OPTIMIZED REGRESSION DISCONTINUITY DESIGNS

Guido Imbens and Stefan Wager*

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 nonparametric regression models on both sides of a treatment assignment boundary and then reporting plug-in estimates for the effect of interest. In applications, however, it is often difficult to tune the nonparametric regressions in a way that is well calibrated for the specific target of inference; for example, the model with the best global in-sample fit may provide poor estimates of the discontinuity parameter, which depends on the regression function at boundary points. 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 provides uniform confidence intervals for the regression discontinuity parameter with both discrete and continuous running variables. The method also naturally extends to the case of multiple running variables.

I. Introduction

REGRESSION discontinuity designs, first developed in the 1960s (Thistlethwaite & Campbell, 1960), often allow for simple and transparent identification of treatment effects from observational data (Hahn, Todd, & Van der Klaauw, 2001; Imbens & Lemieux, 2008; Trochim, 1984), and their statistical properties have been the subject of recent interest (Armstrong & Kolesár, 2018; Calonico, Cattaneo, & Titiunik, 2014; Cheng, Fan, & Marron, 1997; Kolesár & Rothe, 2018). The sharp regression discontinuity design assumes a treatment assignment generated by a running variable $X \in \mathbb{R}^k$, such that individuals get treated if and only $X \in \mathcal{A}$ for some set $\mathcal{A} \subset \mathbb{R}^k$. For example, in epidemiology, $X \in \mathbb{R}$ could be a severity index (e.g., age or CD4 count), and patients are assigned a medical intervention whenever $X \geq c$ for some threshold c (i.e., $\mathcal{A} = \{x \in \mathbb{R} : x \geq c\}$). In educational settings, $X \in \mathbb{R}$ could be a test score that has to exceed a threshold c, or, in political science, $X \in \mathbb{R}^2$ could denote the latitude and longitude of a household, and $\mathcal{A}$ could be an administrative region that has enacted a specific policy.

Given appropriate assumptions, we can identify a causal effect by comparing subjects i with X_i barely falling within the treatment region $\mathcal{A}$ to those with X_i just outside it. Variants of this identification strategy have proven to be useful in education (Angrist & Lavy, 1999; Black, 1999; Jacob & Lefgren, 2004), political science (Caughey & Sekhon, 2011; Keele & Titiunik, 2014; Lee, 2008), criminal justice (Berk & Rauma, 1983), program evaluation (Lalive, 2008; Ludwig & Miller, 2007), and other areas. As we discuss will in more detail, standard methods for inference in the regression discontinuity design rely on plug-in estimates from local linear regression.

In this paper, motivated by a large literature on minimax linear estimation (Armstrong & Kolesár, 2018; Cai & Low, 2003; Donoho, 1994; Donoho & Liu, 1991; Ibragimov & Khas'minskii, 1985; Johnstone, 2011; Juditsky & Nemirovski, 2009), we study an alternative approach based on directly minimizing finite sample error bounds via numerical optimization, under an assumption that the second derivative of the response surface is bounded away from the boundary of the treatment region.[1] This approach has several advantages relative to local regression. Our estimator is well defined regardless of the shape of the treatment region $\mathcal{A}$, whether it be a half line, as in the standard univariate regression discontinuity specification, or an oddly shaped region, as might appear with a geographic regression discontinuity; moreover, our implementation is not affected by potential discreteness of the running variable. Finally, even with univariate designs, our approach strictly dominates local linear regression in terms of minimax mean-squared error. We start by presenting our method in the context of classical univariate regression discontinuity designs with a single treatment cutoff: with $X_i \in \mathbb{R}$ and $\mathcal{A} = \{x \in \mathbb{R} : x \geq c\}$. A solution to the more general problem will then follow by direct extension. A software implementation, `optrdd` for `R`, is available on `CRAN`.

A. *Optimized Inference with Univariate Discontinuities*

We start with the simple setting where we have access to $i = 1, \ldots, n$ independent pairs (X_i, Y_i) where $X_i \in \mathbb{R}$ is the running variable and $Y_i \in \mathbb{R}$ is our outcome of interest; the treatment is assigned as $W_i = \mathbf{1}(\{X_i \geq c\})$. Following the potential outcomes model (Imbens & Rubin, 2015; Neyman, 1923; Rubin, 1974), we 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

[1] Of these papers, our work is most closely related to that of Armstrong and Kolesár (2018), who consider minimax linear estimation in the regression discontinuity model for an "approximately linear" model in the sense of Sacks and Ylvisaker (1978) that places restrictions on second differences relative to the response surface at the threshold. In contrast, we assume bounded second derivatives away from the threshold. An advantage of their approach is that it allows a closed-form solution for the weights. However, a disadvantage is that they allow for jumps in the response surface away from the threshold, which implies that given the same value for our bound on the second derivative and their bound on second differences, our confidence intervals can be substantially shorter (moreover, allowing for discontinuities in the response surface does not seem conceptually attractive given that the assumption of continuity of the conditional expectation at the threshold is fundamental to the regression discontinuity design). We discuss this comparison further in section IC.

Received for publication August 4, 2017. Revision accepted for publication April 2, 2018. Editor: Yuriy Gorodnichenko.

*Imbens and Wager: Stanford University.

We are grateful for helpful comments from Joshua Angrist, Timothy Armstrong, Max Farrell, Michal Kolesár, Christoph Rothe, and Cun-Hui Zhang; seminar participants at Berkeley, Heidelberg, Munich, Stanford and the University of Chicago; as well as the editor and four anonymous referees.

A supplemental appendix is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/rest_a_00793.

The Review of Economics and Statistics, May 2019, 101(2): 264–278

doi:10.1162/rest_a_00793

the conditional average treatment effect $\tau(x)$ in terms of the conditional response functions $\mu_w(x) = \mathbb{E}[Y_i(w) \mid X_i = x]$, such that $\tau(x) = \mu_1(x) - \mu_0(x)$. Provided the functions $\mu_w(x)$ are both continuous at c, the regression discontinuity identifies the conditional average treatment effect at the threshold c, which is the estimand in this setting:[2]

$$\tau(c) = \lim_{x \downarrow c} \mathbb{E}[Y_i \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i \mid X_i = x]. \tag{1}$$

Given this setup, local linear regression estimates $\tau(c)$ as (Hahn et al., 2001; Porter, 2003),

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

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; typically $K(\Delta)$ is taken to be 0 for Δ outside a bounded interval. When we do not observe data right at the boundary c (e.g., when X_i has discrete support), then $\tau(c)$ is not point identified. However, given smoothness assumptions on $\mu_w(x)$, Kolesár and Rothe (2018) propose an approach to local linear regression that can still be used to construct partial identification intervals for $\tau(c)$ in the sense of Imbens and Manski (2004); see section IIA for a discussion.

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 et al. (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 et al. (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 $|\mu_w''(x)| \leq B$, Armstrong and Kolesár (2018) and Kolesár and Rothe (2018) 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 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.[3] 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. Regardless of the shape of the kernel $K(\cdot)$ in equation (2), local linear regression yields a linear estimator[4] for τ, that is, one of the form $\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i$ for weights $\hat{\gamma}_i$ that depend only on the distances $X_i - c$ (the weights $\hat{\gamma}_i$ underlying local linear regression can be written out using the closed-form solution to ordinary least squares regression). Here, we find that if we are willing to rely on numerical optimization tools, we can derive better weights.

In order to derive the optimal estimator of the form $\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i$, we first consider the general properties of such estimators for a fixed set of weights $\hat{\gamma}_i$. The expected value of any such estimator, conditionally on the X_i and W_i, can be written as

$$\mathbb{E}\left[\sum_{i=1}^{n} \hat{\gamma}_i Y_i \mid \{X_i,\ W_i\}_{i=1}^{n} \right] = \sum_{\{i:W_i=1\}} \hat{\gamma}_i \mu_1(X_i) + \sum_{\{i:W_i=0\}} \hat{\gamma}_i \mu_0(X_i), \tag{3}$$

resulting in bias $\sum_{i:W_i=1} \hat{\gamma}_i \mu_1(X_i) + \sum_{i:W_i=0} \hat{\gamma}_i \mu_0(X_i) - (\mu_1(c) - \mu_0(c))$. Furthermore, given a set of weights $\hat{\gamma}_i$, the worst-case absolute bias over a class of functions $\mathcal{K}$ is

$$I_{\mathcal{K}}(\hat{\gamma}) = \sup_{\mu_0(\cdot), \mu_1(\cdot) \in \mathcal{K}} \left\{ \left| \sum_{\{i:W_i=1\}} \hat{\gamma}_i \mu_1(X_i) + \sum_{\{i:W_i=0\}} \hat{\gamma}_i \mu_0(X_i) - (\mu_1(c) - \mu_0(c)) \right| \right\}. \tag{4}$$

The conditional variance of any such estimator is $\sum_{i=1}^{n} \hat{\gamma}_i^2 \sigma_i^2$, with $\sigma_i^2 = \text{Var}\ Y_i \mid X_i$. Then, because the expected squared

[2] In the fuzzy regression discontinuity design where the probability of receiving the treatment changes discontinuously at $x = c$, but not necessarily from 0 to 1, the estimand can be written as the ratio of two such differences. The issues we address in this paper also arise in that setting, and our discussion extends naturally to it. See section V for a discussion.

[3] One inconvenience that can arise in local linear regression with discrete running variables is that if we use a data-driven rule to pick the bandwidth $\hat{h}$ (e.g., the one of Imbens & 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.

[4] We note the unfortunate terminological overlap between "local linear regression" estimators of τ and "linear" estimators of type $\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i$. The word *linear* in these two contexts refers to different things. All "local linear regression" estimators are "linear" but not vice versa.

error of an estimator depends on only its variance and squared bias, the minimax linear estimator for τ can be derived by minimizing the variance and worst-case bias terms above. In this paper, we focus on the case where $\mu_0(\cdot)$ and $\mu_1(\cdot)$ belong to the class of functions with the second derivative bounded by B, in which case the minimax linear estimator conditionally on the X_i and W_i is

$$\hat{\tau} = \sum_{i=1}^{n} \hat{\gamma}_i Y_i, \quad \hat{\gamma} = \text{argmin}_\gamma \left\{ \sum_{i=1}^{n} \gamma_i^2 \sigma_i^2 + I_B^2(\gamma) \right\},$$

$$I_B(\gamma) := \sup_{\mu_0(\cdot),\mu_1(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_{W_i}(X_i) - (\mu_1(c) - \mu_0(c)) : |\mu_w''(x)| \le B \text{ for all } w,\ x \right\}. \tag{5}$$

We note that because no limitations are placed on $\mu_w(c)$ or $\mu_w'(c)$, the optimization in equation (5) 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 $\sum_i (1 - W_i)(X_i - c)\hat{\gamma}_i = 0$. At values of γ for which these constraints are not all satisfied, we can choose $\mu_1(x)$ and $\mu_0(x)$ with second derivative bounded by B such as to make the conditional bias arbitrarily bad, that is, $I_B(\gamma) = +\infty$; thus, the solution $\hat{\gamma}$ to equation (5) must satisfy the constraints. The problem, equation (5) is a convex program and can be efficiently solved using readily available software described in, for example, Boyd and Vandenberghe (2004).

Because the estimator 5 is minimax among the class of linear estimators and local linear regression is also a linear estimator, our estimator dominates local linear regression estimator in a minimax sense over all problems where we only know that $\text{Var}[Y_i \mid X_i] = \sigma_i^2$ and $|\mu_w''(x)| \le B$. For further discussion of related estimators, see Armstrong and Kolesár (2018), Cai and Low (2003), Donoho (1994), Donoho and Liu (1991), and Juditsky and Nemirovski (2009).

In practice, of course, we need methods for choosing the values of σ_i^2 and B to run our method with. Tuning the noise scale σ_i^2 is not too difficult (it is comparable to estimating the irreducible noise in any regression problem); however, obtaining a good value for B usually requires problem-specific insight. We discuss some approaches to choosing B in the context of applications in sections III and IV, and recommend performing a sensitivity analysis for different choices of B. Specifying B is closely related to choice of bandwidth, for example, in existing methods such as those of Calonico et al. (2014) or Imbens and Kalyanaraman (2012); the difference is that choosing B directly reflects a quantitative belief about the world—in terms of regularity of the functions $\mu_w(x)$)—whereas a choice of bandwidth interacts with the fundamental parameters of the problem in a more indirect way, which depends on the shape of the kernel $K(\cdot)$ and the distribution of the running variable X_i.

When X_i has a discrete distribution, the parameter $\tau(c)$ is usually not point identified because there may not be any observations X_i in a small neighborhood of c. However, we can get meaningful partial identification of $\tau(c)$ thanks to our bounds on the second derivative of $\mu_w(x)$. Moreover, because our approach controls for bias in finite samples, the estimator (5) is still justified in the partially identified setting and, as discussed further in section IIA, provides valid confidence intervals for $\tau(c)$ in the sense of Imbens and Manski (2004). We view the fact that our estimator can seamlessly move between the point and partially identified settings as an important feature.

The top panel of figure 1 compares the weights $\hat{\gamma}_i$ obtained via equation (5) in two different settings: one with a discrete, asymmetric running variable X depicted the lower left panel of the figure, 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.

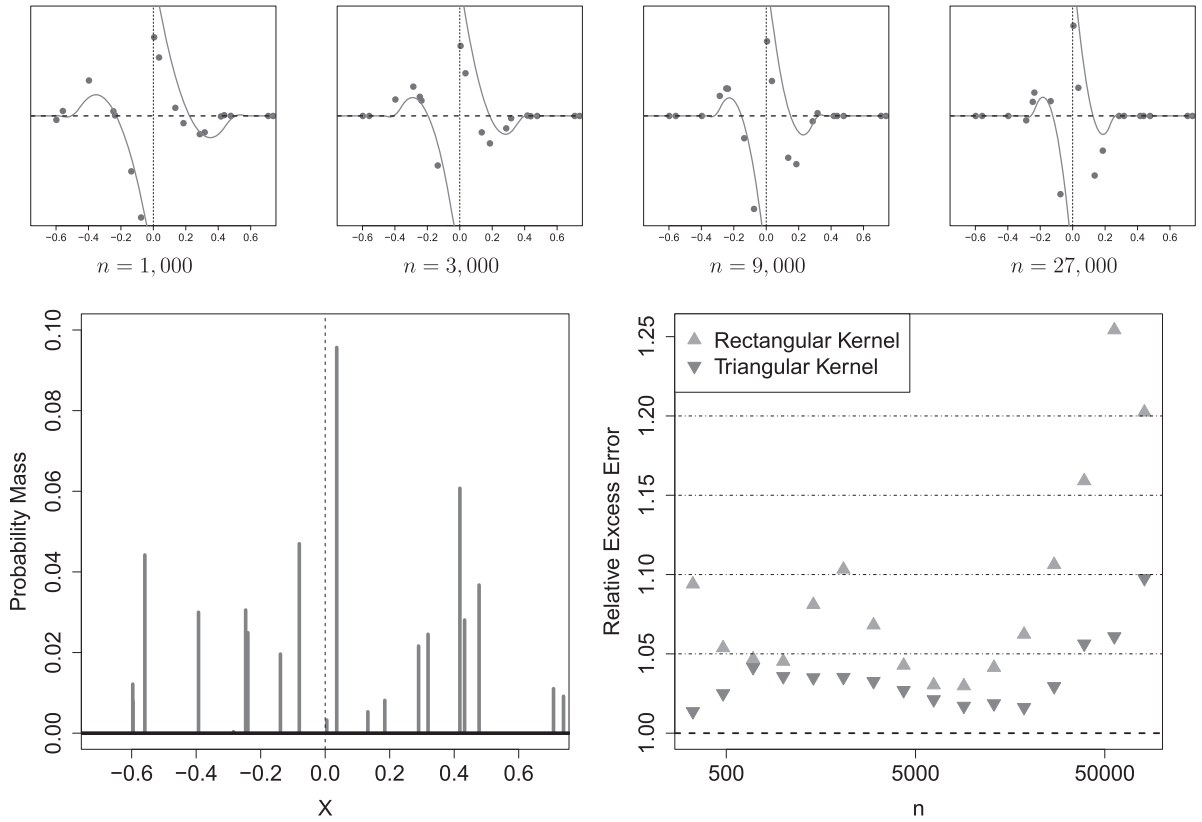
In the lower right panel of figure 1, we also compare the worst-case conditional mean-squared error of our method relative to that of optimally tuned local linear regression, both with a rectangular and triangular kernel; For the smallest sample size we consider, $n = 333$, the discreteness of the running variable has a fairly mild effect on estimation and, as one 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.[5]

B. *Optimized Inference with Generic Discontinuities*

The methods we have presented extend naturally to the general case, where $X_i \in \mathbb{R}^k$ may be multivariate and $\mathcal{A}$ is unrestricted. 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 along the boundary of the treated region $\mathcal{A}$. Most of the literature on this setup, including Papay, Willett, and Murnane (2011), Reardon and Robinson (2012), and Wong, Steiner, and Cook (2013), have

[5] 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 (in the above example, we used the same bandwidth for local linear regression on both sides of the boundary). However, from a practical perspective, estimator, equation (5), removes the need to consider such questions in applied data analysis and automatically adapts to the structure of the data at hand.

FIGURE 1.—EXAMPLE WITH DISCRETE RUNNING VARIABLE



(Top) Optimized regression discontinuity design obtained via equation (5) 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 lower left panel). 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 $\sigma^2 = 1$. (Bottom left) Probability mass function of the running variable. (Bottom right) Comparison of our procedure, equation (5), 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, equation (5).

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 multivariate setting by, for example, using shortest distance to the boundary of the treatment region $\mathcal{A}$ as a univariate running variable (Black, 1999), or considering only a subset of the data where univariate methods are appropriate (Jacob & Lefgren, 2004; Matsudaira, 2008).

Here, we show how our optimization-based method can be used to sidestep the problem of choosing a multivariate kernel function by hand. In addition to providing a simple-to-apply algorithm, our method lets us explicitly account for the curvature of the mean-response function $\mu_w(x)$ for statistical inference, thus strengthening formal guarantees relative to prior work.

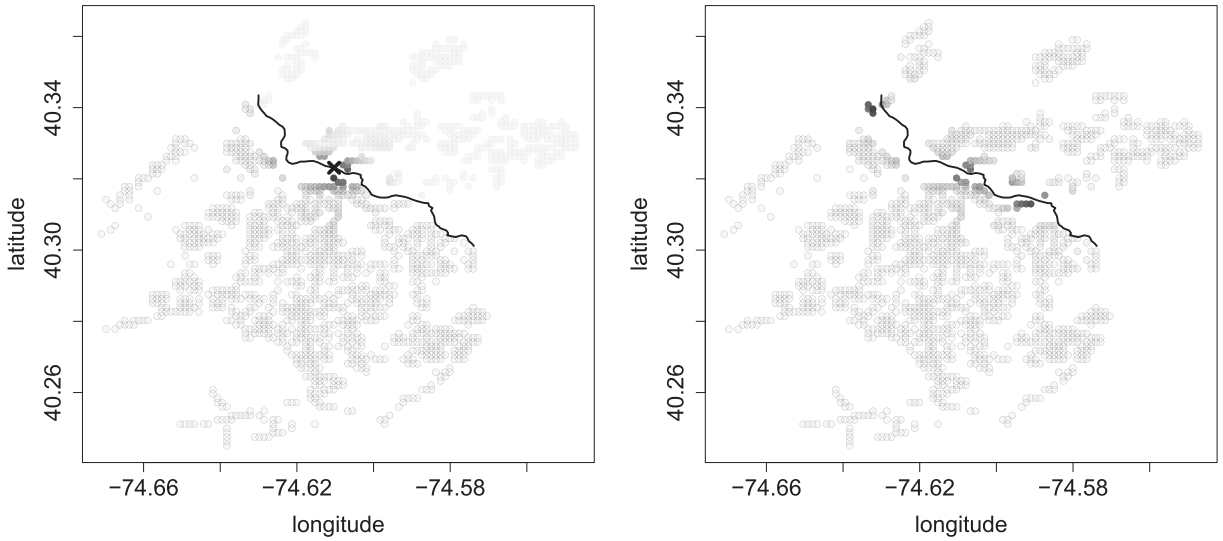
Relative to the univariate case, the multivariate case has two additional subtleties we need to address. First, in equation (5), it is natural to impose a constraint $|\mu_w''(x)| \leq B$ to ensure smoothness; in the multivariate case, however, we have more choices to make. For example, do we constrain $\mu_w(x)$ to be an additive function, or do we allow for interactions? Here, we opt for the more flexible specification and simply require that $||\nabla^2 \mu_w(x)|| \leq B$, where $||\cdot||$ denotes the operator norm (i.e., the largest absolute eigenvalue of the second derivative).

Moreover, as Papay et al. (2011), emphasized, 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 treatment effect functionals. Here, we focus on the following two causal estimands. First, writing c for a focal point of interest, we can directly generalize the estimator, equation (5), as $\hat{\tau}_c = \sum_{i=1}^n \hat{\gamma}_{c,i} Y_i$ with

$$\hat{\gamma}_c = \text{argmin}_\gamma \left\{ \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + \left(\sup_{||\nabla^2 \mu_w(x)|| \leq B} \left\{ \sum_{i=1}^n \gamma_i \mu_{W_i}(X_i) - (\mu_1(c) - \mu_0(c)) \right\} \right)^2 \right\}. \quad (6)$$

This is the minimax linear estimator for the conditional average treatment effect at c. The upside of this approach is that it

FIGURE 2.—WEIGHTING FUNCTION FOR A GEOGRAPHIC REGRESSION DISCONTINUITY DESIGN



Points depict potential voters within a single school district, and the solid black line is a media market boundary. The left panel depicts an optimal weighting function for the conditional average treatment effect at the point c marked with a bold $\times$ as in equation (6), while the right one allows for a weighted treatment effect as in equation (7) or, equivalently, shows the optimal weighting function for a constant effect. Households below the line are treated (i.e., in the Philadelphia media market), whereas those above it are controls (i.e., in the New York media market). The color of the point depicts the γ-weight: red points receive positive weight, and blue points receive negative weight; the shading indicates the absolute value of the weight (darker is larger).

gives us an estimand that is easy to interpret; the downside is that the when curvature is nonnegligible, equation (6) can effectively make use of only data near the specified focal point c, thus resulting in relatively long confidence intervals.

In order to potentially improve precision, we also study weighted conditional average treatment effect estimation with weights greedily chosen such as to make the inference as precise as possible. In the spirit of Crump et al. (2009), Li, Morgan, and Zaslavsky (2018), or Robins et al. (2008), we consider $\hat{\tau}_* = \sum_{i=1}^n \hat{\gamma}_{*,i} Y_i$, with

$$\hat{\gamma}_* = \text{argmin}_\gamma \left\{ \sum_{i=1}^n \gamma_i^2 \sigma_i^2 + \left(\sup_{\|\nabla^2 \mu_0(x)\| \le B} \left\{ \sum_{i=1}^n \gamma_i \mu_0(X_i) \right\} \right)^2 : \sum_{i=1}^n \gamma_i W_i = 1 \right\}. \quad (7)$$

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

$$|\mathbb{E}[\hat{\tau}_* \mid \{X_i\}] - \bar{\tau}(\hat{\gamma}_*)| \le \sup_{\|\nabla^2 \mu_0(x)\| \le B} \left\{ \sum_{i=1}^n \hat{\gamma}_{*,i} \mu_0(X_i) \right\},$$

$$\bar{\tau}(\hat{\gamma}_*) := \sum_{i=1}^n W_i \hat{\gamma}_{*,i} \tau(X_i). \quad (8)$$

Because $\sum W_i \hat{\gamma}_{*,i} = 1$, we see that $\bar{\tau}(\hat{\gamma}_*)$ 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 $x_* = \sum \hat{\gamma}_{*,i} W_i X_i$.

In some cases, it is helpful to consider other interpretations of the estimand underlying equation (7). If we are willing to assume a constant treatment effect $\tau(x) = \tau$, then $\bar{\tau} = \tau$, and $\hat{\tau}_*$ is the minimax linear estimator for τ. Relatedly, we can always use the confidence intervals from section IIA built around $\hat{\tau}_*$ to test the global null hypothesis $\tau(x) = 0$ for all x.

To gain intuition for the multivariate version of our method, we outline a simple example building on the work of Keele and Titiunik (2014) on the effect of television advertising on voter turnout in presidential elections. To estimate this effect, Keele and Titiunik (2014) examine a school district in New Jersey, half of which belongs to the Philadelphia media market and the other half to the New York media market. Before the 2008 presidential elections, the Philadelphia half was subject to heavy campaign advertising, whereas the New York half was not, thus creating a natural experiment for the effect of television advertising assuming the media market boundary did not coincide with other major boundaries within the school district. Keele and Titiunik (2014) use this identification strategy to build a regression discontinuity design, comparing sets of households straddling the media market boundary.

However, despite the multivariate identification strategy, Keele and Titiunik (2014) then reduce the problem to a univariate regression discontinuity problem for estimation. They first compute Euclidean distances $D_i = \|X_i - c_2\|$ to a focal point c and then use D_i as a univariate running variable. In contrast, our approach allows for transparent inference without needing to rely on such a reduction. Figure 2 depicts $\hat{\gamma}$-weights generated by our optimized approach; the resulting

treatment effect estimator is then $\sum_i \hat{\gamma}_i Y_i$. Qualitatively, we replicate the "no measurable effect" finding of Keele and Titiunik (2014) while directly and uniformly controlling for spatial curvature effects. We discuss details, including placebo diagnostics and the choice of tuning parameter, in our working paper (Imbens & Wager, 2017).

We also see that, at least here, standard heuristics used to reduce the multivariate regression discontinuity problem to a univariate one are not sharp. In the setup of the left panel of figure 2, where we seek to estimate the treatment effect at a focal point c, some treated points due west of c get a positive weight, whereas points the same distance south from c get a mildly negative weight, thus violating the heuristic of Keele and Titiunik (2014) that weights should depend only on $D_i = \|X_i - c_2\|$. Meanwhile, we can compare the approach in the right panel of figure 2, where we allow for averaging of treatment effects along the boundary, to the popular heuristic of using shortest distance to the boundary of the treatment region as a univariate running variable (Black, 1999). But this reduction does not capture the behavior of our optimized estimator. There are some points at the eastern edge of the treated region that are very close to the boundary but get essentially 0 weight (presumably because there are no nearby units on the control side of the boundary).

C. Related Work

The idea of constructing estimators of the type 5 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). For a broader discussion of minimax linear estimation over nonparametric function classes, see Cai and Low (2003), Donoho (1994), Ibragimov and Khas'minskii (1985), Johnstone (2011), Juditsky and Nemirovski (2009), and references in them. An important result in this literature is that for many problems of interest, minimax linear estimators are within a small explicit constant of being minimax among all estimators (Donoho & Liu, 1991).

Armstrong and Kolesár (2018) apply these methods to regression discontinuity designs, resulting in an estimator of the form 5, except with weights:[6]

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

$$A_B(\gamma) = \sup_{\mu_0(\cdot),\mu_1(\cdot)} \left\{ \sum_{i=1}^{n} \gamma_i \mu_{W_i}(X_i) - \tau(c) : |\mu_w(x) - \mu_w(c) - \mu_w'(c)(x-c)| \le \frac{B}{2}(x-c)^2 \right\}. \quad (9)$$

Now, although this class of functions is cosmetically quite similar to the bounded-second-derivative class used in equation (5), we note that the class of weights allowed for in equation (9) 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 in fact can have jumps of magnitude $B(x-c)^2/2$. Given that the key assumption underlying regression discontinuity designs 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 jumps through the condition (9) 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 (9) rather than our proposed one, equation (5), appears to be that the optimization problem induced by equation (9) is substantially easier and allows for closed-form solutions for $\hat{\gamma}_i$. Conversely, we are aware of no closed-form solution for equation (5) and instead need to rely on numeric convex optimization.

In the special case where the running variable X is assumed to have a continuous density around the threshold c, there have been a considerable number of recent proposals for asymptotic confidence intervals while imposing smoothness assumptions on $\mu_w(x)$. Calonico et al. (2014) propose a bias correction to the local linear regression estimator that allows for valid inference, and Calonico, Cattaneo, and Farrell (2018) provide further evidence that such bias corrections may be preferable to undersmoothing. Meanwhile, Armstrong and Kolesár (2016) show that when $\mu_w(x)$ is twice differentiable and X has a continuous density around c, we can use local linear regression with a bandwidth chosen to optimize mean-squared error as the basis for bias-adjusted confidence intervals, provided we inflate confidence intervals by an appropriate, universal constant (e.g., to build 95% confidence intervals, one should use a critical threshold of 2.18 instead of 1.96). Gao (2017) characterizes the asymptotically optimal kernel for the regression discontinuity parameter under the bounded second derivative assumption with a continuous running variable. As discussed above, the value of our approach relative to this literature is that we allow for considerably more generality in the specification of the regression discontinuity design: the running variable X may be discrete or multivariate, and the treatment boundary may be irregularly shaped.

Optimal inference with multiple running variables is less developed than in the univariate case. Papay et al. (2011) and Reardon and Robinson (2012) study local linear regression

[6] Armstrong and Kolesár (2018) also consider a more general setting where we assume accuracy of the kth order Taylor expansion of $\mu_w(x)$ around c; 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.

with a "small" bandwidth, but do not account for finite sample bias due to curvature. 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. Keele, Titiunik, and Zubizarreta (2015) consider an approach to geographic regression discontinuity designs based on matching. To our knowledge, the approach we present is the first to allow for uniform, bias-adjusted inference in the multivariate regression discontinuity setting.

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 $\mathbb{E}[Y_i \mid X_i = x]$ using a global polynomial expansion along with a jump at c (see Lee & Lemieux, 2010, for a review and examples. However, as Gelman and Imbens (forthcoming) argued, this approach is not recommended, as the model with the best in-sample fit may provide poor estimates of the discontinuity parameter. For example, high-order polynomials may give large influence to samples i for which X_i is far from the decision boundary c, and thus lead to unreliable performance.

Another approach to regression discontinuity designs (including in the discrete case) builds on randomization inference (see Cattaneo, Frandsen, & Titiunik, 2015; Cattaneo, Idrobo, & Titiunik, forthcoming; and Li, Mattei, & Mealli, 2015, for a discussion). The problem of specification testing for regression discontinuity designs is considered by Cattaneo, Jansson, & Ma (2016), Frandsen (2017), and McCrary (2008).

II. Formal Considerations

A. *Uniform Asymptotic Inference*

Our main result verifies that optimized designs can be used for valid asymptotic inference about $\tau(c)$. We here consider the problem of estimating a conditional average treatment effect at a point c as in equation (5) or (6); similar arguments extend directly to the averaging case as in (7). Following, for example, Robins and van der Vaart (2006), we seek confidence intervals $\mathcal{I}_\alpha$ that attain uniform coverage over the whole regularity set under consideration:

$$\liminf_{n\to\infty} \inf\{\mathbb{P}\left[\mu_1(c) - \mu_0(c) \in \mathcal{I}_\alpha\right] : \|\nabla^2 \mu_w(x)\| \le B \text{ for all } w,\ x\} \ge 1 - \alpha. \tag{10}$$

As in Armstrong and Kolesár (2018), 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 (all proofs are available in our working paper: Imbens & Wager, 2017):

Theorem 1. *Suppose that we have a moment bound* $\mathbb{E}[(Y_i - \mathbb{E}[Y_i \mid X_i])^q \mid X_i = x] \le C$ *uniformly over all* $x \in \mathbb{R}^k$*, for some exponent* $q > 2$ *and constant* $C \ge 0$*. Suppose, moreover, that* $0 < \sigma_{min} \le \sigma_i$ *for all* $i = 1, \ldots, n$ *for a deterministic value* σ_{min}*, and that none of the weights* $\hat{\gamma}_i$ *derived in equations (5) or (6) dominates all the others:*[7]

$$\max_{1\le i\le n} \left\{\hat{\gamma}_i^2\right\} \Big/ \sum_{i=1}^n \hat{\gamma}_i^2 \to_p 0. \tag{11}$$

Then our estimator $\hat{\tau}$ *from equations (5) or (6) is asymptotically gaussian,*

$$(\hat{\tau} - b(\hat{\gamma}) - \tau(c)) / s(\hat{\gamma}) \Rightarrow \mathcal{N}(0,\ 1),$$
$$b(\hat{\gamma}) = \sum_{i=1}^n \hat{\gamma}_i \mu_{W_i}(X_i) - \tau(c), \quad s^2(\hat{\gamma}) := \sum_{i=1}^n \hat{\gamma}_i^2 \sigma_i^2, \tag{12}$$

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

In solving the optimization problem, equation (5), we also obtain an explicit bound $\hat{t}$ on the conditional bias, $b(\hat{\gamma}) \le \hat{t}$, and so can use the following natural construction to obtain confidence intervals for (Imbens and Manski, 2004). In large samples, $\mathcal{I}_\alpha$ is a uniform level-α confidence interval for $\tau(c)$:

$$\mathcal{I}_\alpha = \hat{\tau} \pm l_\alpha, \quad l_\alpha = \min\{l : \mathbb{P}\left[|b + s(\hat{\gamma})Z| \le l\right] \ge \alpha \text{ for all } |b| \le \hat{t}\}, \quad Z \sim \mathcal{N}(0,\ 1). \tag{13}$$

These confidence intervals are asymptotically uniformly valid in the sense of equation (10): for any $\alpha' < \alpha$, there is a threshold $n_{\alpha'}$ for which, if $n \ge n_{\alpha'}$, the confidence intervals, equation (13), achieve α'-level coverage of $\tau(c)$ for any functions $\mu_w(\cdot)$ in our regularity class.

Finally, whenever X_i does not have support arbitrarily close to c (e.g., in the case where X_i has a discrete distribution), the parameter $\tau(c)$ is not point identified. Rather, even with infinite data, the strongest statement we could make is that

$$\tau(c) \in \mathcal{I}^*,$$
$$\mathcal{I}^* = \text{range}\{\mu_{(1)}(c) - \mu_{(0)}(c) : \|\nabla^2 \mu_{(w)}(x)\| \le B, \text{ and}$$
$$\mu_{(w)}(x) = \mathbb{E}\left[Y_i \mid X_i = x,\ W_i = w\right] \text{ for all } (x, w) \in \text{supp}\{X_i,\ W_i\}, \tag{14}$$

where $\text{supp}\{X_i,\ W_i\}$ denotes the support of $(X_i,\ W_i)$. In this case, our confidence intervals, equation (13), remain valid for

[7] The bound on the relative contribution of any single $\hat{\gamma}_i$ is needed to obtain a Gaussian limit distribution for $\hat{\tau}$. In related literature, Armstrong and Kolesár (2018) follow Donoho (1994) and sidestep this issue by assuming Gaussian errors $Y_i(w) - \mu_w(X_i)$, in which case no central limit theorem is needed. Conversely, Athey, Imbens, & Wager (2018) adopt an approach more similar to ours and explicitly bound $\hat{\gamma}_i$ from above during the optimization.

$\tau(c)$; however, they may not cover the whole optimal identification interval $\mathcal{I}^*$. In partially identified settings, these types of confidence intervals (ones that cover the parameter of interest but not necessarily the whole identification interval) are advocated by Imbens and Manski (2004). From the perspective of the practitioner, an advantage of our approach is that intervals for $\tau(c)$ have the same interpretation whether or not $\tau(c)$ is point identified, that is, uniformly in large samples, $\tau(c)$ will be covered with probability $1-\alpha$. Then, asymptotically, intervals (13) will converge to a point if and only if $\tau(c)$ is point identified. (For a further discussion of regression discontinuity inference with discrete running variables, see Kolesár & Rothe, 2018).

B. *Implementation via Convex Optimization*

In our presentation so far, we have discussed several nonparametric convex optimization problems and argued that solving them was feasible given advances in the numerical optimization literature over the past few decades (Boyd & Vandenberghe, 2004). Here, we present a concrete solution strategy via quadratic programming over a discrete grid and show that the resulting discrete solutions converge uniformly to the continuous solution as the grid size becomes small.[8]

To do so, we start by writing the optimization problems underlying equations (5), (6), and (7) in a unified form. Given a specified focal point c, we seek to solve

$$\begin{aligned}
&\underset{\gamma,\,t}{\text{minimize}} \ \sum_{i=1}^{n}\gamma_i^2\sigma_i^2 + B^2t^2 \text{ subject to}\\
&\sum_{i=1}^{n}\gamma_i(f_0(X_i)+\psi\, w(X_i)(f_1(X_i)-f_0(X_i))) \le t,\\
&\quad \text{for all } f_w(c)=0,\ \nabla f_w(c)=0,\\
&\quad \|\nabla^2 f_w(x)\| \le 1 \text{ with } w\in\{0,\,1\}\\
&\sum_{i=1}^{n} w(X_i)\gamma_i = 1,\ \ \sum_{i=1}^{n}(1-w(X_i))\gamma_i = -1,\\
&\sum_{i=1}^{n}\gamma_i(X_i-c)=0,\\
&\psi\sum_{i=1}^{n}(2w(X_i)-1)\gamma_i(X_i-c)=0, \qquad (15)
\end{aligned}$$

where $w(x)$ denotes the treatment assignment function and ψ lets us toggle between different problem types. If we want to estimate the conditional average treatment effect (CATE) at c as in equation (6) we set $\psi=1$, whereas if we want an optimally weighted CATE estimator as in equation (7), we set $\psi=0$.

To further characterize the solution to this problem, we can use Slater's constraint qualification (e.g., Ponstein, 2004, theorem 3.11.2) to verify that strong duality holds and that the optimum of equation (15) matches the optimum of the following problem:

$$\begin{aligned}
&\underset{f(\cdot),\,\lambda}{\text{maximize}} \ \inf_{\gamma,\,t}\Bigg\{\sum_{i=1}^{n}\gamma_i^2\sigma_i^2 + B^2t^2\\
&\quad+\lambda_1\Bigg(\sum_{i=1}^{n}\gamma_i(f_0(X_i)+\psi\, w(X_i)(f_1(X_i))-f_0(X_i))-t\Bigg)\\
&\quad+\lambda_2\Bigg(\sum_{i=1}^{n}\gamma_i w(X_i)-1\Bigg)+\lambda_3\Bigg(\sum_{i=1}^{n}\gamma_i(1-w(X_i))+1\Bigg)\\
&\quad+\sum_{i=1}^{n}\gamma_i(\lambda_4+\psi\,\lambda_5(2w(X_i)-1))(X_i-c)\Bigg\}\\
&\text{subject to } f_w(c)=0,\ \nabla f_w(c)=0,\\
&\quad\|\nabla^2 f_w(x)\|\le 1 \text{ for } w\in\{0,\,1\},\\
&\quad\lambda_1,\ge 0,\ \lambda_2,\ \lambda_3\in\mathbb{R},\ \lambda_4,\ \lambda_5\in\mathbb{R}^k, \qquad (16)
\end{aligned}$$

where k is the number of running variables. Here, we also implicitly used von Neumann's minimax theorem to move the maximization over f outside the $\inf_{\gamma,t}\{\}$ statement.

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 and t in the dual objective, for example,

$$\begin{aligned}
-2\sigma_i^2\hat{\gamma}_i = {}&\hat{\lambda}_1(\hat{f}_0(X_i)+\psi\, w(X_i)(\hat{f}_1(X_i)-\hat{f}_0(X_i)))\\
&+\hat{\lambda}_2 w(X_i)+\ldots, \qquad (17)
\end{aligned}$$

where $\hat{f}_0(\cdot)$, $\hat{f}_1(\cdot)$, $\hat{\lambda}_1$, and so forth are the maximizers of equation (16). Carrying out the substitution results in a more tractable optimization problem over the space of twice-differentiable functions f, along with a finite number of Lagrange parameters λ_j:

$$\begin{aligned}
&\underset{\tilde{f}(\cdot),\,\lambda}{\text{minimize}} \ \frac{1}{4}\sum_{i=1}^{n}\frac{G_i^2}{\sigma_i^2}+\frac{1}{4}\frac{\lambda_1^2}{B^2}+\lambda_2-\lambda_3\\
&\text{subject to } G_i=\tilde{f}(X_i)+\lambda_2 w(X_i)+\lambda_3(1-w(X_i))\\
&\qquad\qquad+\lambda_4(X_i-c)+\psi\,\lambda_5(2w(X_i)\\
&\qquad\qquad-1)(X_i-c)\\
&\tilde{f}(x)=\tilde{f}_0(x)+\psi\, w(x)\big(\tilde{f}_1(x)-\tilde{f}_0(x)\big),\\
&\lambda_1\ge 0,\ \lambda_2,\ \lambda_3\in\mathbb{R},\ \lambda_4,\ \lambda_5\in\mathbb{R}^k,
\end{aligned}$$

[8] In the case where X is univariate, the resulting optimization problem is a classical one, and arguments made by Karlin (1973) imply that the weights $\hat{\gamma}_i$ can be written as $\hat{\gamma}_i = g(X_i)$, where g is a perfect spline; our proposed optimization strategy reflects this fact. However, in the multivariate case, we are not aware of a similar simple characterization.

$$\tilde{f}_w(c) = 0,\ \nabla\tilde{f}_w(c) = 0,$$
$$\|\nabla^2\tilde{f}_w(x)\| \le \lambda_1 \text{ for } w \in \{0,\ 1\}, \tag{18}$$

where $\tilde{f}_w(x)$ in the above problem corresponds to $\lambda_1 f_w(x)$ in equation (16), and we can recover our weights of interest via $\hat{\gamma}_i = -\sigma_i^{-2}\widehat{G}_i/2$ and $\hat{t} = \hat{\lambda}_1/(2B^2)$. The upshot of these manipulations is that equation (18) can be approximated via a finite-dimensional quadratic program. In our software implementation `optrdd`, we use this type of a finite-dimensional approximation to obtain $\hat{\gamma}_i$ following the construction described in the proof of proposition 2, available in our working paper (Imbens & Wager, 2017).

Proposition 2. *Suppose that $X_i \in \mathcal{X}$ belong to some compact, convex set $\mathcal{X} \subset \mathbb{R}^k$. Then for any tolerance level $\eta > 0$, there exists a finite-dimensional quadratic program that can recover the solution $\hat{\gamma}$ to equation (18) with L_2-error at most η.*

C. Minimizing Confidence Interval Length

As formulated in equation (5), 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 (13) as short as possible; and our approach can easily be adapted for this objective. To do so, consider the minimization objective in equation (15). Writing $\hat{\upsilon}^2 = \sum_{i=1}^n \hat{\gamma}_i^2\sigma_i^2$, we see that both the worst-case mean-squared error, $\hat{\upsilon}^2 + B^2\hat{t}^2$, and the confidence interval length in equation (13) are monotone increasing functions of $\hat{\upsilon}$ and $\hat{t}$; the only difference is in how they weight these two quantities at the optimum.

Now, to derive the full Pareto frontier of pairs $(\hat{\upsilon},\ \hat{t})$, we can simply rerun equaiton (15) 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 equation (15) and then using the value of λ that yields the best intervals; this construction provides minimax linear fixed-length confidence intervals (Donoho, 1994). 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 equation (15), and so did not pursue this line of investigation further. This observation is in line with the analytic results of Armstrong and Kolesár (2016), who showed that when X has a continuous density and $\mu_w(x)$ is twice differentiable, using the mean-squared error optimal bandwidth for local linear regression is over 99% efficient relative to using a bandwidth that minimizes the length of bias-adjusted confidence intervals.

Finally, although it is beyond the scope of this paper, it is interesting to ask whether we can generalize our approach to obtain asymptotically quasi-minimax estimators for $\tau(c)$ when the per observation noise scale σ_i needs to be estimated from the data. The resulting question is closely related to the classical issue of when feasible generalized least squares can emulate generalized least squares (see Romano and Wolf (2017) for a recent discussion).

III. Univariate Optimized Designs in Practice

To use this result in practice, we need to estimate the sum $\sum \hat{\gamma}_i^2\sigma_i^2$ and choose a 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 equation (5), 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:

$$\hat{s}^2(\hat{\gamma}) = \sum_{i=1}^n \hat{\gamma}_i^2\left(Y_i - \hat{\mu}_{W_i}(X_i)\right)^2,$$
$$\hat{s}^2(\hat{\gamma})\Big/\sum \hat{\gamma}_i^2\sigma_i^2 \ge 1 - o_P(1). \tag{19}$$

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 (White, 1980). We summarize the resulting method as procedure 1. We always encourage plotting the weights $\hat{\gamma}_i$ against X_i when applying our method:

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 equation (5). This procedure is implemented in our `R` package `optrdd`.[9]

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 $|X_i - c| \le r$. Let $\hat{\sigma}^2$ be the residual error from this regression.

[9]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, for example, some pairs outcomes Y_i are actually aggregated over many observations, then we should run steps 2 and 3 using appropriate inverse-variance weights. Our software allows for such weighting.

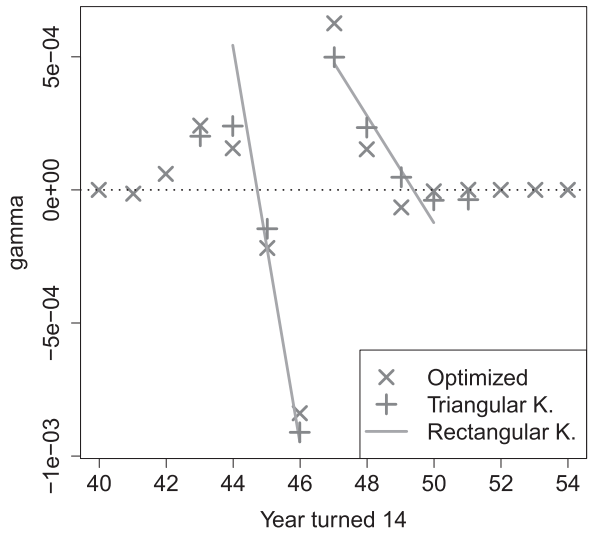
3. Obtain $\hat{\gamma}$ via the quadratic program, equation (15), with σ_i set to $\hat{\sigma}$ and weights outside the range $|X_i - c| \leq r$ set to 0.
4. Confirm that the optimized weights $\hat{\gamma}_i$ are small for $|X_i - c| \approx r$. If not, start again with a larger value of r.[10]
5. 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 equation (13).

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 priori bounds for B is not always possible (see Armstrong & Kolesár, 2018 and Bertanha & Moreira, 2016, and references there). In applications, we recommend considering a range of plausible values of B that could be obtained, for example, 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. In practice, it is often helpful to conduct a sensitivity analysis for the robustness of confidence intervals to changing B (see figure 4 for an example).[11]

A. *Application: The Effect of Compulsory Schooling*

In our first application, we consider a data set from Oreopoulos, 2006, who studied the effect of raising the minimum school-leaving age in the United Kingdom on earnings as an adult. The effect is identified by the change in the minimum school-leaving age from 14 to 15 in 1947, and the response is log-earnings among those with nonzero earnings (in 1998 pounds sterling). This data set exhibits notable discreteness in its running variable and was used by Kolesár and Rothe (2018) to illustrate the value of their bias-adjusted confidence intervals for discrete regression discontinuity designs. For our analysis, we preprocess our data exactly as in Kolesár and Rothe (2018). We refer readers to their paper and to Oreopoulos, 2006 for a more in-depth discussion of the data.

FIGURE 3.—COMPARISON OF WEIGHTING FUNCTIONS



Weighting functions $\hat{\gamma}(X_i)$ produced explicitly by our estimator, equation (5), 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 $B = 0.012$.

As in Kolesár and Rothe (2018), 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 (2018) consider analysis using local linear regression with a rectangular kernel and a bandwidth chosen such as 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.[12]

In order to obtain confidence intervals, it remains to choose a bound B. Following the above discussion, 3, a second-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); thus, we tried $B = 0.006$ and $B = 0.012$. We also consider the more extreme choices $B = 0.003$ and $B = 0.03$. For σ_i^2, we proceed as discussed above. Figure 3 shows the effective $\hat{\gamma}(X_i)$ weighting functions for all three considered methods, with $B = 0.012$.

We present results in the top panel of table 1. Overall, these results are in line with those presented in figure 1. The optimized method yields materially shorter confidence intervals than local linear regression with a rectangular kernel;

[10] 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 nonparametric function over the whole range of x, and being able to reduce the relevant range of x a priori reduces the required computation. 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 IIIA, the optimal bandwidth for local linear regression is around three or six years depending on the amount of assumed smoothness (and choosing a good bandwidth is very important); conversely, using plausibility intervals extending ten, fifteen, or twenty years on both sides of c appears to work reasonably well. When running the method (5), one should always 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.

[11] An interesting wrinkle is that if we are able to bound B in large samples—but not uniformly—then confidence intervals built using estimated values of B will have asymptotic but not uniform coverage.

[12] Oreopoulos (2006) analyzes the data set using a global polynomial specification with clustered random variables, following Lee and Card (2008). However, as Kolesár and Rothe (2018) discussed in detail, this approach does not yield valid confidence intervals.

TABLE 1.—NUMERICAL ESTIMATES OF TREATMENT EFFECTS

A. Effect of Raising Minimum School-Leaving Age

B	Rectangular Kernel	Triangular Kernel	Optimized
0.003	0.0213 ± 0.0761	0.0321 ± 0.0737	0.0302 ± 0.0716
0.006	0.0578 ± 0.0894	0.0497 ± 0.0867	0.0421 ± 0.0841
0.012	0.0645 ± 0.1085	0.0633 ± 0.1037	0.0557 ± 0.1003
0.03	0.0645 ± 0.1460	0.0710 ± 0.1367	0.0710 ± 0.1329

B. Effect of Summer School on Math and Reading Scores

	Estimator	Unweighted CATE Equation (6)			Weighted CATE Equation (7)		
Subject	B	Confidence Interval	Maximum Bias	Sample Error	Confidence Interval	Maximum Bias	Sample Error
Math	0.5×40^{-2}	0.037 ± 0.093	0.030	0.038	0.076 ± 0.037	0.009	0.017
Math	1.0×40^{-2}	0.013 ± 0.126	0.041	0.052	0.068 ± 0.043	0.011	0.019
Reading	0.5×40^{-2}	0.014 ± 0.098	0.030	0.041	0.044 ± 0.037	0.009	0.017
Reading	1.0×40^{-2}	-0.015 ± 0.130	0.040	0.054	0.047 ± 0.043	0.011	0.019

A. 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, equation (5). The confidence intervals account for curvature effects, provided the second derivative is bounded by B. B. 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 $\{X_i\}$.

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 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, equation 5.

Finally, we note that a bound B on the second derivative also implies that the quadratic approximation, equation (9), holds with the same bound B. Thus, we could in principle also use the method of Armstrong and Kolesár (2018) to obtain uniform asymptotic confidence intervals here. However, the constraint, equation (9), is weaker than the actual assumption we were willing to make—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 data set gives confidence intervals of 0.0518 ± 0.0969 with $B = 0.006$ and 0.0682 ± 0.1760 with $B = 0.03$; these intervals are not only noticeably longer than our intervals, but are also longer than the best uniform confidence intervals we can get using local linear regression with a rectangular kernel as in Kolesár and Rothe (2018). Thus, the use of numerical convex optimization tools that let us solve equation (5) instead of equation (9) can be of considerable value in practice.

IV. A Discontinuity Design with Two Cutoffs

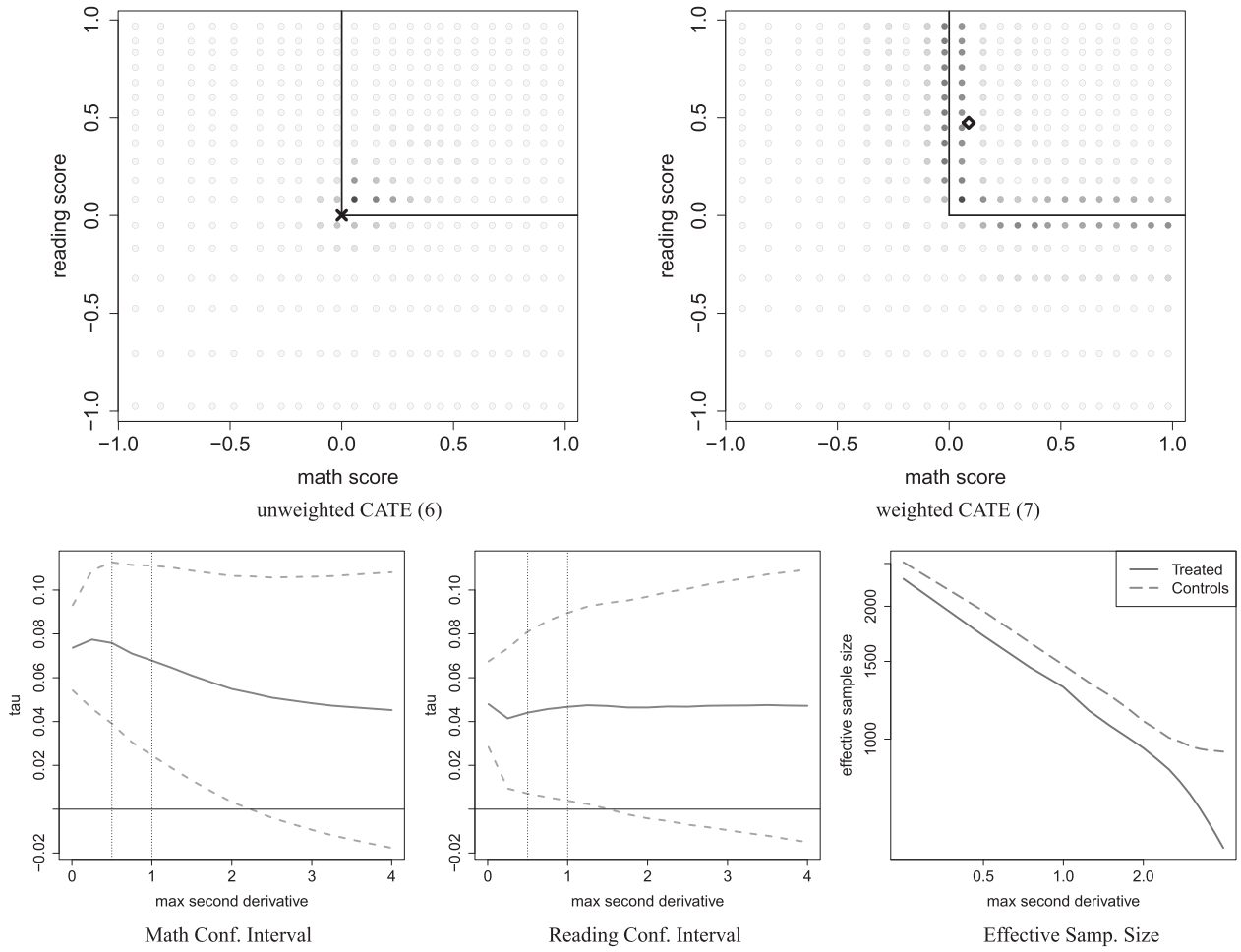
We next consider the behavior of our method in a specific variant of the multiple running variable problem motivated by a common inference strategy in education. Some school districts require students to attend summer school if they fail a year-end standardized test in either math or reading (Jacob & Lefgren, 2004; Matsudaira, 2008), and it is of course important to understand the value of such summer schools. The fact that students are mandated to summer school based on a sharp test score cutoff suggests a natural identification strategy via regression discontinuities; however, standard univariate techniques cannot directly be applied as the regression 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.

We illustrate our approach using Matsudaira's (2008) data set. 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 take tests in math and reading; those failing either of these tests must attend summer school. Here, we focus on the 2001 class of graduating fifth graders and filter the sample to include only the $n = 30{,}741$ students whose fifth-grade math and reading scores fall between 40 points of the passing threshold; this represents 44.7% of the full sample. Matsudaira (2008) analyzed this data set with univariate methods by using reading score as a running variable and considering only the subset of students who passed the math or reading exam. This allows for a simple analysis, but may also result in a loss of precision.[13] 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; 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 noncompliance. We consider both of our optimized estimators, equations 6 and 7.

In order to proceed with our approach, we again need to choose a value for B. Running a second-order polynomial regression on the next year's math and reading scores for both

[13] In order to make a formal power comparison, we need to compare two estimators that target the same estimand. In the simplest case where $\tau(x) = \tau$ is constant, our estimator, equation (7), presents an unambiguous gain in power over those considered in Matsudaira (2008).

FIGURE 4.—ESTIMATING THE EFFECT OF SUMMER SCHOOL ON TEST SCORES



(Top row) Weights $\hat{\gamma}$ underlying treatment effect estimates of the effect of summer school on the following year's reading scores, using both equation (6), which seeks to estimate the conditional average treatment effect (CATE) at $c = (0,\ 0)$, and the estimator, equation (7), 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). In the right panel, the diamond marks the weighted mean of the treated X_i-values: $\sum \hat{\gamma}_i W_i X_i$. These plots were generated with a maximum second derivative bound of $B = 0.5 \times 40^{-2}$. (Bottom row) The first two panels depict a sensitivity analysis for our weighted CATE result, for the math and reading outcomes respectively. We plot point estimates along with bias-aware 95% confidence intervals for different choices of B; the choices of B used in the bottom panel of table 1 are indicated with dotted lines. The third panel depicts effective sample sizes used by the procedure, $\text{ESS}_w = 1/\sum_{\{i:W_i=w\}} \hat{\gamma}_i^2$. For a given value of B, the $\hat{\gamma}_i$ used for the math and reading outcomes, are the same. In all cases, B is multiplied by 40^2 for readability.

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 $B = 1.0 \times 40^{-2}$.

The top row of figure 4 compares weight functions $\hat{\gamma}$ learned by both methods. The estimator $\hat{\tau}_c$ is in fact quite conservative and gives large weights only 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 use more data while cancelling out curvature effects due to $\mu_0(x)$. 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.

The bottom panel of table 1 displays corresponding estimates for the treatment effect. As expected, the confidence intervals using the weighted method, equation (7), are much shorter than those obtained using equation (6), allowing for a 0.95 level significant detection in the first case but not in the second. Since the weighting method allows for shorter confidence intervals and in practice seems to yield a matching-like estimator, we expect it to be more often applicable than the unweighted estimator, equation (6).

The bottom row of figure 4 presents some further diagnostics for our result. The first two panels depict a sensitivity analysis for our weighted CATE result. We vary the maximum bound B on the second derivative and examine how our confidence intervals change.[14] The result on the effect of

[14] We note that if the CATE function $\tau(\cdot)$ is not constant, then our weighted CATE estimand $\bar{\tau}_* = \sum_i W_i \hat{\gamma}_i \tau(X_i)$ may vary with B. This result should

summer school on math scores appears to be fairly robust, as we still get significant bias-aware 95% confidence intervals at $B = 2 \times 40^{-2}$, which is four times the largest apparent curvature observed in the data. The third panel plots a measure of the effective size of the treated and control samples used by the algorithm, $\text{ESS}_w = 1/\sum_{\{i:W_i=w\}} \hat{\gamma}_i^2$. Although our analysis sample has almost exactly the same number of treated and control units ($\bar{W} = 0.501$), it appears that our algorithm is able to make use of more control than treated samples, perhaps because the treated units are "wrapped around" the controls.

Finally, it is natural to ask whether the bivariate specification considered here gave us anything in addition to the simpler approach used by Matsudaira (2008) estimating the treatment effect of summer school on the next year's math exam by running a univariate regression discontinuity analysis on only students who passed the reading exam, and vice versa to the effect on the reading exam.[15] We ran both of these analyses using our method, equation (15), 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, equation (7), in the bottom panel of table 1 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.[16]

V. 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 automatic. The proposed algorithm is the same whether the running variable is continuous or discrete and whether $\tau(c)$ is point identified. Moreover, it 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.

A. Fuzzy Regression Discontinuities

In this paper, we considered only sharp regression discontinuities, where the treatment assignment W_i is a deterministic function of X_i. However, there is also considerable interest in fuzzy discontinuities, where W_i is random but $\mathbb{P}[W_i = 1 \mid X_i = x]$ has a jump at the threshold c (see Imbens & Lemieux, 2008, for a review). In this case, it is common to interpret the indicator $\mathbf{1}(\{X_i \geq c\})$ as an instrument and then to estimate a local average treatment effect via two-stage local linear regression (Imbens & Angrist, 1994). By analogy, we can estimate treatment effects with fuzzy regression discontinuity via two-stage optimized designs as $\hat{\tau}_{LATE} = \sum_{i=1}^n \hat{\gamma}_i Y_i / \sum_{i=1}^n \hat{\gamma}_i W_i$, where the $\hat{\gamma}_i$ are obtained as in equation (15) 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 Donoho (1994) for uniform asymptotic inference is left for future work.

B. 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, for example, 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, that is, to add a constraint $\sum_{i=1}^n \hat{\gamma}_i Z_{ij} = 0$ for all $j = 1, \ldots, p$ to the optimization problem, equation (5). 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 to hold approximately even if we do not enforce them; however, explicitly enforcing such balance may improve robustness.[17]

thus formally be interpreted as either a sensitivity analysis for the constant treatment effect parameter τ if we are willing to assume constant treatment effects or as a robustness check for our rejection of the global null hypothesis $\tau(x) = 0$ for all x.

[15] Matsudaira's (2008) estimator is not exactly comparable to the two we consider. For example, when only focusing on students who passed the reading exam, his estimator effectively averages treatment effects over the math pass/fail boundary but not the reading pass/fail boundary. In contrast, we either estimate the conditional average treatment effect at a point c, equation (6), or allow for averaging over the full boundary, equation (7). It is unclear whether the restriction of Matsudaira (2008) to averaging over only one of the two boundary segments targets a meaningfully more interpretable estimand than equation (7).

[16] 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 fifth 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 & Imbens, forthcoming).

[17] 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). Their strategy is to obtain weights $\hat{\gamma}_i$ without looking at the Z_i, and then to reject the modeling strategy if balance does not hold approximately.

If we have an additive, linear dependence of Y_i on Z_i, then enforcing balance would also result in variance reduction, as the conditional variance of our estimator $\hat{\tau}$ would now depend on $\text{Var}[Y_i \mid X_i,\ Z_i]$, which is always smaller or equal to $\text{Var}[Y_i \mid X_i]$.

C. 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 $\mu_w(\cdot)$ (Kolesár & Rothe, 2018; Imbens & Kalyanaraman, 2012). However, our conceptual framework can also be applied with higher-order smoothness assumptions via bounds on the kth 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

Angrist, J. D., and V. Lavy, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* 114:2 (1999), 533–575, 1999.

Armstrong, T. B., and M. Kolesár, "Simple and Honest Confidence Intervals in Nonparametric Regression," arXiv:1606.01200 (2016).

——— "Optimal Inference in a Class of Regression Models," *Econometrica* 8:2 (2018), 655–683.

Athey, S., G. W. Imbens, and S. Wager, "Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions," *Journal of the Royal Statistical Society: Series B*, 80:4 (2018), 597–623.

Berk, R. A., and D. Rauma, "Capitalizing on Nonrandom Assignment to Treatments: A Regression-Discontinuity Evaluation of a Crime-Control Program," *Journal of the American Statistical Association* 78:381, (1983), 21–27.

Bertanha, M., and M. J. Moreira, "Impossible Inference in Econometrics: Theory and Applications to Regression Discontinuity, Bunching, and Exogeneity Tests," arXiv:1612.02024 (2016).

Black, S. E., "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quarterly Journal of Economics* 114:2 (1999), 577–599.

Boyd, S., and L. Vandenberghe, *Convex Optimization* (Cambridge: Cambridge University Press, 2004).

Cai, T. T., and M. G. Low, "A Note on Nonparametric Estimation of Linear Functionals," *Annals of Statistics* 31:4 (2003), 1140–1153.

Calonico, S., M. D. Cattaneo, and M. H. Farrell, "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association* 113:522 (2018), 1–13. doi:10.1080/01621459.2017.1285776.

Calonico, S., M. D. Cattaneo, and R. Titiunik, "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82:6 (2014), 2295–2326.

Cattaneo, M. D., B. R. Frandsen, and R. Titiunik, "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate," *Journal of Causal Inference*, 3:1 (2015), 1–24. 2015.

Cattaneo, M. D., N. Idrobo, and R. Titiunik, *A Practical Introduction to Regression Discontinuity Designs: Part II* (Cambridge: Cambridge University Press, forthcoming).

Cattaneo, M. D., M. Jansson, and X. Ma, "Simple Local Regression Distribution Estimators with an Application to Manipulation Testing. Unpublished, University of Michigan and University of California Berkeley working paper (2016).

Caughey, D., and J. S. Sekhon, "Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942–2008," *Political Analysis* 19:4 (2011), 385–408.

Cheng, M.-Y., J. Fan, and J. S. Marron, "On Automatic Boundary Corrections," *Annals of Statistics* 25:4, (1997), 1691–1708.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika* 96:1 (2009), 187–199.

Donoho, D. L., "Statistical Estimation and Optimal Recovery," *Annals of Statistics* 22:1 (1994), 238–270.

Donoho, D. L., and R. C. Liu, "Geometrizing Rates of Convergence, III," *Annals of Statistics*, 19:2 (1991), 668–701.

Frandsen, B. R., "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable Is Discrete," (pp. 281–315), in Matias D. Cattaneo and Juan Carlos Escanciano, eds., *Regression Discontinuity Designs: Theory and Applications* (Bingley, UK: Emerald Publishing, 2017.

Gao, W. Y., "Minimax Linear Estimation at a Boundary Point," *Journal of Multivariate Analysis* 165 (2018), 262–269.

Gelman, A., and G. Imbens, "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business and Economic Statistics* (forthcoming).

Hahn, J., P. Todd, and W. Van der Klaauw, "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica* 69:1 (2001), 201–209.

Heckman, J., and Hotz, J., "Alternative Methods for Evaluating the Impact of Training Programs," *Journal of the American Statistical Association* 84:408 (1989), 862–880.

Ibragimov, I. A., and R. Z. Khas'minskii, "On Nonparametric Estimation of the Value of a Linear Functional in Gaussian White Noise," *Theory of Probability and Its Applications* 29:1 (1985), 18–32.

Imbens, G. W., and J. D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica* 62:2 (1994), 467–475.

Imbens, G. W., and K. Kalyanaraman, "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies* 79:3 (2012), 933–959.

Imbens, G. W., and T. Lemieux, "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics* 142:2 (2008), 615–635.

Imbens, G. W., and C. F. Manski, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72:6 (2004), 1845–1857.

Imbens, G. W., and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge: Cambridge University Press, 2015).

Imbens, G., and S. Wager, "Optimized Regression Discontinuity Designs," arXiv:1705.01677 (2017).

Jacob, B. A., and L. Lefgren, "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," this REVIEW 86:1 (2004), 226–244.

Johnstone, I. M., "Gaussian Estimation: Sequence and Wavelet Models," unpublished manuscript (2011).

Juditsky, A. B., and A. S. Nemirovski, "Nonparametric Estimation by Convex Programming, *Annals of Statistics* 37:5A (2009), 2278–2300.

Karlin, S., "Some Variational Problems on Certain Sobolev Spaces and Perfect Splines," *Bulletin of the American Mathematical Society* 79:1 (1973), 124–128.

Keele, L. J., and R. Titiunik, "Geographic Boundaries as Regression Discontinuities," *Political Analysis* 23:1 (2014), 127–155.

Keele, L., R. Titiunik, and J. Zubizarreta, "Enhancing a Geographic Regression Discontinuity Design through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout," *Journal of the Royal Statistical Society, Series A* 178:1 (2015), 223–239.

Kolesár, M., and C. Rothe, "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review* 108:8 (2018), 2277–2304.

Lalive, R., "How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach," *Journal of Econometrics* 142:2 (2008), 785–806.

Lee, D. S., "Randomized Experiments from Non-random Selection in US House Elections," *Journal of Econometrics* 142:2 (2008), 675–697.

Lee, D. S., and D. Card, "Regression Discontinuity Inference with Specification Error," *Journal of Econometrics* 142:2 (2008), 655–674.

Lee, D. S., and T. Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48:2 (2010), 281–355.

Legostaeva, I., and A. Shiryaev, "Minimax Weights in a Trend Detection Problem of a Random Process," *Theory of Probability and Its Applications* 16:2 (1971), 344–349.
Li, F., A. Mattei, and F. Mealli, "Evaluating the Causal Effect of University Grants on Student Dropout: Evidence from a Regression Discontinuity Design Using Principal Stratification," *Annals of Applied Statistics* 9:4, (2015), 1906–1931.
Li, F., K. L. Morgan, and A. M. Zaslavsky, "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association* 113:521 (2017), 390–400.
Ludwig, J., and D. L. Miller, "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122:1, (2007), 159–208.
Matsudaira, J. D., "Mandatory Summer School and Student Achievement," *Journal of Econometrics*, 142:2 (2008), 829–850.
McCrary, J., Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics* 142:2 (2008), 698–714.
Neyman, J., "Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes," *Roczniki Nauk Rolniczych*, 10 (1923), 1–51.
Oreopoulos, P., "Estimating Average and Local Average Treatment Effects of Education When Compulsory Schooling Laws Really Matter," *American Economic Review* 96:1 (2006), 152–175.
Papay, J. P., J. B. Willett, and R. J. Murnane, "Extending the Regression-Discontinuity Approach to Multiple Assignment Variables," *Journal of Econometrics* 161:2 (2011), 203–207.
Ponstein, J., *Approaches to the Theory of Optimization* (Cambridge: Cambridge University Press, 2004).
Porter, J., Estimation in the Regression Discontinuity Model," unpublished manuscript (2003), http://citeseerx.ist.psu.edu/viewdoc/download?
Reardon, S. F., and J. P. Robinson, "Regression Discontinuity Designs with Multiple Rating-Score Variables," *Journal of Research on Educational Effectiveness*, 5:1 (2012), 83–104.
Robins, J., and A. van der Vaart, "Adaptive Nonparametric Confidence Sets," *Annals of Statistics* 34:1 (2006), 229–253.
Robins, J., L. Li, E. Tchetgen, and A. van der Vaart, "Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals," (pp. 335–426), in Deborah Nolan and Terry Speed, eds., *Probability and Statistics: Essays in Honor of David A. Freedman* (Bethesda, MD: Institute of Mathematical Statistics, 2008).
Romano, J. P., and M. Wolf, "Resurrecting Weighted Least Squares," *Journal of Econometrics* 197:1 (2017), 1–19.
Rubin, D. B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66:5 (1974), 688.
Sacks, J., and D. Ylvisaker, "Linear Estimation for Approximately Linear Models," *Annals of Statistics* 6:5 (1978), 1122–1137.
Thistlethwaite, D. L., and D. T. Campbell, "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51:6 (1960), 309–317.
Trochim, W. M., *Research Design for Program Evaluation: The Regression-Discontinuity Approach* (Thousand Oaks, CA: Sage, 1984).
White, H., "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica* 48:4 (1980), 817–838.
Wong, V. C., P. M. Steiner, and T. 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 (2013), 107–141.
Zajonc, T., *Essays on Causal Inference for Public Policy*, PhD dissertation, Harvard University. https://dash.harvard.edu/handle/1/9368030.