hat{\mu}_{\gamma,-}\right)^2}{n\left(\frac{1}{n}\sum_i \gamma_-(Z_i)\right)^2}, \tag{18}$$

where $\hat{\mu}_{\gamma,+}, \hat{\mu}_{\gamma,-}$ are defined in (9).

Second, we need to account for the potential bias $|b_\gamma| = |\theta_\gamma - \tau_w|$. Here, we will not assume that the bias is negligible (i.e., we do not assume "undersmoothing"). Rather, we will derive an upper bound $\widehat{B}_\gamma$ for the bias $|b_\gamma|$. A challenge is that we do not know the expectations in Corollary 3 precisely since they involve integrals over the latent variable U_i and the unknown functions G, $\tau(\cdot)$ and $\alpha_{(0)}(\cdot)$. 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 for the running variable Z_i. To this end, define the marginal distribution function $F_G(\cdot)$ of Z_i when $U_i \sim G$, $F_G(t) = \int \mathbf{1}(\{z \leq t\}) \int p(z \mid u) dG(u) d\lambda(z)$. Then let $\mathcal{G}_n$ be the class of latent variable distributions that lead to marginal distributions F_G that lie within the Dvoretzky–Kiefer–Wolfowitz band [Massart, 1990] of the empirical measure $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left(Z_i \leq t\right)$, i.e.,[11]

$$\mathcal{G}_n = \left\{ G \text{ distrib.} : \sup_{t \in \mathbb{R}} \left| F_G(t) - \widehat{F}_n(t) \right| \leq \sqrt{\frac{\log\left(2/\alpha_n\right)}{2n}} \right\},\ \alpha_n = \min\left\{0.05, n^{-\frac{1}{4}}\right\}. \tag{19}$$

We also consider the following sensitivity model for the CATE:

Sensitivity Model (Treatment effect heterogeneity)**.** For $M \in [0,\, 1]$, we define

$$\mathcal{T}_M = \left\{ \tau(\cdot) \;\middle|\; \tau(u) = \bar{\tau} + \Delta(u) \text{ for } \bar{\tau} \in \mathbb{R} \text{ and } \Delta(\cdot) \text{ s.t. } |\Delta(u)| \leq M \right\}. \tag{20}$$

We note that $\mathcal{T}_0 = \{\text{constant CATE}\}$ and under Assumption 4,

$$\mathcal{T}_1 = \{\text{all CATE functions } \tau(\cdot)\},\ \mathcal{T}_{1/2} \supset \{\text{all CATE functions } \tau(\cdot) \geq 0\},$$

and so the choice $M = 1$ may be used to avoid imposing any additional assumptions on heterogeneity, while $M = 1/2$ is a conservative choice if one is willing to impose a monotonicity restriction.

Proposition 6. *Assume the conditions from Theorem 4 are satisfied, as well as Assumptions 1-4. Furthermore suppose that $\tau(\cdot) \in \mathcal{T}_M$ and that we upper bound the bias as,*

$$\widehat{B}_{\gamma,M} = \sup\left\{ \left| \text{Bias}\left[\gamma_\pm,\, \tau_w;\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right] \right| \,:\, G \in \mathcal{G}_n,\, \alpha_{(0)}(\cdot) \in [0,1],\, \tau(\cdot) \in \mathcal{T}_M \right\}. \tag{21}$$

Then $\mathbb{P}[|b_\gamma| \leq \widehat{B}_{\gamma,M}] \to 1$ as $n \to \infty$.

[10] For example, the local difference-in-means estimator with $\gamma_+(z) = \mathbf{1}(z \in [c, c+h_n])$, $\gamma_-(z) = \mathbf{1}(z \in [c-h_n, c))$ with $h_n \to 0$ will satisfy the condition when $h_n^{-1} = O(n^\beta)$ for $\beta \in (0, 1/2)$ and the running variable has a continuous Lebesgue density at the cutoff.

[11] More generally, any class of distributions $\mathcal{G}_n$ such that $\mathbb{P}\left[G \in \mathcal{G}_n\right] \to 1$ could be used instead, see Ignatiadis and Wager [2022] for further examples of such constructions.

We explain in Supplement B.1 how to compute this bound on the bias. Finally, we build confidence intervals for τ that are robust to estimation bias up to $\widehat{B}_{\gamma,M}$ following Imbens and Manski [2004], Armstrong and Kolesár [2018], and Imbens and Wager [2019].

Corollary 7 (Valid confidence intervals)**.** *Assume the conditions from Theorem 4 are satisfied. Furthermore, suppose that Assumptions 1-4 are satisfied and that $\tau(\cdot) \in \mathcal{T}_M$. Consider the confidence intervals:*

$$\tau_w \in \hat{\tau}_\gamma \pm \ell_\alpha, \quad \ell_\alpha = \min\left\{\ell : \mathbb{P}\left[\left|b + n^{-1/2}\widehat{V}_\gamma^{1/2}\widetilde{Z}\right| \leq \ell\right] \geq 1 - \alpha \text{ for all } |b| \leq \widehat{B}_{\gamma,M}\right\}, \tag{22}$$

where $\widetilde{Z}$ is a standard Gaussian random variable, $\alpha \in (0,1)$ is the significance level, and $\widehat{V}_\gamma$ is an estimate of the sampling variance V_γ. Then, $\liminf_{n\to\infty} \mathbb{P}\left[\tau_w \in \hat{\tau}_\gamma \pm \ell_\alpha\right] \geq 1 - \alpha$.

Formally, our inference builds on the partial identification result stated in Corollary 3. In general, we will consider sequences of weight functions in (9), that make the bias progressively smaller. As discussed further in Section 4, the choice of weighting functions is governed by a bias–variance tradeoff, whereby reducing the worst-case bias entails increasing the variance of the estimator (9). In some settings, e.g., when $Z_i|U_i$ has a binomial distribution, treatment effects are only partially identified, and so it is not possible to get zero bias—even asymptotically. For a further discussion of point versus partial identification in regression discontinuity designs, see Section II.A of Imbens and Wager [2019].

3.1 Robustness to misspecification of CATE heterogeneity

Applying our result requires a specification of the sensitivity model (20). While one can adopt the unrestrictive model $\mathcal{T}_1$, in this section we explore the robustness of our approach to misspecification of the sensitivity model. More concretely, we suppose that the CATE $\tau(\cdot)$ is not constant as a function of u, yet we conduct inference using $\mathcal{T}_0$. In that case, our intervals attain the correct coverage for the convenience-weighted treatment effect:

$$\tau_{h,+} := \int \frac{h(u,\ \gamma_+)}{\mathbb{E}_G\left[h(U,\ \gamma_+)\right]} \tau(u)\, dG(u). \tag{23}$$

This estimand 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]. We formalize this result in the following corollary:

Corollary 8 (Valid confidence intervals for the convenience-weighted treatment effect)**.** *Assume the conditions from Theorem 4 are satisfied, as well as Assumptions 1-4. Suppose that $\tau(\cdot) \in \mathcal{T}_M$ but we construct confidence intervals as in (22) using $M' \geq 0$ instead of M (say, $M' < M$). Then these confidence intervals satisfy:*

$$\liminf_{n\to\infty} \mathbb{P}\left[\tau_{h,+} \in \hat{\tau}_\gamma \pm \ell_\alpha\right] \geq 1 - \alpha,$$

where $\tau_{h,+}$ is defined in (23).

If we are interested in the null hypothesis of no treatment effects:

$$H_0 : \tau(u) = 0 \text{ for all } u,$$

then we can form a valid test by forming confidence intervals for τ_w under the sensitivity model $\mathcal{T}_0$ and rejecting the null hypothesis when the resulting confidence interval does not include 0.

4 Designing estimators via quadratic programming

Given a choice of weighting functions γ for (9), Propositions 5, 6 and Corollary 7 provided a complete recipe for building valid confidence intervals. As discussed above, at this point, one could already take weighting functions implied by various regression discontinuity estimators, and use these results to build valid confidence intervals for τ that are directly justified by noise-induced randomization. Existing weighting functions γ, however, were not designed for this purpose, and so may not yield particularly short confidence intervals. Hence we now turn to the problem of deriving weighting functions γ_+, γ_- with an eye towards making confidence intervals obtained via Corollary 7 short.

Our strategy is to choose γ_+, γ_- by minimizing an approximate bound on the worst-case mean-squared error of the estimator (9). Let $w(\cdot)$ be the latent weighting of the estimand (10) and suppose we posit the sensitivity model $\mathcal{T}_M$. Furthermore, let $\bar{F}(\cdot)$ be a guess or estimate of the marginal distribution $F_G(\cdot)$ of Z_i under Assumption 2 and let $\bar{w}(\cdot)$ be an estimate of the normalized latent weight $w(\cdot)/\mathbb{E}_G\left[w(U)\right]$. We propose solving the following quadratic program:[12]

$$\min_{\gamma_\pm(\cdot)} \quad \frac{1}{n}\left(\int \gamma_-^2(z)\, d\bar{F}(z) + \int \gamma_+^2(z)\, d\bar{F}(z)\right) \;+\; (t_1+t_2)^2 \tag{24a}$$

$$\text{s.t.} \quad |h(u,\,\gamma_+) - h(u,\,\gamma_-)| \leq t_1,\;\; M\,|h(u,\,\gamma_\diamond) - \bar{w}(u)| \leq t_2 \text{ for } \diamond \in \{\pm\} \text{ and all } u \tag{24b}$$

$$\int \gamma_-(z) d\bar{F}(z) = 1, \qquad \int \gamma_+(z) d\bar{F}(z) = 1 \tag{24c}$$

$$\gamma_-(z) = 0 \text{ for } z \geq c,\;\; \gamma_+(z) = 0 \text{ for } z < c \tag{24d}$$

$$|\gamma_\diamond(z)| \leq C n^{\beta} \text{ for } \diamond \in \{\pm\} \text{ and all } z. \tag{24e}$$

In choosing $\bar{F}(\cdot)$ and $\bar{w}(\cdot)$, we make use of the structure provided by Assumption 2, and estimate G as $\bar{G}$ via nonparametric maximum likelihood [Kiefer and Wolfowitz, 1956] and then we let $\bar{F}(\cdot) = F_{\bar{G}}(\cdot)$ and $\bar{w}(\cdot) = w(\cdot)/\mathbb{E}_{\bar{G}}\left[w(U)\right]$.

We next elaborate on the motivation behind optimization problem (24). The first term in (24a) is a proxy for the variance of our estimator, motivated by the fact that $\mathrm{Var}\left[\gamma_\diamond(Z_i)Y_i\right] \leq \int \gamma_\diamond^2(z)\, dF(z)$ for $\diamond \in \{+,-\}$. The next term, $(t_1+t_2)^2$, seeks to approximately bound the worst-case bias of the estimator. The bias is decomposed through the triangle inequality into the two terms appearing in the bias-decomposition of Corollary 3; t_1 in (24b) bounds the confounding bias and seeks to balance $h(\cdot,\,\gamma_+)$ and $h(\cdot,\,\gamma_-)$, while t_2 bounds the CATE-heterogeneity bias and seeks to balance h with the normalized $w(\cdot)$. (24c) is a normalization constraint, (24d) enforces that γ_+, γ_- assign weight only to treated, resp. control units, and (24e) ensures that no single observation is given excessive influence.[13]

The following results shows that the resulting weights derived from optimization problem (24) satisfy the conditions of Theorem 4 and thus enable valid inference.

Proposition 9 (Sufficient condition for weighting kernels)**.** *Assume we derive $\gamma_\pm = \gamma_\pm^{(n)}$ by solving optimization problem (24) for $M > 0$, where $\bar{F}(\cdot), \bar{w}(\cdot)$ are guesses for $F_G(\cdot)$, $w(\cdot)/\mathbb{E}_G\left[w(U)\right]$ or estimates based on a held-out sample. Furthermore, assume that $\bar{F}$ assigns non-trivial mass to $[c,\infty)$ and that $\bar{w}(\cdot)$ is bounded, i.e., there exists $k > 1$ such that*

[12] An appropriately discretized version of this problem can be solved using standard convex optimization software. In our implementation we use the MOSEK solver [ApS, 2020].

[13] In our implementation, we found that the constraint (24e) was never active, and so we omitted this constraint from our numerical results below.

Algorithm 1: Confidence intervals for treatment effects in regression discontinuity designs identified via noise-induced randomization (NIR).

Input: Samples $Z_i, Y_i, W_i, i = 1, .., n$ and RD cutoff c
Sensitivity model $\mathcal{T}_M$ (20), $M \in [0, 1]$
Estimand of interest τ_w (10)
Nominal significance level $\alpha \in (0, 1)$

1. Form a guess or estimate $\bar{F}$ of the marginal Z-distribution and $\bar{w}(\cdot)$ of the normalized latent weighting function $w(\cdot)/\mathbb{E}_G[w(U)]$.
2. Solve the minimax quadratic program (24) to get γ_+, γ_-.
3. Form the point estimate $\hat{\tau}_\gamma$ as in (9).
4. Estimate the variance of $\hat{\tau}_\gamma$ by $\widehat{V}_\gamma$ as in (18).
5. Estimate the worst-case bias $\widehat{B}_\gamma$ by (21).
6. Form bias-aware confidence intervals at level α as in (22).

$\mathbb{P}[1/k < \bar{F}([c, \infty)) < 1 - 1/k,\ \sup_u |\bar{w}(u)| < k] \to 1$ *as* $n \to \infty$ *and that the expectation of* $\gamma_+^{(n)}, \gamma_-^{(n)}$ *is asymptotically lower bounded by a strictly positive number, i.e., there exists* $\delta > 0$ *such that* $\mathbb{P}[\int \gamma_\diamond^{(n)}(z) dF(z) > \delta, \diamond \in \{\pm\}] \to 1$ *as* $n \to \infty$*. Then the weights derived from optimization problem* (24) *satisfy condition* (16) *from Theorem 4 on an event* A_n *with* $\mathbb{P}[A_n] \to 1$ *as* $n \to \infty$.

Our approach is heuristic, and may not recover the optimal weights, i.e., weights that would make the confidence intervals from Corollary 7 as short as possible, and we leave the topic of minimax-optimal inference under noise-induced randomization to further work. However, as evidenced by our numerical experiments, this heuristic appears to yield weighting functions γ_+, γ_- that yield powerful inference in practice.

In practice, we use the full dataset to also form estimates for $\bar{F}(\cdot)$ and $\bar{w}(\cdot)$; throughout our simulations we have not observed any undercoverage thereby. We summarize our approach to inference in Algorithm 1.

5 Noise-induced versus continuity-based inference?

As mentioned in the introduction, a standard approach to inference in regression discontinuity designs relies on smoothness assumptions for the conditional response functions $\mu_{(w)}(z)$. Now, one can verify that noise in the running variable as in Assumption 2 implies smoothness properties on the conditional response function: Under Assumptions 2–3,

$$\mu_{(w)}(z) = \mathbb{E}\left[Y_i(w) \,\middle|\, Z_i = z\right] = \int \alpha_{(w)}(u) p(z \,|\, u)\, dG(u) \Big/ \int p(z \,|\, u)\, dG(u), \tag{25}$$

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 e.g., Lee [2008, Proposition 2]). Furthermore, higher order differentiability of $p(z \,|\, u)$ implies the same for $\mu_{(w)}(z)$ (e.g., Dong and Kolesár [2021, Lemma A.1.]).

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. Recall that many continuity-based methods, including Armstrong and Kolesár [2020], Imbens and Wager [2019] and Kolesár and Rothe [2018], rely on $\mu_{(w)}(z)$ having a bounded second derivative to drive inference. To build a formal connection between our setting and this line of work, we derive upper bounds on $|\mu''_{(w)}(z)|$ that are justified by (25). Formally, we define the worst-case possible curvature at z among all data-generating distributions satisfying Assumptions 2–4 with conditional density $p(\cdot \mid \cdot)$ such that the marginal density of the running variable at z is lower bounded by $\rho > 0$:

$$\operatorname{Curv}(z,\, \rho,\, p) := \sup \left\{ \left| \frac{d^2 \mu_{(w)}(z)}{dz^2} \right| \,:\, f_G(z) = \int p(z \mid u) dG(u) \geq \rho,\ \alpha_{(w)}(\cdot) \in [0, 1] \right\}. \quad (26)$$

In (26) we constrain ourselves to marginal densities such that $f_G(z) \geq \rho$ for $\rho > 0$, because typically $\operatorname{Curv}(z,\, 0,\, p) = \infty$. In Supplement B.2, we explain how the quantity (26) may be computed numerically for any sufficiently regular p. One can then use the upper bounds on the second derivative of $\mu_{(w)}(z)$ in (26) 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. 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 our applications and simulations below, however, bounds based on (26) appear to be 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.

To provide intuition for (26), we provide analytic lower and upper bounds on (26) in the case of Gaussian measurement error, i.e., with $Z_i \mid U_i \sim \mathcal{N}(U_i,\, \nu^2)$ that quantify dependence on the noise level ν and the lower bound ρ on the density. The lower bound for $\mu''_{(w)}(z)$ below is obtained by considering a distribution G with two point masses symmetrically positioned around the cutoff and a third point mass at the cutoff. Meanwhile, the upper bounds below build on a lemma of Jiang and Zhang [2009].

Proposition 10. *Suppose that Assumptions 2–4 hold with noise model $Z_i \mid U_i \sim \mathcal{N}\left(U_i,\, \nu^2\right)$, where $\nu > 0$. Then, $\mu_{(w)}(z)$ is infinitely differentiable. Furthermore, for any point $z \in \mathbb{R}$ and any $0 < \rho < 1/\sqrt{2\pi\nu^2}$:*

$$\frac{-\log(2\pi\nu^2\rho^2)}{10\nu^2} \leq \operatorname{Curv}(z,\, \rho,\, \mathcal{N}(\cdot,\, \nu^2)) \leq \frac{-18\log(\pi\nu^2\rho^2)}{\nu^2}.$$

6 Applications

6.1 Antiretroviral Therapy (ART) Eligibility and Retention

In this section, we apply our approach to a medical study. 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. Concretely, the response of interest $Y_i \in \{0, 1\}$ is an 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.

According to health guidelines used in South Africa at the time, an HIV-positive patient should receive immediate ART if their measured CD4 count[14] was below 350 cells/μL.

[14] CD4 cells are specialized immune system cells, and low CD4 count is indicative of poor immune function.

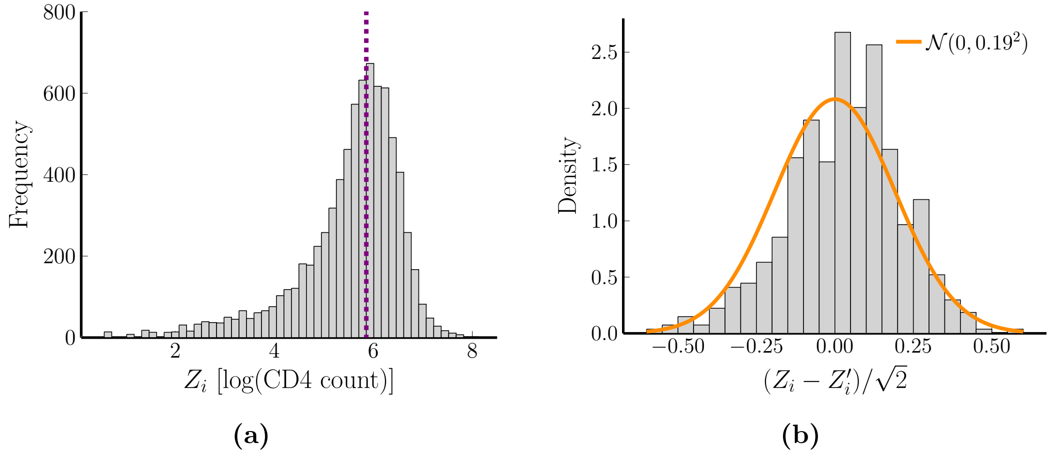

Figure 2: CD4 counts (cells/μL) as a noisy running variable in a regression discontinuity analysis. **(a)** Histogram of the running variable Z_i in the dataset of Bor et al. [2017]. **(b)** Differences $(Z_i - Z_i')/\sqrt{2}$ between repeated measurements in the dataset of Venter et al. [2018], overlaid with a Gaussian probability density function.

This setting can naturally be analyzed as a regression discontinuity design,[15] with running variable Z_i corresponding to the log of the CD4 count (in cells/μL) and a treatment cutoff $c = \log(350)$.[16] Figure 2(a) shows a histogram of Z_i from Bor et al. [2017], with treatment cutoff c denoted by a dashed line.

Bor et al. [2017] emphasize 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.

Here, in contrast, we seek an explicitly randomization-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 2(b) 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 $\nu = 0.19$.[17] Henceforth in applying our approach, we assume that measurement error in the log CD4 counts can be modeled as $Z_i \mid U_i \sim \mathcal{N}(U_i, \nu^2)$, where U_i is the true underlying log CD4 count of patient i.

We apply our noise-induced randomization (NIR) approach, with sensitivity model $\mathcal{T}_0$ to test for the existence of any treatment effects (Section 3.1). We also consider the following strategies to the problem.

[15] The compliance to this gudeline is not perfect, but, following Bor et al. [2017] we consider inference for intention-to-treat effects.

[16] We discard the patients with zero CD4 count.

[17] 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.

Method	*95% Confidence Interval*
NIR ($\mathcal{T}_0$)	0.111 ± 0.102
optrdd, $B = \mathrm{Curv}(c,\, \hat{f}(c),\, \mathcal{N}(\cdot,\, 0.19^2)) = 31.3$	0.071 ± 0.130
optrdd, $B = 1.46$ (heuristic)	0.153 ± 0.080
rdrobust	0.170 ± 0.076

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).

Our first baseline builds on the identifying assumption used for local linear regression, 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 τ under the assumed curvature bound.

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, we consider a randomization-based approach and use the curvature computation in (26) to obtain an upper bound B on curvature that is rigorously justified given our noise model. Concretely, we estimate the marginal density of Z_i at the cutoff ($\hat{f}(c) = 0.57$ using the nonparametric maximum likelihood estimator) and then let $B = \mathrm{Curv}(c,\, \hat{f}(c),\, \mathcal{N}(\cdot,\, 0.19^2))$. Second, 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 the design itself.

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 (6), 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. We run `rdrobust` with all tuning parameters set to the default values.

We present the results in Table 1. While our inference using NIR is purely randomization-based and only relies on the noise in the running variable Z_i, the treatment effect estimate remains significant at the 5% level. In contrast, using an upper-bound on curvature for optrdd that is purely justified by the noise-model, we do not obtain significant results. Noting the difficulty of accurately estimating curvature (especially in small samples), we believe the ability of our method to deliver confidence intervals that are purely justified by randomization to be potentially useful in practice.

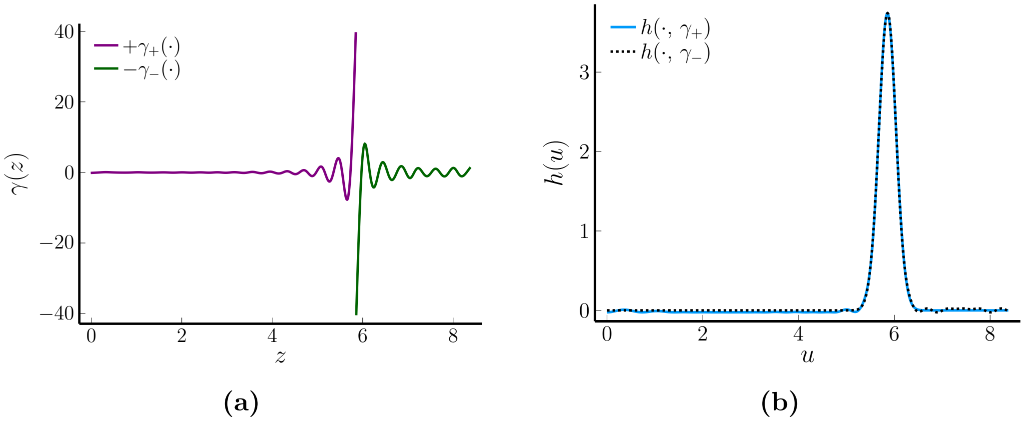

Figure 3: Noise-induced randomization analysis of a regression discontinuity design with CD4 counts as the running variable. **(a)** $\gamma_\pm$ weights as function of the running variable z. **(b)** Implied weighting of the latent variable $h(\cdot, \gamma_+)$, $h(\cdot, \gamma_-)$ as a function of the latent u.

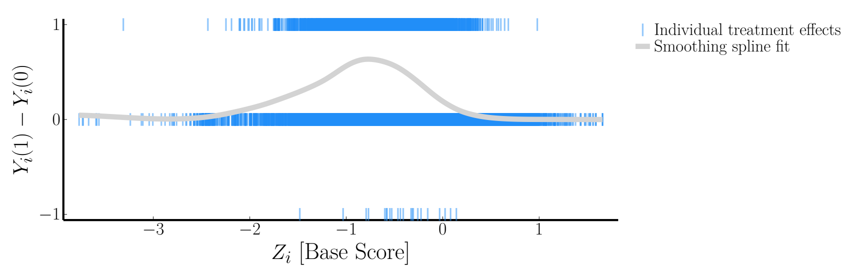

Figure 4: Scatterplot of the individual treatment effects $Y_i(1) - Y_i(0)$ against the running variable Z_i, with data derived from the Early Childhood Longitudinal Study [Tourangeau et al., 2015] as discussed in Section 6.2. We fit the ground truth treatment effect function $\mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, Z_i = z\right]$, shown as a line, using a smoothing spline.

In Figure 3 we show the weights $\gamma_\pm$ selected via quadratic programming and that were used by the NIR approach (Section 4), and the implied weighting of the latent variable $h(\cdot, \gamma_+)$, $h(\cdot, \gamma_-)$ as per (5). Units with Z_i close to the cutoff are strongly upweighted, and so we achieve approximate balance in terms of the latent U_i.

6.2 Test Scores in Early Childhood

We next consider the behavior of our method in a semi-synthetic regression discontinuity design built using data from the Early Childhood Longitudinal Study [Tourangeau et al., 2015]. This dataset has scaled mathematics 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 $i = 1, \ldots, n$ is built using the sequence of test scores from a single

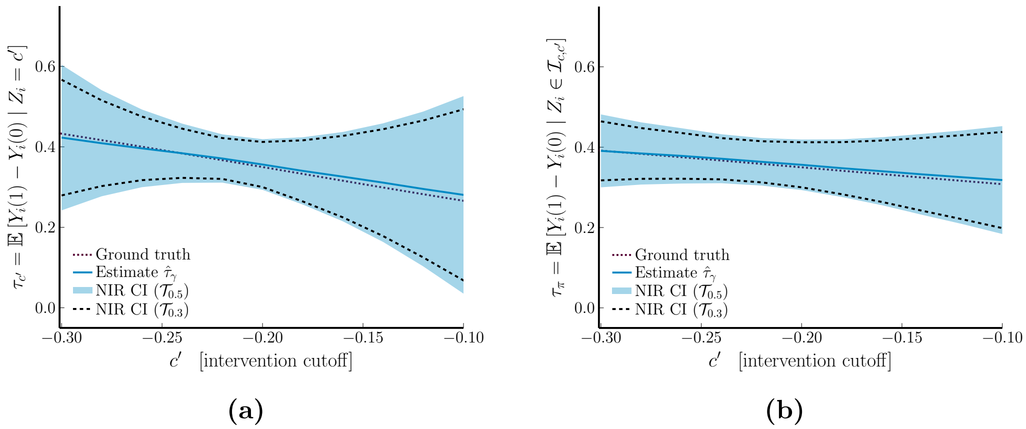

Figure 5: Regression discontinuity inference using our method (NIR) with data generated as in Figure 4. **(a)** Estimates and 95% confidence intervals of the regression discontinuity parameter (13), and **(b)** of the policy relevant parameter (14). $\mathcal{I}_{c,c'}$ is the left-closed, right-open interval between c and c' when $c \neq c'$, and $\mathcal{I}_{c,c'} = \{c\}$ for $c = c' = -0.2$.

child. We set the running variable Z_i to be the 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 4, there is considerable heterogeneity in the regression discontinuity parameter $\tau_{c'} = \mathbb{E}\left[Y_i(1) - Y_i(0) \,\middle|\, Z_i = c'\right]$ 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 construction of this data 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 (13) at c' away from the cutoff, and second we consider the policy-relevant parameter (14) 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 (2) and, following the remark in the last paragraph of Section 1.1, we set $\nu = 0.2043$ to match the lowest noise estimate provided in the Early Childhood Longitudinal Study dataset [Tourangeau et al., 2015]. To derive the point estimate and confidence intervals we run our method using sensitivity model $\mathcal{T}_{0.5}$. In the setting of this application, the monotonicity restriction $\tau(\cdot) \geq 0$ appears plausible, since the treatment effect measures the child's improvement between first and second grades, and in that case, as explained after (20), the sensitivity model $\mathcal{T}_{0.5}$ does not place any further restrictions on treatment effect heterogeneity. We also construct confidence intervals centered at the same point estimates under sensitivity model $\mathcal{T}_{0.3}$; this sensitivity model is plausible based on the treatment heterogeneity in the ground truth individual treatment effects (Figure 4).

Results for both targets are shown in Figure 5. We see that our method is able to recover heterogeneity. 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 away from the cutoff.

7 Simulation Study

In order to complement 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. 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, with $n \in \{1000, 2000, 10000\}$, we generate for $i = 1, \dots, n$:

$$\begin{aligned} &U_i \sim G = \text{Uniform}([0.5, 0.9]), \quad Z_i \,\big|\, U_i \sim \text{Binomial}(K, U_i), \\ &W_i = \mathbb{1}\left(\{Z_i \geq 0.6K\}\right), \quad Y_i(w) \sim \text{Bernoulli}\left(\mathbb{E}\left[Y_i(w) \,\big|\, U_i\right]\right), \end{aligned} \tag{27}$$

where the number of trials K is a simulation parameter that we vary. We consider two different choices for $\mathbb{E}\left[Y_i(w) \mid U_i\right]$ with null treatment effects $\tau(u) = 0$:

$$\mathbb{E}\left[Y_i(w) \,\big|\, U_i = u\right] = 0.25 \cdot \mathbb{1}\left(\{u < 0.6\}\right) + 0.75 \cdot \mathbb{1}\left(\{u \geq 0.6\}\right), \tag{28}$$

$$\mathbb{E}\left[Y_i(w) \,\big|\, U_i = u\right] = \sin(9u)/3 + 0.4. \tag{29}$$

We compare the following point estimates and 95% confidence intervals for the (null) treatment effect.

- **Noise-induced randomization (NIR)** with $p(\cdot \mid u) = \text{Binomial}(K, u)$ and using the sensitivity class $\mathcal{T}_0$ (cf. justification in Section 3.1).
- **optrdd** with curvature upper bound B specified as $\text{Curv}(c,\ f_G(c),\ p)$ (26), where $f_G(c)$ is the true marginal pmf at c.[18]
- **rdrobust** as implemented in the `R` package `rdrobust` of Calonico, Cattaneo, and Titiunik [2015] with default specification and taking the debiased estimate as the point estimate.

We evaluate methods by computing the confidence interval coverage, the expected half-length of confidence intervals and the mean absolute error (MAE). These metrics are computed by averaging over 1,000 Monte Carlo replications.

[18] $p(z \mid u)$ and $\mu_{(w)}(z)$ are only defined at $z \in \{0, \dots, K\}$ and so $\mu''(c)$ and $\text{Curv}(c,\ f_G(c),\ p)$ are ill-defined. However, as explained by Kolesár and Rothe [2018] and Imbens and Wager [2019], inference using optrdd with bound B is valid as long as there exists any function interpolating $\mu_{(w)}(\cdot)$ at $z \in \{0, \dots, K\}$ that is twice differentiable and whose worst-case curvature is upper bounded by B. In our computation of $\text{Curv}(c,\ f_G(c),\ p)$ we interpolate $p(z \mid u)$ for $z \in (0, K)$ (and consequently $\mu_{(w)}(z)$ through (25)) as $p(z \mid u) = p_B(u; z+1,\ K-z+1) - \log(K+1)$, where $p_B(u; \alpha, \beta)$ is the density of the $\text{Beta}(\alpha, \beta)$ distribution at $u \in (0, 1)$.

		K	5	10	25	50	100	200
$n = 1,000$	coverage	optrdd	100.0%	99.8%	98.8%	98.0%	97.1%	96.5%
		rdrobust	–	–	94.2%	94.5%	93.4%	93.4%
		NIR	100.0%	97.1%	97.2%	97.6%	98.1%	98.6%
	length	optrdd	0.452	0.383	0.347	0.344	0.370	0.423
		rdrobust	–	–	0.352	0.376	0.353	0.342
		NIR	**0.433**	**0.220**	**0.228**	**0.257**	**0.303**	0.398
	MAE	optrdd	0.089	0.095	0.113	0.119	0.130	0.153
		rdrobust	–	–	0.148	0.164	0.155	0.151
		NIR	0.068	0.076	0.084	0.091	0.105	0.126
$n = 2,000$	coverage	optrdd	100.0%	100.0%	98.8%	98.5%	97.4%	96.6%
		rdrobust	–	–	94.0%	94.0%	93.9%	92.8%
		NIR	100.0%	95.4%	95.9%	96.8%	97.1%	97.5%
	length	optrdd	0.396	0.325	0.280	0.273	0.287	0.322
		rdrobust	–	–	0.244	0.261	0.246	0.239
		NIR	**0.333**	**0.161**	**0.160**	**0.178**	**0.207**	0.258
	MAE	optrdd	0.067	0.072	0.083	0.092	0.102	0.118
		rdrobust	–	–	0.105	0.117	0.110	0.106
		NIR	0.052	0.063	0.061	0.065	0.076	0.093
$n = 10,000$	coverage	optrdd	100.0%	100.0%	100.0%	99.1%	98.3%	98.1%
		rdrobust	–	–	94.0%	94.6%	94.4%	94.4%
		NIR	100.0%	96.4%	95.9%	96.3%	96.0%	96.4%
	length	optrdd	0.322	0.249	0.184	0.167	0.167	0.177
		rdrobust	–	–	0.104	0.115	0.107	0.108
		NIR	**0.220**	**0.078**	**0.074**	**0.081**	**0.093**	0.111
	MAE	optrdd	0.030	0.033	0.038	0.047	0.056	0.063
		rdrobust	–	–	0.044	0.050	0.047	0.048
		NIR	0.021	0.030	0.029	0.031	0.036	0.043

Table 2: Simulation results in the binomial noise setting (27) with conditional response functions (28) for different choices of sample size n and number of trials K. We compare three methods (NIR, optrdd, rdrobust) and report the coverage of confidence intervals ("coverage"), the expected half-length of the confidence intervals ("length") and the mean absolute error ("MAE"). The best method in terms of expected half-length for each simulation setting, when it also has at least 95% coverage, is shown in bold.

		K	5	10	25	50	100	200
$n = 1,000$	coverage	optrdd	100.0%	100.0%	99.1%	99.3%	97.9%	98.6%
		rdrobust	–	–	–	95.0%	91.3%	93.4%
		NIR	100.0%	97.6%	97.9%	98.6%	98.6%	99.5%
	length	optrdd	**0.418**	0.346	0.294	0.280	0.288	0.320
		rdrobust	–	–	–	0.271	0.246	0.234
		NIR	0.428	**0.197**	**0.194**	**0.212**	**0.242**	0.321
	MAE	optrdd	0.081	0.080	0.087	0.088	0.098	0.104
		rdrobust	–	–	–	0.117	0.116	0.098
		NIR	0.077	0.07	0.068	0.072	0.079	0.086
$n = 2,000$	coverage	optrdd	100.0%	100.0%	99.9%	99.0%	98.1%	98.5%
		rdrobust	–	–	94.2%	93.9%	92.5%	92.8%
		NIR	100.0%	98.0%	96.7%	97.9%	97.1%	98.9%
	length	optrdd	0.372	0.298	0.240	0.225	0.227	0.247
		rdrobust	–	–	0.183	0.187	0.171	0.164
		NIR	**0.329**	**0.143**	**0.135**	**0.144**	**0.163**	0.203
	MAE	optrdd	0.062	0.058	0.064	0.07	0.074	0.084
		rdrobust	–	–	0.079	0.085	0.077	0.072
		NIR	0.058	0.052	0.050	0.052	0.059	0.065
$n = 10,000$	coverage	optrdd	100.0%	100.0%	100.0%	100.0%	99.0%	98.6%
		rdrobust	–	–	94.9%	95.5%	93.3%	93.3%
		NIR	100.0%	96.2%	96.6%	96.7%	97.0%	96.3%
	length	optrdd	0.311	0.237	0.167	0.141	0.135	0.138
		rdrobust	–	–	0.078	0.081	0.074	0.072
		NIR	**0.219**	**0.069**	**0.062**	**0.064**	**0.071**	0.084
	MAE	optrdd	0.038	0.033	0.030	0.035	0.040	0.046
		rdrobust	–	–	0.033	0.036	0.033	0.032
		NIR	0.038	0.027	0.024	0.024	0.027	0.031

Table 3: Simulation results in the binomial noise setting (27) with conditional response functions (29). The results shown are analogous to the results of Table 2 (with a different data generating process).

The results of the simulation study are shown in Table 2 (response function (28)) and Table 3 (response function (29)). All methods have approximately correct coverage, with optrdd and NIR always achieving the nominal 95% level and rdrobust slightly undercovering. We note that rdrobust and its distributional theory have been developed under the assumption of a continuous rather than discrete random variable, but it nevertheless performs reasonably well and the software furthermore detects the discreteness.[19] NIR yields the shortest confidence intervals in most settings. K determines the noise level; the smaller K is the more effective noise there is in the running variable, and so the better our method does (with the exception of the smallest K, namely $K \leq 10$). This is in contrast to rdrobust, whose performance improves as K increases and the running variable becomes less discrete, until at $K = 200$ it leads to shorter confidence intervals than NIR. As expected, the confidence interval length decreases for all methods as the sample size n increases.

At a high level, this simulation experiment corroborates the claim that our method, NIR, can flexibly turn assumptions about exogenous noise in the running variable Z_i into a practical, randomization-based procedure for inference in regression discontinuity designs. We achieve nominal coverage across a wide variety of simulation settings. Our results also point to the possibility that NIR may in fact result in improved power in settings where running variables are discrete with known noise. This would not be unreasonable, as continuity-based approaches were not necessarily designed for this setting,[20] whereas NIR can directly exploit structure of the binomial distribution. However, a detailed study of the power (as opposed to feasibility) of randomization-based 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. Our results suggest that the pursuit of randomization-based inference in regression discontinuity designs 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 randomization-based inference for conceptual reasons.

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

[19] For small K and n, rdrobust sometimes return an error, in which case we do not report its performance in the tables. Furthermore, even when rdrobust does not return an error, it provides the user with the informative warning "Mass points detected in the running variable".

[20] Although, as discussed in Kolesár and Rothe [2018] they can rigorously be used in this setting given appropriate interpretation.

testing [e.g., Alatas, Banerjee, Hanna, Olken, and Tobias, 2012] may also readily admit analysis with the noise-induced randomization approach. Even data ostensibly arising from a complete census of a population may have measurement error in population totals or characteristics [cf. Fraga and Merseth, 2016]. 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 is does not apply to all regression discontinuity designs, as some running variables are not readily interpretable as having measurement error or other exogenous noise. For example, numerous studies have used geographic boundaries as discontinuities [Keele and Titiunik, 2014, Rischard, Branson, Miratrix, and Bornn, 2021], 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, 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, though perhaps there are other sources of exogenous noise [e.g., weather, Gomez, Hansford, and Krause, 2007, Cooperman, 2017]. 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.

Software

All numerical results in this paper are reproducible with the code in the following Github repository: `https://github.com/nignatiadis/noise-induced-randomization-paper`. There we provide an implementation of the proposed methods as a package in the Julia programming language [Bezanson et al., 2017] that depends, among others, on JuMP.jl [Dunning et al., 2017].

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.

MOSEK ApS. *The MOSEK Optimization Suite Manual, Version 9.2*, 2020. URL `https://www.mosek.com/`.

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 nonparametric 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.

Harold P Benson. A simplicial branch and bound duality-bounds algorithm for the linear sum-of-ratios problem. *European journal of operational research*, 182(2):597–611, 2007.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

Martin Bilodeau. Estimating a multivariate treatment effect under a biased allocation rule. *Communications in Statistics-Theory and Methods*, 26(5):1119–1124, 1997.

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.

Christian N Brinch, Magne Mogstad, and Matthew Wiswall. Beyond LATE with a discrete instrument. *Journal of Political Economy*, 125(4):985–1039, 2017.

Sebastian Calonico, Matias D Cattaneo, and Rocío Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014.

Sebastian Calonico, Matias D. Cattaneo, and Rocío 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.

Matias D Cattaneo, Brigham R Frandsen, and Rocío 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.

Thomas D Cook. Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, 142(2):636–654, 2008.

Alicia Dailey Cooperman. Randomization inference with rainfall data: Using historical weather patterns for variance estimation. *Political Analysis*, 25(3):277–288, 2017.

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.

Yingying Dong and Michal Kolesár. When can we ignore measurement error in the running variable? *arXiv preprint arXiv:2111.07388*, 2021.

Yingying Dong and Arthur Lewbel. Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5):1081–1092, 2015.

Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

Michael O Finkelstein, Bruce Levin, and Herbert Robbins. Clinical and prophylactic trials with assured new treatment for those at greater risk: I. A design proposal. *American Journal of Public Health*, 86(5):691–695, 1996a.

Michael O Finkelstein, Bruce Levin, and Herbert Robbins. Clinical and prophylactic trials with assured new treatment for those at greater risk: II. Examples. *American Journal of Public Health*, 86(5):696–705, 1996b.

Bernard L Fraga and Julie Lee Merseth. Examining the causal impact of the voting rights act language minority provisions. *Journal of Race, Ethnicity, and Politics*, 1(1):31–59, 2016.

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.

Brad T Gomez, Thomas G Hansford, and George A Krause. The Republicans should pray for rain: Weather, turnout, and voting in US presidential elections. *The Journal of Politics*, 69(3):649–663, 2007.

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. Sample selection bias as a specification error. *Econometrica*, 47(1): 153–161, 1979.

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. Confidence intervals for nonparametric empirical Bayes analysis. *Journal of the American Statistical Association*, (forthcoming), 2022.

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.

Zhichao Jiang and Peng Ding. Measurement errors in the binary instrumental variable model. *Biometrika*, 107(1):238–245, 2020.

Nathan Kallus. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020.

Luke J Keele and Rocío Titiunik. Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155, 2014.

Morgan Kelly. The standard errors of persistence. CEPR Discussion Paper DP13783, 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.

Hiroshi Konno and Natsuroh Abe. Minimization of the sum of three linear fractional functions. *Journal of Global Optimization*, 15(4):419–432, 1999.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

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.

Fan Li, Andrea Mercatanti, Taneli Mäkinen, and Andrea Silvestrini. A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds. *The Annals of Applied Statistics*, 15(1):304–322, 2021.

Pascal Massart. The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):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.

Magne Mogstad, Andres Santos, and Alexander Torgovitsky. Using instrumental variables for inference about policy relevant treatment parameters. *Econometrica*, 86(5):1589–1619, 2018.

Monica Morell. *Multilevel Regression Discontinuity Models with Latent Variables.* PhD thesis, University of Maryland, College Park, 2020.

Monica Morell, Ji Seung Yang, and Yang Liu. Latent variable regression discontinuity design with cluster level treatment assignment. *Multivariate Behavioral Research*, 55(1):146–146, 2020.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Judea Pearl. On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432, 2010.

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. Do school districts affect NYC house prices? identifying border differences using a Bayesian nonparametric approach to geographic regression discontinuity designs. *Journal of the American Statistical Association*, 116(534):619–631, 2021.

H Robbins. Comparing two treatments under biased allocation. *La Gazette des Sciences mathématiques du Québec*, 15:35–41, 1993.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1956.

Herbert Robbins and Cun-Hui Zhang. Estimating a treatment effect under biased sampling. *Proceedings of the National Academy of Sciences*, 85(11):3670–3672, 1988.

Herbert Robbins and Cun-Hui Zhang. Estimating the superiority of a drug to a placebo when all and only those patients at risk are treated with the drug. *Proceedings of the National Academy of Sciences*, 86(9):3003–3005, 1989.

Herbert Robbins and Cun-Hui Zhang. Estimating a multiplicative treatment effect under biased allocation. *Biometrika*, 78(2):349–354, 1991.

Miikka Rokkanen. Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design. Department of Economics Discussion Paper 1415-03, Columbia University, 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.

Andrew Donald Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951.

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.

Jasjeet S Sekhon and Rocío Titiunik. On interpreting the regression discontinuity design as a local experiment. *Advances in Econometrics*, 38:1–28, 2017.

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 Akpomiemie, 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.

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 1

Proof: Conditioning on the latent variable U, we find that

$$\begin{aligned}\mathbb{E}\left[\gamma_+(Z)Y \,\middle|\, U\right] &\overset{(i)}{=} \mathbb{E}\left[\gamma_+(Z)Y \cdot 1\left(\{Z \geq c\}\right) \,\middle|\, U\right] \\ &\overset{(ii)}{=} \mathbb{E}\left[\gamma_+(Z)Y(1) \cdot 1\left(\{Z \geq c\}\right) \,\middle|\, U\right] \\ &\overset{(iii)}{=} \underbrace{\mathbb{E}\left[Y(1) \,\middle|\, U\right]}_{\alpha_{(1)}(U)} \underbrace{\mathbb{E}\left[\gamma_+(Z)1\left(\{Z \geq c\}\right) \,\middle|\, U\right]}_{h(U,\gamma_+)=\int \gamma_+(z)p(z \mid U)\, d\lambda(z)}\end{aligned}$$

In (i) we used that $\gamma_+(z) = 0$ for $z < c$, in (ii) we used the fact that $Y = Y(1)$ for $Z \geq c$ by Assumption 1 and in (iii) we used exogeneity of the noise (Assumption 3). Finally, the expression for $\mathbb{E}\left[\gamma_+(Z)1\left(\{Z \geq c\}\right) \,\middle|\, U\right] = \mathbb{E}\left[\gamma_+(Z) \,\middle|\, U\right]$ follows from Assumption 2. By iterated expectation we thus find that $\mathbb{E}\left[\gamma_+(Z)Y\right] = \mathbb{E}\left[\alpha_{(1)}(U)h(U, \gamma_+)\right]$. The proof for γ_- is analogous. $\square$

A.2 Proof of Corollary 3

Proof. Noting that $\tau(U) = \alpha_{(1)}(U) - \alpha_{(0)}(U)$, this is proved by direct algebraic manipulation. $\square$

A.3 Proof of Theorem 4

Proof. Notation: We write $\mathbb{E}_n\left[\cdot\right]$ to denote empirical averages, i.e., for a function $h(\cdot)$, we write:

$$\mathbb{E}_n\left[h(Z_i)\right] = \frac{1}{n}\sum_{i=1}^{n} h(Z_i).$$

We omit dependence on n of the weighting kernels. 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 $\sum_i \gamma_+(Z_i)\left(Y_i(1) - \mu_{\gamma,+}\right)$: We seek to prove the following central limit theorem:

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

We first note that the numerator has expectation 0, since:

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

In the last step, we used the fact that by (12):

$$\mu_{\gamma,+} = \frac{\mathbb{E}\left[\alpha_{(1)}(U)h(U,\ \gamma_+)\right]}{\mathbb{E}\left[h(U,\ \gamma_+)\right]}. \tag{S1}$$

But by Proposition 1 we also have that $\mathbb{E}\left[\alpha_{(1)}(U)h(U,\ \gamma_+)\right] = \mathbb{E}\left[\gamma_+(Z_i)Y_i(1)\right]$, while an analogous argument as in the proof of Proposition 1 demonstrates that $\mathbb{E}\left[h(U,\ \gamma_+)\right] = \mathbb{E}\left[\gamma_+(Z_i)\right]$.

Next we will check the condition of Lyapunov's central limit theorem. Let $\underline{\sigma}^2 := \inf_z \operatorname{Var}\left[Y_i \,\middle|\, Z_i = z\right] > 0$.

$$\begin{aligned}
\operatorname{Var}\left[\gamma_+(Z_i)\left(Y_i(1) - \mu_{\gamma,+}\right)\right] &\geq \mathbb{E}\left[\operatorname{Var}\left[\gamma_+(Z_i)\left(Y_i(1) - \mu_{\gamma,+}\right) \,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[\gamma_+(Z_i)^2 \operatorname{Var}\left[Y_i(1) - \mu_{\gamma,+} \,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[\gamma_+(Z_i)^2 \operatorname{Var}\left[Y_i(1) \,\middle|\, Z_i\right]\right] \\
&= \mathbb{E}\left[\gamma_+(Z_i)^2 \operatorname{Var}\left[Y_i \,\middle|\, Z_i\right]\right] \\
&\geq \underline{\sigma}^2 \mathbb{E}\left[\gamma_+(Z_i)^2\right].
\end{aligned} \tag{S2}$$

In the penultimate line we used the fact that $Y_i(1) = Y_i$ on $\{Z_i \geq c\}$ and that $\gamma_+(z) = 0$ for $z < c$. We next bound $\mu_{\gamma,+}$ in (S1). First, since $Y_i \in [0,1]$ by Assumption 4, it also follows that $\alpha_{(1)}(U) \in [0,1]$ almost surely. Thus:

$$|\mu_{\gamma,+}| = \left|\frac{\mathbb{E}\left[\alpha_{(1)}(U)h(U,\ \gamma_+)\right]}{\mathbb{E}\left[\gamma_+(Z_i)\right]}\right| \leq \frac{\mathbb{E}\left[|h(U,\ \gamma_+)|\right]}{\mathbb{E}\left[\gamma_+(Z_i)\right]} \leq \frac{\sup_u |h(u,\ \gamma_+)|}{\mathbb{E}\left[\gamma_+(Z_i)\right]} \leq C',$$

for n large enough. Then, for $q > 0$ (and n large enough) we have that:

$$\begin{aligned}
\mathbb{E}\left[\left|\gamma_+(Z_i)\left(Y_i(1) - \mu_{\gamma,+}\right)\right|^{2+q}\right] &\leq (C'+1)^{2+q}\mathbb{E}\left[\left|\gamma_+(Z_i)\right|^{2+q}\right] \\
&\leq (C'+1)^{2+q} \cdot \sup_z |\gamma_+(z)|^q \cdot \mathbb{E}\left[\gamma_+(Z_i)^2\right].
\end{aligned}$$

So:

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

This proves the central limit theorem.

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

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

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

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

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

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

$\square$

A.4 Proof of Proposition 5

Proof. The proof here continues from the argument used for the proof of Theorem 4. As we did there, we only prove the result for the variance of $\hat{\mu}_{\gamma,+}$, the result for $\hat{\tau}_\gamma$ follows analogously. In the proof of Theorem 4 we already showed that $\mathbb{E}_n\left[\gamma_+(Z_i)\right] / \mathbb{E}\left[\gamma_+(Z_i)\right] = 1 + o_{\mathbb{P}}(1)$. It thus suffices to show that:

$$\mathbb{E}_n\left[\gamma_+(Z_i)^2\left(Y_i(1) - \hat{\mu}_{\gamma,+}\right)^2\right] / \mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right] = 1 + o_{\mathbb{P}}(1). \tag{S3}$$

We start by arguing that:

$$\mathbb{E}_n\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right] / \mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right] = 1 + o_{\mathbb{P}}(1). \tag{S4}$$

First:

$$\begin{aligned}
\operatorname{Var}\left[\frac{\mathbb{E}_n\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right]}{\mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right]}\right] &= \frac{\operatorname{Var}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right]}{n \cdot \mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right]^2} \\
&\leq \frac{\mathbb{E}\left[\left\{\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right\} \cdot \left\{\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right\}\right]}{n \cdot \mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right]^2} \\
&\leq \frac{(C'+1)^2 \sup_z \left|\gamma_+(z)\right|^2}{n\mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1) - \mu_{\gamma,+}\right)^2\right]} \to 0 \text{ as } n \to \infty.
\end{aligned}$$

Note that we verified that the last expression converges to 0 as $n \to \infty$ during the verification of Lyapunov's condition in the proof of Theorem 4. It follows that the asymptotic convergence in (S4) holds in L^2, thus also in probability. It remains to show that the feasible estimator in (S3) is asymptotically equivalent. We have the decomposition:

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

From the CLT of Theorem 4, we know that:

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

and so:

$$\frac{\mathbb{E}_n\left[\gamma_+(Z_i)^2\right]}{\mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1)-\mu_{\gamma,+}\right)^2\right]}\cdot\left(\widehat{\mu}_{\gamma,+}-\mu_{\gamma,+}\right)^2=O_{\mathbb{P}}(1)\cdot o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1).$$

The fact that the first term is $O_{\mathbb{P}}(1)$ follows by arguing with Chebyshev's inequality. First note that from (S2), we know that $\mathbb{E}[\gamma_+(Z_i)^2\left(Y_i(1)-\mu_{\gamma,+}\right)^2]\geq\underline{\sigma}^2\mathbb{E}[\gamma_+(Z_i)^2]$ and so it suffices to show that $\mathbb{E}_n\left[\gamma_+(Z_i)^2\right]/\mathbb{E}\left[\gamma_+(Z_i)^2\right]$ is $O_{\mathbb{P}}(1)$. Indeed this term is also $1+o_{\mathbb{P}}(1)$, since for any $\varepsilon>0$:

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

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

$$\frac{\mathbb{E}_n\left[\gamma_+(Z_i)^2\left(Y_i(1)-\mu_{\gamma,+}\right)\right]}{\mathbb{E}\left[\gamma_+(Z_i)^2\left(Y_i(1)-\mu_{\gamma,+}\right)^2\right]}\cdot\left(\widehat{\mu}_{\gamma,+}-\mu_{\gamma,+}\right)=O_{\mathbb{P}}(1)\cdot o_{\mathbb{P}}(1)=o_{\mathbb{P}}(1),$$

and it remains to prove that the first term is indeed $O_{\mathbb{P}}(1)$. By Cauchy-Schwarz

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

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

A.5 Proof of Proposition 6

Proof. Consider the event $\{G\in\mathcal{G}_n\}$. On this event, by definition we have $|b_\gamma|\leq\widehat{B}_{\gamma,M}$ This implies that $\{G\in\mathcal{G}_n\}\subset\{|b_\gamma|\leq\widehat{B}_{\gamma,M}\}$ and so $\mathbb{P}[|b_\gamma|\leq\widehat{B}_{\gamma,M}]\geq\mathbb{P}\left[G\in\mathcal{G}_n\right]$. It thus suffices to show that the RHS converges to 1 as $n\to\infty$. By construction of $\mathcal{G}_n$ in (19) and Massart's tight constant for the DKW inequality [Massart, 1990], it holds that

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

Since $\alpha_n\to 0$, we conclude the proof. □

A.6 Proof of Corollary 7

By Theorem 4, Proposition 5, and Slutsky, we have that

$$\frac{\sqrt{n}\left(\hat{\tau}_\gamma-\tau_w-b_\gamma\right)}{\widehat{V}_\gamma^{1/2}}\Rightarrow\mathcal{N}\left(0,1\right),$$

where $b_\gamma = \theta_\gamma - \tau_w$ by definition. So, letting $\tilde{Z} \sim \mathcal{N}(0,1)$ independent of everything else:

$$\begin{aligned}
\mathbb{P}\left[\tau_w \in \hat{\tau}_\gamma \pm \ell_\alpha\right] &= \mathbb{P}\left[-\ell_\alpha - b_\gamma \le \hat{\tau}_\gamma - \tau_w - b_\gamma \le \ell_\alpha - b_\gamma\right] \\
&= \mathbb{P}\left[-\sqrt{n}\widehat{V}_\gamma^{-1/2}\left(\ell_\alpha + b_\gamma\right) \le \sqrt{n}\widehat{V}_\gamma^{-1/2}\left(\hat{\tau}_\gamma - \tau_w - b_\gamma\right) \le \sqrt{n}\widehat{V}_\gamma^{-1/2}\left(\ell_\alpha - b_\gamma\right)\right] \\
&\overset{(i)}{=} \mathbb{E}\left[\mathbb{P}\left[-\sqrt{n}\widehat{V}_\gamma^{-1/2}\left(\ell_\alpha + b_\gamma\right) \le \tilde{Z} \le \sqrt{n}\widehat{V}_\gamma^{-1/2}\left(\ell_\alpha - b_\gamma\right) \;\middle|\; \widehat{V}_\gamma, \widehat{B}_{\gamma,M}, \hat{\tau}_\gamma\right]\right] + o(1) \\
&= \mathbb{E}\left[\mathbb{P}\left[-\ell_\alpha \le b_\gamma + n^{-1/2}\widehat{V}_\gamma^{1/2}\tilde{Z} \le \ell_\alpha \;\middle|\; \widehat{V}_\gamma, \widehat{B}_{\gamma,M}, \hat{\tau}_\gamma\right]\right] + o(1) \\
&\overset{(ii)}{\ge} \mathbb{E}\left[1-\alpha\right] + o(1) \\
&= 1 - \alpha + o(1).
\end{aligned}$$

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 ℓ_α in (22) and the fact that $\mathbb{P}[|b_\gamma| \le \widehat{B}_{\gamma,M}] \to 1$ as $n \to \infty$.

A.7 Proof of Corollary 8

Let $\tau_{h,+}$ be defined as in (23). We will show that

$$\left|\text{Bias}\left[\gamma_\pm,\, \tau_{h,+};\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right]\right| \le \widehat{B}_{\gamma,M'},$$

where $\widehat{B}_{\gamma,M'}$ is defined as in (21) for the estimand τ_w (rather than $\tau_{h,+}$). We make this dependence explicit by writing $\widehat{B}_{\gamma,M'} = \widehat{B}_{\gamma,M',\tau_w}$. The "CATE heterogeneity bias" in Corollary 3 vanishes for $\tau_{h,+}$, i.e.,

$$\text{Bias}\left[\gamma_\pm,\, \tau_{h,+};\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right] = \int \left(\frac{h(u,\ \gamma_+)}{\mathbb{E}_G\left[h(U,\ \gamma_+)\right]} - \frac{h(u,\ \gamma_-)}{\mathbb{E}_G\left[h(U,\ \gamma_-)\right]}\right)\alpha_{(0)}(u)\, dG(u).$$

However, the "CATE heterogeneity bias" also vanishes when $\tau(\cdot) \in \mathcal{T}_0$ (for any choice of estimand τ_w), so that:

$$\left|\text{Bias}\left[\gamma_\pm,\, \tau_{h,+};\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right]\right| = \left|\text{Bias}\left[\gamma_\pm,\, \tau_w;\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right]\right|, \text{ when } \tau(\cdot) \in \mathcal{T}_0.$$

Taking the supremum on the RHS over $\alpha_{(0)}(\cdot) \in [0,1]$, $\tau(\cdot) \in \mathcal{T}_0$ and $G \in \mathcal{G}_n$ demonstrates that:

$$\left|\text{Bias}\left[\gamma_\pm,\, \tau_{h,+};\, \alpha_{(0)}(\cdot),\, \tau(\cdot),\, G\right]\right| \le \widehat{B}_{\gamma,0,\tau_w} \le \widehat{B}_{\gamma,M',\tau_w}.$$

The conclusion follows as in the proofs of Proposition 6 and Corollary 7.

A.8 Proof of Proposition 9

Proof. We provide the argument for $\gamma_+ = \gamma_+^{(n)}$; the argument for $\gamma_- = \gamma_-^{(n)}$ is analogous. First note that $\max_z |\gamma_+(z)| > 0$ must hold, otherwise constraint (24c) of the optimization problem would not be satisfied. For convenience we define the events:

$$B_{n,1} = \left\{1/k < \bar{F}([c,\infty)) < 1 - 1/k,\ \sup_u |\bar{w}(u)| < k\right\},\ B_{n,2} = \left\{\int_{[c,\infty)} \gamma_+(z) dF(z) > \delta\right\}.$$

Next, let C be as in (24e) and $\tilde{C} = C/\delta$. Then, on the event $B_{n,2}$ we have that:

$$\sup_z |\gamma_+(z)| \leq Cn^\beta \leq \tilde{C}n^\beta \cdot \delta < \tilde{C}n^\beta \int_{[c,\infty)} \gamma_+(z) dF(z).$$

Next, we will bound $|h(u,\ \gamma_+)|$ on the event $B_{n,2}$. Consider the weighting kernel $\tilde{\gamma}_+(z) = \mathbf{1}(\{z \geq c\})/\bar{F}([c,\infty))$ and $\tilde{\gamma}_-(z) = \mathbf{1}(\{z < c\})/(1 - \bar{F}([c,\infty)))$. This is a feasible solution under the constraints of optimization problem (24), since $\sup_u |\tilde{\gamma}_\diamond(u)| \leq 1/k \leq Cn^\beta$, $\diamond \in \{+,-\}$ on $B_{n,2}$ (and n large enough), $\int \tilde{\gamma}_-(z) d\bar{F}(z) = 1$ and $\int \tilde{\gamma}_+(z) d\bar{F}(z) = 1$. We upper bound the objective of the optimization problem for that choice of weighting kernel. First, the variance-proxy term in the objective is equal to $\left[\bar{F}([c,\infty))^{-1} + (1 - \bar{F}([c,\infty)))^{-1}\right]/n$ which is $\leq 2k/n \leq 1$ for n large enough. On the other hand, note that we also have that:

$$|h(u, \tilde{\gamma}_+)| = \left| \bar{F}([c,\infty))^{-1} \cdot \int_{[c,\infty)} p(z \mid u) d\lambda(z) \right| \leq \bar{F}([c,\infty))^{-1} \leq k,$$

and similarly $|h(u, \tilde{\gamma}_-)| \leq k$. Hence by the triangle inequality we may bound the "t_1" term of the variance objective as $2k$, and similarly for the "t_2" term (recall that $|\bar{w}(u)| \leq k$ on $B_{n,2}$). Thus the objective of the whole optimization problem is upper bounded by $1 + 16k^2$. Thus, the objective for the optimal $\gamma_\pm$ in (24b) must be $\leq 1 + 16k^2$, which implies in particular for the optimal $\gamma_\pm$:

$$M\,|h(u, \gamma_+) - \bar{w}(u)| \leq \sqrt{1 + 16k^2} \leq 5k.$$

Thus:

$$\sup_u |h(u, \gamma_+)| \leq \sup_u \{|h(u, \gamma_+) - \bar{w}(u)| + |\bar{w}(u)|\} \leq 5k/M + k \leq 6k/M.$$

We conclude that for $\tilde{C}$ as above and $\tilde{C}' = 6k/(M\delta)$, it holds that:

$$\begin{aligned} &\mathbb{P}\left[\sup_z \left|\gamma_+^{(n)}(z)\right| < \tilde{C}n^\beta \mathbb{E}\left[\gamma_+^{(n)}(Z_i)\right],\ \sup_u \left|h(u,\ \gamma_+^{(n)})\right| < \tilde{C}'\mathbb{E}\left[\gamma_+^{(n)}(Z_i)\right]\right] \\ \geq\ &\mathbb{P}\left[B_{n,1} \cap B_{n,2}\right] \to 1 \text{ as } n \to \infty. \end{aligned}$$

□

A.9 Proof of Proposition 10

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 $Z_i \,\big|\, U_i \sim \mathcal{N}\left(U_i, \nu^2\right)$. Then $\tilde{Z}_i = Z_i/\nu_i \,\big|\, U_i \sim \mathcal{N}\left(U_i/\nu_i, 1\right)$ 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 $\mu_{(0)}(z)$; the arguments for $\mu_{(1)}(z)$ are identical.

Upper bound: Fix $\tilde{c} > 0$. Let $\tilde{\alpha}_{(0)}(u) = \tilde{c} + \alpha_{(0)}(u) \in [\tilde{c}, 1 + \tilde{c}]$. Let $H \ll G$ be the probability measure with

$$\frac{dH}{dG}(u) = \frac{\tilde{\alpha}_{(0)}(u)}{\int \tilde{\alpha}_{(0)}(u) dG(u)},$$

and write $h(z) = \int \varphi(z-u) dH(u)$. Then:

$$\mu_{(0)}(z) = \mathbb{E}\left[\alpha_{(0)}(U_i) \,\middle|\, Z_i = z\right] = \mathbb{E}\left[\tilde{\alpha}_{(0)}(U_i) \,\middle|\, Z_i = z\right] - \tilde{c} = \frac{h(z) \cdot \int \tilde{\alpha}_{(0)}(u) dG(u)}{f(z)} - \tilde{c}.$$

Taking the derivative:

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

We next bound the three terms appearing in the expression above. First, we already saw that $\int \tilde{\alpha}_{(0)}(u) dG(u) \cdot \frac{h(z)}{f(z)} = \mu_{(0)}(z) + \tilde{c}$ with $\mu_{(0)}(z) \in [0,1]$ and so this term is upper bounded in absolute value by $1 + \tilde{c}$. Next, by Lemma A.1. in Jiang and Zhang [2009] (which we state and prove at the end of this section for self-containedness) 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 $h(z)/f(z)$:

$$h(z) = \frac{\int \tilde{\alpha}_{(0)}(u) \varphi(z-u) dG(u)}{\int \tilde{\alpha}_{(0)}(u) dG(u)} \geq \frac{\tilde{c}}{1+\tilde{c}} \cdot \int \varphi(z-u) dG(u) = \frac{\tilde{c}}{1+\tilde{c}} \cdot f(z).$$

Applying the triangle inequality and putting everything together:

$$\left|\frac{d}{dz}\mu_{(0)}(z)\right| \leq \inf_{\tilde{c}>0} \left\{(1+\tilde{c}) \cdot \left(\sqrt{-\log(2\pi f^2(z))} + \sqrt{-\log\left(\frac{2\pi\tilde{c}^2}{(1+\tilde{c})^2} f^2(z)\right)}\right)\right\}.$$

Taking $\tilde{c} = 1 + \sqrt{2}$ and noting that $2(1+\tilde{c}) < 7$ leads to the bound:

$$\left|\frac{d}{dz}\mu_{(0)}(z)\right| \leq 7\sqrt{-\log(\pi f^2(z))}.$$

Continuing, the second derivative of $\mu_{(0)}(z)$ is equal to:

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

Applying Lemma A.1. in Jiang and Zhang [2009] a second time 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 $\left|\mu_{(0)}(z) + \tilde{c}\right| \leq 1 + \tilde{c}$, that we already bounded $|\mu'_{(0)}(z)|$, $f'(z)/f(z)$ above, and the triangle inequality we conclude.

Lower bound: Without loss of generality, we consider the case that $z = 0$. Let δ_u denote the point mass measure at $\{u\}$. We take $G = \frac{1-w}{2} \cdot (\delta_{-t} + \delta_t) + w \cdot \delta_0$ for parameters $w \in [0, 1]$, $t > 0$ which we will specify later and $\alpha_{(0)}(u) = \mathbf{1}(u = 0)$. Then:

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

To simplify notation we write $\mu(\cdot) = \mu_{(0)}(\cdot)$. By direct calculation we can verify that

$$\mu''(0) = -w\frac{\varphi(0)f(0) + \varphi(0)f''(0)}{f^2(0)}, \quad f''(0) = (1-w)(t^2-1)\varphi(t) - w\varphi(0).$$

Next choose $w = \varphi(t)$, so that $f(0) = (1 + \varphi(0) - \varphi(t))\varphi(t)$ and

$$\mu''(0) = -\varphi(0)\frac{(1-\varphi(t)) \cdot t^2}{(1+\varphi(0)-\varphi(t))^2}.$$

Finally, we pick t so that $\varphi(t) = \rho$. It then holds in particular that $f(0) \geq \rho$ and using the fact that $\varphi(t) \in (0, 1/\sqrt{2\pi}]$, we get:

$$|\mu''(0)| \geq \frac{1}{10}t^2 = \frac{1}{10}(-\log(2\pi\rho^2)).$$

□

A.9.1 Lemma A.1. in Jiang and Zhang [2009]

Lemma A.1. (Jiang and Zhang [2009])**.** *Let G be a distribution on $\mathbb{R}$, let φ be the standard normal density function and let $f_G(z) = \int \varphi(z-u)dG(u)$ be the density of the convolution $G \star \varphi$. Then:*

$$0 \ \leq \ \left(\frac{f_G'(z)}{f_G(z)}\right)^2 \ \leq \ \frac{f_G''(z)}{f_G(z)} + 1 \ \leq \ -\log\left(2\pi f_G^2(z)\right).$$

Proof. Let $U \sim G$ and $Z \mid U \sim \mathcal{N}(U, 1)$. We may verify the following three equalities:

$$\mathbb{E}\left[U - z \mid Z = z\right] = \frac{f_G'(z)}{f_G(z)}, \ \mathbb{E}\left[(U-z)^2 \mid Z = z\right] = \frac{f_G''(z)}{f_G(z)} + 1,$$
$$\mathbb{E}\left[\sqrt{2\pi}\exp\left((U-z)^2/2\right) \mid Z = z\right] = 1/f_G(z).$$

Then, by Jensen's inequality:

$$\left(\frac{f_G'(z)}{f_G(z)}\right)^2 = \mathbb{E}\left[U - z \mid Z = z\right]^2 \leq \mathbb{E}\left[(U-z)^2 \mid Z = z\right] = \frac{f_G''(z)}{f_G(z)} + 1.$$

Next, define the convex function $h(x) = \sqrt{2\pi}\exp(x/2)$ with inverse $h^{-1}(y) = \log(y^2/(2\pi))$. Applying Jensen's inequality again, we see that:

$$\frac{f_G''(z)}{f_G(z)} + 1 \leq h^{-1}\left(\mathbb{E}\left[h\left((U-z)^2\right) \mid Z = z\right]\right) \ = \ h^{-1}(1/f_G(z)) = -\log\left(2\pi f_G^2(z)\right).$$

□

B Computational details

B.1 Computation of worst-case bias

B.1.1 Notation

In this section we explain how to compute the worst-case bias in (21). The main idea behind our optimization algorithm is to define $A_{(0)}$, resp. T as the signed measure that is absolutely continuous with respect to G with density $\alpha_{(0)}(u)$, resp. $\tau(u)$. We will parameterize the optimization problem in terms of optimization variables that represent G, $A_{(0)}$ and T. To simplify notation, we define the following linear functionals:

$$\begin{aligned} L_{h,+}(H) &= \int h(u,\ \gamma_+)dH(u), \\ L_{h,-}(H) &= \int h(u,\ \gamma_-)dH(u), \\ L_w(H) &= \int w(u)dH(u). \end{aligned}$$

Then we can write:

$$\text{Bias}\left[\gamma_\pm,\ \tau_w;\ \alpha_{(0)}(\cdot),\ \tau(\cdot),\ G\right] = \frac{L_{h,+}(A_{(0)}) + L_{h,+}(T)}{L_{h,+}(G)} - \frac{L_{h,-}(A_{(0)})}{L_{h,-}(G)} - \frac{L_w(T)}{L_w(G)}, \tag{S5}$$

a sum-of-ratios of linear functionals. We propose solving:

$$\sup_{G, A_{(0)}, T} \quad \frac{L_{h,+}(A_{(0)}) + L_{h,+}(T)}{L_{h,+}(G)} - \frac{L_{h,-}(A_{(0)})}{L_{h,-}(G)} - \frac{L_w(T)}{L_w(G)} \tag{S6a}$$

$$\text{s.t.} \quad G \in \mathcal{G}_n, \tag{S6b}$$

$$0 \leq \frac{dA_{(0)}}{dG}(u) \leq 1 \text{ for all } u, \tag{S6c}$$

$$0 \leq \frac{dT}{dG}(u) \leq 2M \text{ for all } u. \tag{S6d}$$

We explain why this problem is equivalent to the problem we care about solving in (21). There are two observations:

1. We first show that it suffices to reduce attention to $T(\cdot)$ satisfying (S6d) instead of more general $T(\cdot)$ that satisfy the heterogeneity constraint in (20). Fix G, $A_{(0)}$ and T that are feasible for (21). Let $\bar{\tau}$ be such that $|dT(u)/dG - \bar{\tau}| \leq M$ and define $\check{T} = \bar{\tau} \cdot G + T$. Then $d\check{T}(u)/dG = dT(u)/dG + \bar{\tau} \in [0,\ 2M]$. Hence $G, A_{(0)}, \check{T}$ are also feasible. Furthermore, we may check that $(G, A_{(0)}, T)$ and $(G, A_{(0)}, \check{T})$ lead to the same value of the objective $\text{Bias}\left[\gamma_\pm,\ \tau_w;\ \cdot, \cdot, \cdot\right]$ in (S5).

2. We next show that we may ignore the absolute value in (21). Fix feasible G, $A_{(0)}$ and T. Suppose we replace $A_{(0)}$ and T by $\check{A}_{(0)} = G - A_{(0)}$ and $\check{T} = 2M \cdot G - T$. Then $d\check{A}_{(0)}(u)/dG = 1 - dA_{(0)}(u)/dG \in [0, 1]$, and so the constraint (S6c) will continue to be satisfied and similarly, $d\check{T}(u)/dG = 2M - dT(u)/dG \in [0, 2M]$, and so the constraint (S6d) will also continue to be satisfied. Hence G, $\check{A}_{(0)}$ and $\check{T}$ also are feasible and we may further check that the objective value switches sign compared to its original value and retains its absolute value. Thus optimization problem (S6) is implicitly maximizing the absolute value of the objective.

B.1.2 Optimizing the sum-of-ratios objective

We now explain how (S6) may be solved numerically.[S1] First, we may reduce the number of ratios in the objective by the Charnes and Cooper [1962] transformation, as follows:

$$\sup_{\check{G},\check{A}_{(0)},\check{T},\xi} \quad L_{h,+}(\check{A}_{(0)}) + L_{h,+}(\check{T}) - \frac{L_{h,-}(\check{A}_{(0)})}{L_{h,-}(\check{G})} - \frac{L_w(\check{T})}{L_w(\check{G})} \tag{S7a}$$

$$\text{s.t.} \quad \check{G} \in \left\{\xi \cdot \tilde{G} : \tilde{G} \in \mathcal{G}_n\right\}, \tag{S7b}$$

$$0 \leq \frac{d\check{A}_{(0)}}{d\check{G}}(u) \leq 1 \text{ for all } u, \tag{S7c}$$

$$0 \leq \frac{d\check{T}}{d\check{G}}(u) \leq 2M \text{ for all } u, \tag{S7d}$$

$$L_{h,+}(\check{G}) = 1, \tag{S7e}$$

$$\xi \geq 0. \tag{S7f}$$

The optimization variables are $\xi \geq 0$, $\check{G}$, $\check{A}_{(0)}$, and $\check{T}$. Their interpretation is as follows: $\xi = 1/L_{h,+}(G)$, $\check{G} = \xi \cdot G$, $\check{A}_{(0)} = \xi \cdot A_{(0)}$, and $\check{T} = \xi \cdot T$. Next consider solving (S7) subject to the additional linear constraints that

$$L_{h,-}(\check{G}) = \zeta, \quad L_w(\check{G}) = \kappa,$$

for fixed values of ζ, κ. In more detail, let:

$$\mathcal{L}(\zeta,\kappa) = \sup_{\check{G},\check{A}_{(0)},\check{T},\xi} \quad L_{h,+}(\check{A}_{(0)}) + L_{h,+}(\check{T}) - \frac{1}{\zeta}L_{h,-}(\check{A}_{(0)}) - \frac{1}{\kappa}L_w(\check{T}) \tag{S8a}$$

$$\text{s.t.} \quad \text{(S7b), (S7c), (S7d), (S7e), (S7f)}, \tag{S8b}$$

$$L_{h,-}(\check{G}) = \zeta, \tag{S8c}$$

$$L_w(\check{G}) = \kappa. \tag{S8d}$$

For fixed values of ζ, κ, the above is a linear program. Thus we may solve (S7) by profiling over ζ and κ and repeatedly solving (S8). Formally, let:

$$\begin{aligned} \underline{\zeta} &= \inf_{\check{G},\check{A}_{(0)},\check{T},\xi} \left\{L_{h,-}(\check{G})\right\} \text{ s.t. (S7b), (S7c), (S7d), (S7e), (S7f)}, \\ \text{and } \bar{\zeta} &= \sup_{\check{G},\check{A}_{(0)},\check{T},\xi} \left\{L_{h,-}(\check{G})\right\} \text{ s.t. (S7b), (S7c), (S7d), (S7e), (S7f)}. \end{aligned} \tag{S9}$$

Furthermore, for $\zeta \in [\underline{\zeta}, \bar{\zeta}]$, let:

$$\begin{aligned} \underline{\kappa}(\zeta) &= \inf_{\check{G},\check{A}_{(0)},\check{T},\xi} \left\{L_w(\check{G})\right\} \text{ s.t. } L_{h,-}(\check{G}) = \zeta, \text{ (S7b), (S7c), (S7d), (S7e), (S7f)}, \\ \text{and } \bar{\kappa}(\zeta) &= \sup_{\check{G},\check{A}_{(0)},\check{T},\xi} \left\{L_w(\check{G})\right\} \text{ s.t. } L_{h,-}(\check{G}) = \zeta, \text{ (S7b), (S7c), (S7d), (S7e), (S7f)}. \end{aligned} \tag{S10}$$

Then the worst-case bias we are interested in is equal to:

$$\sup\left\{ \sup\left\{\mathcal{L}(\zeta,\kappa) \ : \ \kappa \in [\underline{\kappa}(\zeta),\ \bar{\kappa}(\zeta)]\right\} \ : \ \zeta \in \left[\underline{\zeta},\ \bar{\zeta}\right] \right\}. \tag{S11}$$

[S1] Such sum-of-ratios optimization problems have been studied in the optimization literature, see e.g., Benson [2007], Konno and Abe [1999] and references therein.

B.1.3 Discretization considerations

To turn the above construction into a practical computational algorithm, we need to solve optimization problems (S8), (S9), and (S10), as well as solve the profiling task (S11). We will achieve this by finely discretizing. We refer to Ignatiadis and Wager [2022, Supplement D] for a more detailed discussion regarding discretization considerations and describe our implementation choices here.

Instead of optimizing over the space of all distributions for the latent variable U, we optimize over all distributions supported on the equidistant finite grid from $a_{\min}$ to $a_{\max}$ with B points:

$$\mathcal{K}(B, a_{\min}, a_{\max}) = \left\{a_{\min},\, a_{\min} + \frac{a_{\max} - a_{\min}}{B},\, a_{\min} + 2\frac{a_{\max} - a_{\min}}{B},\, \ldots,\, a_{\max}\right\}. \quad \text{(S12)}$$

Our default choice uses $B = 499$, $a_{\min} = \min\{Z_1, \ldots, Z_n\}$, $a_{\max} = \max\{Z_1, \ldots, Z_n\}$ for Gaussian noise distributions and $B = 399$, $a_{\min} = 10^{-4}$ and $a_{\max} = 1 - 10^{-4}$ for Binomial noise (the latter choice of $a_{\min}$, $a_{\max}$ for the Binomial empirical Bayes problem is used by default in Koenker and Gu [2017]).

By enumerating the grid elements as $\mathcal{K}(B, a_{\min}, a_{\max}) = \{u_1, \ldots, u_{B+1}\}$, we may represent every distribution G supported on this set by the probabilities $g_j = \mathbb{P}_G[U = u_j]$ assigned to u_j. The g_j lie on the probability simplex. Furthermore, we may represent $\check{G}$ by $\check{g}_j$, which satisfy:

$$\sum_{j=1}^{B+1} \check{g}_j = \xi,\ \check{g}_j \geq 0.$$

Analogously, we may represent $\check{T}, \check{A}_{(0)}$ by $(B+1)$-dimensional vectors and we only need to apply the constraints in (S7c) and (S7d) for $u \in \mathcal{K}(B, L, U)$. Hence, after the aforementioned discretization, all of (S8), (S9), and (S10) turn into finite-dimensional linear programs that we optimize using the MOSEK solver [ApS, 2020].

To solve the profiling problem (S11), instead of considering all $\zeta \in [\underline{\zeta}, \bar{\zeta}]$, we only consider $\zeta \in \mathcal{K}(49, \underline{\zeta}, \bar{\zeta})$. Meanwhile, for each such ζ we discretize $[\underline{\kappa}(\zeta), \bar{\kappa}(\zeta)]$ as an equidistant grid with distance between grid points of at most $\underline{\zeta}/5$. Hence we solve the discretized (S11) by solving a finite number of linear programs.

B.2 Computation of worst-case curvature

The construction for optimizing the worst-case curvature is very similar to the construction in Supplement B.1, i.e., after profiling we reduce the optimization problem to a sequence of linear programming tasks. We provide a sketch here.

Our starting point is the ratio representation (25) of $\mu(\cdot) = \mu_{(w)}(\cdot)$ (omitting the subscript $_{(w)}$ henceforth), which we may write as:

$$\mu(z) = \frac{N(z)}{D(z)},$$

where $N(\cdot)$, resp. $D(\cdot)$ are the numerator, resp. denominator in (25). Then, assuming $N(\cdot), D(\cdot)$ are twice differentiable, we get by the chain rule that:

$$\mu''(z) = \frac{N''(z)}{D(z)} - 2\frac{N'(z)D'(z)}{D^2(z)} - \frac{N(z)D''(z)}{D^2(z)} + 2\frac{N(z)(D'(z))^2}{D(z)^3}.$$

The absolute value of the above is the quantity we seek to maximize. The critical observation now is that all of $D(z), D'(z), D''(z)$ are linear functionals of G. Similarly, $N(z), N'(z), N''(z)$ are linear functionals of A, defined as the measure $\ll G$ with $dA(u)/dG = \alpha_{(w)}(u)$. Applying the Charnes and Cooper [1962] transformation (as we did in Supplement B.1) we may rescale the optimization variables G and A as $\check{G}, \check{A}$, such that $\int p(z \mid u) d\check{G}(u) = 1$. Writing $\check{N}(\cdot), \check{D}(\cdot)$ for the corresponding numerator and denominator

$$\check{N}(\cdot) = \int p(\cdot \mid u) d\check{A}(u), \;\; \check{D}(\cdot) = \int p(\cdot \mid u) d\check{G}(u),$$

we get by the above transformation that $\check{D}(z) = 1$, and hence:

$$\mu''(z) = \check{N}''(z) - 2\check{N}'(z)\check{D}'(z) - \check{N}(z)\check{D}''(z) + 2\check{N}(z)(\check{D}'(z))^2. \tag{S13}$$

To conclude we use a profiling argument as in Supplement B.1. Concretely, fix κ, ζ and consider the linear (in the optimization variables) constraints:

$$\check{D}'(z) = \zeta, \; \check{D}''(z) = \kappa.$$

Under these constraints, we have that:

$$\mu''(z) = \check{N}''(z) - 2\zeta\check{N}'(z) - \kappa\check{N}(z) + 2\zeta^2\check{N}(z).$$

This objective is linear in the optimization variables (for fixed values of ζ, κ) and so we can maximize/minimize it with respect to the constraints by linear programming.