Journal of Econometrics 218 (2020) 216–241

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Regression discontinuity design with many thresholds

Marinho Bertanha

Department of Economics, University of Notre Dame, 3060 Jenkins Nanovic Halls, Notre Dame, IN 46556, United States of America

ARTICLE INFO

Article history:
Received 11 November 2017
Received in revised form 29 March 2019
Accepted 19 September 2019
Available online 27 February 2020

JEL classification:
C14
C21
C52
I21

Keywords:
Regression discontinuity
Multiple cutoffs
Average treatment effect
Peer-effects

ABSTRACT

Numerous empirical studies employ regression discontinuity designs with multiple cutoffs and heterogeneous treatments. A common practice is to normalize all the cutoffs to zero and estimate one effect. This procedure identifies the average treatment effect (ATE) on the observed distribution of individuals local to existing cutoffs. However, researchers often want to make inferences on more meaningful ATEs, computed over general counterfactual distributions of individuals, rather than simply the observed distribution of individuals local to existing cutoffs. This paper proposes a consistent and asymptotically normal estimator for such ATEs when heterogeneity follows a non-parametric function of cutoff characteristics in the sharp case. The proposed estimator converges at the minimax optimal rate of root-n for a specific choice of tuning parameters. Identification in the fuzzy case, with multiple cutoffs, is impossible unless heterogeneity follows a finite-dimensional function of cutoff characteristics. Under parametric heterogeneity, this paper proposes an ATE estimator for the fuzzy case that optimally combines observations to maximize its precision.

1. Introduction

Applications of regression discontinuity design (RDD) have become increasingly popular in economics since the late 1990s (Black, 1999; Angrist and Lavy, 1999; Van der Klaauw, 2002). One of RDD's main advantages is identification of a local causal effect under minimal functional form assumptions. More recently, with increasing availability of richer data sets, there have been many applications with multiple cutoffs and treatments (for example, Black et al. (2007), Egger and Koethenbuerger (2010), De La Mata (2012) and Pop-Eleches and Urquiola (2013)). Existing one-cutoff RDD methods applied to each individual cutoff produce many local effects that are estimated using only a few observations near each cutoff. Researchers often prefer one takeaway summary effect that is more precisely estimated by pooling all the data. The meaning of a summary effect crucially depends on heterogeneity assumptions and weights imposed on the different local effects.

Applied studies with multiple cutoffs often normalize all cutoffs to zero and use the one-cutoff estimator. This normalization procedure estimates an average of local treatment effects weighted by the relative density of individuals near each of the cutoffs (Cattaneo et al. (2016), Proposition 3). Such an average effect would be a meaningful summary measure only in two cases: (i) local treatment effects are all identical and the weighting scheme does not matter; or (ii) local treatment effects are heterogeneous but the researcher is only interested in the average effect on the individuals

E-mail address: mbertanha@nd.edu.
URL: http://www.nd.edu/~mbertanh.

https://doi.org/10.1016/j.jeconom.2019.09.010
0304-4076/© 2020 Elsevier B.V. All rights reserved.

near the existing cutoffs. However, researchers are often interested in combining observed data with assumptions weaker than (i) to make inferences on counterfactual scenarios more general than (ii).[1]

This paper proposes a novel estimation procedure for average treatment effects (ATE). These ATEs are more valuable summary measures than the average effect estimated by the normalization procedure described above for two reasons. First, the researcher explicitly chooses the counterfactual distribution of the ATE, and this distribution may include individuals at or between existing cutoffs. Second, the researcher does not need to assume any specific functional form for the heterogeneity of treatment effects across different cutoffs. As an example of an application, suppose we are interested in estimating the effect of Medicaid benefits on health care utilization. Medicaid eligibility is triggered by income cutoffs that vary across states. Existing one-cutoff RDD methods identify the average effect on individuals with income equal to the income cutoffs. However, most interesting policy questions require the average effect over the entire range of income values in the data.

The framework for RDD with many thresholds is introduced here using a simple example based on the work of Pop-Eleches and Urquiola (2013), PU from now on. Using a wealth of variation of cutoffs from high school assignments in Romania, PU provide rigorous evidence of the impacts of attending a better school on students' academic performance. The economic logic of this application is briefly summarized as follows. A central planner assigns students to high schools based on their scores from a placement test. High schools have limited capacities and are ranked by their qualities. The central planner ranks students by their scores and assigns each of them to the best school available. Each student i submits her score X_i (forcing variable) to the central planner who, based on the entire distribution of scores, determines a minimum test score c_j (cutoff) for admission to each high school j. The quality of high school j is denoted d_j (treatment dose).

The RDD assignment is assumed sharp for now. That is, students attend the best high school available to them based on their score and the cutoffs that apply to them. As the test score crosses an admission threshold c_j, the quality of the school the student attends changes from d_{j-1} to d_j. Local average effects are denoted by $\mathbb{E}[Y_i(d_j)-Y_i(d_{j-1})|X_i = c_j] = \beta(c_j, d_{j-1}, d_j)$, where $Y_i(d)$ is the potential academic achievement student i has if attending a high school of quality d, and $\beta(c, d, d')$ is the treatment effect function. Heterogeneity of local effects comes from values of cutoffs and treatment doses that change across the different cutoffs. PU give a particularly illustrative application, because it exhibits sufficient variation in cutoff and treatment doses to generate ATEs with substantially greater economic relevance than the typical average based on normalizing all of the cutoffs to zero.

Numerous other examples of RDD with multiple cutoffs and treatments exist in different fields of economics. For instance, Egger and Koethenbuerger (2010) study the effect of the size of city government councils on municipal expenditures, where council size is determined by population cutoffs. De La Mata (2012) estimates the effects of Medicaid benefits on health care utilization, where Medicaid eligibility is triggered by income cutoffs that vary across states. Agarwal et al. (2017) and De Giorgi et al. (2017) look at multiple cutoffs on credit scores, used by banks to make credit decisions. Education economics also provides a variety of applications. Angrist and Lavy (1999) and Hoxby (2000) use class size rules to estimate the impact of class size on student achievement. Hoxby (2000) utilizes variation in cutoff values from specific school district class size rules. Several researchers exploit different school starting dates to estimate the impact of educational attainment on various outcomes, for example, Dobkin and Ferreira (2010), and McCrary and Royer (2011). Duflo et al. (2011) analyze school cohorts that are split into low and high-achieving classes based on test scores, where each school has its own cutoff score. Garibaldi et al. (2012) look at different income cutoffs that determine tuition subsidies to study the impact of tuition payment on the probability of late graduation from university. In short, despite many applications with variation in cutoffs and treatment doses, a lack of theory on how to combine observations from all cutoffs impedes our ability to estimate economically-relevant average effects.

Whether local effects can be combined into an average effect depends on how comparable the researcher believes these effects are. The comparability of local treatment effects essentially depends on the heterogeneity of treatment doses and on the heterogeneity of the treatment effect function $\beta(c, d, d')$. This paper considers two types of assumptions regarding these two aspects of heterogeneity. The first heterogeneity assumption says that treatment doses are credibly quantifiable by some variable d. For example, PU find behavioral evidence that average student performance at each school is a good summary measure for school quality. Another example is the case of a single treatment being triggered by varying cutoffs, as when each state has its own income threshold for Medicaid coverage. The second heterogeneity assumption specifies a parametric functional form for $\beta(c, d, d')$ guided by economic theory or *a priori* knowledge of the researcher. For example, in a class size application like Hoxby's (2000), a functional form based on Lazear's (2001) model of achievement can be derived as a function of class size. Another example is given by Bajari et al. (2017) who present a principal–agent model to study how insurers reimburse hospitals. The marginal reimbursement rate is discontinuous on health expenditures.

This paper proposes a consistent and asymptotically normal estimator for the ATE of a counterfactual distribution of treatment assignments specified by the researcher. A counterfactual policy scenario specifies the distribution of (c, d, d'), and the ATE is the integral of $\beta(c, d, d')$ weighted by such a distribution. The ability to predict effects of counterfactual

[1] In a RDD setting with multiple cutoffs and treatments, it is unreasonable to expect that different local treatment effects are always identical. For example, Pop-Eleches and Urquiola (2013) find that the impact of going to a better high school on academic achievement is heterogeneous across students with different ability levels. Another example is De La Mata (2012), who finds that the eligibility for Medicaid benefits decreases the probability of having private health insurance more strongly for lower income individuals. Although I allow for heterogeneous effects across cutoffs, counterfactual analysis requires a pooling and a policy invariance assumption (Section 2).

policies depends crucially on assuming that the distribution of potential outcomes $Y_i(d)$ does not depend on the initial schedule of cutoff-dose values. This policy invariance assumption, along with the first heterogeneity assumption, allows the researcher to choose counterfactual distributions with support more general than the discrete set of cutoff-dose values observed in the data.

The estimator proposed in this paper approximates the ATE integral by averaging estimates of $\beta(c, d, d')$ at existing cutoffs using a proper weighting scheme. Under the first heterogeneity assumption with $\beta(c, d, d')$ non-parametric, the proposed ATE estimator is shown to be consistent and asymptotically normal. This result is novel, because estimation of the non-parametric function $\beta(c, d, d')$ is only possible at deterministic points of the domain, and that creates an additional source of bias. Asymptotic normality requires both the number of observations and cutoffs to grow to infinity, and I provide sufficient conditions on their rate of growth. I demonstrate that the minimax rate of ATE estimation in this setting is root-n, and that the proposed estimator attains the minimax optimal rate for a specific choice of tuning parameters. This extends the previous literature on minimax optimality of non-parametric estimation of regression functions at a boundary point to estimation of averages of these regression functions.

Many applications of RDD with multiple cutoffs are, in fact, fuzzy rather than sharp. In the high school assignment example, a student may choose to attend a high school other than the school she is originally eligible to attend. Multiple treatments result in multiple compliance behaviors, and one-cutoff identification results do not apply. Building on classic definitions of compliance behaviors (Imbens and Rubin, 1997), I define compliance groups in terms of *changes* in treatment eligibility and receipt. "Ever-compliers" are those whose treatment received changes if and only if it changes to the treatment dose for which they become eligible for. I assume that individuals never change into a treatment dose different from the dose of eligibility, a "no-defiance" condition. In the high school example, if the test score of a student currently in school B increases so as to grant her access to school A, no-defiance implies she either chooses to attend school A or stay at school B, and that she is not triggered to attend some other school C.

This paper shows that even local identification in fuzzy RDD with finite multiple treatments is impossible unless the class of treatment effect functions of ever-compliers is restricted to a finite-dimensional class. Important empirical analyses of fuzzy RDD with multiple treatments include those of Angrist and Lavy (1999), Chen and Van der Klaauw (2008), and Hoekstra (2009); nevertheless, this is the first paper to define compliance and study causal identification in a general framework for multi-cutoff fuzzy RDD. This framework lays out conditions for the interpretation of two-stage least squares (2SLS) estimates in applications of multi-cutoff fuzzy RDD, a common practice in applied work. The second heterogeneity assumption states that the treatment effect function is of a parametric class. This assumption allows for consistent and asymptotically normal estimation of ATEs on ever-compliers. It also results in efficiency gains, because observations are optimally combined across cutoffs to minimize the mean squared error (MSE) of the ATE estimator.

The rapid growth in the number of applications of RDD in economics in the late 1990s was accompanied by substantial theoretical contributions for inference in the one-cutoff case. Identification and estimation in the sharp and fuzzy cases were formalized by Hahn et al. (2001). Fan and Gijbels (1996) and Porter (2003) demonstrated low-order bias and rate optimality of the local polynomial estimator. Recent theoretical contributions have addressed the optimal bandwidth choice (Imbens and Kalyanaraman, 2012), alternative asymptotic approximations with better finite sample properties (Calonico et al., 2014), quantile treatment effects (Frandsen et al., 2012), kink treatment effects (Dong, 2018b), and the difficulty of uniform inference (Bertanha and Moreira, 2019).

The contribution of this paper is more closely related to the study of treatment effect extrapolation of Angrist (2004), Bertanha and Imbens (2019), Dong and Lewbel (2015), Angrist and Rokkanen (2015), and Rokkanen (2015). These last two authors use observations on additional covariates. They restrict the relationship between the heterogeneity of treatment effects after conditioning on these covariates to obtain identification away from the cutoff. This paper differs from these other contributions, because the variation of multiple cutoffs and doses identify ATEs over distributions of individuals both between and at cutoffs, without additional covariates.

The remainder of this paper is organized as follows. Section 2 presents the notation and lays out basic assumptions. Section 3 describes the ATE estimator for the sharp case and proves asymptotic normality. It is divided into two sub-sections. Section 3.1 treats ATEs of discrete counterfactual distributions, which is a straightforward generalization of one-cutoff RDD. Section 3.2 is novel; it studies ATEs of continuous counterfactual distributions under the first heterogeneity assumption. Section 4 analyzes the fuzzy case. Appendix A contains all proofs. Supplemental Appendix B collects auxiliary results to the proofs in Appendix A.[2]

2. Setup

This section sets up the framework for RDD with multiple cutoffs. There are P sub-populations of individuals indexed by $p = 1, \ldots, P$. An example of a sub-population may be a town-year in the high school application, or a state in the Medicaid example. Each individual i in sub-population p is fully characterized by a vector of random variables $(X_{i,p}, U_{i,p})$ drawn iid across i from each sub-population. The forcing variable $X_{i,p}$ is a scalar score that governs eligibility for treatment, and it lives in a compact interval $\mathcal{X} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}]$; $U_{i,p}$ is a vector of unobserved heterogeneity. Individual (i, p) receives a

[2] Appendix B is available online at https://doi.org/10.1016/j.jeconom.2019.09.010.

treatment dose $D_{i,p}$ from a set of possible treatments $\mathcal{D}$. The outcome variable $Y_{i,p}$ is determined by a function $\mathbb{Y}$ of the individual characteristics and treatment,

$$Y_{i,p} = \mathbb{Y}(X_{i,p}, D_{i,p}, U_{i,p}). \tag{1}$$

I start with the simpler sharp RDD setting and defer the fuzzy RDD case to Section 4. In the sharp case, the treatment received by the individual is a deterministic function of the forcing variable. For an individual with forcing variable $X_{i,p}$ close to a cutoff c, the treatment dose is d if $X_{i,p} < c$, or d' if $X_{i,p} \geq c$. Hahn et al. (2001) demonstrate that continuity of the conditional mean of outcomes is sufficient to identify average causal effects for individuals local to the cutoff c.

Lemma 1. *Assume that $\mathbb{E}[\mathbb{Y}(X_{i,p}, d, U_{i,p})|X_{i,p} = x]$ is a continuous function of x for the treatment doses d and d' in the neighborhood of the cutoff c. Then, the average causal effect for individuals with $X_{i,p} = c$ is identified:*

$$\begin{aligned}\mathbb{E}\left[\mathbb{Y}(X_{i,p}, d', U_{i,p}) - \mathbb{Y}(X_{i,p}, d, U_{i,p}) \;\middle|\; X_{i,p} = c\right] \\ = \lim_{e\downarrow 0}\left\{ \mathbb{E}[Y_{i,p} \mid X_{i,p} = c + e] - \mathbb{E}[Y_{i,p} \mid X_{i,p} = c - e] \right\}.\end{aligned} \tag{2}$$

Lemma 1 generalizes to the case of multiple cutoffs and treatments under the assumption of continuity of $\mathbb{E}[\mathbb{Y}(X_{i,p}, d, U_{i,p})|X_{i,p} = x]$ as a function of x for every $d \in \mathcal{D}$. Many cutoffs arise because data sets may have many sub-populations with few cutoffs (e.g. Medicaid benefit with one cutoff per state, many states); or few sub-populations with many cutoffs (e.g. Romanian high schools with one town and many schools). The ability to exploit variation in cutoff-dose values relies on the following pooling assumption.

Assumption 1 (*Pooling*)**.** For any $\{d, d'\} \subset \mathcal{D}$, the conditional expectation

$$\mathbb{E}\left[\mathbb{Y}(X_{i,p}, d', U_{i,p}) - \mathbb{Y}(X_{i,p}, d, U_{i,p}) \mid X_{i,p} = x\right]$$

as a function of x does not depend on p.

Assumption 1 does not restrict average outcomes to be the same across different sub-populations. It is less restrictive than common specifications for pooling data in applied work, for example, time-trends and sub-population fixed effects. The pooling assumption says that individuals with the same forcing variable that undergo the same change in treatment have the same average response across different sub-populations. The rest of the paper builds on Assumption 1, and it becomes irrelevant to distinguish sub-populations. Thus, I drop the subscript p and focus on the case of one population with multiple cutoffs.

The cutoffs are ordered such that $c_1 < c_2 < \cdots < c_K$. Sharp RDD means that an individual with forcing variable X_i is deterministically assigned to a treatment dose $D_i = D(X_i)$ according to the following rule:

$$D(x) = \begin{cases} d_0 & \text{if } c_0 \leq x < c_1 \\ d_1 & \text{if } c_1 \leq x < c_2 \\ \vdots & \\ d_K & \text{if } c_K \leq x \leq c_{K+1} \end{cases} \tag{3}$$

where $c_0 = \underline{\mathcal{X}}$, and $c_{K+1} = \overline{\mathcal{X}}$. Each cutoff is characterized by three variables: the scalar threshold c_j; the treatment dose d_{j-1} the individual receives if $c_{j-1} \leq X_i < c_j$; and the treatment dose d_j the individual receives if $c_j \leq X_i < c_{j+1}$. Let $\mathbf{c}_j = (c_j, d_{j-1}, d_j)$. The schedule of cutoffs and treatment doses is given by the non-random set $\mathcal{C}_K = \left\{\mathbf{c}_j\right\}_{j=1}^{K}$. The richness of set $\mathcal{C}_K$ increases as the researcher collects more data.[3]

The data generating process is summarized as follows. Values for the forcing variable X_i and heterogeneity U_i are drawn iid $i = 1, \ldots, n$ from a joint distribution. Given $D(x)$, these n individuals are assigned to different treatment doses $D_i = D(X_i)$. The observed outcome is determined by $Y_i = \mathbb{Y}(X_i, D_i, U_i)$. The econometrician observes the schedule of cutoffs and treatment doses $D(x)$ and (Y_i, X_i, D_i) for $i = 1, \ldots, n$. Following Rubin's model of potential outcomes, let $Y_i(d) = \mathbb{Y}(X_i, d, U_i)$, and assume continuity of $\mathbb{E}[Y_i(d)|X_i = x]$ for every $d \in \mathcal{D}$. A simple extension of Lemma 1 identifies average effects at every cutoff $\mathbf{c} \in \mathcal{C}_K$,

$$\begin{aligned}\beta(\mathbf{c}) =& \mathbb{E}[Y_i(d') - Y_i(d)|X_i = c] \\ =& \lim_{e\downarrow 0}\left\{ \mathbb{E}[Y_i \mid X_i = c + e] - \mathbb{E}[Y_i \mid X_i = c - e] \right\}.\end{aligned} \tag{4}$$

Data with multiple cutoff-dose values allow the researcher to learn the causal effect of a variety of dose changes applied to individuals at various levels of the forcing variables. This fact opens the possibility of using observed data to

[3] The validity of the RDD depends crucially on exogeneity of cutoffs and no manipulation of the forcing variable X by individuals. See McCrary (2008) for a test of forcing variable manipulation. Bajari et al. (2017) present a modified RDD estimator that is consistent under forcing variable manipulation in a class of structural models.

estimate the effect of new policy changes. The individual response function $\mathbb{Y}$ may well depend on the initial assignment of treatments D_i, and it could potentially change under counterfactual policies. Unless such dependence is restricted, it becomes impossible to use existing data to infer the effect of new policies. The remainder of this paper relies on the following policy-invariance assumption.

Assumption 2 (*Policy Invariance*)**.** Regardless of the distribution of (X_i, D_i, U_i) in a counterfactual policy, individual outcomes are always generated by a fixed response function $\mathbb{Y}$, that is, $Y_i = \mathbb{Y}(X_i, D_i, U_i)$.

The methods of this paper leverage RDD variation in cutoff-dose values to make inferences on average effects of policy changes. A policy change is a counterfactual distribution of changes in treatment doses that are randomly applied to individuals, conditional on the forcing variable. An individual i is assigned to a change in treatment dose from D_i^* to D_i^{**}, where the distribution of (D_i^*, D_i^{**}) is independent of U_i after conditioning on X_i. Under Assumption 2, the average causal effect of such an experiment is

$$\begin{aligned}\mu &= \mathbb{E}\left[\ \mathbb{Y}(X_i, D_i^{**}, U_i) - \mathbb{Y}(X_i, D_i^*, U_i)\ \right]\\ &= \mathbb{E}\left[\ \mathbb{E}\left(\mathbb{Y}(X_i, D_i^{**}, U_i) - \mathbb{Y}(X_i, D_i^*, U_i)\ \middle|\ D_i^{**}, D_i^*, X_i\right)\ \right]\\ &= \mathbb{E}\left[\ \mathbb{E}\left(\mathbb{Y}(X_i, D_i^{**}, U_i) - \mathbb{Y}(X_i, D_i^*, U_i)\ \middle|\ X_i\right)\ \right]\\ &= \mathbb{E}\left[\ \beta(X_i, D_i^*, D_i^{**})\ \right] \end{aligned} \tag{5}$$

where the last equality uses the definition of β in Eq. (4). The average effect μ equals an average of the β function over the counterfactual distribution of (X_i, D_i^*, D_i^{**}). The inference methods of this paper first identify β from RDD with many cutoffs, then identify the average of β under a counterfactual distribution pre-specified by the researcher. In a similar setting, Cattaneo et al. (2016) study identification under conditions equivalent to Assumptions 1 and 2 (respectively, their Assumptions 5a and 5b).

The definition of μ captures both the direct effect of changing D, and the composition effect of a change in the distribution of D conditional on X. To investigate the direct effects of D, Rothe (2012) proposes methods for inference on partial policy effects that preserve the distribution of ranks of (D, X) unchanged, thus controlling for composition effects. Although not the focus of this paper, Rothe's methods may be combined with the RDD identification strategy to study partial policy effects.

3. Average treatment effects in the sharp case

This section investigates estimation and inference of averages of the non-parametric function β under sharp RDD with many cutoffs. First, I treat the case of qualitative treatment doses. This is a straightforward extension of single-cutoff RDDs which identify ATEs of discrete counterfactual distributions, with support contained in $\mathcal{C}_K$. Second, I treat the case of quantitative treatment doses, that is, the first heterogeneity assumption. Substantial variation in cutoff-dose values allows for novel methods that estimate ATEs with support more general than $\mathcal{C}_K$.

3.1. Discrete counterfactuals

Consider applications of RDD where the treatment dose variable has a qualitative nature, and is *not* credibly summarized by a real-valued metric. For example, Hastings et al. (2013) study the assignment of students into different degree programs in universities in Chile. There are multiple cutoffs on a test score, but different cutoffs switch students to completely different programs, e.g. physics, engineering, economics, etc. This limits the ability to combine local effects across cutoffs, which restricts ATEs to counterfactual distributions with discrete support contained in $\mathcal{C}_K$. In this section, it is not possible to identify effects of policies that places weight on cutoff-dose combinations (c, d, d') that are not in $\mathcal{C}_K$.

The focus is on discrete counterfactual distributions with probability mass function $\omega^d(\mathbf{c})$ where $\omega_j^d = \omega^d(\mathbf{c}_j)$ for every j. For example, in the high school assignment application, a new policy may reallocate students with test scores marginally across the existing cutoffs. The weight ω_j^d represents the probability mass of students with test score equal to c_j that undergo a change in school quality from d_{j-1} to d_j in the reallocation policy.

The parameter of interest is the average effect on these students, which is a weighted average of local effects at the existing cutoffs:

$$\mu^d = \sum_{j=1}^{K} \omega_j^d\ \beta(\mathbf{c}_j).$$

Identification follows from Eq. (4).[4] Estimation is conducted in two steps. The first step uses local polynomial regressions (LPR) near each cutoff c_j to non-parametrically estimate

$$B_j = \lim_{e \downarrow 0} \left\{\ \mathbb{E}[Y_i|X_i = c_j + e] - \mathbb{E}[Y_i|X_i = c_j - e]\ \right\}. \tag{6}$$

[4] The common practice of normalizing all cutoffs to zero and estimating only one effect produces an estimator consistent for μ^d with weights $\omega_j^d = f(c_j)/\sum_l f(c_l)$ where f is the probability density function of X.

The researcher chooses a bandwidth parameter $h_{1j} > 0$ for each cutoff, a kernel density function $k(.)$, and the order of the polynomial regression $\rho_1 \in \mathbb{Z}_+$. A polynomial in X is fitted on each side of the cutoff, and the estimator $\hat{B}_j$ is the difference between the intercepts of these two polynomial regressions:

$$\hat{B}_j = \hat{a}_j^+ - \hat{a}_j^- \tag{7}$$

$$(\hat{a}_j^+, \hat{\mathbf{b}}_j^+) = \underset{(a,\mathbf{b})}{\text{argmin}} \sum_{i=1}^n \left\{ k\left(\frac{X_i - c_j}{h_{1j}}\right) v_i^{j+} \left[Y_i - a - b_1(X_i - c_j) - \cdots - b_{\rho_1}(X_i - c_j)^{\rho_1}\right]^2 \right\} \tag{8}$$

$$(\hat{a}_j^-, \hat{\mathbf{b}}_j^-) = \underset{(a,\mathbf{b})}{\text{argmin}} \sum_{i=1}^n \left\{ k\left(\frac{X_i - c_j}{h_{1j}}\right) v_i^{j-} \left[Y_i - a - b_1(X_i - c_j) - \cdots - b_{\rho_1}(X_i - c_j)^{\rho_1}\right]^2 \right\} \tag{9}$$

where

$$v_i^{j+} = \mathbb{I}\{c_j \leq X_i < c_j + h_{1j}\},\ v_i^{j-} = \mathbb{I}\{c_j - h_{1j} < X_i < c_j\}, \tag{10}$$

and $\mathbf{b} = (b_1, \ldots, b_{\rho_1})$. The estimator $\widehat{B}_j$ uses observations with X_i in the estimation window $[c_j - h_{1j}, c_j + h_{1j}]$. The choice of bandwidths may allow the windows to overlap at consecutive cutoffs. However, it must be the case that $c_j + h_{1j} < c_{j+1}$ and $c_j \leq c_{j+1} - h_{j+1}$ for $j = 1, \ldots, K-1$. This ensures that $Y_i = Y_i(d_j)$ for $X_i \in [c_j, c_j + h_{1j}]$, and $Y_i = Y_i(d_{j-1})$ for $X_i \in [c_j - h_{1j}, c_j)$.[5]

In the second step, the researcher averages out $\hat{B}_j$ to obtain the estimator $\widehat{\mu}^d$:

$$\widehat{\mu}^d = \sum_{j=1}^K \omega_j^d \hat{B}_j. \tag{11}$$

For the case of one cutoff, Hahn et al. (2001) and Porter (2003) derive the asymptotic normal distribution of the LPR estimator $\hat{B}_j$. I build on their arguments to derive the asymptotic distribution of $\widehat{\mu}^d$ under the assumptions listed below.

Assumption 3. The kernel density function $k : \mathbb{R} \to \mathbb{R}$ is symmetric around zero, has compact support $[-M, M]$ for some $M \in (0, \infty)$, and is Lipschitz continuous.

Assumption 4. **(a)** The distribution of X_i has probability density function $f(x)$ that is continuous and has bounded support $\mathcal{X} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}]$; **(b)** $f(x)$ is differentiable with bounded derivative $\nabla_x f(x)$.

Assumption 5. Let $\rho_1 \in \mathbb{Z}_+$ be the order of the first-step LPR. For arbitrary $d \in \mathcal{D}$, **(a)** $R(x, d) = \mathbb{E}[Y_i(d)|X_i = x]$ is $\rho_1 + 1$ times continuously differentiable wrt x; its $(\rho_1 + 1)$th partial derivative wrt x is denoted as $\nabla_x^{\rho_1+1} R(x, d)$; **(b)** $\sigma^2(x, d) = \mathbb{V}[Y_i(d)|X_i = x]$ where $\mathbb{V}$ is the variance operator; $\sigma^2(x, d)$ is continuously differentiable wrt x; its partial derivative wrt x is denoted as $\nabla_x \sigma^2(x, d)$; $\sigma^2(x, d)$ is bounded away from zero, and $\mathbb{E}[|Y_i(d) - R(X_i, d)|^3 |X_i]$ is bounded.

Theorem 1. *Suppose Assumptions 3–5 hold. Let $\underline{h}_1 = \min_j h_{1j}$ and $\overline{h}_1 = \max_j h_{1j}$. As $n \to \infty$, assume that $\overline{h}_1 \to 0$, $\overline{h}_1 / \underline{h}_1 = O(1)$, $n\overline{h}_1 \to \infty$, and $(n\overline{h}_1)^{1/2} \overline{h}_1^{\rho_1+1} = O(1)$. Then,*

$$\frac{\widehat{\mu}^d - \mathcal{B}_n^d - \mu^d}{\left(\mathcal{V}_n^d\right)^{1/2}} \xrightarrow{d} N(0, 1)$$

where the bias $\mathcal{B}_n^d$ and variance $\mathcal{V}_n^d$ terms are characterized as follows:

$$\mathcal{B}_n^d = \frac{1}{(\rho_1 + 1)!} \sum_{j=1}^K h_{1j}^{\rho_1+1} f(c_j) \left[\nabla^{\rho_1+1} R(c_j, d_j) e_1' G_n^{j+} - \nabla^{\rho_1+1} R(c_j, d_{j-1}) e_1' G_n^{j-}\right] \gamma^* \tag{12}$$

$$\mathcal{V}_n^d = n\mathbb{E}\left\{ \varepsilon_i^2 \left[\sum_{j=1}^K \frac{\omega_j^d}{nh_{1j}} k\left(\frac{X_i - c_j}{h_{1j}}\right) e_1' \left(v_i^{j+} \mathbb{E}[G_n^{j+}] - v_i^{j-} \mathbb{E}[G_n^{j-}] \right) \widetilde{H}_i^j \right]^2 \right\}, \tag{13}$$

with $\varepsilon_i = Y_i - \mathbb{E}[Y_i|X_i]$; $H(u) = [u^0, u^1, \ldots, u^{\rho_1}]'$ is a $(\rho_1 + 1) \times 1$ vector-valued function, $\widetilde{H}_i^j = H(h_{1j}^{-1}(X_i - c_j))$, and $G_n^{j\pm} = (nh_{1j})^{-1} \sum_{i=1}^n v_i^{j\pm} k(h_{1j}^{-1}(X_i - c_j)) \widetilde{H}_i^j \widetilde{H}_i^{j'}$ is a $(\rho_1 + 1) \times (\rho_1 + 1)$ matrix; $v_i^{j\pm}$ are defined in Eq. (10); $\gamma^* = [\gamma_{\rho_1+1} \ \ \ldots \ \ \gamma_{2\rho_1+1}]'$,

[5] This is the first-step estimation procedure for one sub-population with K cutoffs. In many settings, the data have many sub-populations $p = 1, \ldots, P$ with one or more cutoffs $j = 1, \ldots, K(p)$ in each sub-population. In that case, the researcher first estimates $\widehat{B}_{j,p}$ for every j in each sub-population p. Then, Assumption 1 allows for pooling of $\widehat{B}_{j,p}$ across p in the second step.

for $\gamma_d = \int_0^1 k(u)u^d du$; *and* e_1 *is the* $(\rho_1 + 1 \times 1)$ *vector with one in its first coordinate and zero otherwise. Furthermore,* $\left(\mathcal{V}_n^d\right)^{-1/2} = O\left(\left(n\bar{h}_1\right)^{1/2}\right)$, *and* $\left(\mathcal{V}_n^d\right)^{-1/2} \mathcal{B}_n^d = O_P\left(\left(n\bar{h}_1\right)^{1/2} \bar{h}_1^{\rho_1+1}\right)$.

The variance of $\widehat{\mu}^d$ is consistently estimated by

$$\widehat{\mathcal{V}}_n^d = \sum_{i=1}^n \left\{ \widehat{\varepsilon}_i^2 \left[\sum_{j=1}^K \frac{\omega_j^d}{nh_{1j}} k\left(\frac{X_i - c_j}{h_{1j}}\right) e_1' \left(v_i^{j+} G_n^{j+} - v_i^{j-} G_n^{j-}\right) \widetilde{H}_i^j \right]^2 \right\}. \tag{14}$$

The squared residuals $\widehat{\varepsilon}_i^2$ are computed by a nearest-neighbor matching estimator, as suggested by Calonico et al. (2014) (CCT from now on):

$$\widehat{\varepsilon}_i^2 = \frac{3}{4}\left(Y_i - \frac{1}{3}\sum_{l=1}^3 Y_{\ell(i,l)}\right)^2, \tag{15}$$

and $\ell(i, l)$ is the index of the lth closest X to X_i that lies within the same cutoffs c_j and c_{j+1} that X_i does. CCT's Theorem A3 demonstrates that $\widehat{\mathcal{V}}_n^d / \mathcal{V}_n^d \xrightarrow{p} 1$ in the case of one cutoff, and a straightforward generalization yields the same conclusion for a finite number of cutoffs. If the bandwidth choices are such that the standardized bias term $\left(\mathcal{V}_n^d\right)^{-1/2} \mathcal{B}_n^d$ differs from zero asymptotically, then inference must be done using a bias-corrected estimator. A practical way of doing bias correction is to increase the order of the polynomial from ρ_1 to $\rho_1 + 1$ and compute $\widehat{\mu}^{d'}$ and $\widehat{\mathcal{V}}_n^{d'}$ using the same bandwidth choices as $\widehat{\mu}^d$ and $\widehat{\mathcal{V}}_n^d$. It follows that $(\widehat{\mathcal{V}}_n^{d'})^{-1/2}(\widehat{\mu}^{d'} - \mu^d) \xrightarrow{d} N(0, 1)$.

The multi-cutoff setup of Theorem 1 allows for choices of bandwidths that produce overlapping estimation windows in finite samples. For example, if $c_1 + h_{11} > c_2 - h_{12}$, the estimator $\widehat{B}_2$ uses some of the same observations that the estimator $\widehat{B}_1$ does. In theory, a finite number of cutoffs with shrinking bandwidths leads to non-overlapping estimation windows in large samples. As a consequence, the asymptotic variance of $\sqrt{n\bar{h}_1}(\widehat{\mu}^d - \mathcal{B}_n^d - \mu^d)$ may not approximate its finite-sample variance well in case of overlap. Instead, the variance term in (13) takes into account overlap because its formula is constructed based on the finite-sample variance.

In practice, implementation of $\widehat{\mu}^d$ requires the researcher to choose bandwidths $h_{1j} > 0$, the polynomial order $\rho_1 \in \mathbb{Z}_+$, and a kernel density function $k(\cdot)$. In the one-cutoff case, common choices in applied work include the edge kernel $k(u) = \mathbb{I}\{|u| \leq 1\}(1 - u)$, local linear regression $\rho_1 = 1$, and a bandwidth choice that minimizes the mean squared error (MSE) of estimation. Recent work by Imbens and Kalyanaraman (2012) (IK from now on) provides a practical data-driven rule for choosing the bandwidth in the case of one cutoff. With multiple cutoffs, an interesting aspect of the optimal bandwidth problem is the variance reduction from overlapping estimation windows.[6] A formal investigation on optimal bandwidths in the multi-cutoff case is deferred to future work.

A simple recommendation to implement Theorem 1 is to use the IK bandwidth based on local linear regressions with the edge kernel applied to the sub-sample pertaining to each cutoff. These bandwidths produce asymptotic bias, and valid inference must use a bias-corrected estimator and its variance. Use local quadratic regressions ($\rho_1 = 2$) with the edge kernel and the same bandwidths as before to compute the consistent bias-corrected estimator $\widehat{\mu}^{d'}$ and its variance $\widehat{\mathcal{V}}_n^{d'}$. Calonico et al. (2018) propose shrinking MSE-optimal bandwidths as a rule of thumb to improve finite sample coverage of confidence intervals. As means of a robustness check, the researcher may shrink the IK bandwidths by multiplying them by $n^{-1/20}$, and examine the resulting confidence intervals (Section 4.1, Calonico et al. (2018)).

3.2. Continuous counterfactuals

The first heterogeneity assumption allows the researcher to identify counterfactual ATEs with support more general than $\mathcal{C}_K$. An empirical application satisfies the first heterogeneity assumption if the treatment dose is credibly quantifiable in a real-valued variable d. For example, in the high school assignment of PU, the treatment dose is a quality measure for each school. Possible measures of school quality include the average test score of peers, the average number of teachers, or funding per student. An infinite amount of data gives rise to a countably-infinite set of cutoff-dose values $\mathcal{C}_\infty$. In terms of the high school assignment example, a large number of towns and years produce substantial variation in cutoff-dose values. Define $\mathcal{C}$ to be the convex hull of $\mathcal{C}_\infty$. If variation in cutoff-dose values is sufficiently rich, then ATEs with counterfactual distributions supported in $\mathcal{C}$ are identified (Lemma 2).

I focus on scalar treatment doses d and counterfactual distributions with continuous probability density function $\omega^c(\mathbf{c})$. Minor changes to the setup can accommodate multivariate d and discrete or mixed counterfactual distributions. The ATE is defined as

$$\mu^c = \int_{\mathcal{C}} \omega^c(\mathbf{c})\beta(\mathbf{c})\, d(\mathbf{c}). \tag{16}$$

[6] Following Eq. (7), $COV(\widehat{B}_j, \widehat{B}_{j+1}) = COV(\widehat{a}_j^+ - \widehat{a}_j^-, \widehat{a}_{j+1}^+ - \widehat{a}_{j+1}^-) = COV(\widehat{a}_j^+, -\widehat{a}_{j+1}^-) < 0$ because $\widehat{a}_j^+$ and $\widehat{a}_{j+1}^-$ use some of the same observations in the case of overlap.

Lemma 2. *Assume that an infinite amount of data has sufficient variation such that (i) $\mathcal{C}_\infty$ is dense in its convex hull $\mathcal{C}$; and that (ii) $\beta(\mathbf{c})$ is a continuous function over $\mathcal{C}$. Then, μ^c is identified.*

The researcher may impose further heterogeneity restrictions to reduce the dimension of $\beta(\mathbf{c})$ and increase the set of possible counterfactual distributions. For instance, linear returns to school quality say that $\beta(c, d, d')$ depends on $(c, d'-d)$ instead of (c, d, d'). This implies that $\beta(\mathbf{c}) = \phi(c)(d' - d)$ for a smooth function $\phi(c)$, and changes the dimension of set $\mathcal{C}_K$. See Fig. 2 in Section 6 for an empirical illustration. Medicaid coverage is an example of binary treatment that is triggered by various income cutoffs across states. In the case of binary treatment, the treatment effect function depends only on the cutoff value, that is, $\beta(c, d, d') = \phi(c)$. Identification of averages of $\beta(\mathbf{c})$ requires identification of averages of $\phi(c)$, which relies on infinitely many cutoff values that cover a compact interval on the real line. For example, such variation identifies the average effect of giving Medicaid benefits to an entire neighborhood of individuals within the range of income cutoffs seen in the data.[7]

The parameter μ^c is estimated in two steps. The first step is identical to the procedure described in Eqs. (7)–(9). That is, LPRs produce estimates $\widehat{B}_j$, $j = 1, \dots, K$. The second step computes a weighted average of the first-step estimates, using specially designed weights $\{\Delta_j\}_{j=1}^K$ that I call "correction weights",

$$\widehat{\mu}^c = \sum_{j=1}^{K} \Delta_j \hat{B}_j. \tag{17}$$

Unlike the intuition of the discrete case, the correction weight Δ_j is not necessarily equal or proportional to $\omega_j^c = \omega^c(\mathbf{c}_j)$. An analytical expression for Δ_j is given below in Eq. (22), and constructed as follows. The correction weight Δ_j is the contribution of estimate $\widehat{B}_j$ to the integral $\int_{\mathcal{C}} \omega^c(\mathbf{c})\widehat{\beta}(\mathbf{c})\, d(\mathbf{c})$, where $\widehat{\beta}(\mathbf{c})$ is a non-parametric estimate of $\beta(\mathbf{c})$. A weighted regression of $\widehat{B}_j$ on polynomial functions of $\mathbf{c}_j$ centered at $\mathbf{c}$ produces the estimate $\widehat{\beta}(\mathbf{c})$. The researcher specifies the order of the polynomials $\rho_2 \in \mathbb{Z}_+$, and a bandwidth $h_2 > 0$ that defines an estimation neighborhood around $\mathbf{c} \in \mathcal{C}$. The estimate $\widehat{\beta}(\mathbf{c})$ is the intercept of the following weighted least squares regression:

$$\widehat{\boldsymbol{\eta}} = \underset{\eta}{\text{argmin}} \left(\widehat{\mathbf{B}} - \mathbf{E}(\mathbf{c})\boldsymbol{\eta}\right)' \boldsymbol{\Omega}(\mathbf{c}; h_2) \left(\widehat{\mathbf{B}} - \mathbf{E}(\mathbf{c})\boldsymbol{\eta}\right) \tag{18}$$

where

$$\widehat{\mathbf{B}} = \left[\widehat{B}_1, \; \dots, \; \widehat{B}_K\right]' \text{ is a } K \times 1 \text{ vector;} \tag{19}$$

$$\boldsymbol{\Omega}(\mathbf{c}; h_2) = \text{diag}\left\{\Omega_j(\mathbf{c}; h_2)\right\}_{j=1}^K \textit{ is a } K \times K \textit{ matrix, with} \tag{20}$$

$$\Omega_j(\mathbf{c}; h_2) = k\left(\frac{c_j - c}{h_2}\right) k\left(\frac{d_{j-1} - d}{h_2}\right) k\left(\frac{d_{j-1} - d'}{h_2}\right);$$

$$\mathbf{E}(\mathbf{c}) = [E_1(\mathbf{c}), \; \dots, \; E_K(\mathbf{c})]' \textit{ is a } K \times J \textit{ matrix, where} \tag{21}$$

$$E_j(\mathbf{c}) \textit{ is a } J \times 1 \textit{ vector with all polynomials of the form}$$

$$p_{\boldsymbol{\gamma}}(\mathbf{c}_j - \mathbf{c}) = (c_j - c)^{\gamma_1}(d_{j-1} - d)^{\gamma_2}(d_j - d')^{\gamma_3}$$

$$\textit{for } \boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3) \in \mathbb{Z}_+^3, \; \gamma_1 + \gamma_2 + \gamma_3 \leq \rho_2, \; \min\{\gamma_2, \gamma_3\} = 0,$$

$$J = 2\frac{(\rho_2 + 2)!}{2!\rho_2!} - (\rho_2 + 1), \textit{ where } ! \textit{ denotes factorial,}$$

$$\textit{and the first element of } E_j(\mathbf{c}) \textit{ is } 1.$$

The formula for Δ_j comes from integrating $\omega^c(\mathbf{c})\widehat{\beta}(\mathbf{c})$:

$$\begin{aligned}\int_{\mathcal{C}} \omega^c(\mathbf{c})\widehat{\beta}(\mathbf{c})\, d\mathbf{c} &= \int_{\mathcal{C}} \omega^c(\mathbf{c}) e_1' \left(\mathbf{E}(\mathbf{c})'\boldsymbol{\Omega}(\mathbf{c}; h_2)\mathbf{E}(\mathbf{c})\right)^{-1} \sum_j \Omega_j(\mathbf{c}; h_2) E_j(\mathbf{c}) \widehat{B}_j \;\; d(\mathbf{c}) \\ &= \sum_j \int_{\mathcal{C}} \omega^c(\mathbf{c}) e_1' \left(\mathbf{E}(\mathbf{c})'\boldsymbol{\Omega}(\mathbf{c}; h_2)\mathbf{E}(\mathbf{c})\right)^{-1} \Omega_j(\mathbf{c}; h_2) E_j(\mathbf{c})\, d(\mathbf{c})\, \widehat{B}_j\end{aligned}$$

[7] For the Medicaid example, De La Mata (2012) has many income cutoffs that differ by state, age, and year. De La Mata's Table 1 suggests variation between US$ 21,394 and US$36,988. Other examples of rich variation in cutoff values include: (i) Agarwal et al. (2017) who have 714 credit-score cutoffs distributed between 620 and 800 (see their Figure II(E)); and (ii) Hastings et al. (2013) who have at least 1100 cutoffs on admission scores varying between 529.15 and 695.84 (refer to their online appendix's Table A.I.I). Although Angrist and Lavy (1999) have few cutoff values, the pattern of their Figure I suggests variation in dose changes across grades and schools. In such cases, non-parametric identification of β is possible for a range of dose changes at a few cutoff values.

Table 1
Precision of estimators - Choice of h_1.

n	K	Bias				Variance				MSE			
		$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widetilde{\mu}$	$\widetilde{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widetilde{\mu}$	$\widetilde{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widetilde{\mu}$	$\widetilde{\mu}^{bc}$
Overlap													
1789	20	0.0617	−0.0015	−0.2504	−0.2390	0.0079	0.0164	0.0073	0.0110	0.0117	0.0164	0.0700	0.0681
10 120	40	0.0206	0.0003	−0.1245	−0.1216	0.0013	0.0022	0.0012	0.0017	0.0017	0.0022	0.0167	0.0165
27 886	60	0.0095	−0.0002	−0.0834	−0.0821	0.0005	0.0007	0.0004	0.0006	0.0005	0.0007	0.0074	0.0074
57 244	80	0.0056	−0.0001	−0.0625	−0.0618	0.0002	0.0003	0.0002	0.0003	0.0003	0.0003	0.0041	0.0041
100 000	100	0.0038	−0.0001	−0.0499	−0.0496	0.0001	0.0002	0.0001	0.0002	0.0001	0.0002	0.0026	0.0026
No Overlap													
1789	20	0.0698	−0.0019	−0.2423	−0.2402	0.0139	0.0386	0.0122	0.0288	0.0188	0.0386	0.0710	0.0865
10 120	40	0.0227	−0.0002	−0.1225	−0.1221	0.0021	0.0051	0.0020	0.0043	0.0026	0.0051	0.0170	0.0192
27 886	60	0.0106	−0.0006	−0.0824	−0.0825	0.0007	0.0017	0.0007	0.0016	0.0009	0.0017	0.0075	0.0084
57 244	80	0.0062	0.0000	−0.0620	−0.0617	0.0004	0.0008	0.0003	0.0008	0.0004	0.0008	0.0042	0.0046
100 000	100	0.0039	−0.0003	−0.0498	−0.0497	0.0002	0.0005	0.0002	0.0004	0.0002	0.0005	0.0027	0.0029

Notes: The table reports simulated bias, variance, and mean squared error (MSE) for four estimators ($\widehat{\mu}, \widehat{\mu}^{bc}, \widetilde{\mu}, \widetilde{\mu}^{bc}$), two choices of first-step bandwidth (overlap and no overlap), and five sample sizes n and respective numbers of cutoffs K. The second-step bandwidth is set to $h_2 = 3/(K+1)$, which minimizes MSE of $\widehat{\mu}$. Refer to Table 2 for different choices of h_2. The number of simulations is 10,000.

Table 2
Precision of estimators - Choice of h_2.

$h_2 \cdot (K+1)$	$(n, K) = (1789, 20)$						$(n, K) = (10120, 40)$					
	Bias		Variance		MSE		Bias		Variance		MSE	
	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$
3	0.0617	−0.0015	0.0079	0.0164	0.0117	0.0164	0.0206	0.0003	0.0013	0.0022	0.0017	0.0022
4	0.1128	−0.0012	0.0079	0.0143	0.0206	0.0143	0.0376	0.0004	0.0013	0.0020	0.0027	0.0020
5	0.1708	−0.0014	0.0079	0.0133	0.0370	0.0133	0.0584	0.0005	0.0013	0.0019	0.0047	0.0019
6	0.2322	−0.0014	0.0079	0.0128	0.0618	0.0128	0.0826	0.0004	0.0013	0.0019	0.0081	0.0019
7	0.2935	−0.0015	0.0079	0.0126	0.0940	0.0126	0.1097	0.0004	0.0013	0.0019	0.0133	0.0019
8	0.3513	−0.0015	0.0079	0.0124	0.1313	0.0124	0.1392	0.0004	0.0013	0.0019	0.0207	0.0019
9	0.4019	−0.0015	0.0080	0.0122	0.1695	0.0122	0.1707	0.0004	0.0013	0.0018	0.0304	0.0018
10	0.4420	−0.0015	0.0080	0.0122	0.2034	0.0122	0.2036	0.0004	0.0013	0.0018	0.0427	0.0018
11	0.4680	−0.0015	0.0081	0.0122	0.2272	0.0122	0.2375	0.0004	0.0013	0.0018	0.0577	0.0018
12	0.4773	−0.0015	0.0082	0.0121	0.2360	0.0121	0.2720	0.0004	0.0013	0.0018	0.0753	0.0018

Notes: The table reports simulated bias, variance, and mean squared error (MSE) for two estimators ($\widehat{\mu}, \widehat{\mu}^{bc}$), ten choices of second-step bandwidth ($h_2 \in \{3/(K+1), \ldots, 12/(K+1)\}$), and the two smallest sample sizes n and respective numbers of cutoffs K. The first-step bandwidth is set to $h_1 = 1/(K+1)$ (overlap). Naive estimators ($\widetilde{\mu}, \widetilde{\mu}^{bc}$) are not in this table because they are not affected by the choice of h_2. The number of simulations is 10,000.

$$= \sum_j \underbrace{\int_{\mathcal{C}} \omega^c(\mathbf{c}) \frac{det\left(\mathbf{E}(\mathbf{c})'\boldsymbol{\Omega}(\mathbf{c}; h_2)\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})\right)}{det\left(\mathbf{E}(\mathbf{c})'\boldsymbol{\Omega}(\mathbf{c}; h_2)\mathbf{E}(\mathbf{c})\right)} d(\mathbf{c})}_{\equiv \Delta_j} \widehat{B}_j \tag{22}$$

$$= \sum_j \Delta_j \widehat{B}_j, \tag{23}$$

where the third equality uses the Cramer rule, and $\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})$ is a $K \times J$ matrix equal to $\mathbf{E}(\mathbf{c})$ except for the first column, which is replaced by the $K \times 1$ vector e_j. The vector e_j has one in its jth entry and zero otherwise.

The main contribution of this paper concerns inference on μ^c where $\beta(\mathbf{c})$ is estimated non-parametrically and then averaged across cutoffs. This is not the first paper to study estimation of averages of non-parametric functions; for example, see Newey (1994). The novelty here is that the non-parametric estimation step only occurs at K fixed boundary points $\mathbf{c}_j$. A necessary condition for consistency of $\hat{\mu}^c$ is an "infill type of asymptotics", that is, K grows large with the sample size n, and $\mathcal{C}_K$ becomes dense in its convex hull $\mathcal{C}$. Assumption 6 makes the dependence of K, h_{1j}, h_2, and c_j on n explicit with a subscript. The main text omits the subscript n whenever possible to simplify notation.

Assumption 6. **(a)** The schedule of cutoffs and doses comes from a triangular array of fixed constants $\mathcal{C}_{K_n} = \left\{\mathbf{c}_{j,n}\right\}_{j=1}^{K_n}$ that depends on the sample size n; $\mathcal{C}_{K_n}$ becomes a countably infinite set $\mathcal{C}_\infty$ as $n \to \infty$; $\mathcal{C}_\infty$ is dense in its convex hull $\mathcal{C}$; **(b)** given the first-step bandwidth sequences $h_{1j,n}$, assume that $c_{j,n} + h_{1j,n} < c_{j+1,n}$ and $c_{j,n} \leq c_{j+1,n} - h_{1j+1,n}$ for all $j = 1, \ldots, K_n - 1$; and **(c)** given the second-step bandwidth sequence $h_{2,n}$ and polynomial order ρ_2, define $\mathbf{E}_n(\mathbf{c})$ and

$\mathbf{\Omega}_n(\mathbf{c}; h_{2,n})$ as in Eqs. (20)–(21) for each n. Assume there exists a positive definite $J \times J$ matrix $\mathbf{Q}$ such that

$$\sup_{\mathbf{c}\in\mathcal{C}} \left\| K_n h_{2,n}^3 \left[\mathbf{E}_n(\mathbf{c}/h_{2,n})'\mathbf{\Omega}_n(\mathbf{c}; h_{2,n})\mathbf{E}_n(\mathbf{c}/h_{2,n})\right]^{-1} - \mathbf{Q}\right\| = o\,(1)\,.$$

For large K, cutoff-dose values must be uniformly distributed on the domain $\mathcal{C}$ such that $\mathbf{E}(\mathbf{c}/h_2)'\mathbf{\Omega}(\mathbf{x}; h_2)\mathbf{E}(\mathbf{c}/h_2)$ is invertible and of magnitude Kh_2^3, that is, K times the volume of every h_2-neighborhood of $\mathbf{c}$, for every $\mathbf{c}$ in $\mathcal{C}$. These conditions are satisfied in a variety of examples of triangular arrays of points. In Section B.3 of the supplemental appendix, these conditions are verified for one example of a triangular array. Asymptotic normality also relies on additional smoothness conditions on the moments of the data.

Assumption 7. **(a)** $R(x, d) = \mathbb{E}[Y_i(d)|X_i = x]$ is a $\bar{\rho}$ times continuously differentiable function with $\bar{\rho} = \max\{\rho_1{+}2, \rho_2{+}2\}$, where ρ_1 and ρ_2 are polynomial degrees in the first and second steps; the $\bar{\rho}$th partial derivative of $R(x, d)$ with respect to x is denoted $\nabla_x^{\bar{\rho}} R(x, d)$; **(b)** $\sigma^2(x, d) = \mathbb{V}[Y_i(d)|X_i = x]$ is a continuous function bounded away from zero; and **(c)** $\exists M \in (0, \infty)$ such that $\mathbb{P}[|Y_i(d) - R(X_i, d)| < M] = 1$ for $\forall d \in \mathcal{D}$.

Theorem 2 states the rate conditions under which the estimator $\widehat{\mu}^c$ has an asymptotic normal distribution. Estimation of the ATE consists of approximating the integral of the treatment effect function by a weighted sum of the values of such function at a finite number of points in its domain. The approximation error converges to zero as the number of points grows large. Function evaluations B_j are estimated by $\widehat{B}_j$. The correction weights guarantee that the integral approximation error converges to zero faster than the estimation error.

Theorem 2. *Suppose Assumptions 3–7 hold. As $n \to \infty$, assume that $K \to \infty$, $\overline{h}_1 \to 0$, $\overline{h}_1/\underline{h}_1 = O(1)$, and $h_2 \to 0$ such that **(i)** $\left(Kn\overline{h}_1\right)^{1/2} \overline{h}_1^{\rho_1+1} = O(1)$; **(ii)** $K^{1/2} \log n/\left(n\overline{h}_1\right)^{1/2} = o(1)$, and $K\overline{h}_1 = O(1)$; and **(iii)** $\left(Kn\overline{h}_1\right)^{1/2} h_2^{\rho_2+1} = O(1)$, and $1/Kh_2^3 = O(1)$. Then,*

$$\frac{\widehat{\mu}^c - \mathcal{B}_{1n}^c - \mathcal{B}_{2n}^c - \mu^c}{\left(\mathcal{V}_n^c\right)^{1/2}} \xrightarrow{d} N(0, 1). \tag{24}$$

The first-step bias $\mathcal{B}_{1n}^c$ and variance $\mathcal{V}_n^c$ terms are defined as in Eqs. (12)–(13) except that Δ_j replaces ω_j^d; the second-step bias $\mathcal{B}_{2n}^c$ is characterized as follows:

$$\mathcal{B}_{2n}^c = \int_{\mathcal{C}} \omega^c(\mathbf{c}) \sum_{(\gamma_1,\gamma_2,\gamma_3)} \sum_{j=1}^{K} \left\{ \frac{(c_j - c)^{\gamma_1}(d_{j-1} - d)^{\gamma_2}(d_j - d')^{\gamma_3}}{\gamma_1!\gamma_2!\gamma_3!} \nabla_c^{\gamma_1} \nabla_d^{\gamma_2} \nabla_{d'}^{\gamma_3} \beta(c, d, d') \right.$$
$$\left. \frac{det\left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c}; h_2)\mathbf{E}_{\mathbf{0}\leftarrow e_j}(\mathbf{c})\right)}{det\left(\mathbf{E}(\mathbf{c})'\mathbf{\Omega}(\mathbf{c}; h_2)\mathbf{E}(\mathbf{c})\right)} \right\} d\mathbf{c}, \tag{25}$$

where the first sum runs over all triplets $(\gamma_1, \gamma_2, \gamma_3) \in \mathbb{Z}_+^3$ such that $\gamma_1 + \gamma_2 + \gamma_3 = \rho_2 + 1$, and $\min\{\gamma_2, \gamma_3\} = 0$. Furthermore, $\left(\mathcal{V}_n^c\right)^{-1/2} = O\left(\left(Kn\overline{h}_1\right)^{1/2}\right)$, $\left(\mathcal{V}_n^c\right)^{-1/2} \mathcal{B}_{1n}^c = O_P\left(\left(Kn\overline{h}_1\right)^{1/2} \overline{h}_1^{\rho_1+1}\right)$, and $\left(\mathcal{V}_n^c\right)^{-1/2} \mathcal{B}_{2n}^c = O\left(\left(Kn\overline{h}_1\right)^{1/2} h_2^{\rho_2+1}\right)$.

A consistent estimator for $\mathcal{V}_n^c$ is

$$\widehat{\mathcal{V}}_n^c = \sum_{i=1}^{n} \left\{ \widehat{\varepsilon}_i^2 \left[\sum_{j=1}^{K} \frac{\Delta_j}{nh_{1j}} k\left(\frac{X_i - c_j}{h_{1j}}\right) e_1' \left(v_i^{j+} G_n^{j+} - v_i^{j-} G_n^{j-}\right) \widetilde{H}_i^j \right]^2 \right\}, \tag{26}$$

where $\widehat{\varepsilon}_i^2$ is computed using Eq. (15). Lemma B.10 in the supplemental appendix's Section B.4 demonstrates that $\widehat{\mathcal{V}}_n^c/\mathcal{V}_n^c \xrightarrow{p} 1$ under the condition that $(K\underline{h}_1)^{-1} = O(1)$. If the bandwidth choices are such that the standardized bias term $\left(\mathcal{V}_n^c\right)^{-1/2} \left(\mathcal{B}_{1n}^c + \mathcal{B}_{2n}^c\right)$ differs from zero asymptotically, then inference must be done using a bias-corrected estimator. A practical way of performing bias correction is to increase the order of the polynomials from (ρ_1, ρ_2) to $(\rho_1 + 1, \rho_2 + 1)$, and to compute $\widehat{\mu}^{c'}$ and $\widehat{\mathcal{V}}_n^{c'}$ using the same bandwidth choices as $\widehat{\mu}^c$ and $\widehat{\mathcal{V}}_n^c$. It follows that $(\widehat{\mathcal{V}}_n^{c'})^{-1/2}(\widehat{\mu}^{c'} - \mu^c) \xrightarrow{d} N(0, 1)$.

Convexity of $\mathcal{C}$, along with the asymptotic behavior of the schedule of cutoff-doses (Assumption 6), is crucial for the numerical integration error to vanish sufficiently quickly, as required by Theorem 2. Continuity of $\omega^c(\mathbf{c})$ implies that the boundary of $\mathcal{C}$ has zero probability under the counterfactual distribution. Therefore, the convergence rate of $\widehat{\mu}^c$ is not affected by the value of $\omega^c(\mathbf{c})$ over the boundary of $\mathcal{C}$. In finite samples, local polynomial estimates of $\widehat{\beta}(\mathbf{c})$ may be noisy for values of $\mathbf{c}$ at the boundary of the convex-hull of $\mathcal{C}_K$. Researchers should take that into account when specifying the support of the counterfactual distribution $\omega^c(\mathbf{c})$.

A simple example illustrates the three rate conditions of Theorem 2. Suppose $h_{1j} = n^{-\lambda_1}$ for all j, $h_2 = n^{-\lambda_2}$, and $K = n^{\theta}$. The first-step estimation uses local-linear regression ($\rho_1 = 1$), and the second step, local cubic regression ($\rho_2 = 3$). The

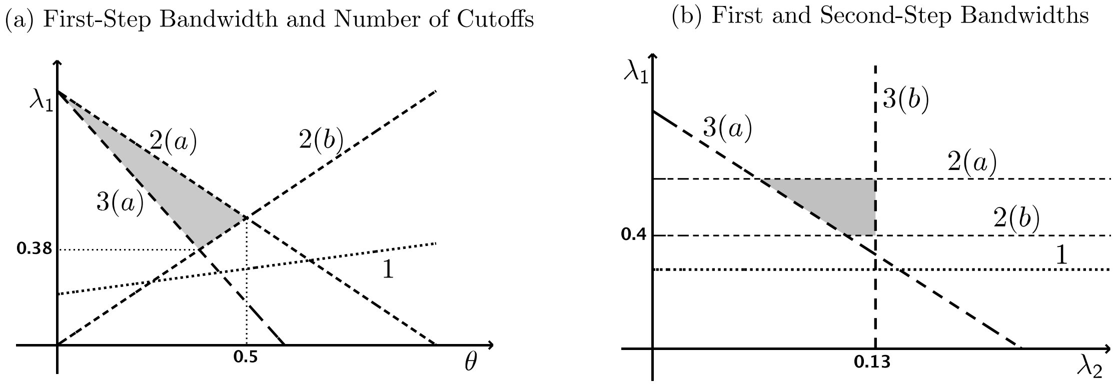

Fig. 1. Rate conditions of Theorem 2. Notes: The diagram shows the rate conditions of Theorem 2 applied to the case where $h_{1j} = n^{-\lambda_1}$ for all j, $h_2 = n^{-\lambda_2}$, and $K = n^{\theta}$. Condition 1, that is $\left(Kn\bar{h}_1\right)^{1/2} \bar{h}_1^{\rho_1+1} = O(1)$, is equivalent to $\lambda_1 \geq (1+\theta)/(3+2\rho_1)$; condition 2(a): $K^{1/2} \log n/\left(n\bar{h}_1\right)^{1/2} = o(1) \Leftrightarrow \lambda_1 < 1-\theta$; condition 2(b): $K\bar{h}_1 = O(1) \Leftrightarrow \lambda_1 \geq \theta$; condition 3(a): $\left(Kn\bar{h}_1\right)^{1/2} h_2^{\rho_2+1} = O(1) \Leftrightarrow \lambda_1 \geq 1+\theta-2\lambda_2(\rho_2+1)$; and condition 3(b): $1/Kh_2^3 = O(1) \Leftrightarrow \lambda_2 \leq \theta/3$. Panel (a) illustrates the rate conditions on the first-step bandwidth and number of cutoffs (λ_1, θ) for $\rho_1 = 1$, $\rho_2 = 3$, and $\lambda_2 = \theta/3$, so that $h_2 = K^{-1/3}$ and condition 3(b) is satisfied. Panel (b) displays the rate conditions on the bandwidths (λ_1, λ_2) given $\theta = 0.4$, $\rho_1 = 1$, and $\rho_2 = 3$.

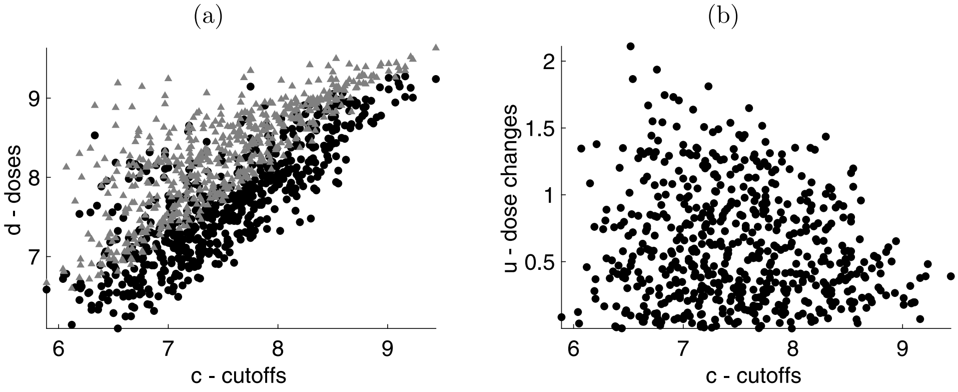

Fig. 2. Variation in cutoff and dose values. Notes: Scatter plot with cutoff values on the x-axis and dose values on the y-axis for the $K = 588$ cutoffs in the Romanian data. Panel (a) shows both doses before and after the cutoff, that is, d_{j-1} in black and d_j in gray. Panel (b) displays dose-change values on the y-axis, that is, $u_j = d_j - d_{j-1}$. The peer-quality of school j (treatment dose d_j) is measured by the average transition score of the students attending that school j. The cutoff c_j for admission into a school j is equal to the minimum transition score among the students assigned to school j.

first rate condition says the first-step bandwidths have to converge to zero fast enough to control the asymptotic bias. That is, $\left(Kn\bar{h}_1\right)^{1/2} \bar{h}_1^{\rho_1+1} = O(1)$; in terms of the example, this condition becomes $\lambda_1 \geq (1+\theta)/(3+2\rho_1)$. The second rate condition restricts how fast the number of cutoffs grows with n. It cannot grow too fast to ensure having enough observations around the cutoffs for uniform consistency of first-step estimates. The second condition has two parts: (a) $K^{1/2} \log n/\left(n\bar{h}_1\right)^{1/2} = o(1) \Leftrightarrow \lambda_1 < 1-\theta$; and (b) $K\bar{h}_1 = O(1) \Leftrightarrow \lambda_1 \geq \theta$. The third rate condition limits how slowly K grows, relative to the sample size, to ensure that the integral approximation error vanishes faster than the estimation variance. Part (a) of the third condition says $\left(Kn\bar{h}_1\right)^{1/2} h_2^{\rho_2+1} = O(1) \Leftrightarrow \lambda_1 \geq 1+\theta-2\lambda_2(\rho_2+1)$; part (b) is $1/Kh_2^3 = O(1) \Leftrightarrow \lambda_2 \leq \theta/3$.

Fig. 1 illustrates these conditions and the feasible set for bandwidth choices (shaded area).[8] Panel (a) shows the conditions in terms of (λ_1, θ) assuming $\lambda_2 = \theta/3$ so that $h_2 = K^{-1/3}$, which satisfies part (b) of the third condition. Panel (b) depicts the same conditions in terms of (λ_1, λ_2), assuming $\theta = 0.4$. The feasible set is well-defined as long as K grows no faster than $\sqrt{n}$, that is, $\theta < 0.5$. In addition, $\rho_2 \geq 3$ because line 3(a) has to be below line 2(a). The maximum rate of convergence of the estimator is $\sqrt{n}$, and it is reached along the dashed line 2(b).

Implementation of Theorem 2 requires the researcher to choose $\rho_1 \in \mathbb{Z}_+$, $h_{1j} > 0$ $\forall j$, $\rho_2 \in \mathbb{Z}_+$, $h_2 > 0$, and $k(\cdot)$. A theory of optimal choice of these tuning parameters is beyond the goals of this paper. Optimal choice of

[8] Section B.3 in the supplemental appendix gives an example of a schedule of cutoff-dose values that satisfies Assumption 6 for feasible choices of (h_1, h_2) in this example.

bandwidths is an interesting topic for future research, because optimality in the multi-cutoff case would account for: (i) the interaction between first and second-stage bandwidths; (ii) the variance reduction from overlapping estimation windows at consecutive cutoffs; and (iii) the recent advances of robust bias-corrected inference and coverage-error optimal bandwidths by Calonico et al. (2018).

The IK bandwidth formula may produce first-step bandwidths with an incorrect rate of convergence. For example, if $\rho_1 = 1$, these bandwidths converge to zero at $n^{-0.2}$, which is not fast enough if $\rho_2 = 3$ (Fig. 1). A simple way to correct this is to adjust the bandwidths by multiplying them by $n^{0.2-\lambda_1}$ for $\lambda_1 \geq \theta$, so that their rate becomes $n^{-\lambda_1}$. Conditions 2(a) and (b) imply that θ is never bigger than 0.5 regardless of ρ_1, ρ_2, and λ_2. Thus, the smallest value for λ_1 consistent with these restrictions is 0.5. The same idea applies to the coverage-error optimal bandwidths by Calonico et al. (2018), which converge to zero at rate $n^{-0.25}$, and need to be adjusted.

In certain cases, the β function may depend on less than the three arguments (c, d, d'). For example, in the Medicaid application, the treatment is binary and β is only a function of c. This is a particular case of the theory in this section. The only rate condition that changes is condition 3(b). It becomes $1/Kh_2 = O(1)$, or $\lambda_2 \leq \theta$ in terms of Fig. 1. A non-empty feasible set of bandwidth choices requires $\rho_2 \geq 1$, as opposed to $\rho_2 \geq 3$ in the general case.

A simple recommendation to implement Theorem 2 is to use the edge kernel, first-step rate-adjusted IK bandwidths for each cutoff, and a second-step bandwidth h_2 that minimizes the MSE of estimation. First, use observations pertaining to each cutoff j, compute the IK bandwidth h_{1j}^{ik} for sharp RD and local-linear regression; adjust the rate of the bandwidths so that $h_{1j} = h_{1j}^{ik} \times n^{-0.3}$. Second, create a grid of possible values for h_2. For each value on the grid, compute $\widehat{\mu}^c(h_2)$ using the edge kernel, the choices of h_{1j} given above, $\rho_1 = 1$, and $\rho_2 = 3$ (or $\rho_2 = 1$ in the binary treatment case). Similarly, compute $\widehat{\mu}^{c'}(h_2)$ using the edge kernel, the choices of h_{1j} given above, $\rho_1 = 2$, and $\rho_2 = 4$ (or $\rho_2 = 2$ in the binary treatment case). Use Eq. (26) to estimate the variance of $\widehat{\mu}^c(h_2)$ and call it $\widehat{\mathcal{V}}_n^c(h_2)$. Evaluate the approximated MSE of $\widehat{\mu}^c(h_2)$ by $(\widehat{\mu}^c(h_2) - \widehat{\mu}^{c'}(h_2))^2 + \widehat{\mathcal{V}}_n^c(h_2)$. Choose the bandwidth value on the grid that minimizes the MSE and call it h_2^*. The bias-corrected estimate is $\widehat{\mu}^{c'}(h_2^*)$, and its variance estimate is $\widehat{\mathcal{V}}_n^{c'}(h_2^*)$.

It may not be immediately clear that root-n is the fastest estimation rate achievable in a setting where both K and n grow large. The double asymptotic setting is conceptually different from the usual asymptotic setting where only $n \to \infty$ and non-parametric averages are estimable at root-n. Estimation rates depend not only on bandwidth choices, but also on how fast K grows, relative to n. Similar examples in econometrics include panels with a large number of observations and time periods, and asymptotics with many instruments. The following theorem demonstrates that the minimax optimal rate of estimation of μ^c is indeed root-n, as long as first-step bandwidths converge to zero at $1/K$ rate.

Theorem 3. *Let $\mathcal{P}$ be the class of models generating potential outcomes $\{Y_i(d)\}_{d\in\mathcal{D}}$ and forcing variables X_i. For a schedule of cutoffs and doses, $\{\mathbf{c}_j\}_{j=1}^K$, observed data (Y_i, X_i, D_i) are generated iid from $P \in \mathcal{P}$ as described in Section 3.1. Assume that (i) each model $P \in \mathcal{P}$ satisfies Assumptions 4–7; (ii) $f(x)$ and $\sigma^2(x, d)$ are bounded away from zero uniformly in $\mathcal{P}$; (iii) the following functions are bounded uniformly in $\mathcal{P}$: $\nabla_x^\rho \sigma^2(x, d)$ $\forall \rho \leq 1$, $\nabla_x^\rho f(x)$ $\forall \rho \leq 1$, $\nabla_x^\rho R(x, d)$ $\forall \rho \leq \bar{\rho}$, $\nabla_d^\rho R(x, d)$ $\forall \rho \leq \bar{\rho}$, where $\bar{\rho} = \max\{\rho_1 + 2, \rho_2 + 2\}$; and (iv) there exists $M \in (0, \infty)$ such that $\mathbb{P}[|Y_i(d) - R(X_i, d)| < M] = 1$ $\forall d \in \mathcal{D}$ uniformly in $\mathcal{P}$. Then, for any $\epsilon > 0$, there exists $\eta > 0$ such that*

$$\inf_{\widetilde{\mu}} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left[\sqrt{n}|\widetilde{\mu} - \mu^c(P)| > \epsilon/2\right] \geq \frac{1}{4\eta} \quad \textit{for large } n. \tag{27}$$

The inf is taken over all estimators $\widetilde{\mu}$ built using the observed data (Y_i, X_i, D_i), $i = 1, \ldots, n$; $\mu^c(P) = \int \omega^c(\mathbf{c})\beta(\mathbf{c}; P)\, d\mathbf{c}$ with $\beta(\mathbf{c}; P) = \mathbb{E}_P[Y_i(d') - Y_i(d)|X_i = c]$; and $\mathbb{P}_P$ and $\mathbb{E}_P$ denote the probability and expectation under model $P \in \mathcal{P}$.

Assume the conditions of Theorem 2, and that first-step bandwidths satisfy $\overline{h}_1 = O(K^{-1})$. Consider the estimator $\widehat{\mu}^c$ defined in Eq. (17). For any small $\delta > 0$, there exists large $\epsilon \in (0, \infty)$ such that

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\left[\sqrt{n}|\widehat{\mu}^c - \mu^c(P)| > \epsilon\right] < \delta \quad \textit{for large } n. \tag{28}$$

Eq. (27) shows that no estimator converges faster than $\sqrt{n}$ uniformly over $\mathcal{P}$. Eq. (28) says the estimator proposed in Theorem 2 converges at root-n uniformly over $\mathcal{P}$ as long as first-step bandwidths converge to zero at $1/K$ rate. Therefore, root-n is the minimax optimal rate of convergence in the non-parametric estimation of ATE in RDD with many thresholds. Authors have previously analyzed minimax optimality of non-parametric estimators of a regression function at a boundary point, for example, Cheng et al. (1997) and Sun (2005). Theorem 3 is novel because it combines boundary points to estimate averages of non-parametric regression functions.

4. Fuzzy case with multiple cutoffs

This section relaxes the sharp assignment mechanism of previous sections and studies the fuzzy RDD case. The analysis focuses on multiple cutoffs, but K is finite as opposed to approaching infinity as in Section 3.2. This makes the exercise more tractable, because the number of compliance cases grows super-exponentially with the number of cutoffs. In contrast to the sharp case, non-parametric identification of local effects in the fuzzy case is impossible. As a result, inference methods in this section rely on a second heterogeneity assumption, namely, the treatment effect function is assumed parametric. Section B.5.2, in the supplemental appendix, provides practical guidelines to compute an MSE-optimal ATE estimator, and demonstrates asymptotic normality.

In the sharp RDD case, all individuals with forcing variable equal to x receive the same treatment $D(x)$ Eq. (3). In the fuzzy RDD case, many of these individuals may receive treatments different from $D(x)$. In the high school assignment example, students may choose to go to a school that is not the best school for which they are eligible. For instance, a student may want to attend the same high school as a certain friend or sibling. Another example is given by Garibaldi et al. (2012). In their study, a schedule of tuition subsidies applies to most students at Bocconi University, but the university reserves the right to grant certain students different subsidies after reassessing their ability to pay.[9]

The fuzzy RDD case is modeled in terms of a potential treatment assignment framework. A potential treatment assignment function $\mathcal{U} : \mathcal{X} \to \mathcal{D}$ describes the treatment received for every value of the forcing variable $x \in \mathcal{X}$. For simplicity, these functions are assumed to belong to the following class:

$$\mathcal{U}^* = \left\{ \mathcal{U} : \mathcal{X} \to \mathcal{D} \ : \ \mathcal{U}(x) = \sum_{j=0}^{K} u_j \mathbb{I}\left\{c_j \le x < c_{j+1}\right\} \text{ for some } u_j \in \{d_0, \ldots, d_K\}, j = 0, \ldots, K \right\}. \tag{29}$$

Sharp RDD is the particular case where the individual potential treatment assignment function $\mathcal{U}_i$ is the same for every individual i, that is, $\mathcal{U}_i(x) = D(x)\ \forall i$ with $D(x)$ defined in Eq. (3). In the fuzzy case, $\mathcal{U}_i$ is sampled iid from a distribution of functions with support in $\mathcal{U}^*$. Potential treatment functions $\mathcal{U}_i(x)$ are unobserved, but the treatments received are observed and given by

$$D_i = \sum_{j=0}^{K} \mathcal{U}_i(c_j)\mathbb{I}\left\{c_j \le X_i < c_{j+1}\right\}.$$

Using classic definitions of compliance behaviors (Imbens and Rubin, 1997), three types of compliance groups are defined in terms of *changes* in treatment eligibility. “Never-changers” are those whose treatment received never changes when eligibility changes. The treatment received by “ever-compliers” or “ever-defiers” changes at least once when eligibility changes. Ever-compliers are those whose treatment received changes if and only if it changes to the treatment dose for which they become eligible. Ever-defiers change to a treatment dose different from the one for which they become eligible. In the case of one cutoff and two treatments, the definition of ever-complier (ever-defier) is equivalent to the classic definition of complier (defier) of Imbens and Lemieux (2008).

The three compliance groups are measurable events that partition the population of individuals with $\mathbf{G}_{nc}$ denoting never-changers, $\mathbf{G}_{ec}$ ever-compliers, and $\mathbf{G}_{ed}$ ever-defiers.[10]

$$\mathbf{G}_{nc} = \left\{\mathcal{U}_i \in \mathcal{U}^* : \left\{j : \ \mathcal{U}_i(c_{j-1}) \neq \mathcal{U}_i(c_j)\right\} = \emptyset\right\} \tag{30}$$

$$\mathbf{G}_{ec} = \left\{\mathcal{U}_i \in \mathcal{U}^* : \left\{j : \ \mathcal{U}_i(c_j) = D(c_j)\right\} \supseteq \left\{j : \ \mathcal{U}_i(c_{j-1}) \neq \mathcal{U}_i(c_j)\right\} \neq \emptyset\right\} \tag{31}$$

$$\mathbf{G}_{ed} = \left\{\mathcal{U}_i \in \mathcal{U}^* : \left\{ \ \left\{j : \ \mathcal{U}_i(c_j) \neq D(c_j)\right\} \cap \left\{j : \ \mathcal{U}_i(c_{j-1}) \neq \mathcal{U}_i(c_j)\right\} \ \right\} \neq \emptyset\right\} \tag{32}$$

where $\emptyset$ denotes empty set.

In the high school assignment case, an example of a never-changer is a student who strongly prefers the high school with the lowest admission cutoff and attends that high school even if she is admitted to better schools. An example of an ever-complier is a student who attends the best school into which she is admitted, or a student who chooses the best school among the nearby schools. Suppose a student has rational preferences and is never indifferent. Assume her choice set is equal to those schools with admission cutoffs that are less than or equal to her test score. Then, such a student is never an ever-defier. In other words, as her test score increases, a new school is added to her choice-set of schools; she either chooses to go to the new school for which she becomes eligible, or she stays at the school which she preferred prior to the increase in her choice-set. Thus, it seems natural to rule out “ever-defiers” in this and other applications.

Never-changers do not produce changes in treatments, so there is no identification on them. For ever-compliers, there are multiple possible changes in treatment at a given cutoff, and ever-compliers may differ in terms of the treatments they comply with. For example, the student who is willing to attend the best school possible complies with all changes in treatment eligibility. On the other hand, the student who is willing to attend the best possible school within a certain distance from home only complies with some of the changes in treatment eligibility. Therefore, besides no-defiance, identification also requires the heterogeneity of ever-compliers to be restricted.

Assumption 8 generalizes the sufficient conditions for identification on compliers in the one-cutoff case (Hahn et al. (2001) and Dong (2018a)). In addition, it restricts the heterogeneity on ever-compliers.

[9] The source of fuzziness varies across applications. One example is the case where the assignment of individuals into different treatments is made through a matching mechanism, and the econometrician does not observe all the individual characteristics used in the matching algorithm. This is the reason why the RDD of PU is fuzzy: based on the entire distribution of test scores and preferences, the central planner ranks students by their test scores and assigns each one to her preferred school among schools with vacancies.

[10] These definitions allow for non-monotonic treatment schedules; for example, the average class-size varies non-monotonically across cutoffs on enrollment (Angrist and Lavy, 1999). Table B.1 in Section B.5.1 of the supplemental appendix illustrates these definitions of compliance groups using a simple example with 3 treatments and 2 cutoffs.

Assumption 8. **(a)** There are no ever-defiers: $\mathbb{P}[\mathbf{G}_{ed}] = 0$; **(b)** for arbitrary $d \in \mathcal{D}$, and $\bar{\mathcal{U}} \in \mathcal{U}^*$, $\mathbb{E}[Y_i(d)|X_i = x, \mathcal{U}_i = \bar{\mathcal{U}}]$ and $\mathbb{P}\left[\mathcal{U}_i = \bar{\mathcal{U}}|X_i = x\right]$ are continuous and bounded functions of x; **(c)** there exists a function $\beta_{ec}(\mathbf{c})$ such that $\mathbb{E}[Y_i(d') - Y_i(d)|X_i = c, \mathcal{U}_i = \bar{\mathcal{U}}] = \beta_{ec}(\mathbf{c})$ for every $\mathbf{c} = (c, d, d') \in \mathcal{C}$ and $\bar{\mathcal{U}} \in \mathbf{G}_{ec}$.

A fuzzy assignment produces several different treatment changes at each cutoff, even after ruling out ever-defiers. The researcher only observes one aggregate change in Y_i at each cutoff, but there are several treatment effects on ever-compliers to be identified at that cutoff. Theorem 4 shows that identification of these effects is not possible without further restricting the class of functions $\beta_{ec}(\mathbf{c})$. Economic theory or *a priori* knowledge guides the choice of a functional form that credibly summarizes the heterogeneity of treatment effects. For example, the principal–agent model of Bajari et al. (2017) yields a functional form to study reimbursement of hospitals by insurers. The second heterogeneity assumption (Assumption 9) restricts the treatment effect function on ever-compliers to a finite-dimensional vector space of functions.

Assumption 9. Let $\mathcal{W}(c, d) = [\mathcal{W}_1(c, d), \ldots, \mathcal{W}_q(c, d)]'$ be a vector-valued function $\mathcal{W} : \mathcal{X} \times \mathcal{D} \to \mathbb{R}^{q\times 1}$ known to the researcher and such that **(a)** $\mathbb{E}_F\left[\mathcal{W}(c, d') - \mathcal{W}(c, d)\right]$ is well-defined for the counterfactual distribution F; and **(b)** $\mathcal{W}_j(c, d') - \mathcal{W}_j(c, d), j = 1, \ldots, q$, are linearly independent functions. The treatment effect function $\beta_{ec}(\mathbf{c})$ is assumed to belong to the following class of functions:

$$\mathcal{H} = \left\{\beta : \mathcal{C} \to \mathbb{R} \ : \ \beta(c, d, d') = \left[\mathcal{W}(c, d') - \mathcal{W}(c, d)\right]' \boldsymbol{\theta}, \text{ for } \boldsymbol{\theta} \in \mathbb{R}^q\right\}.$$

In this case, the ATE on ever-compliers is a linear combination of the true parameter vector $\boldsymbol{\theta}_0^{ec}$. For a counterfactual distribution F chosen by the researcher,

$$\mu^{ec}(F) = \int \beta(\mathbf{c}; \boldsymbol{\theta}_0^{ec})\, dF(\mathbf{c}) = \underbrace{\int \left[\mathcal{W}(c, d') - \mathcal{W}(c, d)\right]' \, dF(\mathbf{c})}_{\equiv \mathbf{Z}(F)}\boldsymbol{\theta}_0^{ec} = \mathbf{Z}(F)\boldsymbol{\theta}_0^{ec}. \tag{33}$$

Theorem 4 shows that the observed change in average outcome at a given cutoff is a weighted average of treatment effects on ever-compliers who switch from various doses into the dose of eligibility at that cutoff. Assumption 9 and variation in cutoff characteristics are sufficient conditions for identification. Conversely, identification on ever-compliers implies that $\beta_{ec}(\mathbf{c})$ belongs to a finite-dimensional class of functions.

Theorem 4. *Under Assumption 8, for $j = 1, \ldots, K$,*

$$B_j = \sum_{l=0, l\neq j}^{K} \omega_{j,l}\beta_{ec}(c_j, d_l, d_j)$$

where B_j is defined in Eq. (6), and

$$\omega_{j,l} = \lim_{e\downarrow 0}\left\{\mathbb{P}[D_i = d_l|X_i = c_j - e] - \mathbb{P}[D_i = d_l|X_i = c_j + e]\right\},$$

for $l = 0, 1, \ldots, K, l \neq j$.

Moreover, suppose β_{ec} belongs to the class of functions $\mathcal{H}$ defined in Assumption 9 with $q \leq K$. Define

$$\widetilde{W}_j = \sum_{l=0, l\neq j}^{K} \omega_{j,l}\left[\mathcal{W}(c_j, d_j) - \mathcal{W}(c_j, d_l)\right] \tag{34}$$

for the vector-valued function $\mathcal{W}(c, d)$ of Assumption 9; build a $K \times q$ matrix $\widetilde{\mathbf{W}}$ by stacking $\widetilde{W}_j$, and $\mathbf{B}$ by stacking B_j. If $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}$ is invertible, then $\beta_{ec}(\mathbf{c})$ is identified and equal to

$$\beta_{ec}(\mathbf{c}) = \left[\mathcal{W}(c, d') - \mathcal{W}(c, d)\right]' \left(\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}}\right)^{-1} \widetilde{\mathbf{W}}'\mathbf{B}.$$

Conversely, suppose β_{ec} belongs to some class of functions $\widetilde{\mathcal{H}}$, and treatment effects on ever-compliers are identified at the $p > K$ cutoff-dose values $\{\tilde{\mathbf{c}} : \tilde{\mathbf{c}} = (c_j, d_l, d_j)$ with $\omega_{\underset{\sim}{j,l}} > 0\}$ of every possible fuzzy assignment generated from the given schedule of cutoffs $\{\mathbf{c}_j\}_{j=1}^{K}$. Then, the class of functions $\widetilde{\mathcal{H}}$ is "finite dimensional" in the sense that

$$\mathcal{G} = \left\{\left(\beta(\tilde{\mathbf{c}}_1), \ldots, \beta(\tilde{\mathbf{c}}_p)\right) : \textit{for } \beta \in \widetilde{\mathcal{H}}\right\} \subseteq \mathbb{R}^p$$

has $\dim\mathcal{G} \leq K$ for every fuzzy assignment $\{\tilde{\mathbf{c}}_j\}_{j=1}^{p}$ generated from $\{\mathbf{c}_j\}_{j=1}^{K}$.

Theorem 4 reveals the requirement of stronger functional form assumptions on $\beta_{ec}(\mathbf{c})$ even for identification of local effects in the fuzzy case with a finite number of multiple cutoffs. For example, identification is not possible when $\widetilde{\mathcal{H}}$ is the class of all smooth functions studied in the non-parametric case of Section 3.2. The result is striking because non-parametric identification of local effects is possible both in the sharp case with a finite number of cutoffs and in the fuzzy case with a single cutoff. It is likely possible to obtain non-parametric identification of $\beta_{ec}(\mathbf{c})$ under a large variation of cutoff-dose values. The function $\beta_{ec}(\mathbf{c})$ may be approximated by a sequence of parametric functions from Assumption 9, where q grows to infinity more slowly than K, so to keep $dim\mathcal{G} \leq K$ as $K \to \infty$. In this paper, the number of cutoffs is kept finite for simplicity, and the case with large K is deferred to future work.

Theorem 4 also clarifies the interpretation of two-stage least squares (2SLS) estimates in applications of fuzzy RD with multiple cutoffs, a common practice in applied work. The practice consists of using $D(X_i)$ as an instrument for D_i in the regression of Y_i on a constant, D_i, and X_i. See Angrist and Pischke (2008) for a discussion. In the single-cutoff case, both the non-parametric RD estimator and 2SLS applied to a neighborhood of the cutoff are consistent to the average treatment effect on compliers (Hahn et al., 2001). To my knowledge, such an equivalence has never been studied in the multiple-cutoff case. Nevertheless, many important applications have multiple-fuzzy cutoffs and use 2SLS; for example, Angrist and Lavy (1999), Chen and Van der Klaauw (2008), and Hoekstra (2009). The 2SLS estimator is consistent for a data-driven weighted average of treatment effects on ever-compliers as long as a sufficiently flexible specification is used; for example, cutoff fixed-effects or varying slopes. The economic meaning of the 2SLS estimands depends crucially on the choice of such a weighting scheme. Unless a parametric functional form is imposed on $\beta_{ec}(\mathbf{c})$, or there is large variation in cutoff-doses, only a data-driven weighted average of $\beta_{ec}(\mathbf{c})$ is identified. In other words, if $\beta_{ec}(\mathbf{c})$ is non-parametric and there are only a few cutoffs, the researcher does not have control over the weighting scheme, and 2SLS estimates do not have a clear interpretation.

Theorem 4 leads to a two-step estimation procedure for $\boldsymbol{\theta}_0^{ec}$ and μ^{ec}. The mechanics are similar to the previous sections, so I omit the details from the main text for brevity. In the first step, the researcher estimates the jump discontinuity of the vector $[Y_i \quad \mathcal{W}(X_i, D_i)']'$ using LPRs at each cutoff to obtain $[\widehat{B}_j \quad \widehat{\widetilde{W}}_j']'$. In the second step, a regression of $\widehat{B}_j$ on $\widehat{\widetilde{\mathbf{W}}}_j'$ obtains $\widehat{\boldsymbol{\theta}}^{ec}$. The ATE estimator is $\widehat{\mu}^{ec} = \mathbf{Z}(F)\widehat{\boldsymbol{\theta}}^{ec}$. Estimation precision varies across cutoffs, and the parametric form of β_{ec} allows us to optimally combine different cutoffs to minimize the MSE of $\widehat{\boldsymbol{\theta}}^{ec}$. The researcher can simply re-weight the second-step regression by the inverse of the MSE matrix of the first-step estimators. Section B.5.2 in the supplemental appendix delineates the estimation and inference procedures of $\boldsymbol{\theta}_0^{ec}$ and μ^{ec} with practical steps.

5. Simulations

In this section, Monte Carlo simulations illustrate the finite sample behavior of the ATE estimator proposed in Section 3.2. The analysis considers estimation precision and coverage of confidence intervals for different choices of tuning parameters and a non-linear specification for β. As predicted by Theorem 2, an incorrect choice of the second-step polynomial degree leads to severe bias and extremely poor coverage of confidence intervals. Moreover, first-step bandwidths that imply overlapping estimation windows produce lower MSE than cases with no overlap, regardless of other tuning parameters.

The DGP draws n iid observations of (X_i, ε_i) where X_i is uniformly distributed over $[0, 1]$, ε_i is normally distributed with zero mean and unit variance, and these variables are independent of each other. There are K cutoffs $c_j = j/(K+1)$, $j = 1, \ldots, K$, on the unit interval $[0, 1]$. The number of cutoffs is $K = \lfloor n^{0.4} \rfloor$, where $\lfloor a \rfloor$ denotes the largest integer smaller than or equal to a. An individual with forcing variable X_i receives a treatment dose equal to $D(X_i)$ as in Eq. (3). The dose increases by one unit at each cutoff, starting at $d_0 = 1$ and ending at $d_K = K + 1$. The outcome variable is $Y_i = \phi(X_i)D(X_i) + \varepsilon_i$ where $\phi(X_i) = 15X_i^3 + 7.5X_i^2 - 18.75X_i + 2.125$. This implies that $\beta(\mathbf{c}) = \phi(c)(d' - d)$, which falls into the binary treatment case. Consider a counterfactual policy that uniformly increases treatment doses by one unit. The ATE parameter μ is the integral of $\phi(c)$ over $c \in [0, 1]$, which equals -1 in this case.

Estimation follows the procedure suggested in Section 3.2. For given bandwidth choices h_1 and h_2, I compare the ATE estimator $\widehat{\mu}$ that uses $\rho_1 = \rho_2 = 1$, to the bias-corrected ATE estimator $\widehat{\mu}^{bc}$ that uses $\rho_1 = \rho_2 = 2$. To emphasize the importance of the second step, I also compute a naive ATE estimator that simply averages the first-step estimates. The naive and bias-corrected naive estimators, respectively $\widetilde{\mu}$ and $\widetilde{\mu}^{bc}$, are constructed as $\widehat{\mu}$ and $\widehat{\mu}^{bc}$ except for the tuning parameters in the second step. Both naive estimators use $\rho_2 = 0$ and $h_2 = \infty$. To examine the effect of overlapping estimation windows in the first step, I compare estimators for two choices of h_1. The first choice is the largest possible bandwidth $h_1 = 1/(K+1)$, which leads to maximum overlap. The second choice is the largest possible bandwidth with no overlap, that is, $h_1 = 0.5/(K+1)$. Finally, I study the effects of ten different choices for the second-step bandwidth, $h_2 \in \{3/(K+1), \ldots, 12/(K+1)\}$. All choices of tuning parameters satisfy the rate conditions of Theorem 2 and produce a convergence rate of root-n for $\widehat{\mu}$ and $\widehat{\mu}^{bc}$. The Monte Carlo experiment simulates 10,000 draws of an iid sample with $n \in \{1789, 10120, 27886, 57244, 100000\}$ and $K \in \{20, 40, 60, 80, 100\}$, respectively. Section B.7 in the supplemental appendix repeats the experiment with data-driven bandwidth choices, following the bandwidth rules proposed in Section 3.2.

The bias and variance of all estimators converge to zero as the sample size increases, regardless of the choice of h_1 (Table 1). The bias-correction of $\widehat{\mu}^{bc}$ eliminates almost all the bias of $\widehat{\mu}$, at the cost of a higher variance. The naive estimator $\widetilde{\mu}$ oversmooths the second step beyond the conditions of Theorem 2. As a result, the bias of $\widetilde{\mu}$ is substantially larger than that of $\widehat{\mu}$. Simply correcting for bias in the first step does not solve the problem, as the difference in bias between $\widetilde{\mu}$ and $\widetilde{\mu}^{bc}$ is small. First-step bandwidths that produce overlap (Table 1, rows 1–5) yield approximately the same bias, but substantially smaller variance, compared to first-step bandwidths that produce no overlap (Table 1, rows 6–10).

Table 3
Coverage of 95% confidence intervals.

n	K	% coverage				Avg. length			
		$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widetilde{\mu}$	$\widetilde{\mu}^{bc}$	$\widehat{\mu}$	$\widehat{\mu}^{bc}$	$\widetilde{\mu}$	$\widetilde{\mu}^{bc}$
1789	20	0.8931	0.9546	0.1686	0.3834	0.3526	0.5135	0.3368	0.4165
10 120	40	0.9087	0.9545	0.0568	0.1788	0.1403	0.1850	0.1373	0.1656
27 886	60	0.9287	0.9489	0.0203	0.1005	0.0834	0.1063	0.0822	0.0988
57 244	80	0.9292	0.9502	0.0113	0.0585	0.0578	0.0724	0.0572	0.0686
100 000	100	0.9359	0.9492	0.0057	0.0353	0.0436	0.0540	0.0432	0.0518

Notes: The table reports simulated percentage of correct coverage and average length of 95% confidence intervals. Confidence intervals are constructed using four estimators ($\widehat{\mu}, \widehat{\mu}^{bc}, \widetilde{\mu}, \widetilde{\mu}^{bc}$). They equal an estimator plus or minus its estimated standard deviation multiplied by 1.96. Coverage and average length are computed for five sample sizes n and respective numbers of cutoffs K. The first-step bandwidth is set to $h_1 = 1/(K+1)$ (overlap), and the second-step bandwidth is set to $h_2 = 3/(K+1)$, which minimizes MSE of $\widehat{\mu}$. The number of simulations is 10,000.

Next, I study how the choice of h_2 affects precision of $(\widehat{\mu}, \widehat{\mu}^{bc})$ for a fixed choice of $h_1 = 1/(K+1)$ (Table 2). The smallest value for h_2 is $3/(K+1)$. This defines a second-step estimation window with at least three cutoffs to ensure invertibility of matrices in the regressions. The bias of $\widehat{\mu}$ is substantially smaller when h_2 is set to its smallest value. All other measures are practically unaffected across different h_2.

The significant bias of the naive ATE estimators $\widetilde{\mu}$ and $\widetilde{\mu}^{bc}$ decreases the coverage of 95% confidence intervals as the sample increases (Table 3). The naive estimators oversmooth in the second step, and Theorem 2 implies the bias grows faster than root-n. For each of the four estimators, the confidence intervals equal the estimator plus or minus 1.96 times its standard error. The variance of estimators are obtained as described in Eq. (26). The bias-corrected ATE estimator $\widehat{\mu}^{bc}$ produces confidence intervals with correct coverage for all samples sizes. Although $\widehat{\mu}$ yields intervals with average length smaller than $\widehat{\mu}^{bc}$, the bias of $\widehat{\mu}$ leads to a slightly lower coverage.

6. Application

In this section, the methods proposed in this paper are illustrated using the data from PU on high school assignments in Romania.[11] Many policy questions demand an ATE of a continuous counterfactual distribution of treatments, and this section provides an example of such a policy question. The estimators designed for the sharp RDD case are consistent for "Intent-to-Treat" (ITT) average effects when applied to the fuzzy data of PU. In this application, the ITT effect measures the impact of being assigned to a better school but not necessarily attending it. The parametric methods of Section 4 yield noticeable efficiency gains in the estimation of the ATE on ever-compliers. Treatment effects for ever-compliers reveal a heterogeneity pattern unlike the heterogeneity of ITT effects.

The administrative data from Romania cover 3 cohorts of 9th grade students for the years 2001, 2002, and 2003, with a total of 334,137 observations. The essential elements of the high school assignment in Romania are described below. The assignment to high school is nationally centralized by the Ministry of Education. At the end of grade 8, students submit a transition score and a complete ranking of preferences for high schools. The transition score is an average of the student's performance on a national exam taken in grade 8 and the student's grade point average during grades 5–8. The Ministry of Education ranks students by their transition score and no other criteria. The mechanism assigns the student ranked first to her most preferred school, the student ranked second to her most preferred school, etc. Students cannot decline their assignment, and they have incentives to truthfully reveal their preference rankings.

The observed variables are the town and year of student i, the transition score X_i, the school the student is assigned to, and the student's score on the "baccalaureate exam". This is an exam taken at the end of high school, and the grade on the exam is the outcome variable Y_i. The quality of school j (treatment dose d_j) is measured by the average transition score of the students attending that school j. The cutoff c_j for admission into a school j is equal to the minimum transition score among the students that are assigned to that school j. The student's preferences in high schools are not observed in the data, which makes the RDD fuzzy. For example, a student may have a score greater than the cutoff for the best school in her town, but still be assigned to a different school because of her personal preferences. For a transition score X_i, the treatment dose of eligibility $D(X_i)$ is equal to the largest d among those schools with admission cutoff c less than X_i. The treatment dose received D_i coincides with the treatment dose of eligibility $D(X_i)$ for 40% of the students in the sample. Thus, the assignment is fuzzy, and causal inference beyond ITT effects requires the methods of Section 4. Following PU, I drop observations with missing values for Y_i. I also drop cutoffs without enough observations around them to carry out the matrix inversions of the local polynomial regressions. The dropping of cutoffs leaves the empirical distribution of outcomes, forcing variable, cutoffs, and treatment doses practically unchanged. The estimation sample has 588 cutoffs with a total of 179,995 individuals from 769 schools in 121 towns and 3 years. The variation of cutoff and dose values is displayed in Fig. 2.

Non-parametric identification of $\beta(\mathbf{c})$ is limited to the set $\mathcal{C}$, which is the convex-hull of $\mathcal{C}_\infty$. The set $\mathcal{C}_\infty$ is not entirely observed, and the researcher relies on the observation of $\mathcal{C}_K$ (Fig. 2(a)). For the sake of simplicity, I restrict β to be a

[11] The data set is available online in the supplemental materials of PU on the website of the *American Economic Review*.

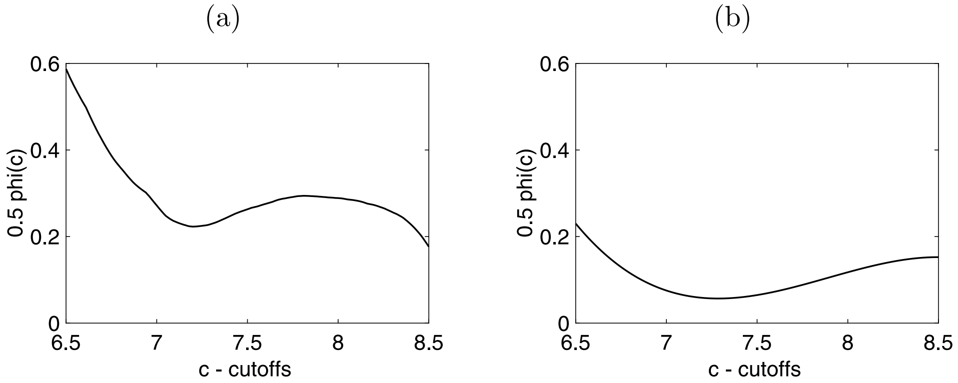

Fig. 3. Treatment effect function. Notes: Estimated average treatment effect function for a 0.5 increase in school quality for students with score equal to c. The figure plots $\widehat{\beta}(c, d, d+0.5) = 0.5\widehat{\phi}(c)$ for $c \in [6.5, 8.5]$. Panel (a) shows the ITT effect of a 0.5 increase in average peer performance for various levels of transition score. The ϕ function is estimated non-parametrically with bias correction following Section 3.2 (sharp case). Panel (b) displays the effect on ever-compliers of the same uniform change in treatment dose. The ϕ function is estimated parametrically with bias correction following the iterated procedure of Section B.5.2 in the supplemental appendix (fuzzy case).

function of dose changes ($u = d' - d$) instead of doses before and after (d and d'). The restriction greatly simplifies the visualization and estimation of $\beta(c, d, d')$, because it implies that $\beta(c, d, d') = \phi(c)(d' - d)$, where ϕ is a continuously differentiable function. Fig. 2(b) illustrates the variation of cutoff and dose-change values and defines the limits on estimation of policy counterfactuals. For example, it is not possible to estimate the effects of randomly assigning students with grades between 8 and 9 to a change in treatment dose of 2. The support of such counterfactual distribution falls outside the observed variation of cutoff and dose-change values. On the other hand, it is possible to estimate the ATE of randomly assigning students with grades between 6.5 and 8.5 to dose increases between 0 and 1.5.

The following policy question illustrates the ATE estimator proposed in this paper. Suppose a new charter school is constructed in one of the towns in Romania. The new charter school has more autonomy and better management than traditional public schools, and admitted students experience an increase in school quality as if they were admitted to a school with better peers. More specifically, the policy counterfactual is to give a 0.5 increase in peer-quality to a uniform distribution of scores between 6.5 and 8.5. The ATE parameter is defined as

$$\mu = \frac{1}{2}\int_{6.5}^{8.5} \phi(c)\, dc. \tag{35}$$

I follow the estimation procedure suggested in Section 3.2 and take into account the restriction $\beta(c, d, d') = \phi(c)(d'-d)$. As in the binary treatment case, the restriction lowers the polynomial degree requirement on the second-step estimation to $\rho_2 = 1$. See Fig. 1 and the discussion that follows it. The grid for h_2 has 32 equally-spaced points between 0.1 and 3.6, respectively, the smallest bandwidth for which the estimator is computable, and the maximum distance between two different cutoffs. The MSE-optimal bandwidth choice is $h_2^* = 1.837$. The new charter school has a bias-corrected ATE of 0.2964 with standard error of 0.1296, and it is statistically significant at 5%. Fig. 3(a) plots $0.5\widehat{\phi}(c)$, that is, the effect of a 0.5 increase in the treatment dose for various levels of c. The graph reveals heterogeneous marginal effects of ability on returns to school quality. Heterogeneity of treatment effects is *a priori* unknown, and the ATE estimator proposed in this paper is consistent for μ regardless of the shape of $\phi(c)$. This highlights the empirical relevance of Theorem 2 and the importance of the second-step estimation. In other words, the common strategy of normalizing all cutoffs to zero and estimating one discontinuity using the pooled data is not consistent for μ when $\phi(c)$ has such heterogeneity.

Estimation of treatment effects on ever-compliers requires a parametric functional form on β_{ec} (Theorem 4). I assume $\beta_{ec}(c, d, d') = \theta_1(d'-d)+\theta_2 c(d'-d)+\theta_3 c^2(d'-d)+\theta_4 c^3(d'-d)$ and carry out the iterative estimation procedure described in the supplemental appendix's Section B.5.2. The algorithm achieves convergence of θs within 30 iterations. The iterated bias-corrected ATE on ever-compliers equals 0.107 with standard error of 0.0109. The precision is substantially greater than the non-parametric case. Fig. 3(b) displays the treatment effect function on ever-compliers for a dose change of 0.5. Compared to ITT effects in Fig. 3(a), the return of better schooling on ever-compliers is also positive, but much less heterogeneous across ability levels.

7. Conclusion

Difficulty in gathering experimental data in many fields within the social sciences makes quasi-experimental techniques such as RDD extremely important to evaluate policies and social programs. RDD has been used in a wide range of applications in economics since the late 1990s. More recently, there has been an increasing number of applications with one forcing variable and multiple cutoffs, assigning individuals to heterogeneous treatments. The demand for multi-cutoff RDD methods is constantly growing, as richer data sets become ever more available.

This paper states conditions under which multiple RDD effects are combined to infer ATE over the entire range of cutoff values. The proposed estimator is consistent and asymptotically normal for ATEs over the entire support of variation in cutoffs and treatment doses. Asymptotic results are derived under a large number of observations and cutoffs in the sharp case of non-parametric treatment effect functions. Sufficient conditions on the rate of growth of the number of cutoffs, relative to the number of observations, are given. These rate conditions determine the feasible choice set of tuning parameters. This paper also shows that non-parametric identification in fuzzy RDD with a finite number of multiple cutoffs is impossible unless the treatment effect function is finite-dimensional. A parametric specification provides an MSE-optimal ATE estimator for the fuzzy case that is consistent and asymptotically normal.

The relevance of the ATE estimators proposed in this paper is illustrated with the data of Pop-Eleches and Urquiola (2013) on high school assignment in Romania. Of interest is the effect of high school quality on academic performance of students. I find strong evidence of non-linearities in the returns to better schooling, as a function of students' ability level. Monte Carlo simulations demonstrate that such non-linearities severely bias a naive average of local effects that does not use the correction weighting scheme proposed in this paper. Applying the fuzzy RDD methods to the Romanian data reveals causal effects on ever-compliers that are smaller and less heterogeneous than ITT effects.

The proposed estimator converges at the minimax optimal rate of root-n, as long as first-step bandwidths converge to zero at $1/K$ rate. It would be interesting to learn about efficiency properties of the ATE estimator. Theoretical tools commonly employed to derive efficiency lower-bounds may not be immediately applicable to the setting of this paper. These tools are designed for regular estimators, and for data drawn from a population where the parameter of interest is identified. In contrast, the fixed-cutoff RDD design relies on an "identification at infinity argument", and I wonder about the sufficient conditions that would obtain regularity of the ATE estimator. A possibility for future work is to the generalize the uniform convergence tools from this paper to arrive at such conditions.

Acknowledgments

I am indebted to Han Hong, Caroline Hoxby, and Guido Imbens for invaluable advice. The paper also benefited from feedback received from seminar participants at Stanford, Boston University, Cambridge, Iowa, Notre Dame, CORE-UcLouvain, UCSD, UCDavis, FGV-EESP, FGV-EPGE, Insper, PUC-Rio, Toulouse, UIUC, and at various conferences. I thank Tim Bresnahan, Arun Chandrasekhar, Michael Dinerstein, Ivan Fernandez-Val, Ivan Korolev, Michael Leung, Huiyu Li, Jessie Li, Petra Moser, Stephen Terry, Xiaowei Yu, and anonymous referees for suggestions and comments. I gratefully acknowledge the financial support received from the B.F. Haley and E.S. Shaw Fellowship at SIEPR-Stanford, United States, CORE-UcLouvain, Belgium, ISLA-Notre Dame, United States, and while visiting the Kenneth C. Griffin Department of Economics at the University of Chicago, United States.

Appendix A

Throughout the appendices, M is used as a generic finite and positive constant in the proofs. For a $p \times q$ matrix A, the norm of A is induced by the Euclidean norm $\| \cdot \|$, i.e., $\|A\| = \max_{x \in \mathbb{R}^q, x \neq 0} \|Ax\| / \|x\|$. The determinant of matrix A is denoted $\det(A)$. References to the online supplemental Appendix B include B in the numbering; for example, Lemma B.1, or Table B.2.

A.1. Proof of Theorem 1

Lemma B.1 derives asymptotic normality of the bias-corrected jump-discontinuity estimator at one cutoff based on local polynomial regressions of a vector $\mathbf{Y}_i$ on a scalar forcing variable X_i. The proof of Theorem 1 is a straightforward generalization of Lemma B.1 in the particular case of a scalar Y_i. As the sample size increases and the number of cutoff remains fixed, the jump-discontinuity estimators are independent across cutoffs. First apply Lemma B.1 to each cutoff individually, and then aggregate over cutoffs. □

A.2. Proof of Lemma 2

Define $\overline{\mathcal{C}} = [\underline{\mathcal{X}}, \overline{\mathcal{X}}] \times [\underline{\mathcal{D}}, \overline{\mathcal{D}}] \times [\underline{\mathcal{D}}, \overline{\mathcal{D}}]$. Consider the partition of $\overline{\mathcal{C}}$ made of the set of non-intersecting cubicles $T_n = \{C_1, \ldots, C_M\}$ with $M = n^3$, $n = \{1, 2, \ldots\}$. Each C_j is a half-open cubicle of the form $[x_{l-1}, x_l) \times [y_{m-1}, y_m) \times [z_{o-1}, z_o)$ with sides of lengths equal to $(\overline{\mathcal{X}} - \underline{\mathcal{X}})/n$, $(\overline{\mathcal{D}} - \underline{\mathcal{D}})/n$, and $(\overline{\mathcal{D}} - \underline{\mathcal{D}})/n$. Define the sub-collection $U_n = \{C \in T_n : C \subset \mathcal{C}\} = \{A_1, \ldots, A_Q\}$. Since $\mathcal{C}_\infty$ is dense in $\mathcal{C}$, for every $A_j \in U_n$, find a point $\mathbf{c}_j \in \mathcal{C}_\infty \cap A_j$ for which $\beta(\mathbf{c}_j)$ is known. The sum $\mu_n = \sum_{j=1}^{Q} \omega(\mathbf{c}_j)\beta(\mathbf{c}_j) \left(\overline{\mathcal{X}} - \underline{\mathcal{X}}\right) \left(\overline{\mathcal{D}} - \underline{\mathcal{D}}\right)^2 / n^3$ converges to μ^c as $n \to \infty$ because $\omega(\mathbf{c})\beta(\mathbf{c})$ is Riemann integrable on $\mathcal{C}$. □

A.3. Proof of Theorem 2

The proof combines arguments from the proof of Lemma B.1 with lemmas on the uniform convergence of empirical processes from Sections B.2 and B.3. Define μ^*, $\widetilde{\mu}$, and μ_n as follows:

$$\begin{aligned}\mu^* &= \sum_{j=1}^{K} \Delta_j \Bigg\{ e_1' \mathbb{E}[G_n^{j+}] \frac{1}{nh_{1j}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_{1j}}\right) v_i^{j+} Y_i \widetilde{H}_i^j \\ &\qquad - e_1' \mathbb{E}[G_n^{j-}] \frac{1}{nh_{1j}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_{1j}}\right) v_i^{j-} Y_i \widetilde{H}_i^j \Bigg\} && \text{(A.1)} \\ &= \sum_{j=1}^{K} \Delta_j \frac{1}{nh_{1j}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_{1j}}\right) Y_i e_1' \left(v_i^{j+} \mathbb{E}[G_n^{j+}] - v_i^{j-} \mathbb{E}[G_n^{j-}] \right) \widetilde{H}_i^j && \text{(A.2)} \\ \widetilde{\mu} &= \sum_{j=1}^{K} \Delta_j \left\{ e_1' G_n^{j+} \mathbb{E}\left[\frac{1}{nh_{1j}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_{1j}}\right) v_i^{j+} Y_i^{j+} \widetilde{H}_i^j \right] \right. \\ &\qquad \left. - e_1' G_n^{j-} \mathbb{E}\left[\frac{1}{nh_{1j}} \sum_{i=1}^{n} k\left(\frac{X_i - c_j}{h_{1j}}\right) v_i^{j-} Y_i^{j-} \widetilde{H}_i^j \right] \right\} \\ \mu_n &= \sum_{j=1}^{K} \Delta_j B_j. && \text{(A.3)}\end{aligned}$$

Write

$$\begin{aligned}\frac{\widehat{\mu} - \mathcal{B}_{1n}^c - \mathcal{B}_{2n}^c - \mu}{(\mathcal{V}_n^c)^{1/2}} &= \frac{\mu^* - \mathbb{E}[\mu^* | \mathcal{X}_n]}{(\mathcal{V}_n^c)^{1/2}} && \text{(A.4)} \\ &+ \frac{\widetilde{\mu} - \mathcal{B}_{1n}^c}{(\mathcal{V}_n^c)^{1/2}} && \text{(A.5)} \\ &+ \frac{\mu_n - \mathcal{B}_{2n}^c - \mu}{(\mathcal{V}_n^c)^{1/2}} && \text{(A.6)} \\ &+ \frac{\widehat{\mu} - \mathbb{E}[\widehat{\mu} | \mathcal{X}_n] - (\mu^* - \mathbb{E}[\mu^* | \mathcal{X}_n])}{(\mathcal{V}_n^c)^{1/2}} && \text{(A.7)} \\ &+ \frac{\mathbb{E}[\widehat{\mu} - \mu_n | \mathcal{X}_n] - \widetilde{\mu}}{(\mathcal{V}_n^c)^{1/2}}. && \text{(A.8)}\end{aligned}$$

The proof in this appendix applies a central limit theorem (CLT) to show that Part (A.4) converges in distribution to a standard normal; it demonstrates that $\mathcal{B}_{1n}$ approximates the first-step bias, that is, that part (A.5) converges in probability to zero; and it shows that $\mathcal{B}_{2n}$ approximates the second-step bias (integration error), that is, that part (A.6) converges to zero. Lemma B.7 shows that parts (A.7) and (A.8) converge in probability to zero.

Part (A.4)

First, find the rate that $(\mathcal{V}_n^c)^{-1/2}$ grows. Define ϕ_n and rewrite $\mathcal{V}_n^c$ as follows:

$$\phi_n(X_i) = \sum_{j=1}^{K} \frac{\Delta_j}{nh_{1j}} k\left(\frac{X_i - c_j}{h_{1j}}\right) e_1' \left(v_i^{j+} \mathbb{E}[G_n^{j+}] - v_i^{j-} \mathbb{E}[G_n^{j-}] \right) \widetilde{H}_i^j \tag{A.9}$$

$$\mathcal{V}_n^c = \sum_{i=1}^{n} \mathbb{E}\left[\varepsilon_i^2 \phi_n(X_i)^2 \right]. \tag{A.10}$$

Choose alternative bandwidths $h_{1j}^*, j = 1, \ldots, K$, such that (i) there exists $\delta > 0$ (independent of n) such that $\delta < h_{1j}^*/h_{1j} \leq 1$ $\forall j$; and (ii) $[c_j - h_{1j}^*, c_j + h_{1j}^*] \cap [c_{j'} - h_{1j'}^*, c_{j'} + h_{1j'}^*] = \emptyset$ for any $j \neq j'$.

$$\mathcal{V}_n^c = n\mathbb{E}\left[\zeta^2(X_i) \phi_n^2(X_i) \right] \geq n \sum_{j=1}^{K} \int_{c_j - h_{1j}^*}^{c_j + h_{1j}^*} \zeta^2(x) \phi_n^2(x) f(x)\, dx \tag{A.11}$$

$$= n\sum_{j=1}^{K}\int_{c_j-h^*_{1j}}^{c_j+h^*_{1j}} \zeta^2(x)\left(\frac{\Delta_j}{nh_{1j}}k\left(\frac{x-c_j}{h_{1j}}\right)e_1'\left(\mathbb{I}\{x\geq 0\}\mathbb{E}[G_n^{j+}]-\mathbb{I}\{x<0\}\mathbb{E}[G_n^{j-}]\right)\right.$$

$$\left. H\left(\frac{x-c_j}{h_{1j}}\right)\right)^2 f(x)\,dx \tag{A.12}$$

$$= \frac{1}{Kn}\frac{1}{K}\sum_{j=1}^{K}\frac{K^2\Delta_j^2}{h_{1j}}\int_{-h^*_{1j}/h_{1j}}^{h^*_{1j}/h_{1j}}\zeta^2(c_j+uh_{1j})\left(k\left(u\right)e_1'\left(\mathbb{I}\{u\geq 0\}\mathbb{E}[G_n^{j+}]-\mathbb{I}\{u<0\}\mathbb{E}[G_n^{j-}]\right)\right.$$

$$\left. H\left(u\right)\right)^2 f(c_j+uh_{1j})\,du \;\geq\; \frac{M}{Kn\bar{h}_1} \tag{A.13}$$

where the first inequality follows from the integrand being positive and $\cup_{j=1}^{K}[c_j-h^*_{1j}, c_j+h^*_{1j}]\subseteq\cup_{j=1}^{K}[c_j-h_{1j}, c_j+h_{1j}]$; the third equality uses a change of variables $u=(x-c_j)/h_{1j}$; and the last inequality follows because (a) $h_{1j}\leq\bar{h}_1$; (b) $K^2\Delta_j^2$ is bounded away from zero uniformly over j (Lemma B.9); and (c) each integral is bounded away from zero over j because the integration limits, $\zeta^2(c_j+uh_{1j})$, $\mathbb{E}[G_n^{j\pm}]$, and $f(c_j+uh_{1j})$ are uniformly close to quantities that are positive definite uniformly over j (see Lemma B.6 and recall that f and ζ are bounded away from zero because of Assumptions 4 and 7). The inequality in (A.13) implies that $(\mathcal{V}_n^c)^{-1}=O(Kn\bar{h}_1)$ where $Kn\bar{h}_1\to\infty$.

Second, write part (A.4) as a weighted sum across i:

$$\mu^* = \sum_{i=1}^{n}Y_i\underbrace{\sum_{j=1}^{K}\frac{\Delta_j}{nh_{1j}}k\left(\frac{X_i-c_j}{h_{1j}}\right)e_1'\left(v_i^{j+}\mathbb{E}[G_n^{j+}]-v_i^{j-}\mathbb{E}[G_n^{j-}]\right)\widetilde{H}_i^j}_{\equiv\phi_n(X_i)} = \sum_{i=1}^{n}Y_i\phi_n(X_i) \tag{A.14}$$

so that

$$\frac{\mu^*-\mathbb{E}[\mu^*|\mathcal{X}_n]}{(\mathcal{V}_n^c)^{1/2}} = \frac{\sum_{i=1}^{n}\left(Y_i-\mathbb{E}[Y_i|X_i]\right)\phi_n(X_i)}{(\mathcal{V}_n^c)^{1/2}} = \frac{\sum_{i=1}^{n}\varepsilon_i\phi_n(X_i)}{(\mathcal{V}_n^c)^{1/2}}. \tag{A.15}$$

Eq. (A.15) is a sum of iid random variables with zero mean, where $\mathcal{V}_n^c$ is the variance of the numerator. The Lindeberg condition is verified next. Take an arbitrary $\delta>0$.

$$\sum_{i=1}^{n}\mathbb{E}\left[(\mathcal{V}_n^c)^{-1}\varepsilon_i^2\phi_n(X_i)^2\mathbb{I}\left\{\left|(\mathcal{V}_n^c)^{-1/2}\varepsilon_i\phi_n(X_i)\right|>\delta\right\}\right] \tag{A.16}$$

$$\leq\sum_{i=1}^{n}\mathbb{E}\left[MKn\bar{h}_1\phi_n(X_i)^2\mathbb{I}\left\{M'\left(Kn\bar{h}_1\right)^{1/2}|\phi_n(X_i)|>\delta\right\}\right] \tag{A.17}$$

$$\leq\sum_{i=1}^{n}\mathbb{E}\left[MKn\bar{h}_1\left(Kn\underline{h}_1\right)^{-2}\mathbb{I}\left\{M'\left(Kn\bar{h}_1\right)^{1/2}\left(Kn\underline{h}_1\right)^{-1}>\delta\right\}\right] \tag{A.18}$$

$$\leq\left(K\underline{h}_1\right)^{-1}\mathbb{I}\left\{M'\left(Kn\underline{h}_1\right)^{-1/2}>\delta\right\}=o(1) \tag{A.19}$$

where the first inequality relies on the fact that ε_i is a.s. bounded (Assumption 7), and that $(\mathcal{V}_n^c)^{-1}=O\left(Kn\bar{h}_1\right)$ Eq. (A.13). The second inequality uses that $\phi_n(x)=O\left(Kn\underline{h}_1\right)^{-1}$ uniformly over x. In fact, $\phi_n(x)$ is a sum of K components of which at most two are non-zero, $\Delta_j=O\left(K^{-1}\right)$ uniformly over j (Lemma B.9), $k(\cdot)$ is bounded (Assumption 3), $\mathbb{E}[G_n^{j\pm}]$ is uniformly close to $G^{j\pm}$ whose norm is bounded away from zero (Lemma B.6). The last inequality relies on the rate condition $\bar{h}_1/\underline{h}_1=O(1)$, and that the indicator becomes zero for large n. The Lindeberg–Feller CLT says that Eq. (A.15), and thus part (A.4), converges in distribution to a standard normal.

Part (A.5)

First consider

$$\mathbb{E}\left[\frac{1}{h_{1j}}k\left(\frac{X_i-c_j}{h_{1j}}\right)v_i^{j+}\widetilde{H}_i^j\mathbb{E}\left[Y_i^{j+}|X_i\right]\right] \tag{A.20}$$

$$=\mathbb{E}\left[\frac{1}{h_{1j}}k\left(\frac{X_i-c_j}{h_{1j}}\right)v_i^{j+}\widetilde{H}_i^j\frac{\nabla^{(\rho_1+1)}R(c_j,d_j)}{(\rho_1+1)!}\left(\frac{X_i-c_j}{h_{1j}}\right)^{\rho_1+1}h_{1j}^{\rho_1+1}\right] \tag{A.21}$$

$$+\,\mathbb{E}\left[\frac{1}{h_{1j}}k\left(\frac{X_i-c_j}{h_{1j}}\right)v_i^{j+}\widetilde{H}_i^j\frac{\nabla^{(\rho_1+2)}R(c_j^*,d_j)}{(\rho_1+2)!}\left(\frac{X_i-c_j}{h_{1j}}\right)^{\rho_1+2}h_{1j}^{\rho_1+2}\right] \tag{A.22}$$

$$=h_{1j}^{\rho_1+1}\frac{\nabla^{(\rho_1+1)}R(c_j,d_j)}{(\rho_1+1)!}f(c_j)\gamma^*+O\left(\bar{h}_1^{\rho_1+2}\right) \tag{A.23}$$

where $\mathbb{E}\left[Y_i^{j+}|X_i\right]$ is the difference between $\mathbb{E}[Y_i|X_i]$ and its ρ_1th order Taylor expansion around $X_i=c_j$ (see Equations B.37 and B.38). The expectations in Eqs. (A.21) and (A.22), without the $h_{1j}^{\rho_1+1}$ and $h_{1j}^{\rho_1+2}$ terms, are bounded over j because the kernel, derivatives, and polynomials are bounded functions of $u=(x-c_j)h_{1j}^{-1}$ (Assumptions 3 and 7). The remainder term $O\left(\bar{h}_1^{\rho_1+2}\right)$ is uniform over j.

Next,

$$\begin{aligned}\frac{\widetilde{\mu}-\mathcal{B}_{1n}}{(\mathcal{V}_n^c)^{1/2}}=&(\mathcal{V}_n^c)^{-1/2}\sum_{j=1}^{K}\Delta_j e_1' G_n^{j+}\mathbb{E}\left[\frac{1}{nh_{1j}}\sum_{i=1}^{n}k\left(\frac{X_i-c_j}{h_{1j}}\right)\upsilon_i^{j+}\mathbb{E}\left[Y_i^{j+}|X_i\right]\widetilde{H}_i^j\right]\\&-(\mathcal{V}_n^c)^{-1/2}\mathcal{B}_{1n}^{+} &\text{(A.24)}\\&-(\mathcal{V}_n^c)^{-1/2}\sum_{j=1}^{K}\Delta_j e_1' G_n^{j-}\mathbb{E}\left[\frac{1}{nh_{1j}}\sum_{i=1}^{n}k\left(\frac{X_i-c_j}{h_{1j}}\right)\upsilon_i^{j-}\mathbb{E}\left[Y_i^{j-}|X_i\right]\widetilde{H}_i^j\right]\\&+(\mathcal{V}_n^c)^{-1/2}\mathcal{B}_{1n}^{-} &\text{(A.25)}\end{aligned}$$

where

$$\mathcal{B}_{1n}=\mathcal{B}_{1n}^{+}-\mathcal{B}_{1n}^{-} \tag{A.26}$$

$$\mathcal{B}_{1n}^{+}=((\rho_1+1)!)^{-1}\sum_{j=1}^{K}h_{1j}^{\rho_1+1}\Delta_j f(c_j)\nabla_x^{\rho_1+1}R(c_j,d_j)e_1' G_n^{j+}\gamma^* \tag{A.27}$$

$$\mathcal{B}_{1n}^{-}=((\rho_1+1)!)^{-1}\sum_{j=1}^{K}h_{1j}^{\rho_1+1}\Delta_j f(c_j)\nabla_x^{\rho_1+1}R(c_j,d_{j-1})e_1' G_n^{j-}\gamma^*. \tag{A.28}$$

Consider part (A.24). Part (A.25) follows a symmetric argument.

$$\begin{aligned}\text{(A.24)}=&(\mathcal{V}_n^c)^{-1/2}\sum_{j=1}^{K}\Delta_j e_1' G_n^{j+}\mathbb{E}\left[\frac{1}{nh_{1j}}\sum_{i=1}^{n}k\left(\frac{X_i-c_j}{h_{1j}}\right)\upsilon_i^{j+}\mathbb{E}\left[Y_i^{j+}|X_i\right]\widetilde{H}_i^j\right]-(\mathcal{V}_n^c)^{-1/2}\mathcal{B}_{1n}^{+} &\text{(A.29)}\\=&(\mathcal{V}_n^c)^{-1/2}\left[\sum_{j=1}^{K}\Delta_j e_1' G_n^{j+}h_{1j}^{\rho_1+1}\frac{\nabla^{(\rho_1+1)}R(c_j,d_j)}{(\rho_1+1)!}f(c_j)\gamma^*-\mathcal{B}_{1n}^{+}\right] &\text{(A.30)}\\&+(\mathcal{V}_n^c)^{-1/2}\left[\sum_{j=1}^{K}\Delta_j e_1' G_n^{j+}O\left(\bar{h}_1^{\rho_1+2}\right)\right] &\text{(A.31)}\\=&0+O\left(\left(Kn\bar{h}_1\right)^{1/2}\right)KO(K^{-1})O_P(1)O\left(\bar{h}_1^{\rho_1+2}\right)=o_P(1) &\text{(A.32)}\end{aligned}$$

where the second equality uses the expansion in Eq. (A.23). The third equality uses the definition of $\mathcal{B}_{1n}^{+}$, that $\Delta_j=O(K^{-1})$ uniformly over j, and that $G_n^{j+}=O_P(1)$. These terms are $o_P(1)$ because of the rate condition $\left(Kn\bar{h}_1\right)^{1/2}\bar{h}_1^{\rho_1+1}=O(1)$.

Part (A.6)

$$\begin{aligned}\frac{\mu_n-\mathcal{B}_{2n}-\mu}{(\mathcal{V}_n^c)^{1/2}}&=O\left(\left(Kn\bar{h}_1\right)^{1/2}\right)\left(\sum_{j=1}^{K}\Delta_j B_j-\mathcal{B}_{2n}-\int_{\mathcal{C}}\omega(\mathbf{c})\beta(\mathbf{c})\,d(\mathbf{c})\right) &\text{(A.33)}\\&=O\left(\left(Kn\bar{h}_1\right)^{1/2}\right)O\left(h_2^{\rho_2+2}\right)=O(1)O\left(h_2\right)=o(1) &\text{(A.34)}\end{aligned}$$

where the first equality uses the rate on $(\mathcal{V}_n^c)^{-1/2}$ Eq. (A.13). The second equality applies Lemma B.9 and relies on Assumption 6 (asymptotic behavior of $\{\mathbf{c}_j\}_j$) and Assumption 7 (smoothness of $\beta(\mathbf{c})$). The third equality uses the rate condition $\left(Kn\bar{h}_1\right)^{1/2}h_2^{\rho_2+1}=O(1)$. Lemma B.9 also shows that $\mathcal{B}_{2n}=O\left(h_2^{\rho_2+1}\right)$, which yields $(\mathcal{V}_n^c)^{-1/2}\mathcal{B}_{2n}=O\left(\left(Kn\bar{h}_1\right)^{1/2}h_2^{\rho_2+1}\right)=O(1)$.

Lemma B.7 shows that parts (A.7) and (A.8) converge in probability to zero, which concludes the proof. □

A.4. Proof of Theorem 3

Part (27)

First, consider the ideal setting where estimators μ^* are functions of data observed from $\{Y_i(d)\}_{d\in\mathcal{D}}$ and X_i. For a choice of loss function $L(\mu, \mu')$, the minimax risk of estimating the parameter $\mu^c(P)$ is defined as $\inf_{\mu^*} \sup_{P\in\mathcal{P}} \mathbb{E}_P\left[L(\mu^*, \mu^c(P))\right]$. Here, the 0–1 loss function is used, that is, $L_n(\mu, \mu') = \mathbb{I}\{n^r|\mu - \mu'| > \epsilon\}$, for a positive rate r and ϵ. In this case, $\mathbb{E}_P\left[L_n(\mu^*, \mu^c(P))\right] = \mathbb{P}_P\left[n^r|\mu^* - \mu^c(P)| > \epsilon\right]$. The minimax risk is the supremum probability over $\mathcal{P}$ of an estimator being farther than ϵn^{-r} from the truth minimized over all possible estimators μ^*. The rate r is an upper bound on the rate of convergence if for small $\epsilon > 0$ there exists a lower bound $L \in (0, 1)$ such that $\inf_{\mu^*} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left[n^r|\mu^* - \mu^c(P)| > \epsilon\right] \geq L$ for large n. The rate r is the minimax optimal rate if it is an upper bound and achievable; that is, if there exists an estimator $\widehat{\mu}$ that converges at rate r uniformly. The estimator $\widehat{\mu}$ converges at rate r uniformly if, for any small $\delta > 0$, there exists large $\epsilon \in (0, \infty)$ such that $\sup_{P\in\mathcal{P}} \mathbb{P}_P\left[n^r|\widehat{\mu} - \mu^c(P)| > \epsilon\right] < \delta$ for large n. See discussion in Chapter 2 of Tsybakov (2009).

One common approach to compute lower bounds for the minimax risk is to use Le Cam's method. For $\epsilon > 0$, choose two models $P, Q \in \mathcal{P}$ such that $|\mu^c(P) - \mu^c(Q)| > \epsilon n^{-r}$. Le Cam's method leads to the following inequality: $\inf_{\mu^*} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left[n^r|\mu^* - \mu^c(P)| > \epsilon/2\right] \geq e^{-nKL(P,Q)}/4$, where $KL(P, Q)$ is the Kullback–Leibler divergence between P and Q. See Equations (2.7), (2.9), and Theorem 2.2(iii) of Tsybakov (2009). This inequality is used to prove part (27) with $r = 1/2$.

Consider the continuous counterfactual density $\omega^c(\mathbf{c})$. The researcher must be choose a counterfactual density such that its marginal densities $\int \omega^c(c, d, d')\, d(d')$ and $\int \omega^c(c, d, d')\, d(d)$ are different functions; otherwise, $\mu^c = 0$. Construct an infinitely differentiable bounded function $g(c, d) \geq 0$ such that $\int [g(c, d') - g(c, d)]\, \omega(\mathbf{c})\, d\mathbf{c} = 1$.

Construct two models $P, Q \in \mathcal{P}$ as follows. Let $\varepsilon_i \sim N(0, 1)$ and $X_i \sim U[0, 1]$ iid and independent of each other. Pick $\xi > 2\sqrt{\pi}\epsilon > 0$. For model P, define $Y_i(d) = \Phi\left(\xi n^{-1/2} g(X_i, d) + \varepsilon_i\right)$, where Φ is the standard normal cdf. For model Q, define $Y_i(d) = \Phi\left(\varepsilon_i\right)$. The expectation of $Y_i(d)$ conditional on $X_i = c$, that is, $R(c, d)$, is an infinitely differentiable function. The variables have bounded support, and models P and Q satisfy all the conditions to be in $\mathcal{P}$. Under model P,

$$
\begin{aligned}
\beta(\mathbf{c}; P) &= \mathbb{E}_P[Y_i(d') - Y_i(d)\ |X_i = c] \\
&= \mathbb{E}_P[\Phi\left(\xi n^{-1/2} g(X_i, d') + \varepsilon_i\right) - \Phi\left(\xi n^{-1/2} g(X_i, d) + \varepsilon_i\right)\ |X_i = c] \\
&= \mathbb{E}_P[\phi\left(\varepsilon_i^*\right) \xi n^{-1/2} \left(g(c, d') - g(c, d)\right)\ |X_i = c] \\
&= \mathbb{E}_P[\phi\left(\varepsilon_i^*\right)]\ \xi n^{-1/2} \left(g(c, d') - g(c, d)\right)
\end{aligned}
$$

where ϕ is the standard normal pdf, and ε_i^* is in between $\varepsilon_i + \xi n^{-1/2} g(c, d')$ and $\varepsilon_i + \xi n^{-1/2} g(c, d)$. As n grows large, $\mathbb{E}_P[\phi\left(\varepsilon_i^*\right)] = \mathbb{E}_P[\phi\left(\varepsilon_i\right)] + o(1) = \frac{1}{2\sqrt{\pi}} + o(1)$ where the $o(1)$ term is uniform over (c, d, d'). Then,

$$
\mu^c(P) = \frac{1}{2\sqrt{\pi}} \xi n^{-1/2} \int \left(g(c, d') - g(c, d)\right)\ \omega(\mathbf{c})\, d\mathbf{c} + o\left(n^{-1/2}\right) = \frac{1}{2\sqrt{\pi}} \xi n^{-1/2} + o\left(n^{-1/2}\right).
$$

Under model Q, $\beta(\mathbf{c}; Q) = 0$. Therefore,

$$
\mu^c(P) - \mu^c(Q) = \frac{1}{2\sqrt{\pi}} \xi n^{-1/2} + o\left(n^{-1/2}\right) > \epsilon n^{-1/2}
$$

for large n, because $\frac{1}{2\sqrt{\pi}}\xi > \epsilon$.

Next, we use the following inequality to show that $r = 1/2$ is an upper bound on the rate of convergence,

$$
\inf_{\mu^*} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left[n^{1/2}|\mu^* - \mu^c(P)| > \epsilon/2\right] \geq e^{-nKL(P,Q)}/4.
$$

Let d^* be such that $g(c, d^*) > 0$ for some c. For simple models like P and Q, any function of the variables $\{Y_i(d)\}_{d\in\mathcal{D}}$ and X_i can be rewritten as functions of $Y_i(d^*)$ and X_i because $Y_i(d)$ is a deterministic function of $Y_i(d^*)$ and X_i for any d. It suffices to look at the distribution of $Y_i(d^*)$ and X_i instead of the distribution of $\{Y_i(d)\}_{d\in\mathcal{D}}$ and X_i. Consider the Kullback–Leibler divergence for the distributions P and Q of $(Y_i(d^*), X_i)$,

$$
KL(P, Q) = \int \log\left[\frac{p(y, x)}{q(y, x)}\right] p(y, x)\, dydx,
$$

where $p(y, x)$ and $q(y, x)$ are the pdfs of $(Y_i(d^*), X_i)$ under P and Q respectively. Define $\widetilde{Y}_i = \xi n^{-1/2} g(X_i, d^*) + \varepsilon_i$ under P, and $\widetilde{Y}_i = \varepsilon_i$ under Q. It follows that $(Y_i(d^*), X_i) = (\Phi(\widetilde{Y}_i), X_i)$ under both P and Q. The Kullback–Leibler divergence is invariant to such a transformation of variables.

$$
KL(P, Q) = \int \log\left[\frac{\widetilde{p}(y, x)}{\widetilde{q}(y, x)}\right] \widetilde{p}(y, x)\, dydx,
$$

where $\widetilde{p}(y,x) = \phi\left(y - \xi n^{-1/2} g(x,d^*)\right)$ and $\widetilde{q}(y,x) = \phi(y)$ are the pdfs of $(\widetilde{Y}_i, X_i)$ under P and Q respectively.

$$\begin{aligned}
KL(P,Q) &= \int \log\left[\frac{\exp\left\{-(1/2)\left(y - \xi n^{-1/2} g(x,d^*)\right)^2\right\}}{\exp\left\{-(1/2)y^2\right\}}\right] \widetilde{p}(y,x)\,dydx \\
&= \int \log\left[\exp\left\{y\xi n^{-1/2} g(x,d^*) - (1/2)\xi^2 n^{-1} g(x,d^*)^2\right\}\right] \widetilde{p}(y,x)\,dydx \\
&= \int \left[y\xi n^{-1/2} g(x,d^*) - (1/2)\xi^2 n^{-1} g(x,d^*)^2\right] \widetilde{p}(y,x)\,dydx \\
&= \int (1/2)\xi^2 n^{-1} g(x,d^*)^2\,dx \\
&= (1/2)\xi^2 n^{-1} \int g(x,d^*)^2\,dx > 0
\end{aligned}$$

Pick $\eta > 1$ such that $(1/2)\xi^2 \int g(x,d^*)^2\,dx < \log(\eta)$. Then, $e^{-nKL(P,Q)}/4 > 1/(4\eta) > 0$, and

$$\inf_{\mu^*} \sup_{P \in \mathcal{P}} \mathbb{P}_P\left[n^{1/2}|\mu^* - \mu^c(P)| > \epsilon/2\right] \geq \frac{1}{4\eta}.$$

This is a minimax lower bound for estimators μ^* that are functions of an ideal sample of $\{Y_i(d)\}_{d\in\mathcal{D}}$ and X_i. In practice, only part of these variables are observed according to the schedule of cutoff-doses $\{\mathbf{c}_j\}_{j=1}^K$. The set of all estimators $\widetilde{\mu}$ that are functions of the observed variables (Y_i, X_i) is a subset of the set of all estimators μ^*. Therefore, the lower bound above is also a minimax lower bound for all estimators $\widetilde{\mu}$:

$$\inf_{\widetilde{\mu}} \sup_{P \in \mathcal{P}} \mathbb{P}_P\left[n^{1/2}|\widetilde{\mu} - \mu^c(P)| > \epsilon/2\right] \geq \frac{1}{4\eta}.$$

Part (28)

Let $\widehat{\mu}$ denote $\widehat{\mu}^c$ and $\mu = \mu^c(P)$ for notational ease. The goal is to show that, for any small $\delta > 0$, there exists large $\epsilon \in (0,\infty)$ such that $\sup_{P\in\mathcal{P}} \mathbb{P}_P\left[n^{1/2}|\widehat{\mu} - \mu| > \epsilon\right] < \delta$ for large n. The choice of $\overline{h}_1$ plus the discussion preceding Eq. (A.13) lead to $(\mathcal{V}_n^c)^{-1/2} \geq Mn^{1/2}$ for large n. Thus, $\mathbb{P}_P\left[n^{1/2}|\widehat{\mu} - \mu| > \epsilon/M\right] \leq \mathbb{P}_P\left[(\mathcal{V}_n^c)^{-1/2}|\widehat{\mu} - \mu| > \epsilon\right]$ uniformly over $\mathcal{P}$ for large n. Theorem 2 breaks $(\mathcal{V}_n^c)^{-1/2}|\widehat{\mu} - \mu|$ into four components: the CLT component N_n, that converges in distribution to a standard normal (part (A.4)); the first-step bias component B_n, that converges in probability to zero (part (A.5)); the integration error component I_n, that converges in probability to zero (part (A.6)); and the remainder terms R_n, that converge in probability to zero (parts (A.7) and (A.8)). It is true that

$$\mathbb{P}_P\left((\mathcal{V}_n^c)^{-1/2}|\widehat{\mu} - \mu| > \epsilon\right) \leq \mathbb{P}_P\left(|N_n| > \epsilon/4\right) + \mathbb{P}_P\left(|B_n| > \epsilon/4\right) + \mathbb{P}_P\left(|I_n| > \epsilon/4\right) + \mathbb{P}_P\left(|R_n| > \epsilon/4\right).$$

Hence, for each of the four components, it suffices to show that for a choice of $\delta > 0$ small, there exist large n and large $\epsilon > 0$ such that the supremum probability over $\mathcal{P}$ is less than δ. The restrictions placed in the class of models $\mathcal{P}$ along with the proof of Theorem 2 give the result.

N_n-**term**: part (A.4) has zero mean and unit variance (see Eq. (A.15)). Chebyshev's inequality implies that the supremum probability of the absolute value of part (A.4) being greater than $\epsilon/4$ is smaller than $16/\epsilon^2$ uniformly over $\mathcal{P}$.

B_n-**term**: B_n is the sum of B_n^+ (part (A.24)), and B_n^- (part (A.25)). B_n^+ converges in probability to zero uniformly over $\mathcal{P}$ because the approximations of Lemma B.6, the bounds on the derivatives of $R(x,d)$, on $f(x)$, on $\sigma^2(x,d)$, and on the rate of $(\mathcal{V}_n^c)^{-1/2}$ hold uniformly over $\mathcal{P}$. The weights Δ_j do not depend on P. The same idea applies to B_n^-. Thus, for $\epsilon > 0$, $\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(|B_n| > \epsilon/4\right)$ converges to zero.

I_n-**term**: uniform bounds on the partial derivatives of $\beta(\mathbf{c})$ yield a uniform bound on the approximation error of the numerical integral. See Lemma B.9. The bounds on the rate of $(\mathcal{V}_n^c)^{-1/2}$ also hold uniformly over $\mathcal{P}$. For every $\epsilon > 0$ there exists a large n for which $|I_n| \leq \epsilon/4$ holds uniformly over $\mathcal{P}$.

R_n-**term**: R_n is the sum of R_n^a (part (A.7)) and R_n^b (part (A.8)). Lemma B.7 shows that both converge in probability to zero. They also converge in probability to zero uniformly over $\mathcal{P}$ for the same reasons that the B_n-term above does. Therefore, for $\epsilon > 0$, $\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(|R_n| > \epsilon/4\right)$ converges to zero. □

A.5. Proof of Theorem 4

Define $\delta_{j,l} = \mathbb{I}\{\mathcal{U}_i(c_j) = d_l\}$. Assumption 8 (no ever-defiers) implies the following facts: (i) $\mathbb{P}\left[\delta_{j-1,l} = 0,\ \delta_{j,l} = 1\right] = 0$ for $\forall l \neq j$; (ii) $\mathbb{P}\left[\delta_{j-1,l} = 1,\ \delta_{j,l} = 0\right] = 0$ for $l = j$;
(iii) $\mathbb{P}\left[\delta_{j-1,l} = 1,\ \delta_{j,l} = 0,\ \delta_{j,u} = 1\right] = 0$ for $\forall u \neq j$ and $u \neq l$.

Fix a small $e > 0$ and use fact (i) to obtain

$$\begin{aligned}
&\mathbb{E}[Y_i|X_i = c_j + e] = \sum_{l=0}^{K} \mathbb{E}\left[\delta_{j,l} Y_i(d_l)|X_i = c_j + e\right] \\
&= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j + e,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j + e\right] \\
&+ \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j + e,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 0\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 0|X_i = c_j + e\right] \\
&= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j + e,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j + e\right] \\
&+ \mathbb{E}\left[Y_i(d_j)|X_i = c_j + e,\ \delta_{j,j} = 1,\ \delta_{j-1,j} = 0\right] \mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,j} = 0|X_i = c_j + e\right].
\end{aligned}$$

Take the limit as $e \downarrow 0$. Use that $\{\delta_{j,l} = 1,\ \delta_{j-1,l} = 1\}$ and $\{\delta_{j,j} = 1,\ \delta_{j-1,j} = 0\}$ are finite unions of measurable sets of the form $\{\mathcal{U}_i = \bar{\mathcal{U}}\}$, $\bar{\mathcal{U}} \in \mathcal{U}^*$. The conditional expectation and probability are continuous functions of x conditional on these sets (Assumption 8).

$$\begin{aligned}
&\lim_{e\downarrow 0} \mathbb{E}[Y_i|X_i = c_j + e] \\
&= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] \\
&+ \mathbb{E}\left[Y_i(d_j)|X_i = c_j,\ \delta_{j,j} = 1,\ \delta_{j-1,j} = 0\right] \mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,j} = 0|X_i = c_j\right] \\
&= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] \\
&+ \sum_{l=0, l\neq j}^{K} \mathbb{E}\left[Y_i(d_j)|X_i = c_j,\ \delta_{j,j} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right]
\end{aligned}$$

Similarly, use fact (ii) for the left-hand-side limit, $\lim_{e\downarrow 0} \mathbb{E}[Y_i|X_i = c_j - e]$

$$\begin{aligned}
&= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] \\
&+ \sum_{l=0, l\neq j}^{K} \mathbb{E}\left[Y_i(d_j)|X_i = c_j,\ \delta_{j,l} = 0,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 0,\ \delta_{j-1,l} = 1|X_i = c_j\right]
\end{aligned}$$

Use fact (iii) to get

$$\begin{aligned}
&= \sum_{l=0}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,l} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,l} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] \\
&+ \sum_{l=0, l\neq j}^{K} \mathbb{E}\left[Y_i(d_l)|X_i = c_j,\ \delta_{j,j} = 1,\ \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right].
\end{aligned}$$

The difference between right and left hand side limits is B_j

$$\begin{aligned}
&= \sum_{l=0, l\neq j}^{K} \mathbb{E}\left[Y_i(d_j) - Y_i(d_l)|X_i = c_j, \delta_{j,j} = 1, \delta_{j-1,l} = 1\right] \mathbb{P}\left[\delta_{j,j} = 1, \delta_{j-1,l} = 1|X_i = c_j\right] \\
&= \sum_{l=0, l\neq j}^{K} \beta_{ec}(c_j, d_l, d_j) \mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right].
\end{aligned}$$

Next, it is shown that $\mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] = \omega_{j,l}$, for $l \neq j$.

$$\begin{aligned}
&\mathbb{P}\left[\delta_{j,j} = 1,\ \delta_{j-1,l} = 1|X_i = c_j\right] = \mathbb{P}[\delta_{j,l} = 0, \delta_{j-1,l} = 1|X_i = c_j] \\
&= \mathbb{P}[\delta_{j,l} = 0|X_i = c_j] - \mathbb{P}[\delta_{j-1,l} = 0|X_i = c_j]
\end{aligned}$$

$$= \lim_{e \downarrow 0} \left\{ \mathbb{P}[\mathcal{U}_i(c_j) \neq d_l | X_i = c_j + e] - \mathbb{P}[\mathcal{U}_i(c_{j-1}) \neq d_l | X_i = c_j - e] \right\}$$

$$= \lim_{e \downarrow 0} \left\{ \mathbb{P}[D_i = d_l | X_i = c_j - e] - \mathbb{P}[D_i = d_l | X_i = c_j + e] \right\}$$

where facts (i) and (ii) are used. This proves the first part of the theorem.

If β_{ec} belongs to the class of functions of Assumption 9, then $B_j = \widetilde{W}_j \boldsymbol{\theta}_0$. If the matrix $\widetilde{\mathbf{W}}'\widetilde{\mathbf{W}} = \sum_j \widetilde{W}_j \widetilde{W}_j'$ is invertible, then the second part of the theorem follows.

Conversely, suppose that the $p > K$ elements in $\{\beta_{ec}(c_j, d_l, d_j)$ for $(j, l) : \omega_{j,l} > 0\}$ are identified for every fuzzy assignment $\tilde{\mathbf{c}}_1 = (c_1, d_0, d_1)$, ..., $\tilde{\mathbf{c}}_p = (c_K, d_{K-1}, d_K)$. Identification means that there is a unique solution to the following constrained linear system:

$$\begin{bmatrix} B_1 \\ \vdots \\ B_K \end{bmatrix} = \begin{bmatrix} \omega_{1,0} & \dots & \omega_{1,K} & 0 & 0 & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \omega_{2,0} & \dots & \omega_{2,K} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & & & & & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & \omega_{K,0} & \dots & \omega_{K,K-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

such that $(\beta_1, \dots, \beta_p) \in \mathcal{G}$.

The $K \times p$ matrix of coefficients has rank equal to K because the assignment is fuzzy. Since $p > K$, the unconstrained system has infinitely many nonzero solutions of the form $\mathbf{b} = \mathbf{b}^p + \sum_{m=1}^{p-K} \lambda_m \mathbf{b}_m^s$ for any $(\lambda_1, \dots, \lambda_{p-K}) \in \mathbb{R}^{p-K}$, where $\{\mathbf{b}_m^s\}_{m=1}^{p-K}$ are the basis vectors of the null-space of the unconstrained system, and $\mathbf{b}^p$ is a particular solution. By assumption, the constrained system has one unique solution $\mathbf{b}^* \in \mathcal{G}$, so $\mathbf{b}^* + \mathbf{b}_m^s \notin \mathcal{G} \ \forall m$. This implies that $\mathbf{b}_m^s \notin \mathcal{G} \ \forall m$ because $\mathcal{G}$ is a vector subspace of $\mathbb{R}^p$. This is a set of $p - K$ linearly independent vectors in $\mathbb{R}^p$ not in $\mathcal{G}$. Therefore, the $dim\mathcal{G} \leq p - (p - K) = K$, and the third part of the theorem follows. □

Appendix B. Supplemental appendix

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.09.010.

References

Agarwal, Sumit, Chomsisengphet, Souphala, Mahoney, Neale, Stroebel, Johannes, 2017. Do banks pass through credit expansions to consumers who want to borrow? Q. J. Econ. 133 (1), 129–190.

Angrist, J.D., 2004. Treatment effect heterogeneity in theory and practice. Econom. J. 114, C52–C83.

Angrist, J.D., Lavy, V., 1999. Using maimonides' rule to estimate the effect of class size on scholastic achievement. Q. J. Econ. 114 (2), 533–575.

Angrist, Joshua D., Pischke, Jörn-Steffen, 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Angrist, J.D., Rokkanen, Miikka, 2015. Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. J. Amer. Statist. Assoc. 110 (512), 1331–1344.

Bajari, Patrick, Hong, Han, Park, Minjung, Town, Robert, 2017. Estimating price sensitivity of economic agents using discontinuity in nonlinear contracts. Quant. Econ. 8 (2), 397–433.

Bertanha, Marinho, Imbens, Guido, 2019. External validity in fuzzy regression discontinuity designs. J. Bus. Econ. Stat. forthcoming.

Bertanha, Marinho, Moreira, Marcelo J., 2019. Impossible inference in econometrics: Theory and applications. J. Econometrics forthcoming.

Black, S.E., 1999. Do better schools matter? Parental valuation of elementary education. Q. J. Econ. 114 (2), 577–599.

Black, Dan A., Galdo, Jose, Smith, Jeffrey A., 2007. Evaluating the worker profiling and reemployment services system using a regression discontinuity approach. Amer. Econ. Rev. 97 (2), 104–107.

Calonico, Sebastian, Cattaneo, Matias D., Farrell, Max H., 2018. Optimal bandwidth choice for robust bias corrected inference in regression discontinuity designs. arXiv preprint arXiv:1809.00236.

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

Cattaneo, Matias D., Titiunik, Rocio, Vazquez-Bare, Gonzalo, Keele, Luke, 2016. Interpreting regression discontinuity designs with multiple cutoffs. J. Politics 78 (4), 1229–1248.

Chen, Susan, Van der Klaauw, Wilbert, 2008. The work disincentive effects of the disability insurance program in the 1990s. J. Econometrics 142 (2), 757–784.

Cheng, Ming-Yen, Fan, Jianqing, Marron, James S., 1997. On automatic boundary corrections. Ann. Statist. 25 (4), 1691–1708.

De Giorgi, Giacomo, Drenik, Andres, Seira, Enrique, 2017. Sequential Banking: Direct and Externality Effects on Delinquency. CEPR Discussion Paper No. DP12280.

De La Mata, Dolores, 2012. The effect of medicaid eligibility on coverage, utilization, and children's health. Health Econ. 21 (9), 1061–1079.

Dobkin, Carlos, Ferreira, Fernando, 2010. Do school entry laws affect educational attainment and labor market outcomes? Econ. Educ. Rev. 29 (1), 40–54.

Dong, Yingying, 2018a. Alternative assumptions to identify LATE in fuzzy regression discontinuity designs. Oxford Bull. Econ. Stat. 80 (5), 1020–1027.

Dong, Yingying, 2018b. Jump or Kink? Regression Probability Jump and Kink Design for Treatment Effect Evaluation. Working Paper, University of California, Irvine.

Dong, Yingying, Lewbel, Arthur, 2015. Identifying the effect of changing the policy threshold in regression discontinuity models. Rev. Econ. Stat. 97 (5), 1081–1092.

Duflo, E., Dupas, P., Kremer, M., 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. Amer. Econ. Rev. 101 (5), 1739–1774.

Egger, Peter, Koethenbuerger, Marko, 2010. Government spending and legislative organization: Quasi-experimental evidence from Germany. Amer. Econ. J.: Appl. Econ. 2 (4), 200–212.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. In: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.

Frandsen, R., Frölich, M., Melly, B., 2012. Quantile treatment effects in the regression discontinuity design. J. Econometrics 168 (2), 382–395.

Garibaldi, P., Giavazzi, F., Ichino, A., Rettore, E., 2012. College cost and time to obtain a degree: Evidence from tuition discontinuities. Rev. Econ. Stat. 94 (3), 699–711.

Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica 69 (1), 201–209.

Hastings, Justine S., Neilson, Christopher A., Zimmerman, Seth D., 2013. Are Some Degrees Worth More Than Others? Evidence from College Admission Cutoffs in Chile. NBER Working Paper 19241.

Hoekstra, Mark, 2009. The effect of attending the flagship state university on earnings: a discontinuity-based approach. Rev. Econ. Stat. 91 (4), 717–724.

Hoxby, C.M., 2000. The effects of class size on student achievement: New evidence from population variation. Q. J. Econ. 115 (4), 1239–1285.

Imbens, Guido, Kalyanaraman, Karthik, 2012. Optimal bandwidth choice for the regression discontinuity estimator. Rev. Econom. Stud. 79 (3), 933–959.

Imbens, Guido, Lemieux, Thomas, 2008. Regression discontinuity designs: a guide to practice. J. Econometrics 142 (2), 615–635.

Imbens, Guido, Rubin, Donald B., 1997. Estimating outcome distributions for compliers in instrumental variables models. Rev. Econom. Stud. 64 (4), 555–574.

Van der Klaauw, Wilbert, 2002. Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. Internat. Econom. Rev. 43 (4), 1249–1287.

Lazear, E., 2001. Educational production. Q. J. Econ. 116 (3), 777–803.

McCrary, J., 2008. Manipulation of the running variable in the regression discontinuity design: a density test. J. Econometrics 142 (2), 698–714.

McCrary, Justin, Royer, Heather, 2011. The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. Amer. Econ. Rev. 101 (1), 158–195.

Newey, Whitney K., 1994. Kernel estimation of partial means and a general variance estimator. Econometric Theory 10 (02), 1–21.

Pop-Eleches, C., Urquiola, M., 2013. Going to a better school: Effects and behavioral responses. Amer. Econ. Rev. 103 (4), 1289–1324.

Porter, J., 2003. Estimation in the Regression Discontinuity Model. University of Wisconsin, Madison, Unpublished Manuscript.

Rokkanen, Miikka, 2015. Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design. Columbia University, Unpublished Manuscript.

Rothe, Christoph, 2012. Partial distributional policy effects. Econometrica 80 (5), 2269–2301.

Sun, Yixiao, 2005. Adaptive Estimation of the Regression Discontinuity Model. Working Paper Available at SSRN: 739151.

Tsybakov, Alexandre B., 2009. Introduction to Nonparametric Estimation: Translated from French by Vladimir Zaiats. Springer Series in Statistics, New York.