Optimized Regression Discontinuity Designs

Guido Imbens
imbens@stanford.edu

Stefan Wager
swager@stanford.edu

Draft version May 2017

Abstract

The increasing popularity of regression discontinuity methods for causal inference in observational studies has led to a proliferation of different estimating strategies, most of which involve first fitting non-parametric regression models on both sides of a threshold and then reporting plug-in estimates for the discontinuity parameter. In applications, however, it is often difficult to tune the non-parametric regressions in a way that is well calibrated for this target of inference; for example, the model with the best global in-sample fit may provide poor estimates of the discontinuity parameter. In this paper, we propose an alternative method for estimation and statistical inference in regression discontinuity designs that uses numerical convex optimization to directly obtain the finite-sample-minimax linear estimator for the regression discontinuity parameter, subject to bounds on the second derivative of the conditional response function. Given a bound on the second derivative, our proposed method is fully data-driven and does not rely on further tuning by the practitioner (e.g., no explicit bandwidth choice is needed). Our approach provides honest confidence intervals for the regression discontinuity parameter with both discrete and continuous running variables, and it naturally extends to the case of multiple running variables.

Keywords: Convex optimization, discrete running variable, multiple running variables, uniform asymptotic inference.

1 Introduction

Regression discontinuities often allow for simple and transparent identification of treatment effects from observational data [Hahn et al., 2001, Imbens and Lemieux, 2008, Thistlethwaite and Campbell, 1960, Trochim, 1984]. In the sharp regression discontinuity design, we assume the existence of a running variable $X \in \mathbb{R}$ such that individuals get treated if and only X is larger than some threshold c. For example, in epidemiology, X could be a severity index according to which patients are assigned a medical intervention whenever $X \geq c$, whereas in education, X_i could be a (negative) test score, and students who fail the test ($X \geq c$) need to retake the class or attend summer school. Then, given appropriate assumptions, we can identify a causal effect by comparing subjects i with X_i just above c to those with X_i just below c. Variants of this identification strategy have proven to be useful in education [Angrist and Lavy, 1999, Black, 1999, Jacob and Lefgren, 2004], political science [Caughey and Sekhon, 2011, Lee, 2008], criminal justice [Berk and Rauma, 1983], the evaluation of active labor market programs [Lalive, 2008, Ludwig and Miller, 2007], and other areas.

More formally, suppose that we have access to $i = 1, ..., n$ independent pairs $(X_i, Y_i) \in \mathbb{R} \times \mathbb{R}$, where X_i is the running variable and Y_i is our outcome of interest. We denote the treatment assignment as W_i, which in the sharp regression discontinuity design satisfies $W_i = \mathbf{1}(\{X_i \geq x\})$ and, following the potential outcomes model [Imbens and Rubin, 2015, Neyman, 1923, Rubin, 1974], posit potential outcomes $Y_i(w)$, for $w \in \{0, 1\}$, corresponding to the outcome subject i would have experienced had they received treatment w, and define the conditional average treatment effect $\tau(x)$ in terms of the conditional response functions $\mu_w(x)$:

$$\mu_w(x) = \mathbb{E}\left[Y_i(w) \,\middle|\, X_i = x\right], \quad \tau(x) = \mu_1(x) - \mu_0(x). \tag{1}$$

Then, provided the functions $\mu_w(x)$ are both continuous at 0,

$$\tau(c) = \lim_{x \downarrow c} \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i \,\middle|\, X_i = x\right], \tag{2}$$

i.e., the regression discontinuity identifies the conditional average treatment effect at the threshold c. In the fuzzy regression discontinuity design where the probability of receiving the treatment changes discontinuously at $x = c$, but not necessarily from zero to one, the estimand can be written as the ratio of two such differences. The issues we address in this paper also arise in that setting; however, in the interest of space we do not explicitly discuss it here.

Given this setup, local linear regression is a popular strategy for estimating $\tau(c)$ [Hahn et al., 2001, Porter, 2003]:

$$\hat{\tau} = \operatorname{argmin} \left\{ \frac{1}{nh_n} \sum_{i=1}^{n} K\left(1 - \frac{|\Delta_i|}{h_n}\right) \left(Y_i - a - \tau W_i - \beta_- (\Delta_i)_- - \beta_+ (\Delta_i)_+\right)^2 \right\}, \tag{3}$$

where $K(\cdot)$ is some weighting function, h_n is a bandwidth, $\Delta_i = X_i - c$, and a and $\beta_\pm$ are nuisance parameters. The behavior of regression discontinuity estimation via local linear regression is fairly well understood. When the running variable X is continuous (i.e., X has a continuous positive density at c) and $\mu_w(x)$ is twice differentiable with a bounded second derivative in a neighborhood of c, Cheng, Fan, and Marron [1997] show that the triangular kernel $K(t) = (1 - t)_+$ minimizes worst-case asymptotic mean-squared error among all possible choices of K, Imbens and Kalyanaraman [2012] provide a data-adaptive choice of h_n to minimize the mean-squared error of the resulting estimator, and Calonico, Cattaneo, and Titiunik [2014] propose a method for removing bias effects due to the curvature of $\mu_w(x)$ to allow for asymptotically unbiased estimation. Meanwhile, given a second-derivative bound $|\mu''_w(x)| \leq B$, Kolesár and Rothe [2016] construct confidence intervals centered at the local linear estimator $\hat{\tau}$ that attain uniform asymptotic coverage, even when the running variable X may be discrete.

Despite its ubiquity, however, local linear regression still has some shortfalls. First of all, under the bounded second derivative assumption often used to justify local linear regression (i.e., that $\mu_w(x)$ is twice differentiable and $|\mu''_w(x)| \leq B$ in a neighborhood of c), local linear regression is not the minimax optimal linear estimator for $\tau(c)$—even with a continuous running variable. Second, and perhaps even more important, all the motivating theory for local linear regression relies on X having a continuous distribution; however, in practice, X often has a discrete distribution with a modest number of points of support. When the running variable is discrete there is no compelling reason to expect local linear regression to be particularly effective in estimating the causal effect of interest.[1]

[1] One practical inconvenience that can arise in local linear regression with discrete running variables is

In spite of these limitations, local linear regression is still the method of choice, largely because of its intuitive appeal. The goal of this paper is to show that we can systematically do better: if we are willing to rely on numerical optimization tools, then minimax linear estimation of $\tau(c)$ is both simple and methodologically transparent.

We focus on linear estimators, i.e., estimators of the form $\hat{\tau} = \sum_{i=1}^{n} \gamma_i Y_i$, for some weights γ_i that depend only on the distances $X_i - c$; note that local linear regression belongs to this class. If we know that $\text{Var}\left[Y_i \,\middle|\, X_i\right] = \sigma_i^2$ and that $|\mu_w''(x)| \leq B$, we propose estimating τ_i as follows:

$$\begin{aligned} &\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i, \quad \hat{\gamma} = \text{argmin}_\gamma \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + I_B^2(\gamma) \right\}, \\ &I_B\left(\gamma\right) := \sup_{\mu_0(\cdot), \mu_1(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_{W_i}(X_i) - (\mu_1(c) - \mu_0(c)) : |\mu_w''(x)| \leq B \text{ for all } w,\, x \right\}. \end{aligned} \tag{4}$$

To motivate this estimator, note note that the first term in the minimization problem corresponds to the conditional variance of $\hat{\tau}$ given $\{X_i\}$, while the second term is the worst-case conditional squared bias given that $|\mu_w''(x)| \leq B$; thus, given our regularity assumptions, the estimator $\hat{\tau}$ minimizes the worst-case conditional mean-squared error among all linear estimators. Because no constraints are placed on $\mu_w(c)$ or $\mu_w'(c)$, the optimization in (4) also automatically enforces the constraints $\sum_i W_i \hat{\gamma}_i = 1$, $\sum_i (1 - W_i)\hat{\gamma}_i = -1$, $\sum_i W_i (X_i - c)\hat{\gamma}_i = 0$, and $\sum_i (1 - W_i)(X_i - c)\hat{\gamma}_i = 0$. This is a convex program, and so can be routinely solved using off-the-shelf software as described in, e.g., Boyd and Vandenberghe [2004].

Although this estimator depends explicitly on knowledge of σ_i^2 and B, we note that all practical methods for estimation in the regression discontinuity model, including Calonico et al. [2014], Imbens and Kalyanaraman [2012] and Kolesár and Rothe [2016], require estimating these quantities in order to tune the algorithm. Then, once estimators for these parameters have been specified, the procedure (4) is fully automatic—in particular, there is no need to ask whether the running variable is discrete or continuous, as the optimization is conditional on $\{X_i\}$—whereas the baseline procedures still have other choices to make, e.g., what weight function $K(\cdot)$ to use, or whether to debias the resulting $\hat{\tau}$-estimator.

Figure 1 compares the weights $\hat{\gamma}_i$ obtained via (4) in two different settings: one with a discrete, asymmetric running variable X depicted in the left panel of Figure 2, and the other with a standard Gaussian running variable. We see that, for $n = 1,000$, the resulting weighting functions look fairly similar, and are also comparable to the implicit weighting function generated by local linear regression with a triangular kernel. However, as n grows and the discreteness becomes more severe, our method changes both the shape and the scale of the weights, and the discrepancy between the optimized weighting schemes for discrete versus continuous running variables becomes more pronounced.

In the right panel of Figure 2, we also compare the worst-case conditional mean-squared error of our method relative to that of optimally tuned local linear regression, both with a rectangular and triangular kernel. For the smallest sample size we consider, $n = 333$, the discreteness of the running variable has a fairly mild effect on estimation and—as one

that, if we use a data-driven rule to pick the bandwidth $\hat{h}$ (e.g., the one of Imbens and Kalyanaraman [2012]), we may end up with no data inside the specified range (i.e., there may be no observations with $|X_i - c| \leq h$); the practitioner is then forced to select a different bandwidth ad-hoc. Ideally, methods for regression discontinuity analysis should be fully data-driven, even when X is discrete.

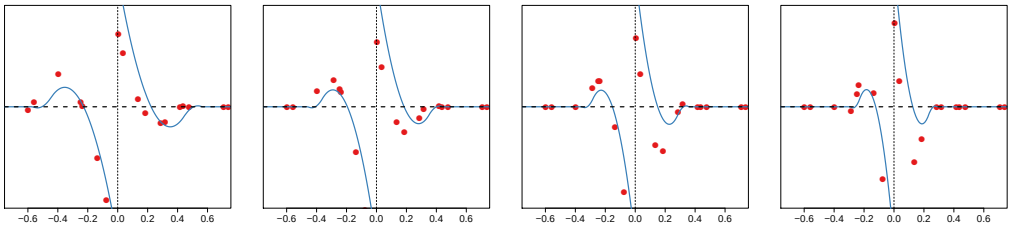


Figure 1: Optimized regression discontinuity design obtained via (4), for different values of n and two different X distributions. The red dots show the learned weighting function in a case where the running variable X is discrete, and different support points are sampled with different probabilities (the probability mass function is shown in the left panel of Figure 2); the blue line shows $\gamma(X_i)$ for standard Gaussian X. We plot $n^{4/5}\hat{\gamma}_i$, motivated by the fact that, with a continuous running variable, the optimal bandwidth for local linear regression scales as $h_n \sim n^{-1/5}$. The weights $\hat{\gamma}_i$ were computed with $B = 5$ and $\sigma^2 = 1$.

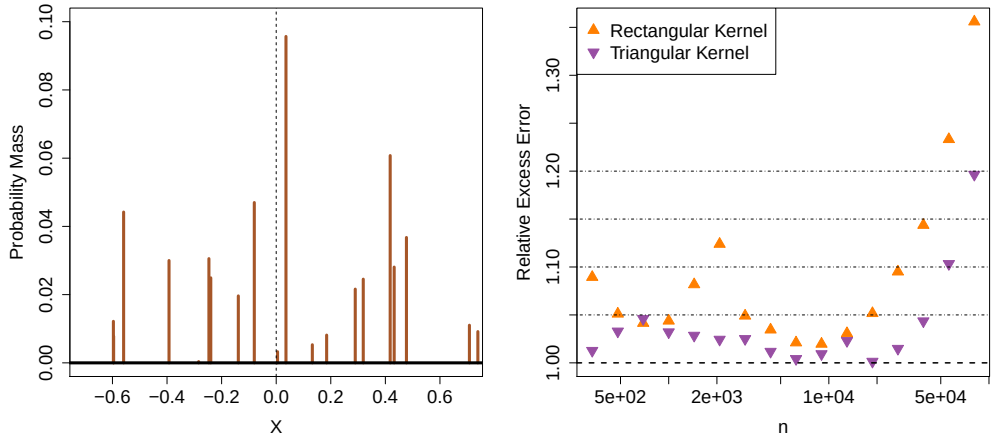


Figure 2: Left panel: Probability mass function underlying the example in Figure 1 and the right panel of the present figure. Right panel: Comparison of our procedure (4) with local linear regression, both using a rectangular ($K(t) = 1(\{t \leq 1\})$) and triangular ($K(t) = (1-t)_+$) kernel. We compare methods in terms of their worst-case mean-squared error conditional on $\{X_i\}$; for local linear regression, we always chose the bandwidth to make this quantity as small as possible. We depict performance relative to our estimator (4).

might have expected—the triangular kernel is noticeably better than the rectangular kernel, while our method is slightly better than the triangular kernel. However, as the sample size increases, the performance of local linear regression relative to our method ebbs and flows rather unpredictably. Perhaps most surprisingly, when $n = 700$ and $19,000$, local linear regression performs slightly better with a rectangular kernel than with a triangular one. As

a matter of intellectual curiosity, it is intriguing to ask whether there exist discrete distributions for which the rectangular kernel may work substantially better than the triangular kernel, or whether additional algorithmic tweaks—such as using different bandwidths on different sides of the threshold—may have helped. However, from a practical perspective, the estimator (4) removes the need to consider such questions in applied data analysis, and automatically adapts to the structure of the data at hand.

We begin our analysis in Section 2 by briefly characterizing our estimator in the conceptual setting of Kolesár and Rothe [2016] that is agnostic as to whether the running variable X_i is continuous or discrete. We find that our estimates provided by our optimized regression discontinuity designs (4) have all the same theoretical guarantees as local linear regression, and, in particular, we can use the construction of Kolesár and Rothe [2016] to build uniformly valid asymptotic confidence intervals for $\tau(c)$; meanwhile, the confidence intervals provided by our method can be made to be strictly shorter than those induced by local linear regression.

The main advantage of our estimator, however, comes from the flexibility of our optimization-based approach that lets us easily adapt to quirks that invariably arise in real-world applications. In Section 3, we illustrate our method using a dataset of Oreopoulos [2006] that has a severely discrete running variable X_i, and compare the performance of our estimator (4) to that of local linear regression. Finally, in Section 4, we consider a dataset of Matsudaira [2008] that has two running variables (a student needs to attend summer school if they fail year-end achievement tests in either math or reading), and show how our method can be adapted to the case where $\mu_w(x)$ is a multivariate function and we have uniform operator-norm bounds for $\nabla^2\mu_w(x)$.

1.1 Related Work

The idea of constructing estimators of the type (4) that are minimax with respect to a regularity class for the underlying data-generating process has a long history in statistics. In early work, Legostaeva and Shiryaev [1971] and Sacks and Ylvisaker [1978] independently studied inference in "almost" linear models that arise from taking a Taylor expansion around a point; see also Cheng et al. [1997]. Armstrong and Kolesár [2016] apply these methods to regression discontinuity designs, resulting in an estimator of the form (4), except with weights[2]

$$\hat{\gamma} = \operatorname{argmin}_\gamma \left\{ \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + A_B(\gamma)^2 \right\}, \tag{5}$$

$$A_B(\gamma) = \sup_{\mu_0(\cdot),\mu_1(\cdot)} \left\{ \sum_{i=1}^n \gamma_i \mu_{W_i}(X_i) - \tau(c) : |\mu_w(x) - \mu_w(c) - \mu_w'(c)(x-c)| \le \frac{B}{2}(x-c)^2 \right\}.$$

Now, although this class of functions is cosmetically quite similar to the bounded-second-derivative class used in (4), we note that the class of weights allowed for in (5) is substantially larger, even if the value of B is the same. This is because the functions $\mu_w(\cdot)$ underlying the above weighting scheme need not be continuous, and can in fact have jumps of magnitude $B(x-c)^2/2$. Given that the key assumption underlying regression discontinuity designs

[2] Armstrong and Kolesár [2016] also consider a more general setting where we assume accuracy of the k-th order Taylor expansion of $\mu_w(x)$ around c; and, in fact, our method also extends to this setting. Here, however, we focus on second-derivative bounds, which are by far the most common in applications.

is continuity of the conditional means of the potential outcomes at the threshold for the running variable, it would appear to be reasonable to impose continuity away from the threshold as well. Allowing for discontinuities through the condition in (5) can make the resulting confidence intervals for $\tau(c)$ substantially larger than they are under the smoothness condition with bounded second derivatives. One key motivation for the weighting scheme (5) rather than our proposed one (4) appears to be that the optimization problem induced by (5) is substantially easier, and allows for closed-form solutions for $\hat{\gamma}_i$. Conversely, we are aware of no closed-form solution for (4), and instead need to rely on numeric convex optimization.

From a practical perspective, our paper fits into a recent trend of using optimization tools to directly minimize feasible error bounds to attain more powerful statistical inference. Similar ideas have also been applied for high-dimensional inference [Javanmard and Montanari, 2014], average treatment effect estimation under unconfoundedness [Athey et al., 2016, Zubizarreta, 2015], matching [Rosenbaum, 1989], and online learning [Duchi et al., 2011].

Finally, although local methods for inference in the regression discontinuity design have desirable theoretical properties, many practitioners also seek to estimate $\tau(c)$ by fitting $\mathbb{E}\left[Y_i \,\middle|\, X_i = x\right]$ using a global polynomial expansion, along with a jump at c; see Lee and Lemieux [2010] for a review and examples. However, as argued by Gelman and Imbens [2014], this approach is not recommended, as it can give large influence to samples i for which X_i is far from the decision boundary c and thus lead to unreliable performance.

2 Optimizing Regression Discontinuity Designs

2.1 Uniform Asymptotic Inference

We start by verifying that our optimized designs can be used for valid asymptotic inference about $\tau(c)$. Following, e.g., Robins and van der Vaart [2006], we seek honest confidence intervals $\mathcal{I}_\alpha$ that attain uniform coverage over the whole regularity set under consideration:

$$\liminf_{n\to\infty} \inf\left\{\mathbb{P}\left[\mu_1(c) - \mu_0(c) \in \mathcal{I}_\alpha\right] : |\mu_w''(x)| \leq B \text{ for all } w,\, x\right\} \geq 1 - \alpha. \tag{6}$$

As in Kolesár and Rothe [2016], our approach to building such confidence intervals relies on an explicit characterization of the bias of $\hat{\tau}$ rather than on undersmoothing. Our key result is as follows.

Theorem 1. *Suppose that we have a moment bound* $\mathbb{E}[(Y_i - \mathbb{E}\left[Y_i \,\middle|\, X_i\right])^q \,\middle|\, X_i = x] \leq C$ *uniformly over all* $x \in \mathbb{R}$*, for some exponent* $q > 2$ *and constant* $C \geq 0$*. Suppose, moreover, that* $0 < \sigma_{\min} \leq \sigma_i$ *for all* $i = 1, ..., n$ *for a deterministic value* $\sigma_{\min}$*, and that none of the weights* $\hat{\gamma}_i$ *derived in* (4) *dominates all the others, i.e.,*

$$\max_{1\leq i\leq n}\left\{\hat{\gamma}_i^2\right\} \Big/ \sum_{i=1}^n \hat{\gamma}_i^2 \to_p 0. \tag{7}$$

Then, our estimator $\hat{\tau}$ *from* (4) *is asymptotically Gaussian,*

$$\left(\hat{\tau} - b\left(\hat{\gamma}\right)\right) \Big/ s\left(\hat{\gamma}\right) \Rightarrow \mathcal{N}\left(0,\, 1\right), \quad b(\hat{\gamma}) = \sum_{i=1}^n \hat{\gamma}_i \mu_{W_i}(X_i) - \tau(c), \quad s^2\left(\hat{\gamma}\right) := \sum_{i=1}^n \hat{\gamma}_i^2 \sigma_i^2, \tag{8}$$

where $b\left(\hat{\gamma}\right)$ *denotes the conditional bias, and* $s^2\left(\hat{\gamma}\right) \to_p 0$.

Now, in solving the optimization problem (4), we also obtain an explicit bound $\hat{t}$ on the conditional bias, $b(\hat{\gamma}) \leq \hat{t}$, so we can use the construction of Kolesár and Rothe [2016] to obtain confidence intervals for $\tau(c)$ as follows:

$$\tau(c) \in \hat{\tau} \pm l_\alpha, \quad l_\alpha = \min\left\{l : \mathbb{P}\left[|b + s\left(\hat{\gamma}\right) Z| \leq l\right] \geq \alpha \text{ for all } |b| \leq \hat{t}\right\}, \quad Z \sim \mathcal{N}\left(0, 1\right), \quad (9)$$

where α is the significance level. These confidence intervals are asymptotically uniformly valid in the sense of (6), i.e., for any $\alpha' < \alpha$, there is a threshold $n_{\alpha'}$ for which, if $n \geq n_{\alpha'}$, the confidence intervals (9) achieve α'-level coverage for any functions $\mu_w(\cdot)$ in our regularity class.

2.2 Setting Problem Parameters

To use this result in practice, we of course need to estimate sum $\sum \hat{\gamma}_i^2 \sigma_i^2$ and the bound B on curvature. Estimating the former is relatively routine; and we recommend the following. First, we estimate $\mu_w(x)$ globally, or over a large plausible relevant interval around the threshold, and average the square of the residuals $R_i = Y_i - \hat{\mu}_{W_i}(X_i)$ to obtain an estimate $\hat{\sigma}^2$ of the average value of σ_i^2. Then, we optimize weights $\hat{\gamma}_i$ using (4), with $\sigma_i^2 \leftarrow \hat{\sigma}^2$. Finally, once we have chosen the weights γ_i, we estimate the sampling error of $\hat{\tau}$ as below, noting that the estimator will be consistent under standard conditions

$$\hat{s}^2\left(\hat{\gamma}\right) = \sum_{i=1}^{n} \hat{\gamma}_i^2 \left(Y_i - \hat{\mu}_{W_i}(X_i)\right)^2, \quad \hat{s}^2\left(\hat{\gamma}\right) \Big/ \sum \hat{\gamma}_i^2 \sigma_i^2 \gtrsim_p 1. \quad (10)$$

Conceptually, this strategy is comparable to first running local linear regression without heteroskedasticity adjustments to get a point estimate, but then ensuring that the uncertainty quantification is heteroskedasticity-robust. We summarize the resulting method as Procedure 1.

Conversely, obtaining good bounds on the curvature B is more difficult, and requires problem specific insight. In particular, adapting to the true curvature $\mu_w(x)$ without a-prior bounds for B is not always possible; see Armstrong and Kolesár [2016] and references therein. In applications, we recommend considering a range of plausible values of B that could be obtained, e.g., from subject-matter expertise or from considering the mean-response function globally. For example, we could estimate $\mu_w(x)$ using a quadratic function globally, or over a large plausible relevant interval around the threshold, and then multiply maximal curvature of the fitted model by a constant, e.g., 2 or 4. The larger the value of B we use the more conservative the resulting inference.

Finally, only considering data over an a-priori specified "large plausible relevant interval" around c that safely contains all the data relevant to fitting $\tau(c)$ can also be of computational interest. Our method relies on estimating a smooth non-parametric function over the whole range of x; and being able to reduce the relevant range of x a-priori can reduce the required computation by an order of magnitude. Although defining such plausibility intervals is of course heuristic, our method ought not be too sensitive to how exactly the interval was chosen. For example, in the setup considered in Section 3, the optimal bandwidth for local linear regression is around 3 or 6 years depending on the amount of assumed smoothness (and choosing a good bandwidth is very important); conversely, using plausibility intervals extending 10, 15, or 20 years on both sides of c appears to work reasonably well. In any case, when running the method (4), one should make sure that the weights $\hat{\gamma}_i$ get very small near the edge of the plausibility interval; if not, the interval should be made larger.

Procedure 1. Optimized Regression Discontinuity Inference

This algorithm provides confidence intervals for the conditional average treatment effect $\tau(c)$, given an a-priori bound B on the second derivative of the functions $\mu_w(x)$. We assume that the conditional variance parameters σ_i^2 are unknown; if they are known, they should be used as in (4). This procedure is implemented in our `R` package `optrdd`.[3]

1. Pick a large window r, such that data with $|X_i - c| > r$ can be safely ignored without loss of efficiency. (Here, we can select $r = \infty$, but this may result in unnecessary computational burden.)
2. Run ordinary least-squares regression of Y_i on the interaction of X_i and W_i over the window $|X_i - c| \leq r$. Let $\hat{\sigma}^2$ be the residual error from this regression.
3. Obtain $\hat{\gamma}$ via the quadratic program (11), with σ_i set to $\hat{\sigma}$ and weights outside of the range $|X_i - c| \leq r$ set to 0.
4. Confirm that the optimized weights $\hat{\gamma}_i$ are small for $|X_i - c| \approx r$. If not, start again with a larger value of r.
5. Estimate $\hat{\tau} = \sum_{i=1}^n \hat{\gamma}_i Y_i$ and $\hat{s}^2 = \sum_{i=1}^n \hat{\gamma}_i^2 (Y_i - \hat{\mu}_{W_i}(X_i))^2$, where the $\hat{\mu}_{W_i}(X_i)$ are predictions from the least squares regression from step 1.
6. Build confidence intervals as in (9).

2.3 Solving the Optimization Problem

In order to leverage off-the-shelf convex optimization software, it is helpful to write the program (4) as a standard quadratic program,[4]

$$\begin{aligned} &\text{minimize } \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + B^2 t^2 \text{ subject to} \\ &\sup\left\{ \sum_{i=1}^n \gamma_i f(X_i) : f(c) = 0,\ f'(c) = 0,\ |f''(x)| \leq 1 \right\} \leq t, \\ &\sum_{\{i:X_i<c\}} \gamma_i = -1, \quad \sum_{\{i:X_i \geq c\}} \gamma_i = 1, \\ &\sum_{i=1}^n \gamma_i (X_i - c) = 0, \quad \sum_{i=1}^n \left(\mathbf{1}\left(\{X_i \geq c\}\right) - 1/2\right) \gamma_i (X_i - c) = 0. \end{aligned} \tag{11}$$

[3] Here, the algorithm assumes that all observations are of roughly the same quality (i.e., we do not know that σ_i^2 is lower for some observations than others). If we have a-priori information about the relative magnitudes of the conditional variances of different observations, e.g., some pairs outcomes Y_i are actually aggregated over many observations, then we should run steps 2 and 3 below using appropriate inverse-variance weights. Our software allows for such weighting.

[4] This representation is valid when X is univariate. When X is multivariate, we need to explicitly consider two different f-functions: one for the treated, and one for the controls. The reason this issue does not arise in the univariate case is that, then, $f_0(x)$ and $f_1(x)$ only touch at c, and we already know that $f_0(c) = f_0'(c) = f_1(c) = f_1'(c) = 0$.

Given this form, the only subtlety is in dealing with the functional constraint that $\sum \gamma_i f(X_i) \leq t$ for all functions with $f(c) = 0$, $f'(c) = 0$, $|f''(\cdot)| \leq 1$.

When the running variable X is univariate, we can get a simple, primal solution via integration by parts; procedurally, this amounts to representing $f(\cdot)$ in terms of its second derivative. Let $\zeta(\cdot)$ denote the second derivative $f''(\cdot)$, and let M_2 be the linear operator that sends $\zeta(\cdot)$ to $f(\cdot)$, i.e., such that $(M_2\zeta)(c) = 0$, $(M_2\zeta)'(c) = 0$, $(M_2\zeta)''(x) = \zeta(x)$ for measurable functions $\zeta(x)$. Then, our constraint of interest becomes

$$\sup\left\{\sum_{i=1}^{n}\gamma_i(M_2\zeta)(X_i) : |\zeta(x)| \leq 1,\, x \in \mathbb{R}\right\} \leq t \iff \int_{\mathbb{R}}\left|\sum_{i=1}^{n}\gamma_i(M_2^*X_i)(z)\right| dz \leq t, \quad (12)$$

where $(M_2^*X_i)(\cdot)$ is the functional for which $(M_2\zeta)(X_i) = \int_{\mathbb{R}}(M_2^*X_i)(z)\zeta(z)\,dz$. In this form, the problem can be passed directly to a standard quadratic programming package such as `quadprog` for `R` [Turlach and Weingessel, 2013]; as usual, we represent the continuous variable X using a fine, discrete grid.

When the running variable it multivariate, we need a more general approach based on convex duality. Because the problem (11) has strictly feasible solutions, we can use Slater's theorem to verify that strong duality holds, and that the optimum of (11) matches the optimum of the following one:[5]

$$\begin{aligned}\text{maximize } \inf_{\gamma,\,t}\Bigg\{&\sum_{i=1}^{n}\gamma_i^2\sigma_i^2 + B^2t^2 + \lambda_1\left(\sum_{i=1}^{n}\gamma_i f(X_i) - t\right) + \lambda_2\left(\sum_{\{i:X_i<c\}}\gamma_i + 1\right)\\ &+ \lambda_3\left(\sum_{\{i:X_i\geq c\}}\gamma_i - 1\right) + \sum_{i=1}^{n}\left(\lambda_4 + \lambda_5\left(\mathbf{1}\left(\{X_i \geq c\}\right) - 1/2\right)\right)\gamma_i(X_i - c)\Bigg\}\\ \text{subject to } &f(c) = 0,\ f'(c) = 0,\ |f''(x)| \leq 1,\ \lambda_1 \geq 0,\ \lambda_2, \ldots, \lambda_5 \in \mathbb{R}.\end{aligned} \quad (13)$$

The advantage of this dual representation is that, by examining first order conditions in the $\inf_{\gamma,\,t}\{\}$ term, we can analytically solve for γ in the dual objective,

$$-2\sigma_i^2\hat{\gamma}_i = \hat{\lambda}_1\hat{f}(X_i) + \hat{\lambda}_2\mathbf{1}\left(\{X_i < c\}\right) + \hat{\lambda}_3\mathbf{1}\left(\{X_i \geq c\}\right) + \ldots, \quad (14)$$

where $\hat{f}(\cdot)$, $\hat{\lambda}_1$, etc., are the maximizers of (13). This results in an optimization problem over the space of twice differentiable functions f (along with a finite number of Lagrange parameters λ_j), and again be solved using, e.g., `quadprog`; see Appendix B for details.

2.4 Minimizing Confidence Interval Length

As formulated in (11), our estimator seeks to minimize the worst-case mean-squared error over the specified bounded-second-derivative class. However, in some applications, we may be more interested in making the confidence intervals (9) as short as possible. Writing $\hat{v}^2 = \sum_{i=1}^{n}\hat{\gamma}_i^2\sigma_i^2$, we see that both the worst-case mean-squared error, $\hat{v}^2 + B^2\hat{t}^2$ and the confidence interval length in (9) are monotone increasing functions of $\hat{v}$ and $\hat{t}$; the only difference is in how they weight these two quantities at the optimum.

[5] See Section 5.2.3. of Boyd and Vandenberghe [2004]. Here, we also implicitly used von Neumann's minimax theorem to move the maximization over f outside the $\inf_{\gamma,\,t}\{\}$ statement.

Now, to derive the full Pareto frontier of pairs $(\hat{v}, \hat{t})$, we can simply re-run (11) with the term B^2t^2 in the objective replaced with λB^2t^2, for some $\lambda > 0$. A practitioner wanting to minimize the length of confidence intervals could consider computing this whole solution path to (11), and then using the value of λ that yields the best intervals. Since this procedure never looks at the responses Y_i, the inferential guarantees for the resulting confidence intervals remain valid. In our applications, however, we did not find a meaningful gain from optimizing over λ instead of just minimizing worst-case mean-squared error as in (11), and so did not pursue this line of investigation further.

3 Application: The Effect of Compulsory Schooling

In our first application, we consider a dataset from Oreopoulos [2006], who studied the effect of raising the minimum school-leaving age on earnings as an adult. The effect is identified by the UK changing its minimum school-leaving age from 14 to 15 in 1947, and the response is log-earnings among those with non-zero earnings (in 1998 pounds). This dataset exhibits notable discreteness in its running variable, and was used by Kolesár and Rothe [2016] to illustrate the value of their bias-adjusted confidence intervals for discrete regression discontinuity designs. For our analysis, we pre-process our data exactly as in Kolesár and Rothe [2016]; we refer the reader to their paper and to Oreopoulos [2006] for a more in-depth discussion of the data.

As in Kolesár and Rothe [2016], we seek to identify the effect of the change in minimum school-leaving age on average earnings via a local analysis around the regression discontinuity; our running variable is the year in which a person turned 14, with a treatment threshold at 1947. Kolesár and Rothe [2016] consider analysis using local linear regression with a rectangular kernel and a bandwidth chosen such as to make to make their honest confidence intervals as short as possible (recall that we can measure confidence interval length without knowing the point estimate, and so tuning the interval length does not invalidate inference). Here, we also consider local linear regression with a triangular kernel, as well as our optimized design.[6]

In order to obtain confidence intervals, it remains to choose a bound B. Kolesár and Rothe [2016] consider a very optimistic choice of $B = 0.003$ and a less optimistic one of $B = 0.03$. Meanwhile, following the discussion in Section 2.2, a 2nd-order polynomial fit with a "large" bandwidth of either 12 or 18 has a curvature of 0.006 (the estimate is insensitive to the choice of large bandwidth); motivated by this, we also consider $B = 0.006$ and $B = 0.012$. For σ_i^2, we proceed as discussed in Section 2.2. Figure 3 shows the effective $\hat{\gamma}(X_i)$ weighting functions for all 3 considered methods, with $B = 0.012$.

We present results in Table 1. Overall, these results are in line with those presented in Figure 2. The optimized method yields materially shorter confidence intervals than local linear regression with a rectangular kernel: for example, with $B = 0.03$, the rectangular kernel intervals are 11% longer. In comparison, the triangular kernel comes closer to matching the performance of our method, although the optimized method still has shorter confidence intervals. Moreover, when considering comparisons with the triangular kernel, we note that the rectangular kernel is far more prevalent in practice, and that the motivation for using the triangular kernel often builds on the optimality results of Cheng et al. [1997]. And, once

[6] Oreopoulos [2006] analyze the dataset using a global polynomial specification with clustered random variables, following Lee and Card [2008]. However, as discussed in detail by Kolesár and Rothe [2016], this approach does not yield valid confidence intervals.

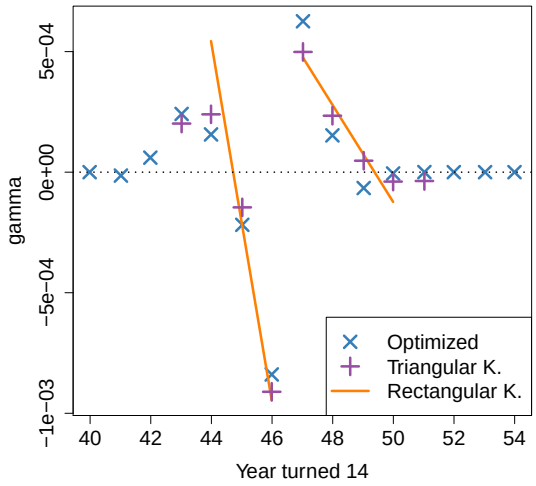


Figure 3: Weighting functions $\hat{\gamma}(X_i)$ produced explicitly by our estimator (4), and implicitly via local linear regression with a rectangular or triangular kernel. Both local linear regression methods have a finite bandwidth, and the effective weights of $\hat{\gamma}(X_i) = 0$ outside this bandwidth are not shown. The weighting functions were generated with $B = 0.012$.

B	rect. kernel	tri. kernel	optimized
0.003	0.0208 ± 0.0762	0.0313 ± 0.0737	0.0291 ± 0.0716
0.006	0.0576 ± 0.0896	0.0484 ± 0.0866	0.0412 ± 0.0840
0.012	0.0642 ± 0.1087	0.0626 ± 0.1039	0.0554 ± 0.1003
0.03	0.0642 ± 0.1475	0.0707 ± 0.1384	0.0707 ± 0.1326

Table 1: Confidence interval ($\alpha = 95\%$) for the effect of raising the minimum school-leaving age on average log-earnings, as given by local linear regression with a rectangular kernel, local linear regression with a triangular kernel, and our optimized method (4). The confidence intervals account for curvature effects, provided the second derivative is bounded by B.

one has set out on a quest for optimal weighting functions, there appears to be little reason to not just use the actually optimal weighting function (4).[7]

Finally, we note that a bound B on the second derivative also implies that the quadratic approximation (5) holds with the same bound B. Thus, we could in principle also use the method of Armstrong and Kolesár [2016] to obtain uniform asymptotic confidence intervals here. However, the constraint (5) is weaker than the actual assumption we were willing to make (i.e., that the functions $\mu_w(\cdot)$ have a bounded second derivative), and so the resulting confidence intervals are substantially larger. Using their approach on this dataset gives confidence intervals of 0.0508 ± 0.0971 with $B = 0.006$ and 0.0682 ± 0.1774 with $B = 0.03$; these intervals are not only noticeably longer than our intervals, but are also longer than the

[7] The results for the rectangular kernel with $B = 0.003$ and $B = 0.03$ replicate those of Kolesár and Rothe [2016] up to the 3rd decimal place; after that, rounding effects from our histogram-based implementation (Section 2.3) become apparent. We use the same binned representation of the data for all 3 methods, so this rounding error does not affect relative comparisons between methods.

best uniform confidence intervals we can get using local linear regression with a rectangular kernel as in Kolesár and Rothe [2016]. Thus, the use of numerical convex optimization tools that let us solve (4) instead of (5) can be of considerable value in practice.

4 Inference with Multiple Running Variables

To illustrate the flexibility of our convex-optimization-based approach, we extend our discussion to the case of treatments determined by multiple running variables. Our discussion is motivated by a common inference strategy in education that arises from standardized testing, built on the fact that some school districts mandate students to attend summer school if they fail a year-end test in either math or reading [Jacob and Lefgren, 2004, Matsudaira, 2008]. The argument for identification via a regression discontinuity is still clear; however, the discontinuity now no longer occurs along a point, but rather along a surface in the bivariate space encoding both a student's math and reading scores.

The problem of regression discontinuity inference with multiple running variables is considerably richer than the corresponding problem with a single running variable, because an investigator could now plausibly hope to identify many different treatment effects, including the average effect of summer school for a typical student who received borderline scores in both math and reading, failed math but passed reading, passed reading but failed math, etc. Most of the existing literature on this setup, including Papay et al. [2011], Reardon and Robinson [2012] and Wong et al. [2013], have focused on these questions of identification, while using some form of local linear regression for estimation.

In the multivariate case, however, questions about how to tune local linear regression are exacerbated, as the problems of choosing the kernel function $K(\cdot)$ and the bandwidth h are now multivariate. Perhaps for this reason, it is still popular to use univariate methods to estimate treatment effects in the bivariate setting and, e.g., only consider students who passed the math test, and then use passing/failing the reading test as a univariate discontinuity. This is the approach taken by both Jacob and Lefgren [2004] and Matsudaira [2008].

Here, we show how our approach can be used to side-step the problem of choosing a multivariate kernel function by hand.[8] In addition to providing a simple-to-apply algorithm, our method also lets us formally account for the curvature of the mean-response function $\mu_w(x)$ in our statistical inference, thus strengthening formal guarantees relative to prior work. Papay et al. [2011] and Reardon and Robinson [2012] study local linear regression with a "small" bandwidth, but do not account for finite sample bias due to curvature. In his thesis, Zajonc [2012] extends the analysis of Imbens and Kalyanaraman [2012] to the multivariate case, and studies optimal bandwidth selection for continuous running variables given second derivative bounds; the inference, however, again requires undersmoothing. To our knowledge, the approach we present below is the first to allow for uniform, bias-adjusted inference in the multivariate regression discontinuity setting.

[8] One particular difficulty with local linear regression in the multivariate case is that different boundary shapes may require different basis representations and/or kernel weighting functions. For example, how should one adapt local linear regression methodologies if the treatment area were a half-space or a Euclidean ball instead of a quadrant? Our optimization-based approach avoids this conceptual concern completely, as the specification of our estimator does not explicitly depend on the shape of the boundary.

4.1 Optimizing Multivariate Discontinuity Designs

Relative to the univariate case, the multivariate case has two additional subtleties we need to address. First, in (4) it is natural to impose a constraint $|\mu_w''(x)| \leq B$ to ensure smoothness; in the multivariate case, however, we have more choices to make. For example, do we constrain $\mu_w(x)$ to be an additive function, or do we allow for interactions? Here, we opt for the more flexible specification, and simply require that $\left\|\nabla^2 \mu_w(x)\right\| \leq B$, where $\|\cdot\|$ denotes the operator norm (i.e., the largest absolute eigenvalue of the second derivative).

Moreover, as emphasized by Papay et al. [2011], whereas the univariate design only enables us to identify the conditional average treatment effect at the threshold c, the multivariate design enables us to potentially identify a larger family of conditional average treatment effects. In our setting, for example, we might ask about the effectiveness of summer school on students who are borderline in both math and reading, those who are borderline in math but easily pass reading, and those who are borderline in reading but easily pass math.

Here, we consider the following two specifications. First, writing c as the bivariate threshold at which a student just barely passes both tests, let $\hat{\tau}_c = \sum_{i=1}^n \hat{\gamma}_{c,i} Y_i$ with

$$\hat{\gamma}_c = \text{argmin}_\gamma \left\{ \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + \left(\sup_{\|\nabla^2 \mu_w(x)\| \leq B} \left\{ \sum_{i=1}^n \gamma_i \mu_{W_i}(X_i) - (\mu_1(c) - \mu_0(c)) \right\} \right)^2 \right\} \quad (15)$$

denote an estimator for the conditional average treatment effect at c. The upside with this approach is that it gives us an estimand that is easy to interpret; the downside is that, when curvature is non-negligible, (15) can effectively only make use of data near the specified test point c, thus resulting in relatively low power.

In order to improve power, we can also consider weighted conditional average treatment effect greedily chosen such as to make the inference as precise as possible, in the spirit of Crump et al. [2009]: $\hat{\tau}_* = \sum_{i=1}^n \hat{\gamma}_{*,i} Y_i$ with

$$\hat{\gamma}_* = \text{argmin}_{\gamma,\, x_0 \in \mathbb{R}^k} \left\{ \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + \left(\sup_{\|\nabla^2 \mu_0(x)\| \leq B} \left\{ \sum_{i=1}^n \gamma_i \mu_0(X_i) \right\} \right)^2 \right\}. \quad (16)$$

In other words, we seek to pick weights γ_i that are nearly immune to bias due to curvature of the baseline response surface $\mu_0(x)$. By construction, this estimator satisfies

$$\left| \mathbb{E}\left[\hat{\tau}_* \,\middle|\, \{X_i\} \right] - \bar{\tau}(\hat{\gamma}_*) \right| \leq \sup_{\|\nabla^2 \mu_0(x)\| \leq B} \left\{ \sum_{i=1}^n \hat{\gamma}_{*,i} \mu_0(X_i) \right\}, \quad \bar{\tau}(\gamma) := \sum_{i=1}^n W_i \, \hat{\gamma}_{*,i} \tau(X_i). \quad (17)$$

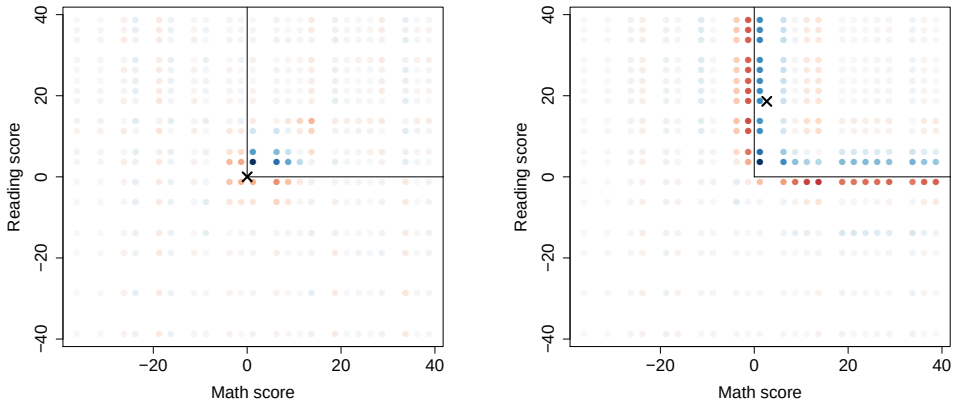
Note that $\sum W_i \hat{\gamma}_{*,i} = 1$, and so $\bar{\tau}(\hat{\gamma}_*)$ is in fact a weighted average of the conditional average treatment effect function $\tau(\cdot)$ over the treated sample. If we ignored the curvature of $\tau(\cdot)$, we could interpret $\hat{\tau}_*$ as an estimate for the conditional average treatment effect at $x_* = \sum \hat{\gamma}_{*,i} W_i X_i$.

4.2 Application

We illustrate our approach using the dataset of Matsudaira [2008]. As discussed above, the goal is to study the impact of summer school on future test scores, and the effect of summer school is identified by a regression discontinuity: At the end of the school year, students need to take year-end tests in math and reading; then, students failing either of these tests

	fail reading		pass reading	
	fail math	pass math	fail math	pass math
number of students	3,586	1,488	10,331	15,336
summer school attendance	69.6%	61%	52.9%	10.6%

Table 2: Summary statistics for a subset of the dataset of Matsudaira [2008].



Unweighted CATE (15) Weighted CATE (16)

Figure 4: Weights $\hat{\gamma}$ underlying treatment effect estimates of the effect of summer school on the following year's reading scores, using both (15) which seeks to estimate the conditional average treatment effect (CATE) at $c = (0,\, 0)$, and the estimator (16) which allows weighted CATE estimation. The size of $\hat{\gamma}_i$ is depicted by the color, ranging from dark red (very positive) to dark blue (very negative); the boldface-$\times$ marks the weighted mean of the treated X_i-values, i.e., $\sum \hat{\gamma}_i W_i X_i$. These plots were generated with a maximum second derivative bound of $B = 0.5 \times 40^{-2}$.

are mandated to attend summer school. Here, we focus on the 2001 class of graduating 5th graders, and filter the sample to only include the $n = 30,741$ students whose 5th-grade math and reading scores both fall between 40 points of the passing threshold; this represents 44.7% of the full sample. Matsudaira [2008] analyzed this dataset via univariate analyses, by using reading score as a running variable and only consider the subset of students who passed the math exam, etc. This allows for a simple analysis, but may also result in a needless loss of power.

We present some summary statistics in Table 2. Clearly, not all students mandated to attend summer school in fact attend, and some students who pass both tests still need to attend for reasons discussed in Matsudaira [2008]. That being said, the effect of passing tests on summer school attendance is quite strong and, furthermore, the treatment effect of being mandated to summer school is interesting in its own right, so here we perform an "intent to treat" analysis without considering non-compliance.

We consider both of our optimized estimators, (15) and (16), and compare weight functions $\hat{\gamma}$ learned by both methods in Figure 4. The estimator $\hat{\tau}_c$ is in fact quite conservative,

	estimator:	unweighted CATE (15)			weighted CATE (16)		
subject	B	conf. int.	m. bias	samp. err	conf. int.	m. bias	s. err
math	0.5×40^{-2}	0.058 ± 0.107	0.034	0.044	0.068 ± 0.036	0.009	0.016
math	1.0×40^{-2}	0.055 ± 0.145	0.051	0.057	0.067 ± 0.043	0.012	0.019
reading	0.5×40^{-2}	-0.006 ± 0.11	0.034	0.046	0.041 ± 0.036	0.009	0.016
reading	1.0×40^{-2}	-0.041 ± 0.147	0.051	0.058	0.046 ± 0.043	0.012	0.019

Table 3: Estimates for the effect of summer school on math and reading scores on the following year's test, using different estimators and choices of B. Reported are bias-adjusted 95% confidence intervals, a bound on the maximum bias given our choice of B, and an estimate of the sampling error conditional on $\{X_i\}$.

and only gives large weights to students who scored close to c. Our choice of estimating the conditional average treatment effect at (0, 0) may have been particularly challenging, as it is in a corner of control-space and so does not have particularly many control neighbors.

In contrast, the weighted method $\hat{\tau}_*$ appears to have effectively learned matching: It constructs pairs of observations all along the treatment discontinuity, thus allowing it to gain in power while canceling out curvature effects due to $\mu_0(x)$. As seen in Table 2, in this sample, it is much more common to fail math and pass reading than vice-versa; thus, the mean of the samples used for "matching" lies closer to the math pass/fail-boundary than the reading one.

In order to get confidence intervals for the treatment effect, we again need to choose a value for B. Running a 2nd order polynomial regression on the next year's math and reading scores for both treated and control students separately, we find the largest curvature effect among the reading score of control students; roughly a curvature of 0.46×40^{-2} along the $(1, 2)$ direction. Thus, we run our algorithm with both an optimistic choice of $B = 0.5 \times 40^{-2}$ and a more conservative choice $B = 1.0 \times 40^{-2}$ (we report curvatures on the "scale" of the plots in Figure 4, such that a curvature of 1.0×40^{-2} results in a worst-case bias of 1 in the corners of the plot).

Results are given in Table 3. As expected, the confidence intervals using the weighted method (16) are much shorter than those obtained using (15), allowing for a 0.95-level significant detection in the first case but not in the second. Since the weighting method is so much more powerful, and in practice seems to yield a matching-like methodology that resembles existing recommendations [e.g., Rosenbaum, 2002], we expect it to be more often applicable than the unweighted estimator (15).

Finally, it is of course natural to ask whether whether the bivariate specification considered here gave us anything in addition to the simpler approach used by Matsudaira [2008], i.e., of estimating the treatment effect of summer school on the next year's math exam by running a univariate regression discontinuity analysis on only those students who passed the reading exam, and vice-versa to the effect on the reading exam. We ran both of these analyses using our method (4), again considering bounds $B = 0.5 \times 40^{-2},\ 1 \times 40^{-2}$ on the second derivative. For math, we obtained 95% confidence intervals of 0.083 ± 0.040 and 0.079 ± 0.047 for the smaller and larger B-bounds respectively; for reading, we obtained 0.037 ± 0.075 and 0.030 ± 0.090. In both cases, the corresponding bounds for the weighted estimator (16) in Table 3 are shorter, despite accounting for the possibility of bivariate curvature effects. The difference is particularly strong for the reading outcome, since our estimator $\hat{\tau}_*$ can also use

students near the math pass/fail-boundary for improved precision.[9]

5 Discussion

In this paper, we introduced an optimization-based approach to statistical inference in regression discontinuity designs. By using numerical convex optimization tools, we explicitly derive the minimax linear estimator for the regression discontinuity parameter under bounds on the second derivative of the conditional response surface. Because any method based on local linear regression is also a linear estimator of this type, our approach dominates local linear regression in terms of minimax mean-squared error. We also show how our approach can be used to build uniformly valid confidence intervals.

A key advantage of our procedure is that, given bounds on the second derivative, estimation of the regression discontinuity parameter is fully data-driven. The proposed algorithm is the same whether the running variable be continuous or discrete, and does not depend on the shape of treatment assignment boundary when X is multivariate. We end our discussion with some potential extensions of our approach.

Fuzzy regression discontinuities In the present paper, we only considered sharp regression discontinuities, where the treatment assignment T_i is a deterministic function of X_i. However, there is also considerable interest in fuzzy discontinuities, where W_i is random but $\mathbb{P}\left[W_i = 1 \,\middle|\, X_i = x\right]$ has a jump at the threshold c; see Imbens and Lemieux [2008] for a review. In this case, it is common to interpret the indicator $\mathbf{1}\left(\{X_i \geq c\}\right)$ as an instrument, and then to estimate a local average treatment effect via two-stage local linear regression [Imbens and Angrist, 1994]. By analogy, we can estimate treatment effects with fuzzy regression discontinuity via two-stage optimized designs as

$$\hat{\tau}_{LATE} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i \Big/ \sum_{i=1}^{n} \hat{\gamma}_i W_i, \tag{18}$$

where the $\hat{\gamma}_i$ are obtained as in (11) with an appropriate choice penalty on the maximal squared imbalance t^2. This approach would clearly be consistent based on results established in this paper; however, deriving the best way to trade off bias and variance in specifying $\hat{\gamma}_i$ and extending the approach of Kolesár and Rothe [2016] for uniform asymptotic inference may require some further consideration.

Balancing auxiliary covariates In many applications, we have access to auxiliary covariates $Z_i \in \mathbb{R}^p$ that are predictive of Y_i but unrelated to the treatment assignment near the boundary c. As discussed in, e.g., Imbens and Lemieux [2008], such covariates are not necessary for identification; but controlling for them can increase robustness to hidden biases. One natural way to use such auxiliary covariates in our optimized designs is to require the weights $\hat{\gamma}_i$ to balance these covariates, i.e., to add a constraint

$$\sum_{i=1}^{n} \hat{\gamma}_i \, Z_{ij} = 0 \text{ for all } j = 1, ..., p \tag{19}$$

[9] The corresponding headline numbers from Matsudaira [2008] are a 95% confidence interval of 0.093 ± 0.029 for the effect on the math score, and 0.046 ± 0.045 for the reading score; see Tables 2 and 3, reduced form estimates for 5th graders. These confidence intervals, however, do not formally account for bias. They estimate the discontinuity parameter using a global cubic fit; such methods, however, do not reliably eliminate bias [Gelman and Imbens, 2014].

to the optimization problem (4).[10] In principle, if the distribution of Z_i is in fact independent of X_i when X_i is near the threshold c, we would expect the balance conditions (19) to hold approximately even if we do not enforce them; however, explicitly enforcing such balance may improve robustness.[11] If we have an additive, linear dependence of Y_i on Z_i, then enforcing balance as in (19) would also result in variance reduction, as the conditional variance of our estimator $\hat{\tau}$ would now depend on $\text{Var}\left[Y_i \,\middle|\, X_i,\, Z_i\right]$, which is always smaller or equal to $\text{Var}\left[Y_i \,\middle|\, X_i\right]$.

Working with generic regularity assumptions Following standard practice in the regression discontinuity literature, we focused on minimax linear inference under bounds on the second derivative of $\mu_w(\cdot)$ [e.g., Kolesár and Rothe, 2016, Imbens and Kalyanaraman, 2012]. However, our conceptual framework can also be applied with higher order smoothness assumptions via bounds on the k-th derivative of $\mu_w(\cdot)$, and can easily be combined with other forms of structural information about the conditional response functions (e.g., perhaps we know from theory that the functions $\mu_w(\cdot)$ must be concave). Thanks to the flexibility of our optimization-based approach, acting on either of these ideas would simply involve implementing the required software using standard convex optimization libraries.

References

Joshua D Angrist and Victor Lavy. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575, 1999.

Timothy B Armstrong and Michal Kolesár. Optimal inference in a class of regression models. 2016.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

Richard A Berk and David Rauma. Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, 78(381):21–27, 1983.

Sandra E Black. Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics*, 114(2):577–599, 1999.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

[10] The constraint (19) is a linear constraint, and so the optimization problem (11) remains a quadratic program with this constraint.

[11] A related idea would be to use the covariates Z_i for post-hoc specification testing as in Heckman and Hotz [1989] or Imbens and Lemieux [2008]: The strategy is to obtain weights $\hat{\gamma}_i$ without looking at the Z_i, and then to reject the modeling strategy if (19) does not hold approximately.

Devin Caughey and Jasjeet S Sekhon. Elections and the regression discontinuity design: Lessons from close us house races, 1942-2008. *Political Analysis*, pages 385–408, 2011.

Ming-Yen Cheng, Jianqing Fan, and James S Marron. On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul): 2121–2159, 2011.

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

Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

James Heckman and Joseph Hotz. Alternative methods for evaluating the impact of training programs. *Journal of the American Statistical Association*, 84(408):862–880, 1989.

Guido W Imbens and Joshua D Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994.

Guido W Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959, 2012.

Guido W Imbens and Thomas Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Brian A Jacob and Lars Lefgren. Remedial education and student achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*, 86(1):226–244, 2004.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

Michal Kolesár and Christoph Rothe. Inference in regression discontinuity designs with a discrete running variable. 2016.

Rafael Lalive. How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics*, 142(2):785–806, 2008.

David S Lee. Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*, 142(2):675–697, 2008.

David S Lee and David Card. Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674, 2008.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

IL Legostaeva and AN Shiryaev. Minimax weights in a trend detection problem of a random process. *Theory of Probability & Its Applications*, 16(2):344–349, 1971.

Jens Ludwig and Douglas L Miller. Does head start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly journal of economics*, 122(1):159–208, 2007.

Jordan D Matsudaira. Mandatory summer school and student achievement. *Journal of Econometrics*, 142(2):829–850, 2008.

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

Philip Oreopoulos. Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *The American Economic Review*, 96(1): 152–175, 2006.

John P Papay, John B Willett, and Richard J Murnane. Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161 (2):203–207, 2011.

Jack Porter. Estimation in the regression discontinuity model. 2003.

Sean F Reardon and Joseph P Robinson. Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness*, 5(1):83–104, 2012.

James Robins and Aad van der Vaart. Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253, 2006.

Paul R Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

Paul R Rosenbaum. *Observational Studies*. Springer, 2002.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Jerome Sacks and Donald Ylvisaker. Linear estimation for approximately linear models. *The Annals of Statistics*, pages 1122–1137, 1978.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, 51(6): 309–317, 1960.

William MK Trochim. *Research design for program evaluation: The regression-discontinuity approach*. Sage Publications, Inc, 1984.

Berwin A Turlach and A Weingessel. *quadprog: Functions to solve Quadratic Programming Problems.*, 2013. URL https://CRAN.R-project.org/package=quadprog. R package version 1.5-5.

Vivian C Wong, Peter M Steiner, and Thomas D Cook. Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, 38(2):107–141, 2013.

Tristan Zajonc. Regression discontinuity design with multiple forcing variables. In *Essays on Causal Inference for Public Policy*, pages 45–81. 2012.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

A Proof of Theorem 1

By construction, we already know that

$$\mathbb{E}\left[\hat{\tau} \,\middle|\, X_1, ..., X_n\right] - \tau(c) = b\left(\hat{\gamma}\right), \quad \text{Var}\left[\hat{\tau} \,\middle|\, X_1, ..., X_n\right] = s^2\left(\hat{\gamma}\right).$$

Thus, to establish our desired result, it suffices to establish asymptotic Gaussianity of

$$\hat{\tau} - \mathbb{E}\left[\hat{\tau} \,\middle|\, X_1, ..., X_n\right] = \sum_{i=1}^{n} \hat{\gamma}_i \varepsilon_i, \quad \varepsilon_i = Y_i - \mu_{W_i}(X_i).$$

To do so, we use the Lyapunov central limit theorem. Thanks to our bound on the q-th moment of the ε_i, we know that

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\gamma}_i \varepsilon_i\right)^q \,\middle|\, \hat{\gamma}_i\right] \Big/ s^q\left(\hat{\gamma}\right) &\leq \sum_{i=1}^{n} C \hat{\gamma}_i^q \Big/ \left(\sigma_{\min}^q \left\|\hat{\gamma}\right\|_2^q\right) \\ &\leq \frac{C}{\sigma_{\min}^q} \sup_{1 \leq i \leq n} \left\{\left|\hat{\gamma}_i\right| / \left\|\hat{\gamma}\right\|_2\right\}^{q-2} \to_p 0\end{aligned}$$

by assumption (7). Thus, in particular, there exists a sequence a_n such that $a_n \to 0$ and

$$\limsup_{n \to \infty} \; \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{\gamma}_i \varepsilon_i\right)^q\right] \Big/ s^q\left(\hat{\gamma}\right) \geq a_n \;\middle|\; \hat{\gamma}\right] = 0.$$

Now, define weights $\tilde{\gamma}_i$ such that $\tilde{\gamma} = \hat{\gamma}$ when the above event holds, and $\tilde{\gamma}_i = 1$ else for all $i = 1, ..., n$. By Lyapunov's central limit theorem, $\sum_{i=1}^{n} \tilde{\gamma}_i \varepsilon_i \,/\, (\sum_{i=1}^{n} \gamma_i^2 \sigma_i^2)^{1/2}$ is asymptotically standard normal. Moreover, we know that $\tilde{\gamma} = \hat{\gamma}$ with probability tending to 1, and so our estimator $\hat{\tau}$ must also be conditionally standard normal as claimed.

B Implementation via Dual Optimization

We start from the dual representation (13); however, to be able to deal with the multivariate case, we use two different f-functions for the treated and the controls separately (see the footnote at (11)). Writing $W(x)$ for whether individuals with $X_i = x$ are treated, we get the following problem:

$$\begin{aligned}\text{maximize} \; \inf_{\gamma,\, t} \Bigg\{ &\sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + B^2 t^2 + \lambda_1 \left(\sum_{i=1}^{n} \gamma_i f(X_i) - t\right) + \lambda_2 \left(\sum_{i=1}^{n} (1 - W(X_i))\gamma_i + 1\right) \\ &+ \lambda_3 \left(\sum_{i=1}^{n} W(X_i)\gamma_i - 1\right) + \lambda_4 \sum_{i=1}^{n} \gamma_i (X_i - c) + \lambda_5 \sum_{i=1}^{n} (W(X_i) - 1/2)\gamma_i (X_i - c)\Bigg\} \\ \text{subject to} \; &f(x) = (1 - W(x)) f_0(x) + W(x) f_1(x), \; \lambda_1 \geq 0, \; \lambda_2, \, \lambda_3 \in \mathbb{R}, \; \lambda_4, \, \lambda_5 \in \mathbb{R}^k, \\ &f_0(c) = 0, \; \nabla f_0(c) = 0, \; \left\|\nabla^2 f_0(x)\right\| \leq 1, \quad f_1(c) = 0, \; \nabla f_1(c) = 0, \; \left\|\nabla^2 f_1(x)\right\| \leq 1,\end{aligned}$$

where k is the number of running variables (i.e., $X_i \in \mathbb{R}^k$) Then, as discussed in Section 2.3, we can explicitly solve for γ and t, resulting in the following (we also turned the problem

into a minimization problem to follow convention):

$$\begin{aligned}
&\text{minimize } \frac{1}{4}\sum_{i=1}^{n}\sigma_i^{-2}G_i^2 + \frac{1}{4}\frac{\lambda_1^2}{B^2} - \lambda_2 + \lambda_3 \\
&\text{subject to } G_i = \lambda_1 f(X_i) + \lambda_2(1 - W(X_i)) + \lambda_3 W(X_i) \\
&\qquad\qquad + \lambda_4(X_i - c) + \lambda_5(W(X_i) - 1/2)(X_i - c) \\
&\quad f(x) = (1 - W(x))f_0(x) + W(x)f_1(x),\ \lambda_1 \geq 0,\ \lambda_2,\, \lambda_3 \in \mathbb{R},\ \lambda_4,\, \lambda_5 \in \mathbb{R}^k, \\
&\quad f_0(c) = 0,\ \nabla f_0(c) = 0,\ \left\|\nabla^2 f_0(x)\right\| \leq 1, \quad f_1(c) = 0,\ \nabla f_1(c) = 0,\ \left\|\nabla^2 f_1(x)\right\| \leq 1.
\end{aligned} \tag{20}$$

The above is not a standard-form quadratic program, since it has a product in one of the constraints, namely $\lambda_1 f(X_i)$. However, this problem can be fixed via mild reorganization,

$$\begin{aligned}
&\text{minimize } \frac{1}{4}\sum_{i=1}^{n}\sigma_i^{-2}G_i^2 + \tilde{\lambda}_1^2 - \lambda_2 + \lambda_3 \\
&\text{subject to } G_i = 2B\tilde{f}(X_i) + \lambda_2(1 - W(X_i)) + \lambda_3 W(X_i) \\
&\qquad\qquad + \lambda_4(X_i - c) + \lambda_5(W(X_i) - 1/2)(X_i - c) \\
&\quad \tilde{f}(x) = (1 - W(x))\tilde{f}_0(x) + W(x)\tilde{f}_1(x),\ \lambda_1 \geq 0,\ \lambda_2,\, \lambda_3 \in \mathbb{R},\ \lambda_4,\, \lambda_5 \in \mathbb{R}^k, \\
&\quad \tilde{f}_0(c) = 0,\ \nabla\tilde{f}_0(c) = 0,\ \left\|\nabla^2\tilde{f}_0(x)\right\| \leq \tilde{\lambda}_1, \quad \tilde{f}_1(c) = 0,\ \nabla\tilde{f}_1(c) = 0,\ \left\|\nabla^2\tilde{f}_1(x)\right\| \leq \tilde{\lambda}_1,
\end{aligned} \tag{21}$$

where we also reparametrized $\tilde{\lambda}_1 \leftarrow \lambda_1/(2B)$ and $\tilde{f}_\cdot \leftarrow \tilde{\lambda}_1 f_\cdot$. The first reparametrization isn't necessary but lets us avoid dividing by B in the objective, which is desirable since B can sometimes be quite small. Finally, we recover weights $\gamma_i = -\sigma_i^{-2}G_i/2$.