Donuts and Distant LATEs: Derivative Bounds for RD Extrapolation

Find the Most Recent Version Here

Connor Dowd[*][†]

July 29, 2020

Abstract

Regression Discontinuity (RD) uses policy thresholds to identify causal treatment effects at the threshold. In most settings, the Local Average Treatment Effect (LATE) at the threshold is not the parameter of interest. I provide high level smoothness conditions under which extrapolation across the threshold is possible. Under these restrictions, both estimation and inference for the LATE in other locations is possible. In some situations, extrapolation may be necessary to merely estimate the LATE at the threshold. RD donuts are one such situation, and I provide results allowing estimation and inference in that setting as well.

[*] PhD Candidate at the University of Chicago's Booth School of Business – cdowd@chicagobooth.edu

[†] I would like to thank Max Farrell, Chris Hansen, Constantine Yannelis, Panos Toulis, Jianfei Cao, Sam Hirshman, Zoë Guengerich, attendees of the Econ Metrics Lunch, and the Tuesday Working Group for valuable assistance and feedback.

1 Intro

We study extrapolation of treatment effects in regression discontinuity designs. The standard sharp regression discontinuity (RD) design can nonparametrically identify treatment effects for units local to the treatment threshold. However, policy makers may care about treatment effects for units away from the policy threshold. Moreover, there are applications, like ”donut” designs, which drop all units local to the threshold, where nonparametric identification is not possible even at the threshold. This paper leverages global smoothness conditions to provide identification and inferential guarantees in these situations.

The distinguishing feature of RD designs is that weak smoothness conditions identify the treatment effect at the cutoff. Under the standard sharp RD design, our focus in this paper, the researcher observes an outcome of interest, Y, and a running variable, X. Units for which X exceeds some known threshold, c, receive treatment, and those with $X < c$ do not recieve treatment. If the conditional expectations of Y given X are continuous on both sides of the threshold, the average effect of the policy on the outcome, for units at the threshold, may be causally identified.

Identifying causal effects at other values of X requires additional assumptions. Our approach is to make the weakest possible assumptions consistent with performing estimation and inference – which will lead to partial identification of the parameter of interest. By using bounds on a given higher-order derivative of the conditional means, which are known a priori or estimated, we can bound the error terms in a Taylor expansion, allowing us to construct identified sets. Figure 1 shows an example of how bounding the second derivative leads to a range of possible values for the $E[Y|X, T = 0]$, and a corresponding range of possible treatment effects. In order to produce point estimates for the treatment effects, we would need to make much stronger assumptions. For instance, we could assume that the second derivative is 0 between the threshold and our point of interest – which would correspond to making a linear projection from the threshold to any point we

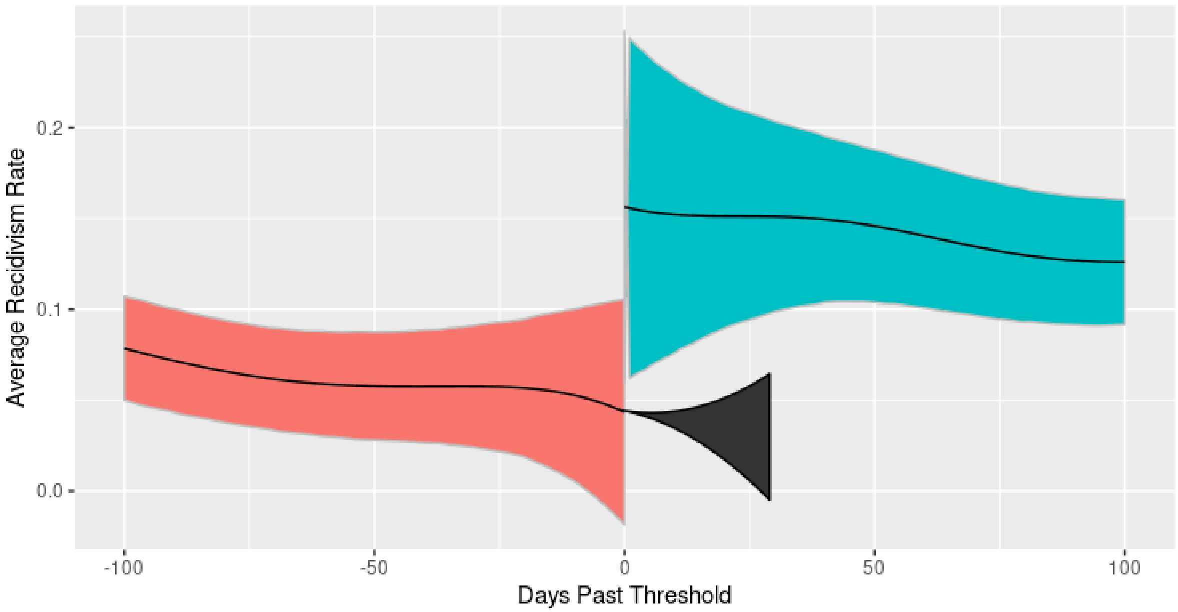

Figure 1: *Recidivism Rate based on time of first offense relative to threshold. Local polynomials are used to estimate the confidence regions and point estimates. The dark region contains $\mu_0(x) = E[Y|X, T = 0]$ when that function has second derivatives between -0.0004 and 0.0006, which are numbers estimated from the data. Building a confidence region on top of that requires accommodating substantial uncertainty in both those estimates and the lower derivatives of the function.*

are interested in. This paper then shows how to construct those identified sets, as well as how to conduct inference on the sets containing the parameter of interest. We also show how the identified set responds to the strength of the assumptions the researcher imposes.

To build an identified set, we use regions of the support where we can obtain point identification of the conditional mean functions and their derivatives, as seen in figure 1. We extrapolate from those points to the desired parameter by bounding the derivative terms in a Taylor expansion of $E[Y|X, T = 0]$. We consider several approaches for bounding those derivatives, which are derivatives of the conditional expectation functions. We allow for bounding an arbitrary derivative k, and show

how any given bound produces an identified set. The simplest bounds involve prior information available to the researcher – for instance about the convexity or concavity of the underlying function, or known limits on the curvature. We also consider estimating bounds for the derivatives under a global smoothness condition. Leveraging recent results of Cattaneo, Farrell, and Feng [2018], we can use the data to estimate bounds on this derivative using data away from the cutoff, allowing us to identify the treatment effects in a fully data-driven manner. The general procedure is described below:

Outline of Estimation Procedure

1. Set a confidence level α and a point of interest, x_0
2. Find a set $\mathbb{C}$ that contains the kth derivative of μ_1 with probability at least $1-\kappa > 1-\alpha$
3. Estimate the first $k-1$ derivatives for μ_1 at the threshold c.
4. Estimate the value of μ_0 at the point of interest, x_0.
5. Estimate μ_1 at x_0, using its first $k-1$ derivatives and a Taylor projection.
6. Estimate $\tau(x_0) = \mu_1(x_0) - \mu_0(x_0)$ and build a $1-\alpha+\kappa$ CI for this projection.

   This is a valid CI for the treatment effect if $\mu_1^{(k)}(x) = 0 \ \ \forall x \in (x_0, c)$.
7. Use the extreme values of $\mathbb{C}$ to find the maximal errors in the Taylor projection above.
8. Add the maximal errors to the $1-\alpha+\kappa$ CI for τ.

   This new range is a conservative $1-\alpha$ CI for a region containing the treatment effect.

If we had precise estimates for the entire infinite sequence of derivatives,[1] we could make precise projections of the mean functions across the threshold to any point, identifying the treatment effect at any point. In the absence of that detailed information, we can make projections based on $k-1$ derivatives, but these projections will have large errors. By bounding the value of the kth derivative, we can bound the size of those projection errors by assuming that the kth derivative is immediately

[1] Along with a few technical conditions like the function being analytic.

and forever at its bounds. Any point estimate outside the bounds created by that assumption would require that the kth derivative be outside its bounds at least briefly. This turns a problem of projection into a problem of bounding the derivatives of a mean function – which is also known as a smoothness condition. This paper will require that there are some bounds on a derivative, which are either known, or estimable somehow. Our primary technique will be to assume that we observe a maximal value for the kth derivative of the mean function, somewhere in the range of the data.

Because this approach relies solely on information which is identified in the support of the data, to obtain information outside the support of the data, we can generalize the approach. This allows us to study the ”donut” design. Donut designs are a modification of sharp RD which attempt to deal with selection issues by dropping observations in some radius of the threshold. The radius around the threshold is dictated by the ability of units to manipulate their observed X value, and thus decide on their treatment status. In an extension of the approach described above, we use the extrapolation techniques to identify the treatment effect at the threshold, after dropping all observations nearby – as is done in donut designs.

To fix ideas, we will revisit the recent application of Tuttle [2019]. Here the running variable is the days after new SNAP policy is implemented, while the outcome of interest is the recidivism behavior of convicted drug dealers. Broadly, convictions for drug dealing after day 0[2] lead to a lifetime ban for SNAP benefits in Florida. We are interested in the effect of the lifetime SNAP ban on recidivism for individuals who are caught after the threshold. The original paper estimates the effect for individuals on day 0, but policy makers may care substantially about the effect people beyond the threshold. In Figure 1, we see the mean functions before and after treatment. That same figure also shows an example taylor projection of one mean function for the control group using a second derivative bound. The area encompassed by the dark region is the identified set – not the confidence interval. The difference between that set and the point estimate for the treated

[2] The drug dealing must have been after day 0, not the conviction.

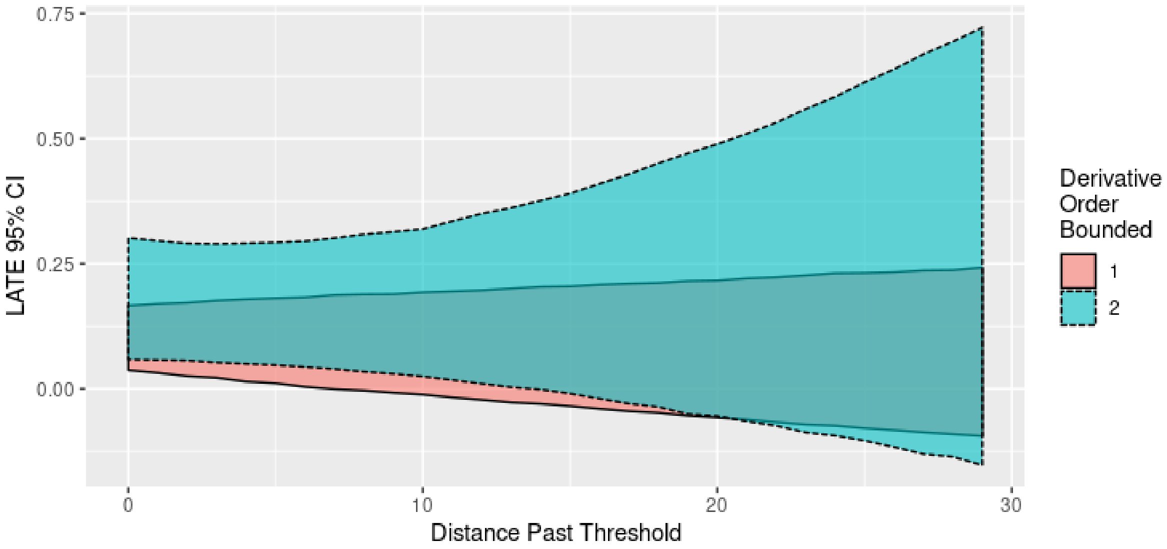

Figure 2: *Local Average Treatment Effects of lifetime SNAP bans on Recidivism as we move away from the treatment threshold. Shown here are two different assumptions about structure of underlying mean functions, either that the first or second derivatives are globally bounded.*

group above it gives our estimate of the LATE for these regions beyond the threshold.

See Figure 2 to see the estimated 95% confidence intervals arising from the assumption of global bounds on either the first or second derivatives. This is the result of actually calculating the difference between the two sets and incorporating our uncertainty around them. Note that neither interval is a subset of the other – a clear demonstration that neither assumption is strictly weaker or stronger than the other. Depending on the DGP, and point of interest, either set can encompass the other.

There is some prior work extrapolating RD treatment effects away from the threshold, differing substantially in the type and strength of assumptions used for idenfication. Angrist and Rokkanen [2012] assume that the researcher has access to additional variables, and that potential outcomes are mean independent of the running variable, given these additional covariates. Modelling the conditional expectation of outcomes given those covariates, they are able to identify causal effects for other values of the running variable. Our assumptions are much weaker, and we thus obtain weaker identification results, while encompassing more general situations. Dong and Lewbel [2015] use estimated derivatives, as we do, to consider causal effects of small changes to the position of the threshold, which is not our focus.

We work within the standard RD framework which uses nonparametric smoothness assumptions to identify the causal effects of interest [Hahn, Todd, and van der Klaauw, 2001]. The RD literature is large and still expanding: for a recent review, and numerous other references, see Cattaneo, Idrobo, and Titiunik [2019a]. For a comparison of the continuity-based and other approaches to RD identification and estimation, see Cattaneo, Titiunik, and Vazquez-Bare [2017].

This paper is organized as follows: Section 2 provides notation and necessary conditions. Section 3 details the main results. Section 4 looks at an example in the world of recidivism and SNAP benefits to examine the relative strengths of assumptions bounding different derivative orders. Section 5 discusses a natural extension to the world of RD donuts. Section 6 concludes.

2 Framework

For each individual, the econometrician observes a variable X, known as the running variable or forcing variable, which has a compact domain $\chi \subset \mathbb{R}$. There is also a known threshold, $c \in \chi$, such that treatment status $T = \mathbb{1}\left[x \geq c\right]$. Without loss of generality, we assume that $c = 0$. We also observe another variable, Y, referred to as the outcome. We will focus on the standard heteroscedastic non-parametric framework.

$$Y = \mu_T(X) + \epsilon \qquad \mathbb{E}\left[\epsilon\right] = 0 \qquad Var(\epsilon) = \sigma_T^2(X) \tag{1}$$

Where μ_t and σ_t are defined as:

$$\mu_t(x) = \mathbb{E}\left[Y|X = x, T = t\right] \qquad \sigma_t^2(x) = Var(Y|X = x, T = t)$$

The parameter of interest is the local average treatment effect (LATE) on the outcome variable at some known point x_0. Thus, our parameter of interest is:

$$\tau = \tau(x_0) = \mu_1(x_0) - \mu_0(x_0) = \mathbb{E}\left[Y|X = x_0, T = 1\right] - \mathbb{E}\left[Y|X = x_0, T = 0\right] \tag{2}$$

Most RD papers require that the point at which we evaluate the LATE, x_0, coincide with the treatment threshold, c. The restriction that $x_0 = c$ simplifies the problem faced by the econometrician substantially, as in principle, $\mu_1(c)$ and $\mu_0(c)$ are both nonparametrically identified by the data. Notably, it is rare that the point of interest for policymakers is actually c.

For notational simplicity, I will assume throughout that the point of interest, $x_0 < c$. This will allow me to denote the mean function being projected as μ_1. This condition is by no means necessary, and will be relaxed in the section on donuts.

In order to learn about the LATE at points less than the threshold, we need to predict Y in the counterfactual where those individuals were treated. Specifically, we need the ability to conduct inference for $\widehat{\mu_1(x_0)}$, which under standard RD assumptions is not possible. Then we can use our information about $\widehat{\mu_1(x_0)}$ to estimate the treatment effect using $\hat{\mu}_0(x_0)$. Learning $\widehat{\mu_1(x_0)}$ is not simple, so this paper finds weak assumptions that will bound the possible values of $\widehat{\mu_1(x_0)}$.

A simple fix to this would be to assume that the function μ_1 is an order k polynomial or some other known parametric function. Under that assumption, the projection becomes quite simple. We can use standard regression confidence intervals, projected over to the point x_0. However, this is not typically a reasonable assumption. Over the past 15 years a large literature developed examining the behavior of nonparametric estimators for the LATE in RD settings. This literature is the direct result of overly strong parametric RD estimates which frequently relied on polynomial regressions. Non-parametric RD does not make strong enough assumptions to identify $\widehat{\mu_1(x_0)}$ when $x_0 \neq c$, but that literature frequently makes strong assumptions in order to select the optimal bandwidth. In nonparametric estimators like local polynomial regressions, bandwidths balance a trade-off between the curvature of the underlying mean functions and the variance of the errors. As the curvature increases and the mean function becomes less smooth, relatively distant observations become less informative, and so dropping them by shrinking the bandwidth is sensible. At the same time, as the error variance increases, nearby observations become noisier and less informative, and so increasing our effective sample size by expanding the bandwidth becomes attractive.

Data-driven bandwidth choice requires making some assumption about the maximal extent of the curvature and the maximum variance, thus bounding the worst case outcome. Those conditions have a tendency to look like placing an upper bound on the kth derivative of the mean function. This paper will strengthen and extend those conditions. Broadly, I will require an assumption of the form:

Condition 1 (Bounded kth Derivative)**.** For each $t = \{0, 1\}$, for some $k > 1$,

$$\partial^{(k)}_{L,t} \;\leq\; \mu^{(k)}_t(x) \;\leq\; \partial^{(k)}_{U,t} \quad \forall X \in \chi \tag{3}$$

where $\mu^{(k)}_t$ indicates the kth derivative of the function μ_t, χ continues to represent the domain of the running variable, and $\partial^{(k)}_{L,t}$ & $\partial^{(k)}_{U,t}$ represent upper and lower bounds. In some settings, researchers may be able to use domain knowledge to state reasonable bounds $\partial_{L,t}$. Results in this paper will show the validity of that approach. Perhaps more frequently, researchers can make the additional assumption that $\mu^{(k)}_t(x)$ attains its extreme values in the region where we observe $T = t$. Under that condition, this paper will show results for data driven methods to estimate the treatment effect.

Bounds of the form in equation 3 are strictly weaker claims than requiring that μ_t be a polynomial of degree k, as that would imply that for some constant C, $\mu_t^{(k)}(x) = C \quad \forall x \in \chi$.

In order to make use of the bounds above, we need to recall the Taylor projection for a function.

$$\mathcal{P}_\infty(\mu_t, x) = \sum_{j=0}^{\infty} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!}$$

Notably, once we have finite bounds on the kth derivative and we know $c = 0$ we can simplify this somewhat to the Taylor projection below:

$$\mathcal{P}(\mu_t, x_0, \partial_{L,t}^{(k)}, \partial_{U,t}^{(k)}) = \sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{x_0^j}{j!} + \begin{pmatrix} \partial_{U,t}^{(k)} \\ \partial_{L,t}^{(k)} \end{pmatrix} \frac{x_0^k}{k!} = \begin{pmatrix} \mu_t(x_0)_U \\ \mu_t(x_0)_L \end{pmatrix} = \Phi_t \tag{4}$$

The vector created by that projection defines an interval, which Lemma 1 shows will contain the true $\mu_t(x_0)$. That interval is the identified set, which I will refer to as Φ. In order to make the projection feasible however, we will have to estimate or know the first $k-1$ derivatives, as well as the two bounds on the kth derivative. The rest of this section will examine the conditions under which we show that estimating those derivatives to build an interval works.

Condition 2 (Regularity Conditions)**.** Technical conditions on the DGP in order for the results below.

(i) (X,Y,T) are i.i.d. observations from a d.g.p. satisfying Eq (1)

(ii) $\mu_t(\cdot)$ has $k+2$ continuous derivatives

(iii) The density of the running variable, f_x is absolutely continuous and bounded away from 0 over χ.

(iv) The kernel function $K(x) = 0.5\mathbb{1}\left[|x| < 1\right]$.

(v) $\sigma_t(\cdot)$ is positive, bounded above, bounded away from 0, and has two continuous derivatives.

(vi) $\underset{x \in \chi}{sup}\ \mathbb{E}\left[|\epsilon_i|^3 exp(|\epsilon_i|)|x_i\right] = x < \infty$ which implies $\mathbb{E}\left[|\epsilon_i|^3 exp(|\epsilon_i|)\right] < \infty$.

(vii) There is no other treatment policy with a discontinuity in χ which affects Y.

Condition 1 and Condition 2(i,ii) are sufficient for us to establish that the taylor projection described in equation 4 is valid and will contain the true value of $\mu_1(x_0)$.

Condition 2 parts (iii), (iv), and (v) are closely related to standard conditions for asymptotic normality of local polynomial estimates.[Fan et al., 1995] This makes them sufficient (with some mild rate conditions) for us to be able to take a known set of bounds on the kth derivative from equation 3 and make the projections feasible. At this point we could build a confidence region for $\mu_1(x_0)$ which is asymptotically valid for the true region. At the same time these conditions allow us to perform inference for $\mu_0(x_0)$. As the these two procedures use independent pieces of data, it is simple to construct a valid confidence region for τ.

Condition 3 (Rate Conditions)**.** For local polynomial estimates of derivatives to be asymptotically normal I will require $h_p(n) \to 0$ such that as $n \to \infty$:

(i) $nh_p^3 \to \infty$

(ii) $nh_p^{2k+3} \to 0$

Further, if we would like to estimate a global bound on the derivatives using b-splines I need the following conditions on a potentially different bandwidth, h_b:

(iii) $\frac{log(n)^{3/2}}{\sqrt{nh_b}} = o_{\mathbb{P}}(1/log(n))$

(iv) $\frac{log(n)^4}{nh_b} = o(1/log(n))$

(v) $nh_b^{1+2k} = o(1/log(n))$

The top conditions are sufficient for asymptotic normality of local polynomial regressions.Fan et al. [1995] The bottom half of these rate conditions will be necessary for us to get a valid uniform confidence band on the kth derivative. For that to be useful, we need the following condition to hold.

Condition 4 (Derivative bounds are observed)**.** Recall that $c = 0$ and $T = \mathbb{1}[x \geq 0]$. Define C_1,...,C_4 as follows

$$\underset{x>0\in\chi}{sup} \frac{\partial^k \mu_1(x)}{\partial x^k} = C_1 \qquad \underset{x>0\in\chi}{inf} \frac{\partial^k \mu_1(x)}{\partial x^k} = C_2$$
$$\underset{x<0\in\chi}{sup} \frac{\partial^k \mu_0(x)}{\partial x^k} = C_3 \qquad \underset{x<0\in\chi}{inf} \frac{\partial^k \mu_0(x)}{\partial x^k} = C_4$$

Then, for known, continuous, weakly monotonic functions $f_1, ..., f_4$

$$\partial_{L,1}^{(k)} = f_1(C_1, C_2) \qquad \partial_{U,1}^{(k)} = f_2(C_1, C_2)$$
$$\partial_{L,0}^{(k)} = f_3(C_3, C_4) \qquad \partial_{U,0}^{(k)} = f_4(C_3, C_4)$$

Broadly, condition 4 says that we observe values which are known functions of the bounds in condition (1). In practice we will usually take these functions to be the identities. The generality allows for the inf and sup to be absolute values of the biggest observed derivative, as well as allowing for other situations – e.g. we know some maximal bound, but may wish to use the data-driven results below to tighten the bounds if possible.

3 Main Results

3.1 Φ contains $\mu_t(x_0)$

To prove that the set Φ contains the true value of the mean function, we will show that the approximation error of a Taylor projection using $k-1$ derivatives, when the kth derivative is bounded, are at most the projection of the bounds of the kth derivative.

Lemma 1 (Φ contains the true value of $\mu_t(x_0)$)**.** *The projection error of a $k-1$ derivative Taylor projection is bounded by the Taylor projection of the kth derivative's bounds.*

Proof. $\mathcal{P}_\infty(\mu_t, x) - \mathcal{P}_{k-1}(\mu_t, x) = \sum_{j=0}^{\infty} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!} - \sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!} = \sum_{j=k}^{\infty} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!}$
If $\mu_t^{(k)}(x) \leq \partial_{U,t}^{(k)} \forall X \in \chi$, then $\sum_{j=k}^{\infty} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!} \leq \partial_{U,t}^{(k)} \frac{(x-c)^k}{k!}$.
If $\mu_t^{(k)}(x) \geq \partial_{L,t}^{(k)} \forall X \in \chi$, then $\sum_{j=k}^{\infty} \frac{\partial^j \mu_t}{\partial x^j} \frac{(x-c)^j}{j!} \geq \partial_{L,t}^{(k)} \frac{(x-c)^k}{k!}$.
Thus $\partial_{L,t}^{(k)} \frac{(x-c)^k}{k!} \leq \mathcal{P}_\infty(\mu_t, x) - \mathcal{P}_{k-1}(\mu_t, x) \leq \partial_{U,t}^{(k)} \frac{(x-c)^k}{k!}$ □

3.2 Results for Φ

In order to actually estimate the values $C_1, ..., C_4$, much less perform inference on functions of them, we will need to rely on the results in Cattaneo et al. [2018]. If we use b-splines with equally sized partitions to estimate the kth derivative, that paper tells us that we can construct uniform confidence intervals for that derivative. Specifically we can find a $q(\alpha)$ such that we can build asymptotically valid uniform $(1-\alpha)$ CIs which are:

$$\left[\hat{\mu_t}^{(k)}(x) \pm q(\alpha)\sqrt{\hat{\Omega}_t(x)/n} : \ x \in \chi\right] \tag{5}$$

This implies that I can make statements like:

$$lim\ \mathbb{P}\left[\underset{x\in\chi}{sup}\ \mu_t(x) \geq C\right] \leq \alpha/2 \tag{6}$$

Where $C = \underset{x \in \mathcal{X}}{max} \left[\hat{\mu}_j(x) + q_j(\alpha)\sqrt{\hat{\Omega}_j(x)/n} \right]$, i.e. C is the upper bound.

The reverse is also true, and so we can make statements about the sup and inf of the kth derivative over compact domains.

These statements are extremely conservative. This is a function of the confidence band construction which relies on fixed critical values to obtain uniformity. For the purposes of inference on extrema, this means that the bounds obtained will not achieve nominal size, even in the limit. Nevertheless, obtaining any valid probability statement for the sup of an unobserved function is a difficult problem. Chernozhukov et al. [2013] provide a direct approach to this problem, however, their bounds are conservative in the opposite direction, and so cannot be used in this paper.

With the ability to conduct inference for the extrema of derivatives, we can turn to conducting inference for the identified set. Recall that the set $\Phi_t(x_0)$ is the set identified by the taylor projection which contains the mean function at x_0. Asymptotically, without stronger assumptions on the DGP, it is impossible to identify as smaller set. Therefore, we will attempt to contain that region with given size.

Theorem 1 (Containing $\Phi_t(x_0)$)**.** *Under conditions 1-4, using local polynomials to learn the 0,...,k-1 derivatives at 0 and using b-splines to learn the sup and inf of the kth derivative, we can build a $1-\alpha$ confidence region CR_g such that:*

$$lim \mathbb{P}\left[\Phi_t(x_0) \subset CR_g\right] \geq 1 - \alpha$$

The result in theorem 1 builds somewhat naturally on well known results about local polynomials, proofs are in the supplemental appendix. The projection $\mathscr{P}$ is linear in the estimated derivatives and extrema, which makes for easy projections once we can make statements like the one in 6. Combining the results of the extrema estimation routine and the local polynomial is more difficult.

For now, this paper relies on a union bound.

Namely, given two statements of the form $\mathbb{P}\left[X_i > q_i\right] \leq \alpha/2$, we can also state that

$$\mathbb{P}\left[X_1 + X_2 > q_1 + q_2\right] \leq \alpha$$

As each estimation routine can return a straightforward confidence region, we can combine those regions upper and lower bounds as above.

3.3 Inference for τ in standard RD

The focus of this section so far has been inference of the region Φ_t. The results above give a region which asymptotically contains Φ with at least given size. In the same way that union bounds let us move from just the CR for extrema to a region for Φ, we can extend to a region around τ. But first we should discuss the identified set.

Once again, the assumptions above are not adequate to identify a point estimate. Rather, the nature of the derivative bounds in 3 is that they allow us to identify a set which will contain the value of interest. In this case, we can identify the following set:

$$\mathcal{T}(x_0) = \Phi_1(x_0) - \mu_0(x_0) = \begin{pmatrix} \mu_1(x_0)_U - \mu_0(x_0) \\ \mu_1(x_0)_L - \mu_0(x_0) \end{pmatrix} \tag{7}$$

Recall that for simplicity, we are relying on $x_0 < 0 \in \chi$. As a result, we know that we can identify the parameter $\mu_0(x_0)$ using standard results for local polynomials. Theorem 1 gives us a region containing Φ, and so we can combine the two for $\hat{\mathcal{T}}$.

Applying the union bound again leads to the following lemma regarding the estimand of interest, τ.

Lemma 2 (CR for τ)**.** *Under all the conditions of theorem 1, we can build a* $1-\alpha$ *confidence region* CR_τ *such that:*

$$lim\mathbb{P}\left[\mathcal{T}(x_0) \subset CR_\tau\right] \geq 1-\alpha$$

This result follows naturally from theorem 1, but in many ways this is the real meat of the paper. Given a point, we can take an RD design and some higher order derivative bounds, and with them we can partially identify treatment effects at that point – even when it doesn't overlap with the threshold. In section 5 the paper I will discuss applications of this idea to the closely related setup that is an RD donut design.

4 Example: Snap benefits and Recidivism

This example is from the paper by Tuttle [2019]. That paper looks at recidivism as affected by a food assistance program. The treatment effect is identified by leveraging a discontinuity in policy which imposed a lifetime ban on SNAP benefits for individuals who engage in drug trafficking after August 23rd, 1996. The high level finding of that paper is that individuals who received a lifetime ban were about 10% more likely to commit more crimes in the future, with the effects predictably concentrated among crimes with financial benefits.

This is an excellent paper. I merely use the setting to demonstrate the extrapolations discussed here, and certainly not because of concerns about the results. A common question around these extrapolations is what derivative order makes the most sense. The notion that higher order derivative bounds are weaker conditions seems quite intuitive to many people. One critical takeaway from this example is those comparisons are not as straightforward as they may seem. Broadly speaking, as we change the derivative order being bounded, the other assumptions we make are also changing, which may make the overall procedure more conservative or not. Moreover, the location of the LATE to be estimated also can affect the relative strength of these assumptions.

To see this, recall that a second derivative bound will grow at $O(x^2)$, while a third derivative bound will grow at $O(x^3)$. For x near the threshold, the third derivative may well be a stronger assumption, while far away, the second derivative can be more restrictive. I will compare the use of several different derivative bounds for extrapolating the treatment effect.

In context, the assumption of bounds on a derivative corresponds to a bound on the changes in probability of recidivism for treatment and control groups. For a second derivative bound, this suggests that the acceleration of the control group's recidivism is restricted. Perhaps more importantly, the assumption that we observe the extrema of the derivative implies that there are not other structural changes on August 23rd which would cause the control group function to change drastically.

Figure 2 shows the CI for a set containing the LATE across a number of different derivative restrictions. As we can see, the first and second derivative bounds each are fairly comparable for the LATE at the threshold. Nevertheless, the second derivative bound starts substantially wider than the first derivative bound. As time goes on, the second derivative bound also grows faster, eventually containing the entire first derivative set.

The difference in starting positions and variances comes down to the additional information and variance associated with estimating more parameters in the local polynomial regression at the threshold. The more rapid growth is the natural result of allowing the first derivative to grow without limit.

The important point here is that these are different assumptions. One is not necessarily weaker or stronger, but rather different.

5 Special Case: Donut designs

A special case of extrapolation in RD settings is that of a Donut. Donut designs are used when we have fairly standard RD settings – that is some sort of policy threshold – but we are worried that individuals have control over where they fall relative to the threshold. If individuals can shift their x position by a bounded amount, then there may be selection across the threshold. Some individuals may choose to cross it, while others do not. In order to retrieve the LATE for individuals who were exogenously at the threshold, we need to eliminate the selection effect. Donut designs do this by dropping all individuals within some distance d of the threshold. In essence, this corresponds to saying that we do not trust those observations.

However, having thrown out the observations near the threshold, we have gotten rid of the very observations that identify the LATE at the threshold under standard assumptions. Currently donut designs deal with this by implicitly projecting polynomials across the region of the donut.[3] This paper presents an alternative – estimate the extrema of some derivative k, and use that to project an identified set across the region of the donut.

In order for this to be useful, we need to make an additional assumption.

Condition 5 (Donuts)**.**

(i) *Donut Exclusion.* There is a known region, $\mathbb{D} = (d_-, d_+)$, with $d_- < d_+$, hereafter referred to as the donut, such that all manipulation ($\mathcal{M} = 0$ in the absence of manipulation) is contained to the donut.

$$\forall i \;\; s.t. \;\; \mathcal{M}_i \neq 0, \;\; x_i', x_i \in \mathbb{D}$$

(ii) *Unique Threshold.* There is one, and only one, policy relevant to the outcome of interest, which has a threshold inside the region defined as the donut and its boundaries, $[d_-, d_+]$.

[3] This is what dropping those observations and re-running your local polynomial RD estimator does.

Condition 5(i) ensures that all manipulation is contained to the interior of the donut. Nobody from outside that region was induced to change their behavior by the presence of the policy. This is critical – without this assumption, we retain the selection problems which we had before we decided to use a donut.

Condition 5(ii) replaces condition 2(iii) in the donut setting. This assumption looks much more like a natural extension of the standard RD assumption that there is no other co-located policy threshold.

Lemma 3 (Donut)**.** *Under Conditions 1-5, we can find a $1 - \alpha$ confidence region CR_d such that*

$$lim \mathbb{P}\left[\mathcal{T}(0) \subset CR_d\right] \geq 1 - \alpha$$

This is the result that we need in order to perform inference for the LATE at the threshold under a donut design. In the absence of this, or some other extrapolation result, the LATE is not asymptotically identified in donut designs. The problem is that in most situations, the ability to manipulate the running variable is unrelated to sample size. Thus asymptotics based on observing data arbitrarily close to the threshold don't work without these extrapolation results.

5.1 An Example Donut

In their paper, Lindo, Sanders, and Oreopoulos [2010] assess the effect of academic probation using a treatment threshold at a GPA of 1.5. They look at a variety of outcomes split on multiple dimensions, but the focus is about whether students' GPAs rise in subsequent semesters. They acknowledge the risk of manipulation – specifically that students may be "convincing teachers to give them a higher grade". After testing for a discontinuity and finding nothing, as well as checking that a number of covariates are smooth across the threshold, the authors move on.[4]

[4] Thanks to Cattaneo et al. [2019a] the data and code needed to replicate these results are widely available.

Assuming that students are able to convince 1/3 professors to raise their grade one partial letter (a max of 0.6 on the GPA scale), that implies that students have the ability to move up to 0.2 units of GPA in a semester (and year). This creates the circumstances in which a donut is reasonable.

In order to make progress, we will rely on the assumptions in section 2.[5] We set the donut $\mathbb{D} = (-0.25, 0.25)$. We will use k=2 – thus it is the second derivative which is bounded. A full plot of the outcome variable against the RV shows that μ_0 and μ_1 may be exactly linear, so assumptions about the second derivative are reasonable. We use the bandwidth the authors selected of 0.6. Together, these assumptions give us the results seen in table 1.

In the original paper, the authors find a treatment effect of 0.233 GPA (95% CI: [0.18,0.285]) points gained by a person on the threshold. Breakdowns across subgroups give results of a similar magnitude. Those results are consistent with the outcomes from a donut. See table **??** for a comparison of the original papers results as replicated using rdrobust, and the outcomes of a donut estimator.

	Estimate	CI Lower	CI Upper
Bias-Corrected	0.213	0.136	0.291
Robust	0.213	0.122	0.304
Derivative Bounds: $\hat{\tau}$	$[0.275, 0.407]$	0.034	0.727

Table 1: Comparison of Estimates from rdrobust and Donut routines

Overall, this example lets us conclude that the treatment effect of academic probation is inside the

[5] Continuity of the density of the RV is somewhat questionable, the RV takes 160 unique values in the region containing the donut and bandwidths, while there are 16,000 observations in that region.

region $[0.07, 0.68]$ with confidence. This is consistent with the results in both the original paper and the replication by Cattaneo et al. [2019a]. The treatment effect on future GPA is not the only outcome of probation which should be considered for policymakes, but if there was no effect, the justification for such a policy would be thin.

6 Conclusion

This paper provides a simple approach to extending the LATE regime of regression discontinuity away from the threshold. I provide asymptotic size control for the partially identified set. An application of this work to the world of RD donuts was discussed. I hope to demonstrate the utility of this work in the future by looking at other example settings, as well as looking into the possibility of estimating the ATE using this design.

References

Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. Estimating marginal returns to medical care: Evidence from at-risk newborns. *The quarterly journal of economics*, 2010.

Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. The role of hospital heterogeneity in measuring marginal returns to medical care: a reply to barreca, guldi, lindo, and waddell. *The quarterly journal of economics*, 2011.

Michael L Anderson. As the wind blows : The effects of long-term exposure to air pollution on mortality *, 2015.

Joshua Angrist and Miikka Rokkanen. Wanna get away? rd identification away from the cutoff, 2012.

Joshua D Angrist, Victor Lavy, Jetson Leder-Luis, and Adi Shany. Maimonides rule redux. *NBER Working Paper Series*, 2017.

Alan I. Barreca, Melanie Guldi, Jason M. Lindo, and Glenn R. Waddell. Saving babies? revisiting the effect of very low birth weight classification. *The quarterly journal of economics*, 2011.

Ylenia Brilli and Brandon J Restrepo. Birth weight , neonatal intensive care units , and infant mortality : Evidence from macrosomic babies, 2017.

Sebastian Calonico, Matias D. Cattaneo, and Roćıo Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 2014.

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

Cattaneo, Idrobo, and Titiunik. A practical introduction to regression discontinuity designs: Foundations, 2019a.

Matias Cattaneo, Brigham Frandsen, and Rocío Titiunik. Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 2014.

Matias D. Cattaneo, Rocío Titiunik, and Gonzalo Vazquez-Bare. Comparing inference approaches for rd designs: A reexamination of the effect of head start on child mortality. *Journal of Policy Analysis and Management*, 36(3):643–681, 2017. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pam.21985.

Matias D. Cattaneo, Max H. Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators, 2018.

Matias D. Cattaneo, Luke Keele, Rocio Titiunik, and Gonzalo Vazquez-Bare. Extrapolating treatment effects in multi-cutoff regression discontinuity designs, 2019b.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: Estimation and inference. *Econometrica*, 2013.

Gordon B. Dahl, Katrine Vellesen Løken, and Magne Mogstad. Peer effects in program participation. *AER*, 2014.

N. Meltem Daysal. Spillover effects of early-life medical interventions, 2015.

Thomas S. Dee and Emily K. Penner. The causal effects of cultural relevance : Evidence from an ethnic studies curriculum, 2016.

Juan Manuel Ospina Díaz, Nicolás Grau, Tatiana Reyes, and Jorge Rivera. The impact of grade retention on juvenile crime, 2016.

Yingying Dong and Arthur Lewbel. Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 2015.

Nikolay Doudchenko and Guido W Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research, 2016.

Arindrajit Dube, Laura Giuliano, Jonathan, and Léonard. Fairness and frictions : The impact of unequal raises on quit behavior, 2015.

Jianqing Fan, Nancy Heckman, and M.P. Wand. Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *Journal of the American Statistical Association*, 1995.

Dirk Foremny and Albert Solé-Ollé. Who ' s coming to the rescue ? revenue-sharing slumps and implicit bailouts during the great recession, 2016.

Romain Gauriot. Winner effect in dynamic contests, 2014.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. Working Paper 20405, National Bureau of Economic Research, August 2014.

François Gerard, Miikka Rokkanen, and Christoph Rothe. Bounds on treatment effects in regression discontinuity designs with a manipulated running variable, 2019.

Ruud Gerards and Pierre M Theunissen. Becoming a mompreneur: Parental leave policies and mothers' propensity for self- employment, 2018.

Sarena Goodman, Adam Isen, and Constantine Yannelis. A day late and a dollar short: Liquidity and household formation among student borrowers, 2018.

Jinyong Hahn, Petra Todd, and Wilbert H van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 2001.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 2004.

David S. Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 2010.

Jason M. Lindo, Nicholas J. Sanders, and Philip Oreopoulos. Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2010.

Jason M. Lindo, Peter M. Siminski, and Oleg Yerokhin. Breaking the link between legal access to alcohol and motor vehicle accidents: Evidence from new south wales. *Health economics*, 2016.

Benjamin Marx. The cost of requiring charities to report financial information, 2018.

Justin McCrary. Manipulation of the running variable in the regression discontinuity design : A density test. *The Journal of Econometrics*, 2008.

Michael George Mueller-Smith and Kevin T. Schnepel. Diversion in the criminal justice system : Regression discontinuity evidence on court deferrals, 2017.

Ben Ost, Weixiang Pan, and D. B. Webber. The returns to college persistence for marginal students : Regression discontinuity evidence from university dismissal policies, 2016.

Jack Porter. Estimation in the regression discontinuity model. *Monograph*, 2003.

Alexandra Roulet. Unemployment insurance and reservation wages : Evidence from administrative data, 2016.

Carlos Zamarrón Sanz. Direct democracy and government size : Evidence from spain, 2015.

Judith Scott-Clayton and Lauren Schudde. Performance standards in need-based student aid, 2016.

Hitoshi Shigeoka. The effect of patient cost sharing on utilization , health , and risk protection. *AER*, 2014.

Cody Tuttle. Snapping back: Food stamp bans and criminal recidivism. *American Economic Journal: Economic Policy*, 2019.

Seth D. Zimmerman. The returns to four-year college for academically marginal students. *Journal of Labor Economics*, 2014.

A Proofs

A.1 Theorem 1

Condition 2(ii) implies that a taylor projection is a valid technique for approximating the functions μ_t. Condition 1 is somewhat unusual in combination with taylor projections – which usually are infinite sums – but in this case, but putting bounds on the extreme values of the derivative, we can say with certainty that $\mu_t(x_0) \in \Phi_t(x_0)$. The issues here arise from the feasibility of estimating Φ_t, and worse, conducting inference.

Conditions 1-4 are substantially stronger than the needed conditions for asymptotic normality of point estimates and derivatives using local polynomial estimators. They are sufficient for the conditions in Section 5.4 of Fan et al. [1995]. This will allow us to conduct inference on the vector $\theta_t(\cdot) = (\mu_t^{(0)}(\cdot), ..., \mu_t^{(k)}(\cdot))^T$ at x=0 for each of t=0,1. This will also allow inference on the point $\mu_t(x_0)$ for whichever treatment status is observed at x_0. This is critical for Lemma 2.

Conditions 1-4 also imply the necessary conditions for Lemma SA-5.1 and Theorems SA-5.1, SA-5.3, and SA-5.7 in Cattaneo et al. [2018]. Many of the rate restrictions and technical conditions come directly from that paper. That paper provides us with the ability to construct a confidence band that contains the entire function $\mu_t^{(k)}(\cdot)$ with given probability. By finding the extreme values of that band, and using the mappings defined in Condition 4, we can learn about the distribution of $\partial_{U,t}^{(k)}$ and $\partial_{L,t}^{(k)}$.

As the projection $\mathcal{P}$ is linear in the derivatives, we are simply taking the parameters we have now built confidence regions for, scaling them as the projection requires, and adding them. The scaling does not affect our size control.

We have several options to add the parameters together and retain a valid confidence region. If we had a full distribution for the extrema, we could think about the joint distribution and the optimal adding of the two. However, in pushing a supremum through the results in Cattaneo et al. [2018], the outcome statement is substantially conservative, and does not correspond to a proper distribution for the true value. Thus we will use union bounds. This means we can take any α_1 and α_2 such that $\alpha_1 + \alpha_2 = \alpha$, and where we know that $\mathbb{P}\left[\partial_{U,t} > C\right] \leq \alpha_1$ (with C as defined in equation 6) and $\mathbb{P}\left[\sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{x_0^j}{j!} > C_5\right] \leq \alpha_2$, and conclude that $\mathbb{P}\left[\partial_{U,t} \frac{x_0^k}{k!} + \sum_{j=0}^{k-1} \frac{\partial^j \mu_t}{\partial x^j} \frac{x_0^j}{j!} > C_5 + C \frac{x_0^k}{k!}\right] \leq \alpha$. As the first part of that probability defines the upper bound of our identified set Φ_t, and the statement is true for the lower bound as well, we can contain the set Φ_t, with whatever probability given. See Imbens and Manski [2004] for details about construction of such a set.

A.2 Lemma 2

With a set containing $\Phi_1(x_0)$ with some probability, and an asymptotically normal estimate of $\mu_0(x_0)$ from Fan et al. [1995], we can again apply the union bounds to build a set $\mathcal{T}$ which contains the value τ with given probability. Because there is no other policy which affects Y with a threshold in χ, the difference here is the LATE.

This is not the most efficient construction of the LATE however. There are substantial power gains to be had from constructing the LATE equation, which can be decomposed into the projection of the extrema, the projection of a normal, and a normal. By combining the normal distributions then using the needed union bound, we manage to limit the power loss associated with union bounds.

A.3 Lemma 3

Donuts are an interesting application of Theorem 1. In order to use them properly, we need to recenter our projection on the boundaries of the donuts. Condition 5 tells us that the donut has successfully gotten rid of all selection issues. Theorem 1 tells us that projections from those boundaries to the threshold will give us something meaningful. Taking the difference between the two set-identified parameters projected from the edges of the donut uses the same union bound procedure as above.