Optimal inference in a class of regression models*

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

November 22, 2017

Abstract

We consider the problem of constructing confidence intervals (CIs) for a linear functional of a regression function, such as its value at a point, the regression discontinuity parameter, or a regression coefficient in a linear or partly linear regression. Our main assumption is that the regression function is known to lie in a convex function class, which covers most smoothness and/or shape assumptions used in econometrics. We derive finite-sample optimal CIs and sharp efficiency bounds under normal errors with known variance. We show that these results translate to uniform (over the function class) asymptotic results when the error distribution is not known. When the function class is centrosymmetric, these efficiency bounds imply that minimax CIs are close to efficient at smooth regression functions. This implies, in particular, that it is impossible to form CIs that are substantively tighter using data-dependent tuning parameters, and maintain coverage over the whole function class. We specialize our results to inference on the regression discontinuity parameter, and illustrate them in simulations and an empirical application.

*We thank Don Andrews, Isaiah Andrews, Matias Cattaneo, Gary Chamberlain, Denis Chetverikov, Yuichi Kitamura, Soonwoo Kwon, Ulrich Müller and Azeem Shaikh for useful discussions. We thank the editor, three anonymous referees, and numerous seminar and conference participants for helpful comments and suggestions. All remaining errors are our own. The research of the first author was supported by National Science Foundation Grant SES-1628939. The research of the second author was supported by National Science Foundation Grant SES-1628878.

†email: timothy.armstrong@yale.edu

‡email: mkolesar@princeton.edu

1 Introduction

In this paper, we study the problem of constructing confidence intervals (CIs) for a linear functional Lf of a regression function f in a broad class of regression models with fixed regressors, in which f is known to belong to some convex function class $\mathcal{F}$. The linear functional may correspond to the regression discontinuity parameter, an average treatment effect under unconfoundedness, or a regression coefficient in a linear or partly linear regression. The class $\mathcal{F}$ may contain smoothness restrictions (e.g. bounds on derivatives, or assuming f is linear as in a linear regression), and/or shape restrictions (e.g. monotonicity, or sign restrictions on regression coefficients in a linear regression). Often in applications, the function class will be indexed by a smoothness parameter C, such as when $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, the class of Lipschitz continuous functions with Lipschitz constant C.

Our main contribution is to derive finite-sample optimal CIs and sharp efficiency bounds that have implications for data-driven model and bandwidth selection in both parametric and nonparametric settings. To derive these results, we assume that the regression errors are normal, with known variance. When the error distribution is unknown, we obtain analogous uniform asymptotic results under high-level regularity conditions. We derive sufficient low-level conditions in an application to regression discontinuity.

First, we characterize one-sided CIs that minimize the maximum β quantile of excess length over a convex class $\mathcal{G}$ for a given quantile β. The lower limit $\hat{c}$ of the optimal CI $[\hat{c}, \infty)$ has a simple form: take an estimator $\hat{L}$ that trades off bias and variance in a certain optimal sense and is linear in the outcome vector, and subtract (1) the standard deviation of $\hat{L}$ times the usual critical value based on a normal distribution and (2) a bias correction to ensure coverage. This bias correction, in contrast to bias corrections often used in practice, is based on the maximum bias of $\hat{L}$ over $\mathcal{F}$, and is therefore non-random.

When $\mathcal{G} = \mathcal{F}$, this procedure yields minimax one-sided CIs. Setting $\mathcal{G} \subset \mathcal{F}$ to a class of smoother functions is equivalent to "directing power" at these smoother functions while maintaining coverage over $\mathcal{F}$, and gives a sharp bound on the scope for adaptation for one-sided CIs. We show that when $\mathcal{F}$ is centrosymmetric (i.e. $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$), the scope for adaptation is severely limited: when $\mathcal{G}$ is a class of functions that are, in a certain formal sense, "sufficiently smooth" relative to $\mathcal{F}$, CIs that are minimax for β quantile of excess length also optimize excess length over $\mathcal{G}$, but at a different quantile. Furthermore, they are also highly efficient at such smooth functions for the same quantile. For instance, a CI for the conditional mean at a point that is minimax over the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$ is asymptotically 95.2% efficient at constant functions relative to a CI that directs all power

at constant functions. For function classes that bound a derivative of higher order, the efficiency is even higher.

Second, we derive a confidence set that minimizes its expected length at a single function g. We compare its performance to the optimal fixed-length CI of Donoho (1994) (i.e. CI of the form $\hat{L} \pm \chi$, where $\hat{L}$ is an affine estimator, and χ, which doesn't depend on the outcome vector and is therefore non-random, is chosen to ensure coverage). Similarly, to the one-sided case, we find that, when $\mathcal{F}$ is centrosymmetric, the optimal fixed-length CIs are highly efficient at functions that are smooth relative to $\mathcal{F}$. For instance, the optimal fixed-length CI for a conditional mean at a point when $f \in \mathcal{F}_{\text{Lip}}(C)$ is asymptotically 95.6% efficient at any constant function g relative to a confidence set that optimizes its expected length at g.

An important practical implication of these results is that explicit a priori specification of the smoothness constant C cannot be avoided: procedures that try to determine the smoothness of f from the data (and thus implicitly estimate C from the data), including data-driven bandwidth or variable selectors, must either fail to substantively improve upon the minimax CIs or fixed-length CIs (that effectively assume the worst case smoothness), or else fail to maintain coverage over the whole parameter space. We illustrate this point through a Monte Carlo study in a regression discontinuity (RD) setting, in which we show that popular data-driven bandwidth selectors lead to substantial undercoverage, even when combined with bias correction or undersmoothing (see Supplemental Appendix C.2). To avoid having to specify C, one has to strengthen the assumptions on f. For instance, one can impose shape restrictions that break the centrosymmetry, as in Cai et al. (2013) or Armstrong (2015), or self-similarity assumptions that break the convexity, as in Giné and Nickl (2010) or Chernozhukov et al. (2014). Alternatively, one can weaken the coverage requirement in the definition of a CI, by, say, only requiring average coverage as in Cai et al. (2014) or Hall and Horowitz (2013).

We apply these results to the problem of inference in RD. We show, in the context of an empirical application from Lee (2008), that the fixed-length and minimax CIs are informative and simple to construct, and we give a detailed guide to implementing them in practice. We also consider CIs based on local linear estimators, which have been popular in RD due to their high minimax asymptotic MSE efficiency, shown in Cheng et al. (1997). Using the same function classes as in Cheng et al. (1997), we show that in the Lee application, when a triangular kernel is used, such CIs are highly efficient relative to the optimal CIs discussed above.

Our finite-sample approach allows us to use the same framework and methods to cover

problems that are often seen as outside of the scope of nonparametric methods. For instance, the same CIs can be used in RD whether the running variable is discrete or continuous; one does not need a different modeling approach, such as that of Lee and Card (2008). Similarly, we do not need to distinguish between "parametric" or "nonparametric" constraints on f; our results apply to inference in a linear regression model that efficiently use a priori bounds and sign restrictions on the regression coefficients. Here our efficiency bounds imply that the scope for efficiency improvements from CIs formed after model selection (Andrews and Guggenberger, 2009; McCloskey, 2017) is severely limited unless asymmetric or non-convex restrictions are imposed, and they also limit the scope for improvement under certain non-convex restrictions such as the sparsity assumptions used in Belloni et al. (2014). We discuss these issues in an earlier version of this paper (Armstrong and Kolesár, 2016a).

Our results and setup build on a large statistics literature on optimal estimation and inference in the nonparametric regression model. This literature has mostly been concerned with estimation (e.g., Stone (1980), Ibragimov and Khas'minskii (1985), Fan (1993), Donoho (1994), Cheng et al. (1997)); the literature on inference has mostly been focused on bounding rates of convergence. The results most closely related to ours are those in Low (1997), Cai and Low (2004a) and Cai et al. (2013), who derive lower bounds on the expected length of a two-sided CI over a convex class $\mathcal{G}$ subject to coverage over a convex class $\mathcal{F}$. These results imply that, when $\mathcal{F}$ is constrained only by bounds on a derivative, one cannot improve the rate at which a two-sided CI shrinks by "directing power" at smooth functions. We contribute to this literature by (1) deriving a sharp lower bound for one-sided CIs, and for two-sided CIs when $\mathcal{G}$ is a singleton, (2) showing that the negative results for "directing power" at smooth functions generalize to the case when $\mathcal{F}$ is centrosymmetric, and deriving the sharp bound on the scope for improvement, (3) deriving feasible CIs under unknown error distribution and showing their asymptotic validity and efficiency, including in non-regular settings; and (4) computing the bounds and CIs in an application to RD.

The remainder of this paper is organized as follows. Section 2 illustrates our results in an application to RD, and gives a detailed guide to implementing our CIs. Section 3 derives the main results under a general setup. Section 4 considers an empirical application. Proofs, long derivations, and additional results are collected in appendices. Appendix A contains proofs for the main results in Section 3. Appendix B discusses extensions to incorporate covariates in the RD application. Supplemental Appendix C compares our CIs to other approaches, and includes a Monte Carlo study. Additional details for constructing CIs studied in Section 3 are in Supplemental Appendix D. Supplemental Appendix E contains

additional details for the RD application. Asymptotic results are collected in Supplemental Supplemental Appendices F, G and H.

2 Application to regression discontinuity

In this section, we explain our results in the context of an application to sharp regression discontinuity (RD). Section 2.1 illustrates the theoretical results, while Section 2.2 gives step-by-step instructions for implementing our confidence intervals (CIs) in practice.

We observe $\{y_i, x_i\}_{i=1}^n$, where the running variable x_i is deterministic, and

$$y_i = f(x_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2(x_i)) \text{ independent across } i, \tag{1}$$

with $\sigma^2(x)$ known.[1] The running variable determines participation in a binary treatment: units above a given cutoff, which we normalize to 0, are treated; units with $x_i < 0$ are controls. Let $f_+(x) = f(x)1(x \geq 0)$ and $f_-(x) = f(x)1(x < 0)$ denote the part of the regression function f above and below the cutoff, so that $f = f_+ + f_-$. The parameter of interest is the jump of the regression function at zero, and we denote it by $Lf = f_+(0) - f_-(0)$, where $f_-(0) = \lim_{x \uparrow 0} f_-(x)$. If the regression functions of potential outcomes are continuous at zero, then Lf measures the average treatment effect for units with $x_i = 0$.

We assume that f lies in the class of functions $\mathcal{F}_{RDT,p}(C)$,

$$\mathcal{F}_{RDT,p}(C) = \big\{ f_+ + f_- \colon f_+ \in \mathcal{F}_{T,p}(C; \mathbb{R}_+),\ f_- \in \mathcal{F}_{T,p}(C; \mathbb{R}_-) \big\},$$

where $\mathcal{F}_{T,p}(C; \mathcal{X})$ consists of functions f such that the approximation error from a $(p-1)$th-order Taylor expansion of $f(x)$ about 0 is bounded by $C|x|^p$, uniformly over $\mathcal{X}$,

$$\mathcal{F}_{T,p}(C; \mathcal{X}) = \big\{ f \colon \big| f(x) - \textstyle\sum_{j=0}^{p-1} f^{(j)}(0) x^j / j! \big| \leq C|x|^p \text{ all } x \in \mathcal{X} \big\}.$$

This formalizes the notion that locally to 0, f is p-times differentiable with the pth derivative at zero bounded by $p!C$. Sacks and Ylvisaker (1978) and Cheng et al. (1997) considered minimax MSE estimation of $f(0)$ in this class when 0 is a boundary point. Their results formally justify using local polynomial regression to estimate the RD parameter. This class does not impose any smoothness of f away from cutoff, which may be too conservative

[1] This assumption is made to deliver finite-sample results—when the distribution of u_i is unknown, with unknown conditional variance, we show in Supplemental Appendix E that these results lead to analogous uniform-in-f asymptotic results.

in applications. We consider inference under global smoothness in Armstrong and Kolesár (2016b), where we show that for the $p = 2$ case, the resulting CIs are about 10% tighter in large samples (see also Supplemental Appendix C.2 for a Monte Carlo study under global smoothness).

2.1 Optimal CIs

For ease of exposition, we focus in this subsection on the case $p = 1$, so that the parameter space is given by $\mathcal{F} = \mathcal{F}_{RDT,1}(C)$, and assume that the errors are homoskedastic, $\sigma^2(x_i) = \sigma^2$. In Section 2.2, we discuss implementation of the CIs in the general case where $p \geq 1$.

Consider first the problem of constructing one-sided CIs for Lf. In particular, consider the problem of constructing CIs $[\hat{c}, \infty)$ that minimize the maximum βth quantile of excess length, $\sup_{f\in\mathcal{F}} q_{f,\beta}(Lf - \hat{c})$, where $q_{f,\beta}$ denotes the βth quantile of the excess length $Lf - \hat{c}$. We show in Section 3.3 that such CIs can be obtained by inverting tests of the null hypothesis $H_0\colon f_+(0)-f_-(0) \leq L_0$ that maximize their minimum power under the alternative $H_1\colon f_+(0) - f_-(0) \geq L_0 + 2b$, where the half-distance b to the alternative is calibrated so that the minimum power of these tests equals β.

To construct such a test, note that if we set $\mu = (f(x_1), \ldots, f(x_n))'$, and $Y = (y_1, \ldots, y_n)'$, we can view the testing problem as an n-variate normal mean problem $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, in which the vector of means μ is constrained to take values in the convex sets $M_0 = \{(f(x_1), \ldots, f(x_n))'\colon f \in \mathcal{F}, f_+(0) - f_-(0) \leq L_0\}$ under the null, and $M_1 = \{(g(x_1), \ldots, g(x_n))'\colon g \in \mathcal{F}, g_+(0) - g_-(0) \geq L_0 + 2b\}$ under the alternative. The convexity of the null and alternative sets implies that this testing problem has a simple solution: by Lemma A.2, the minimax test is given by the uniformly most powerful test of the simple null $\mu = \mu_0^*$ against the simple alternative $\mu = \mu_1^*$, where μ_0^* and μ_1^* minimize the Euclidean distance between the null and alternative sets M_0 and M_1, and thus represent points in M_0 and M_1 that are hardest to distinguish. By the Neyman-Pearson lemma, such test rejects for large values of $(\mu_1^* - \mu_0^*)'Y$. Because by Lemma A.2, this test controls size over all of M_0, the points μ_1^* and μ_0^* are called "least favorable" (see Theorem 8.1.1 in Lehmann and Romano, 2005).

To compute $\mu_0^* = (f^*(x_1), \ldots, f^*(x_n))'$ and $\mu_1^* = (g^*(x_1), \ldots, g^*(x_n))'$, we thus need to find functions f^* and g^* that solve

$$(f^*, g^*) = \operatorname*{argmin}_{f,g\in\mathcal{F}} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \qquad \text{subject to } Lf \leq L_0,\ Lg \geq L_0 + 2b. \tag{2}$$

A simple calculation shows that the least favorable functions solving this minimization problem are given by

$$\begin{aligned} g^*(x) &= 1(x \geq 0)(L_0 + b) + Ch_+ \cdot k_+(x/h_+) - Ch_- \cdot k_-(x/h_-), \\ f^*(x) &= 2 \cdot 1(x \geq 0)(L_0 + b) - g^*(x), \end{aligned} \tag{3}$$

where $k(u) = \max\{0, 1 - |u|\}$ is the triangular kernel, $k_+(u) = k(u)1(u \geq 0)$ and $k_-(u) = k(u)1(u < 0)$, and the "bandwidths" h_+, h_- are determined by a condition ensuring that $Lg^* \geq L_0 + 2b$,

$$h_+ + h_- = b/C, \tag{4}$$

and a condition ensuring that positive and negative observations are equally weighted,

$$h_+ \sum_{i=1}^{n} k_+(x_i/h_+) = h_- \sum_{i=1}^{n} k_-(x_i/h_-). \tag{5}$$

Intuitively, to make the null and alternative hardest to distinguish, the least favorable functions f^* and g^* converge to each other "as quickly as possible", subject to the constraints $Lf^* \leq L_0$ and $Lg^* \geq b + L_0$, and the Lipschitz constraint—see Figure 1.

By working out the appropriate critical value and rearranging, we obtain that the minimax test rejects whenever

$$\hat{L}_{h_+,h_-} - L_0 - \text{bias}_{f^*}(\hat{L}_{h_+,h_-}) \geq \text{sd}(\hat{L}_{h_+,h_-}) z_{1-\alpha}. \tag{6}$$

Here $\hat{L}_{h_+,h_-}$ is a kernel estimator based on a triangular kernel and bandwidths h_+ to the left and h_- to the right of the cutoff

$$\hat{L}_{h_+,h_-} = \frac{\sum_{i=1}^{n}(g^*(x_i) - f^*(x_i))y_i}{\sum_{i=1}^{n}(g^*_+(x_i) - f^*_+(x_i))} = \frac{\sum_{i=1}^{n} k_+(x_i/h_+)y_i}{\sum_{i=1}^{n} k_+(x_i/h_+)} - \frac{\sum_{i=1}^{n} k_-(x_i/h_-)y_i}{\sum_{i=1}^{n} k_-(x_i/h_-)}, \tag{7}$$

$\text{sd}(\hat{L}_{h_+,h_-}) = \left(\frac{\sum_i k_+(x_i/h_+)^2}{(\sum_i k_+(x_i/h_+))^2} + \frac{\sum_i k_-(x_i/h_-)^2}{(\sum_i k_-(x_i/h_-))^2}\right)^{1/2} \cdot \sigma$ is its standard deviation, $z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution, and $\text{bias}_{f^*}(\hat{L}_{h_+,h_-}) = C \sum_i |x_i| \cdot \left(\frac{k_+(x_i/h_+)}{\sum_j k_+(x_j/h_+)} + \frac{k_-(x_i/h_-)}{\sum_j k_-(x_j/h_-)}\right)$ is the estimator's bias under f^*. The estimator $\hat{L}_{h_+,h_-}$ is normally distributed with variance that does not depend on the true function f. Its bias, however, does depend on f. To control size under H_0 in finite samples, it is necessary to subtract the largest possible bias of $\hat{L}_h$ under the null, which obtains at f^*. Since the rejection probability of the test is decreasing in the bias, its minimum power occurs when the bias is minimal under H_1, which

occurs at g^*, and is given by

$$\beta = \Phi\Big(2C\sqrt{h_+^2 \textstyle\sum_i k_+(x_i/h_+)^2 + h_-^2 \sum_i k_-(x_i/h_-)^2}/\sigma - z_{1-\alpha}\Big). \tag{8}$$

Since the estimator, its variance, and the non-random bias correction are all independent of the particular null L_0, the CI based on inverting these tests as H_0 varies over $\mathbb{R}$ is given by

$$[\hat{c}_{\alpha,h_+,h_-}, \infty), \quad \text{where} \quad \hat{c}_{\alpha,h_+,h_-} = \hat{L}_{h_+,h_-} - \text{bias}_{f^*}(\hat{L}_{h_+,h_-}) - \text{sd}(\hat{L}_{h_+,h_-})z_{1-\alpha}. \tag{9}$$

This CI minimizes the βth quantile maximum excess length with β given by the minimax power of the tests (8). Equivalently, given a quantile β that we wish to optimize, let $h_+(\beta)$ and $h_-(\beta)$ solve (5) and (8). The optimal CI is then given by $[\hat{c}_{\alpha,h_+(\beta),h_-(\beta)}, \infty)$, and the half-distance b to the alternative of the underlying tests is determined by (4). The important feature of this CI is that the bias correction is non-random: it depends on the worst-case bias of $\hat{L}_{h_+(\beta),h_-(\beta)}$, rather than an estimate of the bias. Furthermore, it doesn't disappear asymptotically. One can show that the squared worst-case bias of $\hat{L}_{h_+(\beta),h_-(\beta)}$ and its variance are both of the order $n^{-2/3}$. Consequently, no CI that "undersmooths" in the sense that it is based on an estimator whose bias is of lower order than its variance can be minimax optimal asymptotically or in finite samples.

An apparent disadvantage of this CI is that it requires the researcher to choose the smoothness parameter C. Addressing this issue leads to "adaptive" CIs. Adaptive CIs achieve good excess length properties for a range of parameter spaces $\mathcal{F}_{RDT,1}(C_j)$, $C_1 < \cdots < C_J$, while maintaining coverage over their union, which is given by $\mathcal{F}_{RDT,1}(C_J)$, where C_J is some conservative upper bound on the possible smoothness of f. In contrast, a minimax CI only considers worst-case excess length over $\mathcal{F}_{RDT,1}(C_J)$. To derive an upper bound on the scope for adaptivity, consider the problem of finding a CI that optimizes excess length over $\mathcal{F}_{RDT,1}(0)$ (the space of functions that are constant on either side of the cutoff), while maintaining coverage over $\mathcal{F}_{RDT,1}(C)$ for some $C > 0$.

To derive the form of such CI, consider the one-sided testing problem $H_0\colon Lf \leq L_0$ and $f \in \mathcal{F}_{RDT,1}(C)$ against the one-sided alternative $H_1\colon f(0) \geq L_0 + b$ and $f \in \mathcal{F}_{RDT,1}(0)$ (so that now the half-distance to the alternative is given by $b/2$ rather than b). This is equivalent to a multivariate normal mean problem $Y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, with $\mu \in M_0$ under the null as before, and $\mu \in \tilde{M}_1 = \{(f(x_1), \ldots, f(x_n))'\colon f \in \mathcal{F}_{RDT,1}(0), Lf \geq L_0 + b\}$. Since the null and alternative are convex, by the same arguments as before, the least favorable functions minimize the distance between the two sets. The minimizing functions are given by

$\tilde{g}^*(x) = 1(x \geq 0)(L_0 + b)$, and $\tilde{f}^* = f^*$ (same function as before). Since $\tilde{g}^* - \tilde{f}^* = (g^* - f^*)/2$, this leads to the same test and the same CI as before—the only difference is that we moved the half-distance to the alternative from b to $b/2$. Hence, the minimax CI that optimizes a given quantile of excess length over $\mathcal{F}_{RDT,1}(C)$ also optimizes its excess length over the space of constant functions, but at a different quantile. Furthermore, in Section 3.3, we show that the minimax CI remains highly efficient if one compares excess length at the same quantile: in large samples, the efficiency at constant functions is 95.2%. Therefore, it is not possible to "adapt" to cases in which the regression function is smoother than the least favorable function. Consequently, it is not possible to tighten the minimax CI by, say, using the data to "estimate" the smoothness parameter C.

A two-sided CI can be formed as $\hat{L}_{h_+,h_-} \pm (\text{bias}_{f^*}(\hat{L}_{h_+,h_-}) + \text{sd}(\hat{L}_{h_+,h_-})z_{1-\alpha/2})$, thereby accounting for possible bias of $\hat{L}_{h_+,h_-}$. However, this is conservative, since the bias cannot be in both directions at once. Since the t-statistic $(\hat{L}_{h_+,h_-} - Lf)/\text{sd}(\hat{L}_{h_+,h_-})$ is normally distributed with variance one and mean at most $\text{bias}_{f^*}(\hat{L}_{h_+,h_-})/\text{sd}(\hat{L}_{h_+,h_-})$ and least $-\text{bias}_{f^*}(\hat{L}_{h_+,h_-})/\text{sd}(\hat{L}_{h_+,h_-})$, a nonconservative CI takes the form

$$\hat{L}_{h_+,h_-} \pm \text{sd}(\hat{L}_{h_+,h_-})\,\text{cv}_\alpha(\text{bias}_{f^*}(\hat{L}_{h_+,h_-})/\text{sd}(\hat{L}_{h_+,h_-})),$$

where $\text{cv}_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $\mathcal{N}(t,1)$ distribution, which we tabulate in Table 1. The optimal bandwidths h_+ and h_- simply minimize the CI's length, $2\,\text{sd}(\hat{L}_{h_+,h_-}) \cdot \text{cv}_\alpha(\text{bias}_{f^*}(\hat{L}_{h_+,h_-})/\text{sd}(\hat{L}_{h_+,h_-}))$. It can be shown that the solution satisfies (5), so choosing optimal bandwidths is a one-dimensional optimization problem. Since the length doesn't depend on the data Y, minimizing it does not impact the coverage properties of the CI. This CI corresponds to the optimal affine fixed-length CI, as defined in Donoho (1994). Since the length of the CI doesn't depend on the data Y, it cannot be adaptive. In Section 3.4 we derive a sharp efficiency bound that shows that, similar to the one-sided case, these CIs are nonetheless highly efficient relative to variable-length CIs that optimize their length at smooth functions.

The key to these non-adaptivity results is that the class $\mathcal{F}$ is centrosymmetric (i.e. $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$) and convex. For adaptivity to be possible, it is necessary (but perhaps not sufficient) to impose shape restrictions like monotonicity, or non-convexity of $\mathcal{F}$.

2.2 Practical implementation

We now discuss some practical issues that arise when implementing optimal CIs.[2] To describe the form of the optimal CIs for general $p \geq 1$, consider first the problem of constructing CIs based on a linear estimator of the form

$$\hat{L}_{h_+,h_-} = \sum_{i=1}^{n} w_+(x_i, h_+)y_i - \sum_{i=1}^{n} w_-(x_i, h_-)y_i, \tag{10}$$

where h_+, h_- are smoothing parameters, and the weights satisfy $w_+(-x, h_+) = w_-(x, h_-) = 0$ for $x \geq 0$. The estimator $\hat{L}_{h_+,h_-}$ is normally distributed with variance $\mathrm{sd}(\hat{L}_{h_+,h_-})^2 = \sum_{i=1}^{n}(w_+(x_i, h_+) + w_-(x_i, h_-))^2\sigma^2(x_i)$, which does not depend on f. A simple argument (see Supplemental Appendix E) shows that largest possible bias of $\hat{L}_{h_+,h_-}$ over the parameter space $\mathcal{F}_{RDT,p}(C)$ is given by

$$\overline{\mathrm{bias}}_{\mathcal{F}_{RDT,p}(C)}(\hat{L}_{h_+,h_-}) = C\sum_{i=1}^{n}|w_+(x_i, h_+) + w_-(x_i, h_-)| \cdot |x_i|^p, \tag{11}$$

provided that the weights are such that $\hat{L}_{h_+,h_-}$ is unbiased for f that takes the form of a $(p-1)$th order polynomial on either side of cutoff (otherwise the worst-case bias will be infinite). By arguments as in Section 2.1, one can construct one- and two-sided CIs based on $\hat{L}_{h_+,h_-}$ as

$$[c(\hat{L}_{h_+,h_-}), \infty) \qquad c(\hat{L}_{h_+,h_-}) = \hat{L}_{h_+,h_-} - \overline{\mathrm{bias}}_{\mathcal{F}_{RDT,p}(C)}(\hat{L}_{h_+,h_-}) - \mathrm{sd}(\hat{L}_{h_+,h_-})z_{1-\alpha}, \tag{12}$$

and

$$\hat{L}_{h_+,h_-} \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{F}_{RDT,p}(C)}(\hat{L}_{h_+,h_-})/\,\mathrm{sd}(\hat{L}_{h_+,h_-})) \cdot \mathrm{sd}(\hat{L}_{h_+,h_-}). \tag{13}$$

The problem of constructing optimal two- and one- sided CIs can be cast as a problem of finding weights w_+, w_- and smoothing parameters h_+ and h_- that lead to CIs with the shortest length, and smallest worst-case β quantile of excess length, respectively. The solution to this problem follows from a generalization of results in Sacks and Ylvisaker (1978). The optimal weights w_+ and w_- are given by a solution to a system of $2(p-1)$ equations, described in Supplemental Appendix E. When $p = 1$, they reduce to the weights $w_+(x_i, h_+) = k_+(x_i/h_+)/\sum_i k_+(x_i/h_+)$ and $w_-(x_i, h_-) = k_-(x_i/h_+)/\sum_i k_-(x_i/h_+)$, where $k_+(x_i) = k(x_i)1(x_i \geq 0)$ and $k_-(x_i) = k(x_i)1(x_i < 0)$, and $k(u) = \max\{0, 1 - |u|\}$ is a

[2] An R package implementing these CIs is available at `https://github.com/kolesarm/RDHonest`.

triangular kernel. This leads to the triangular kernel estimator (7). For $p > 1$, the optimal weights depend on the empirical distribution of the running variable x_i.

An alternative to using the optimal weights is to use a local polynomial estimator of order $p-1$, with kernel k and bandwidths h_- and h_+ to the left and to the right of the cutoff. This leads to weights of the form

$$w_+(x_i, h_+) = e_1' \left(\sum_i k_+(x_i/h_+) r_i r_i' \right)^{-1} \sum_i k_+(x_i/h_+) r_i, \tag{14}$$

and similarly for $w_-(x_i, h_-)$, where $r_i = (1, x_i, \ldots, x_i^{p-1})$ and e_1 is the first unit vector. Using the efficiency bounds we develop in Section 3, it can be shown that, provided that the bandwidths h_+ and h_- to the right and to the left of the cutoff are appropriately chosen, in many cases the resulting CIs are highly efficient. In particular, for $p = 2$, using the local linear estimator with the triangular kernel turns out to lead to near-optimal CIs (see Section 4).

Thus, given smoothness constants C and p, one can construct optimal or near-optimal CIs as follows:

1. Form a preliminary estimator of the conditional variance $\hat{\sigma}(x_i)$. We recommend using the estimator $\hat{\sigma}^2(x_i) = \hat{\sigma}_+^2(0)1(x \geq 0) + \hat{\sigma}_-^2(0)1(x < 0)$ where $\hat{\sigma}_+^2(0)$ and $\hat{\sigma}_-^2(0)$ are estimates of $\lim_{x \downarrow 0} \sigma^2(x)$ and $\lim_{x \uparrow 0} \sigma^2(x)$ respectively.[3]

2. Given smoothing parameters h_+ and h_-, compute the weights w_+ and w_- using either (14) (for local polynomial estimator), or by solving the system of equations given in Supplemental Appendix E (for the optimal estimator). Compute the worst case bias (11), and estimate the variance as $\widehat{\mathrm{sd}}(\hat{L}_{h_+,h_-})^2 = \sum_i (w_+(x_i, h_+) + w_-(x_i, h_-))^2 \hat{\sigma}^2(x_i)$.

3. Find the smoothing parameters h_+^* and h_-^* that minimize the β-quantile of excess length

$$2\,\overline{\mathrm{bias}}_{\mathcal{F}_{RDT,p}(c)}(\hat{L}_{h_+,h_-}) + \mathrm{sd}(\hat{L}_{h_+,h_-})(z_{1-\alpha} + z_\beta). \tag{15}$$

   for a given β. The choice $\beta = 0.8$, corresponds to a benchmark used in statistical power analysis (see Cohen, 1988). For two-sided CIs, minimize the length

$$2\widehat{\mathrm{sd}}(\hat{L}_{h_+,h_-})\,\mathrm{cv}_\alpha\left(\overline{\mathrm{bias}}_{\mathcal{F}_{RDT,p}(C)}(\hat{L}_{h_+,h_-})/\widehat{\mathrm{sd}}(\hat{L}_{h_+,h_-})\right). \tag{16}$$

[3] In the empirical application in Section 4, we use estimates based on local linear regression residuals.

4. Construct the CI using (12) (for one-sided CIs), or (13) (for two-sided CIs), based on $\hat{L}_{h_+^*,h_-^*}$, with $\widehat{\text{sd}}(\hat{L}_{h_+^*,h_-^*})$ in place of the (infeasible) true standard deviation.

Remark 2.1. The variance estimator in step 1 leads to asymptotically valid and optimal inference even when $\sigma^2(x)$ is non-constant, so long as it is smooth on either side of the cutoff. However, finite-sample properties of the resulting CI may not be good if heteroskedasticity is important for the sample size at hand. We therefore recommend using the variance estimator

$$\widehat{\text{sd}}_{\text{robust}}(\hat{L}_{h_+^*,h_-^*})^2 = \sum_{i=1}^{n}(w_+(x_i,h_+) + w_-(x_i,h_-))^2\hat{u}_i^2 \tag{17}$$

instead of $\widehat{\text{sd}}(\hat{L}_{h_+^*,h_-^*})$ in step 4, where $\hat{u}_i^2$ is an estimate of $\sigma^2(x_i)$. When using local polynomial regression, one can set $\hat{u}_i$ to the ith regression residual, in which case (17) reduces to the usual Eicker-Huber-White estimator. Alternatively, one can use the nearest-neighbor estimator (Abadie and Imbens, 2006) $\hat{u}_i^2 = \frac{J}{J+1}(Y_i - J^{-1}\sum_{\ell=1}^{J} Y_{j_\ell(i)})^2$, where $j_\ell(i)$ is the ℓth closest unit to i among observations on the same side of the cutoff, and $J \geq 1$ (we use $J = 3$ in the application in Section 4, following Calonico et al., 2014). This mirrors the common practice of assuming homoskedasticity to compute the optimal weights, but allowing for heteroskedasticity when performing inference, such as using OLS in the linear regression model (which is efficient under homoskedasticity) along with heteroskedasticity-robust standard errors.

Remark 2.2. If one is interested in estimation, rather than inference, one can choose h_+ and h_- that minimize the worst-case mean-squared error (MSE) $\overline{\text{bias}}_{\mathcal{F}_{RDT,p}(C)}(\hat{L}_{h_+,h_-})^2 + \text{sd}(\hat{L}_{h_+,h_-})^2$ instead of the CI criteria in step 3. One can form a CI around this estimator by simply following step 4 with this choice of h_+ and h_-. In the application in Section 4, we find that little efficiency is lost by using MSE-optimal smoothing parameters, relative to using h_+ and h_- that minimize the CI length (16). Interestingly, we find that smoothing parameters that minimize the CI length actually oversmooth slightly relative to the MSE optimal smoothing parameters. We generalize these findings in an asymptotic setting in Armstrong and Kolesár (2016b).

Remark 2.3. Often, a set of covariates z_i will be available that does not depend on the treatment, but that may be correlated with the outcome variable y_i. If the parameter of interest is still the average treatment effect for units with $x_i = 0$, one can simply ignore these covariates. Alternatively, to gain additional precision, as suggested in Calonico et al. (2017), one can run a local polynomial regression, but with the covariates added linearly. In

Appendix B, we show that this approach is near-optimal if one places smoothness assumptions on the conditional mean of $\tilde{y}_i$ given x_i, where $\tilde{y}_i$ is the outcome with the effect of z_i partialled out. If one is interested in the treatment effect as a function of z (with x still set to zero), one can use our general framework by considering the model $y_i = f(x_i, z_i) + u_i$, specifying a smoothness class for f, and constructing CIs for $\lim_{x\downarrow 0} f(x,z) - \lim_{x\uparrow 0} f(x,z)$ for different values of z. See Appendix B for details.

A final consideration in implementing these CIs in practice is the choice of the smoothness constants C and p. The choice of p depends on the order of the derivative the researcher wishes to bound. Since much of empirical practice in RD is justified by asymptotic MSE optimality results for $\mathcal{F}_{RDT,2}(C)$ (in particular, this class justifies the use of local linear estimators), we recommend $p = 2$ as a default choice. For C, generalizations of the non-adaptivity results described in Section 2.1 show that the researcher must choose C a priori, rather than attempting to use the data to choose C. To assess the sensitivity of the results to different smoothness assumptions on f, we recommend considering a range of plausible choices for C. We implement this approach for our empirical application in Section 4.

3 General characterization of optimal procedures

We consider the following setup and notation, much of which follows Donoho (1994). We observe data Y of the form

$$Y = Kf + \sigma\varepsilon \tag{18}$$

where f is known to lie in a convex subset $\mathcal{F}$ of a vector space, and $K : \mathcal{F} \to \mathcal{Y}$ is a linear operator between $\mathcal{F}$ and a Hilbert space $\mathcal{Y}$. We denote the inner product on $\mathcal{Y}$ by $\langle \cdot, \cdot \rangle$, and the norm by $\| \cdot \|$. The error ε is standard Gaussian with respect to this inner product: for any $g \in \mathcal{Y}$, $\langle \varepsilon, g \rangle$ is normal with $E\langle \varepsilon, g \rangle = 0$ and $\text{var}\left(\langle \varepsilon, g \rangle\right) = \|g\|^2$. We are interested in constructing a confidence set for a linear functional Lf.

The RD model (1) fits into this setup by setting $Y = (y_1/\sigma(x_1), \ldots, y_n/\sigma(x_n))'$, $\mathcal{Y} = \mathbb{R}^n$, $Kf = (f(x_1)/\sigma(x_1), \ldots, f(x_n)/\sigma(x_n))'$, $Lf = \lim_{x\downarrow 0} f(x) - \lim_{x\uparrow 0} f(x)$ and $\langle x, y \rangle$ given by the Euclidean inner product $x'y$. As we discuss in detail in Supplemental Appendix D.1, our setup covers a number of other important models, including average treatment effects under unconfoundedness, the partly linear model, constraints on the sign or magnitude of parameters in the linear regression model, and other parametric models.

3.1 Performance criteria

Let us now define the performance criteria that we use to evaluate confidence sets for Lf. A set $\mathcal{C} = \mathcal{C}(Y)$ is called a $100 \cdot (1-\alpha)\%$ confidence set for Lf if $\inf_{f \in \mathcal{F}} P_f\left(Lf \in \mathcal{C}\right) \geq 1-\alpha$. We denote the collection of all $100 \cdot (1-\alpha)\%$ confidence sets by $\mathcal{I}_\alpha$.

We can compare performance of confidence sets at a particular $f \in \mathcal{F}$ using expected length, $E_f \lambda(\mathcal{C})$, where λ is Lebesgue measure. Allowing confidence sets to have arbitrary form may make them difficult to interpret or even compute. One way of avoiding this is to restrict attention to confidence sets that take the form of a fixed-length confidence interval (CI), an interval of the form $[\hat{L} - \chi, \hat{L} + \chi]$ for some estimate $\hat{L}$ and nonrandom χ (for instance, in the RD model (1), χ may depend on the running variable x_i and $\sigma^2(x_i)$, but not on y_i). Let

$$\chi_\alpha(\hat{L}) = \min\left\{\chi\colon \inf_{f \in \mathcal{F}} P_f\big(|\hat{L} - Lf| \leq \chi\big) \geq 1-\alpha\right\}$$

denote the half-length of the shortest fixed-length $100 \cdot (1-\alpha)\%$ CI centered around an estimator $\hat{L}$. Fixed-length CIs are easy to compare: one simply prefers the one with the shortest half-length. On the other hand, their length cannot "adapt" to reflect greater precision for different functions $f \in \mathcal{F}$. To address this concern, in Section 3.4, we compare the length of fixed-length CIs to sharp bounds on the optimal expected length $\inf_{\mathcal{C} \in \mathcal{I}_\alpha} E_f(\mathcal{C})$.

If $\mathcal{C}$ is restricted to take the form of a one-sided CI $[\hat{c}, \infty)$, we cannot use expected length as a criterion. We therefore measure performance at a particular parameter f using the βth quantile of their excess length $Lf - \hat{c}$, which we denote by $q_{f,\beta}(Lf - \hat{c})$. To measure performance globally over some set $\mathcal{G}$, we use the maximum βth quantile of the excess length,

$$q_\beta(\hat{c}, \mathcal{G}) = \sup_{g \in \mathcal{G}} q_{g,\beta}(Lg - \hat{c}). \tag{19}$$

If $\mathcal{G} = \mathcal{F}$, minimizing $q_\beta(\hat{c}, \mathcal{F})$ over one-sided CIs in the set $\mathcal{I}_\alpha$ gives minimax excess length. If $\mathcal{G} \subset \mathcal{F}$ is a class of smoother functions, minimizing $q_\beta(\hat{c}, \mathcal{G})$ yields CIs that direct power: they achieve good performance when f is smooth, while maintaining coverage over all of $\mathcal{F}$. A CI that achieves good performance over multiple classes $\mathcal{G}$ is said to be "adaptive" over these classes. In Section 3.3, we give sharp bounds on (19) for a single class $\mathcal{G}$, which gives a benchmark for adapting over multiple classes (cf. Cai and Low, 2004a).

3.2 Affine estimators and optimal bias-variance tradeoff

Many popular estimators are linear functions of the outcome variable Y, and we will see below that optimal or near-optimal CIs are based on estimators of this form. In the general framework (18), linear estimators take the form $\langle w, Y\rangle$ for some non-random $w \in \mathcal{Y}$, which simplifies to (10) in the RD model. It will be convenient to allow for a recentering by some constant $a \in \mathbb{R}$, which leads to an affine estimator $\hat{L} = a + \langle w, Y\rangle$.

For any estimator $\hat{L}$, let $\overline{\text{bias}}_{\mathcal{G}}(\hat{L}) = \sup_{f\in\mathcal{G}} E_f(\hat{L}-Lf)$ and $\underline{\text{bias}}_{\mathcal{G}}(\hat{L}) = \inf_{f\in\mathcal{G}} E_f(\hat{L}-Lf)$. An affine estimator $\hat{L} = a + \langle w, Y\rangle$ follows a normal distribution with mean $E_f\hat{L} = a + \langle w, Kf\rangle$ and variance $\text{var}(\hat{L}) = \|w\|^2\sigma^2$, which does not depend on f. Thus, the set of possible distributions for $\hat{L} - Lf$ as f varies over a given convex set $\mathcal{G}$ is given by the set of normal distributions with variance $\|w\|^2\sigma^2$ and mean between $\underline{\text{bias}}_{\mathcal{G}}(\hat{L})$ and $\overline{\text{bias}}_{\mathcal{G}}(\hat{L})$. It follows that a one-sided CI based on an affine estimator $\hat{L}$ is given by

$$[\hat{c}, \infty) \qquad \hat{c} = \hat{L} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - \text{sd}(\hat{L})z_{1-\alpha}, \tag{20}$$

with $z_{1-\alpha}$ denoting the $1-\alpha$ quantile of a standard normal distribution, and that its worst-case βth quantile excess length over a convex class $\mathcal{G}$ is

$$q_\beta(\hat{c}, \mathcal{G}) = \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - \underline{\text{bias}}_{\mathcal{G}}(\hat{L}) + \text{sd}(\hat{L})(z_{1-\alpha} + z_\beta). \tag{21}$$

The shortest fixed-length CI centered at the affine estimator $\hat{L}$ is given by

$$\hat{L} \pm \chi_\alpha(\hat{L}), \qquad \chi_\alpha(\hat{L}) = \text{cv}_\alpha\left(\frac{\max\{|\overline{\text{bias}}_{\mathcal{F}}(\hat{L})|, |\underline{\text{bias}}_{\mathcal{F}}(\hat{L})|\}}{\text{sd}(\hat{L})}\right)\cdot \text{sd}(\hat{L}), \tag{22}$$

where $\text{cv}_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $\mathcal{N}(t, 1)$ random variable, as tabulated in Table 1.

The fact that optimal CIs turn out to be based on affine estimators reduces the derivation of optimal CIs to bias-variance calculations: since the performance of CIs based on affine estimators depends only on the variance and worst-case bias, one simply minimizes worst-case bias subject to a bound on variance, and then trades off bias and variance in a way that is optimal for the given criterion. The main tool for doing this is the ordered modulus of continuity between $\mathcal{F}$ and $\mathcal{G}$ (Cai and Low, 2004a),

$$\omega(\delta; \mathcal{F}, \mathcal{G}) = \sup\{Lg - Lf \colon \|K(g-f)\| \le \delta, f \in \mathcal{F}, g \in \mathcal{G}\}$$

for any sets $\mathcal{F}$ and $\mathcal{G}$ with a non-empty intersection (so that the set over which the supremum is taken is non-empty). When $\mathcal{G} = \mathcal{F}$, $\omega(\delta; \mathcal{F}, \mathcal{F})$ is the (single-class) modulus of continuity over $\mathcal{F}$ (Donoho and Liu, 1991), and we denote it by $\omega(\delta; \mathcal{F})$. The ordered modulus $\omega(\cdot; \mathcal{F}, \mathcal{G})$ is concave, which implies that the superdifferential at δ (the set of slopes of tangent lines at $(\delta, \omega(\delta; \mathcal{F}, \mathcal{G}))$) is nonempty for any $\delta > 0$. Throughout the paper, we let $\omega'(\delta; \mathcal{F}, \mathcal{G})$ denote an (arbitrary unless otherwise stated) element in this set. Typically, $\omega(\cdot; \mathcal{F}, \mathcal{G})$ is differentiable, in which case $\omega'(\delta; \mathcal{F}, \mathcal{G})$ is defined uniquely as the derivative at δ. We use $g^*_{\delta,\mathcal{F},\mathcal{G}}$ and $f^*_{\delta,\mathcal{F},\mathcal{G}}$ to denote a solution to the ordered modulus problem (assuming it exists), and $f^*_{M,\delta,\mathcal{F},\mathcal{G}} = (f^*_{\delta,\mathcal{F},\mathcal{G}} + g^*_{\delta,\mathcal{F},\mathcal{G}})/2$ to denote the midpoint.[4]

We will show that optimal decision rules will in general depend on the data Y through an affine estimator of the form

$$\hat{L}_{\delta,\mathcal{F},\mathcal{G}} = L f^*_{M,\delta,\mathcal{F},\mathcal{G}} + \frac{\omega'(\delta; \mathcal{F}, \mathcal{G})}{\delta} \left\langle K(g^*_{\delta,\mathcal{F},\mathcal{G}} - f^*_{\delta,\mathcal{F},\mathcal{G}}), Y - K f^*_{M,\delta,\mathcal{F},\mathcal{G}} \right\rangle, \tag{23}$$

with δ and $\mathcal{G}$ depending on the optimality criterion. When $\mathcal{F} = \mathcal{G}$, we denote the estimator $\hat{L}_{\delta,\mathcal{F},\mathcal{F}}$ by $\hat{L}_{\delta,\mathcal{F}}$. When the sets $\mathcal{F}$ and $\mathcal{G}$ are clear from the context, we use $\omega(\delta)$, $\hat{L}_\delta$, f^*_δ, g^*_δ and $f^*_{M,\delta}$ in place of $\omega(\delta; \mathcal{F}, \mathcal{G})$, $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$, $f^*_{\delta,\mathcal{F},\mathcal{G}}$, $g^*_{\delta,\mathcal{F},\mathcal{G}}$ and $f^*_{M,\delta,\mathcal{F},\mathcal{G}}$ to avoid notational clutter.

As we show in Lemma A.1 in the Appendix, a useful property of $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ is that its maximum bias over $\mathcal{F}$ and minimum bias over $\mathcal{G}$ are attained at f^*_δ and g^*_δ, respectively, and are given by

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) = -\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) = \frac{1}{2}\left(\omega(\delta; \mathcal{F}, \mathcal{G}) - \delta\omega'(\delta; \mathcal{F}, \mathcal{G})\right). \tag{24}$$

Its standard deviation equals $\text{sd}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) = \sigma\omega'(\delta; \mathcal{F}, \mathcal{G})$, and doesn't depend on f. As remarked by Cai and Low (2004b), no estimator can simultaneously achieve lower maximum bias over $\mathcal{F}$, higher minimum bias over $\mathcal{G}$, and lower variance than the estimators in the class $\{\hat{L}_{\delta,\mathcal{F},\mathcal{G}}\}_{\delta>0}$. Estimators (23) can thus be used to optimally trade off various levels of bias and variance.

A condition that will play a central role in bounding the gains from directing power at smooth functions is *centrosymmetry*. We say that a class $\mathcal{F}$ is *centrosymmetric* if $f \in \mathcal{F} \Longrightarrow -f \in \mathcal{F}$. Under centrosymmetry, the functions that solve the single-class modulus problem

[4]See Supplemental Appendix D.2 for sufficient conditions for differentiability and a discussion of the non-differentiable case. Regarding existence of a solution to the modulus problem, we verify this directly for our RD application in Supplemental Appendix E.2; see also Donoho (1994), Lemma 2 for a general set of sufficient conditions.

can be taken to satisfy $g^*_\delta = -f^*_\delta$, and the modulus is given by

$$\omega(\delta; \mathcal{F}) = \sup \{2Lf \colon \|Kf\| \leq \delta/2, f \in \mathcal{F}\}. \tag{25}$$

Since $f^*_\delta = -g^*_\delta$, $f^*_{M,\delta}$ is the zero function and $\hat{L}_{\delta,\mathcal{F}}$ is linear:

$$\hat{L}_{\delta,\mathcal{F}} = \frac{2\omega'(\delta; \mathcal{F})}{\delta} \langle Kg^*_\delta, Y \rangle. \tag{26}$$

In the RD model (1) the class $\mathcal{F}_{RDT,p}(C)$ is centrosymmetric, and the estimator $\hat{L}_{\delta,\mathcal{F}_{RDT,p}(C)}$ takes the form $\hat{L}_{h_+,h_-}$ given in (10) for a certain class of weights $w_+(x, h_+)$ and $w_-(x, h_-)$, with the smoothing parameters h_+ and h_- both determined by δ (see Supplemental Appendix E).

3.3 Optimal one-sided CIs

Given β, a one-sided CI that minimizes (19) among all one-sided CIs with level $1 - \alpha$ is based on $\hat{L}_{\delta_\beta;\mathcal{F},\mathcal{G}}$, where $\delta_\beta = \sigma(z_\beta + z_{1-\alpha})$.

Theorem 3.1. *Let $\mathcal{F}$ and $\mathcal{G}$ be convex with $\mathcal{G} \subseteq \mathcal{F}$, and suppose that f^*_δ and g^*_δ achieve the ordered modulus at δ with $\|K(f^*_\delta - g^*_\delta)\| = \delta$. Let*

$$\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = \hat{L}_{\delta,\mathcal{F},\mathcal{G}} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta,\mathcal{F},\mathcal{G}}) - z_{1-\alpha}\sigma\omega'(\delta; \mathcal{F}, \mathcal{G}).$$

Then $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ minimizes $q_\beta(\hat{c}, \mathcal{G})$ for $\beta = \Phi(\delta/\sigma - z_{1-\alpha})$ among all one-sided $1-\alpha$ CIs, where Φ denotes the standard normal cdf. The minimum coverage is taken at f^*_δ and equals $1 - \alpha$. All quantiles of excess length are maximized at g^*_δ. The worst case βth quantile of excess length is $q_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}, \mathcal{G}) = \omega(\delta; \mathcal{F}, \mathcal{G})$.

Since the worst-case bias of $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ is given by (24), and its standard deviation equals $\sigma\omega'(\delta; \mathcal{F}, \mathcal{G})$, it can be seen that $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ takes the form given in (20), and its worst-case excess length follows (21). The assumption that the modulus is achieved with $\|K(f^*_\delta - g^*_\delta)\| = \delta$ rules out degenerate cases: if $\|K(f^*_\delta - g^*_\delta)\| < \delta$, then relaxing this constraint does not increase the modulus, which means that $\omega'(\delta; \mathcal{F}, \mathcal{G}) = 0$ and the optimal CI does not depend on the data.

Implementing the CI from Theorem 3.1 requires the researcher to choose a quantile β to optimize, and to choose the set $\mathcal{G}$. There are two natural choices for β. If the objective is to optimize the performance of the CI "on average", then optimizing the median excess length

($\beta = 0.5$) is a natural choice. Since for any CI $[\hat{c}, \infty)$ such that $\hat{c}$ is affine in the data Y, the median and expected excess lengths coincide, and since $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ is affine in the data, setting $\beta = 0.5$ also has the advantage that it minimizes the expected excess length among affine CIs. Alternatively, if the CI is being computed as part of a power analysis, then setting $\beta = 0.8$ is natural, as, under conditions given in Supplemental Appendix D.2, it translates directly to statements about 80% power, a standard benchmark in such analyses (Cohen, 1988).

For the set $\mathcal{G}$, there are two leading choices. First, setting $\mathcal{G} = \mathcal{F}$ yields minimax CIs:

Corollary 3.1 (One-sided minimax CIs)**.** *Let $\mathcal{F}$ be convex, and suppose that f^*_δ and g^*_δ achieve the single-class modulus at δ with $\|K(f^*_\delta - g^*_\delta)\| = \delta$. Let*

$$\hat{c}_{\alpha,\delta,\mathcal{F}} = \hat{L}_{\delta,\mathcal{F}} - \frac{1}{2}\left(\omega(\delta;\mathcal{F}) - \delta\omega'(\delta;\mathcal{F})\right) - z_{1-\alpha}\sigma\omega'(\delta;\mathcal{F}).$$

Then, for $\beta = \Phi(\delta/\sigma - z_{1-\alpha})$, $\hat{c}_{\alpha,\delta,\mathcal{F}}$ minimizes the maximum βth quantile of excess length among all $1-\alpha$ CIs for Lf. The minimax excess length is given by $\omega(\delta;\mathcal{F})$.

The minimax criterion may be considered overly pessimistic: it focuses on controlling the excess length under the least favorable function. This leads to the second possible choice for $\mathcal{G}$, a smaller convex class of smoother functions $\mathcal{G} \subset \mathcal{F}$. The resulting CIs will then achieve the best possible performance when f is smooth, while maintaining coverage over all of $\mathcal{F}$. Unfortunately, there is little scope for improvement for such a CI when $\mathcal{F}$ is centrosymmetric. In particular, suppose that $g^*_{\delta,\mathcal{F},\mathcal{G}}$ is "sufficiently smooth" relative to $\mathcal{F}$, in the sense that

$$f - g^*_{\delta,\mathcal{F},\mathcal{G}} \in \mathcal{F} \quad \text{for all } f \in \mathcal{F}. \tag{27}$$

Since $\mathcal{F}$ is centrosymmetric, this condition is equivalent to the requirement that the sets $\{f - g^*_{\delta,\mathcal{F},\mathcal{G}} \colon f \in \mathcal{F}\}$ and $\mathcal{F}$ are the same.[5] For instance, (27) holds if $\mathcal{G}$ contains the zero function only. In the RD model (1) with $\mathcal{F} = \mathcal{F}_{RDT,p}(C)$, (27) holds if $\mathcal{G} = \mathcal{F}_{RDT,p}(0)$, the class of piecewise polynomial functions.

Corollary 3.2. *Let $\mathcal{F}$ be centrosymmetric, and let $\mathcal{G} \subseteq \mathcal{F}$ be any convex set such that the solution to the ordered modulus problem exists and satisfies* (27) *with $\|K(f^*_{\delta_\beta} - g^*_{\delta_\beta})\| = \delta_\beta$, where $\delta_\beta = \sigma(z_\beta + z_{1-\alpha})$. Then the one-sided CI $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ that is minimax for the βth quantile also optimizes $q_{\tilde{\beta}}(\hat{c};\mathcal{G})$, where $\tilde{\beta} = \Phi((z_\beta - z_{1-\alpha})/2)$. In particular, $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ optimizes*

[5] We thank a referee for pointing this out.

$q_{\tilde{\beta}}(\hat{c}; \{0\})$. *Moreover, the efficiency of* $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ *for the* β*th quantile of maximum excess length over* $\mathcal{G}$ *is given by*

$$\frac{\inf_{\hat{c}:\, [\hat{c},\infty)\in\mathcal{I}_\alpha} q_\beta(\hat{c}, \mathcal{G})}{q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}, \mathcal{G})} = \frac{\omega(\delta_\beta; \mathcal{F}, \mathcal{G})}{q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}, \mathcal{G})} = \frac{\omega(2\delta_\beta; \mathcal{F})}{\omega(\delta_\beta; \mathcal{F}) + \delta_\beta \omega'(\delta_\beta; \mathcal{F})}. \tag{28}$$

The first part of Corollary 3.2 states that minimax CIs that optimize a particular quantile β will also minimize the maximum excess length over $\mathcal{G}$ at a different quantile $\tilde{\beta}$. For instance, a CI that is minimax for median excess length among 95% CIs also optimizes $\Phi(-z_{0.95}/2) \approx 0.205$ quantile under the zero function. Vice versa, the CI that optimizes median excess length under the zero function is minimax for the $\Phi(2z_{0.5} + z_{0.95}) = 0.95$ quantile.

The second part of Corollary 3.2 gives the exact cost of optimizing the "wrong" quantile $\tilde{\beta}$. Since the one-class modulus is concave, $\delta\omega'(\delta) \leq \omega(\delta)$, and we can lower bound the efficiency of $\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}$ given in (28) by $\omega(2\delta_\beta)/(2\omega(\delta_\beta)) \geq 1/2$. Typically, the efficiency is much higher. In particular, in the regression model (1), the one-class modulus satisfies

$$\omega(\delta; \mathcal{F}) = n^{-r/2} A \delta^r (1 + o(1)) \tag{29}$$

for many choices of $\mathcal{F}$ and L, as $n \to \infty$ for some constant A, where $r/2$ is the rate of convergence of the minimax root MSE. This is the case under regularity conditions in the RD model with $r = 2p/(2p+1)$ by Lemma H.6 (see Donoho and Low, 1992, for other cases where (29) holds). In this case, (28) evaluates to $\frac{2^r}{1+r}(1 + o(1))$, so that the asymptotic efficiency depends only on r. Figure 2 plots the asymptotic efficiency as a function of r. Since adapting to the zero function easier than adapting to any set $\mathcal{G}$ that includes it, if $\mathcal{F}$ is convex and centrosymmetric, "directing power" yields very little gain in excess length no matter how optimistic one is about where to direct it.

This result places a severe bound on the scope for adaptivity in settings in which $\mathcal{F}$ is convex and centrosymmetric: any CI that performs better than the minimax CI by more than the ratio in (28) must fail to control coverage at some $f \in \mathcal{F}$.

3.4 Two-sided CIs

A fixed-length CI based on $\hat{L}_{\delta,\mathcal{F}}$ can be computed by plugging its worst-case bias (24) into (22),[6]

$$\hat{L}_{\delta,\mathcal{F}} \pm \chi_\alpha(\hat{L}_{\delta,\mathcal{F}}), \qquad \chi_\alpha(\hat{L}_{\delta,\mathcal{F}}) = \mathrm{cv}_\alpha\left(\tfrac{\omega(\delta;\mathcal{F})}{2\sigma\omega'(\delta;\mathcal{F})} - \tfrac{\delta}{2\sigma}\right) \cdot \sigma\omega'(\delta;\mathcal{F}).$$

The optimal δ minimizes the half-length, $\delta_\chi = \operatorname{argmin}_{\delta>0} \chi_\alpha(\hat{L}_{\delta,\mathcal{F}})$. It follows from Donoho (1994) that this CI is the shortest possible in the class of fixed-length CIs based on affine estimators. Just as with minimax one-sided CIs, one may worry that since its length is driven by the least favorable functions, restricting attention to fixed-length CIs may be costly when the true f is smoother. The next result characterizes confidence sets that optimize expected length at a single function g, and thus bounds the possible performance gain.

Theorem 3.2. *Let $g \in \mathcal{F}$, and assume that a minimizer f_{L_0} of $\|K(g-f)\|$ subject to $Lf = L_0$ and $f \in \mathcal{F}$ exists for all $L_0 \in \mathbb{R}$. Then the confidence set $\mathcal{C}_g$ that minimizes $E_g\lambda(\mathcal{C})$ subject to $\mathcal{C} \in \mathcal{I}_\alpha$ inverts the family of tests ϕ_{L_0} that reject for large values of $\langle K(g-f_{L_0}), Y\rangle$ with critical value given by the $1-\alpha$ quantile under f_{L_0}. Its expected length is*

$$E_g[\lambda(\mathcal{C}_g)] = (1-\alpha)E\left[(\omega(\sigma(z_{1-\alpha}-Z);\mathcal{F},\{g\}) + \omega(\sigma(z_{1-\alpha}-Z);\{g\},\mathcal{F})) \mid Z \le z_{1-\alpha}\right],$$

where Z is a standard normal random variable.

This result solves the problem of "adaptation to a function" posed by Cai et al. (2013), who obtain bounds for this problem if $\mathcal{C}$ is required to be an interval. The theorem uses the observation in Pratt (1961) that minimum expected length CIs are obtained by inverting a family of uniformly most powerful tests of $H_0\colon Lf = L_0$ and $f \in \mathcal{F}$ against $H_1\colon f = g$, which, as shown in the proof, is given by ϕ_{L_0}; the expression for the expected length of $\mathcal{C}_g$ follows by computing the power of these tests. The assumption on the existence of the minimizer f_{L_0} means that Lf is unbounded over $\mathcal{F}$, and it is made to simplify the statement; a truncated version of the same formula holds when $\mathcal{F}$ places a bound on Lf.

Directing power at a single function is seldom desirable in practice. Theorem 3.2 is very useful, however, in bounding the efficiency of other procedures. In particular, suppose $f-g \in \mathcal{F}$ for all f, so that (27) holds with $\mathcal{G} = \{g\}$ (such as when g is the zero function), and

[6] We assume that $\omega'(\delta;\mathcal{F}) = \mathrm{sd}(\hat{L}_{\delta,\mathcal{F}})/\sigma \neq 0$. Otherwise, the estimator $\hat{L}_{\delta,\mathcal{F}}$ doesn't depend on the data, and the only valid fixed-length CI around it is the trivial CI that reports the whole parameter space for Lf.

that $\mathcal{F}$ is centrosymmetric. Then, by arguments in the proof of Corollary 3.2, $\omega(\delta; \mathcal{F}, \{g\}) = \omega(\delta; \{g\}, \mathcal{F}) = \frac{1}{2}\omega(2\delta; \mathcal{F})$, which yields:

Corollary 3.3. *Consider the setup in Theorem 3.2 with the additional assumption that $\mathcal{F}$ is centrosymmetric and g satisfies $f - g \in \mathcal{F}$ for all f. Then the efficiency of the fixed-length CI around $\hat{L}_{\delta_\chi,\mathcal{F}}$ at g relative to all confidence sets is*

$$\frac{\inf_{\mathcal{C}\in\mathcal{I}_\alpha} E_g\lambda(\mathcal{C}(Y))}{2\chi_\alpha(\hat{L}_{\delta_\chi,\mathcal{F}})} = \frac{(1-\alpha)E\left[\omega(2\sigma(z_{1-\alpha}-Z);\mathcal{F}) \mid Z \le z_{1-\alpha}\right]}{2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta_\chi;\mathcal{F})}{2\sigma\omega'(\delta_\chi;\mathcal{F})} - \frac{\delta_\chi}{2\sigma}\right)\cdot\sigma\omega'(\delta_\chi;\mathcal{F})}. \tag{30}$$

The efficiency ratio (30) can easily be computed in particular applications, and we do so in the empirical application in Section 4. When the one-class modulus satisfies (29), then, as in the case of one-sided CIs, the asymptotic efficiency of the fixed-length CI around $\hat{L}_{\delta_\chi}$ can be shown to depend only on r and α, and we plot it in Figure 2 for $\alpha = 0.05$ (see Theorem E.1 for the formula). When $r = 1$ (parametric rate of convergence) and $\alpha = 0.05$, the asymptotic efficiency equals 84.99%, as in the normal mean example in Pratt (1961, Section 5).

Just like with minimax one-sided CIs, this result places a severe bound on the scope for improvement over fixed-length CIs when $\mathcal{F}$ is centrosymmetric. It strengthens the finding in Low (1997) and Cai and Low (2004a), who derive bounds on the expected length of random length $1-\alpha$ CIs. Their bounds imply that when $\mathcal{F}$ is constrained only by bounds on a derivative, the expected length of any CI in $\mathcal{I}_\alpha$ must shrink at the minimax rate $n^{-r/2}$ for any g in the interior of $\mathcal{F}$.[7] Figure 2 shows that for smooth functions g, this remains true whenever $\mathcal{F}$ is centrosymmetric, even if we don't require $\mathcal{C}$ to take the form of an interval. Importantly, the figure also shows that not only is the rate the same as the minimax rate, the constant must be close to that for fixed-length CIs. Since adapting to a single function g is easier than adapting to any class $\mathcal{G}$ that includes it, this result effectively rules out adaptation to subclasses of $\mathcal{F}$ that contain smooth functions.

4 Empirical illustration

In this section, we illustrate the theoretical results in an RD application using a dataset from Lee (2008). The dataset contains 6,558 observations on elections to the US House

[7] One can use Theorem 3.2 to show that this result holds even if we don't require $\mathcal{C}$ to take the form of an interval. For example, in the RD model with $\mathcal{F} = \mathcal{F}_{RDT,p}(C)$ and $g \in \mathcal{F}_{RDT,p}(C_g)$, $C_g < C$, the result follows from lower bounding $E_g[\lambda(\mathcal{C}_g)]$ using $\omega(\delta; \mathcal{F}, \{g\}) + \omega(\delta; \{g\}, \mathcal{F}) \ge \omega(2\delta, \mathcal{F}_{RDT,p}(C - C_g))$.

of Representatives between 1946 and 1998. The running variable $x_i \in [-100, 100]$ is the Democratic margin of victory (in percentages) in election i. The outcome variable $y_i \in [0, 100]$ is the Democratic vote share (in percentages) in the next election. Given the inherent uncertainty in final vote counts, the party that wins is essentially randomized in elections that are decided by a narrow margin, so that the RD parameter Lf measures the incumbency advantage for Democrats for elections decided by a narrow margin—the impact of being the current incumbent party in a congressional district on the vote share in the next election.

We consider inference under the Taylor class $\mathcal{F}_{RDT,p}(C)$, with $p = 2$. We report results for the optimal estimators and CIs, as well as CIs based on local linear estimators, using the formulas described in Section 2.2 (which follow from the general results in Section 3). We use the preliminary estimates $\hat{\sigma}^2_+(x) = 12.6^2$ and $\hat{\sigma}^2_-(x) = 10.8^2$ in Step 1, which are based on residuals form a local linear regression with bandwidth selected using the Imbens and Kalyanaraman (2012) selector. In Step 4, we use the nearest-neighbor variance estimator with $J = 3$.

Let us briefly discuss the interpretation of the smoothness constant C in this application. By definition of the class $\mathcal{F}_{RDT,2}(C)$, C determines how large the approximation error can be if we approximate the regression functions f_+ and f_- on either side of the cutoff by a linear Taylor approximation at the cutoff: the approximation error is no greater than Cx^2. One way of gauging the magnitude of this approximation error is to look at its effect on prediction error when using the Taylor approximation to predict the vote share in the next election, and the margin in the previous election was x_0. If one uses the Taylor approximation, the prediction MSE is at most $C^2x_0^4 + \sigma^2(x_0)$, whereas using the true conditional mean to predict the vote share would lead to prediction MSE $\sigma^2(x_0)$. Thus, using the true conditional mean leads to a MSE reduction in this prediction problem by a factor of at most $C^2x_0^4/(C^2x_0^4 + \sigma^2(x_0))$. If $C = 0.05$ for instance, this implies MSE reductions of at most 13.6% at $x_0 = 10\%$, and 71.5% at $x_0 = 20\%$, assuming that $\sigma^2(x_0)$ equals our estimate of 12.6^2. To the extent that researchers agree that the vote share in the next election varies smoothly enough with the margin of victory in the current election to make such large reductions in MSE unlikely, $C = 0.05$ is quite a conservative choice.

Our adaptivity bounds imply that one cannot use data-driven methods to tighten our CIs, by say, estimating C. It is, however, possible to lower-bound the value of C. We derive a simple estimate of this lower bound in Supplemental Appendix E.3, which in the Lee data yields the lower bound estimate 0.017. As detailed in the appendix, the lower bound estimate can also be used in a model specification test to check whether a given chosen value of C is

too low. To examine sensitivity of the results to different choices of C, we present the results for the range $C \in [0.0002, 0.1]$ that, by the argument in the preceding paragraph, includes most plausible values.

4.1 Optimal and near-optimal confidence intervals

The top panel in Figure 3 plots the optimal one- and two-sided CIs defined in Section 2, as well as estimates based on minimizing the worst-case MSE (see Remark 2.2). The estimates vary between 5.8% and 7.4% for $C \geq 0.005$, which is close to the original Lee estimate of 7.7% that was based on a global fourth degree polynomial. Interestingly, the lower and upper limits $\hat{c}_u$ and $\hat{c}_\ell$ of the one-sided CIs $[\hat{c}_\ell, \infty)$ and $(-\infty, \hat{c}_u]$ are not always within the corresponding limits for the two-sided CIs. The reason for this is that for any given C, the optimal smoothing parameters h_+ and h_- are smaller for one-sided CIs than for two-sided fixed-length CIs. Thus, when the point estimate decreases with the amount of smoothing as is the case for low values of C, then one-sided CIs are effectively centered around a lower estimate, which explains why at first the one-sided CI limits are both below the two-sided limits. This reverses once the point estimate starts increasing with the amount of smoothing. Furthermore, the optimal smoothing parameters for the minimax MSE estimator are slightly *smaller* than those for fixed-length CIs throughout the entire range of Cs, albeit by a small amount. This matches the asymptotic predictions in Armstrong and Kolesár (2016b).

As we discussed in Remark 2.2, it may be desirable to report an estimate with good MSE, with a CI centered at this estimate (without reoptimizing the smoothing parameters). The bottom panel in Figure 3 gives CIs with the smoothing parameters chosen so that the $\hat{L}_{h_+,h_-}$ minimizes the maximum MSE. The limits of the one-sided CIs are now contained within the two-sided CIs, as they are both based on the same estimator, although they are less than $(z_{1-\alpha/2} - z_{1-\alpha})\,\mathrm{sd}(\hat{L}_{h_+,h_-})$ apart as would be the case if $\hat{L}_{h_+,h_-}$ were unbiased. Finally, Figure 4 considers CIs based on local linear estimators with triangular kernel; these CIs are very close to the optimal CIs in Figure 3.

4.2 Efficiency comparisons and bounds on adaptation

We now consider the relative efficiency of the different CIs reported in Figures 3 and 4. To keep the efficiency comparisons meaningful, we assume that the variance is homoskedastic on either side of the cutoff, and equal to the initial estimates.

First, comparing half-length and excess length of CIs based on choosing h_+, h_- to min-

imize the MSE to that of CIs based on optimally chosen h_+ and h_-, we find that over the range of C's considered, for both optimal and local linear estimators, two-sided CIs based on MSE-optimal estimators are at least 99.9% efficient, and one-sided CIs are at least 97.7% efficient. These results are in line with the asymptotic results in Armstrong and Kolesár (2016b), which imply that the asymptotic efficiency of two-sided fixed-length CIs is 99.9%, and it is 98.0% for one-sided CIs.

Second, comparing half-length and excess length of the CIs based on local linear estimates to that of CIs based on optimal estimators, we find that one- and two-sided CIs based on local linear estimators with triangular kernel are at least 96.9% efficient. This is very close to the asymptotic efficiency result in Armstrong and Kolesár (2016b) that the local linear estimator with a triangular kernel is 97.2% efficient, independently of the performance criterion.

Third, since the class $\mathcal{F}_{RDT,2}(C)$ is centrosymmetric, we can use Corollaries 3.2 and 3.3 to bound the scope for adaptation to the class of piecewise linear functions, $\mathcal{G} = \mathcal{F}_{RDT,2}(0)$. We find that the relative efficiency of CIs that minimax the 0.8 quantile is between 96% and 97.4%, and the efficiency of fixed-length two-sided CIs at any $g \in \mathcal{G}$ is between 95.5% and 95.9% for the range of C's considered. This is very close to the asymptotic efficiency predictions, 96.7% and 95.7%, respectively, implied by Figure 2 (with $r = 4/5$). Thus, one cannot avoid choosing C a priori.

Appendix A Proofs for main results

This section contains proofs of the results in Section 3. Appendix A.1 contains auxiliary lemmas used in the proofs. The proofs of the results in Section 3 are given in the remainder of the section. Proofs of Corollaries 3.1 and 3.3 follow immediately from the theorems and arguments in the main text, and their proofs are omitted. We assume throughout this section that the sets $\mathcal{F}$ and $\mathcal{G}$ are convex.

Before proceeding, we recall that $\omega'(\delta; \mathcal{F}, \mathcal{G})$ was defined in Section 3 to be an arbitrary element of the superdifferential. We denote this set by

$$\partial\omega(\delta; \mathcal{F}, \mathcal{G}) = \{d\colon \text{for all } \eta > 0, \omega(\eta; \mathcal{F}, \mathcal{G}) \leq \omega(\delta; \mathcal{F}, \mathcal{G}) + d(\eta - \delta)\}\,.$$

It is nonempty since $\omega(\cdot; \mathcal{F}, \mathcal{G})$ is concave—if f_δ^*, g_δ^* attain the modulus at δ and similarly for $\tilde{\delta}$, then, for $\lambda \in [0,1]$, $f_\lambda = \lambda f_\delta^* + (1-\lambda) f_{\tilde{\delta}}^*$ and $g_\lambda = \lambda g_\delta^* + (1-\lambda) g_{\tilde{\delta}}^*$ satisfy $\|K(g_\lambda - f_\lambda)\| \leq \lambda\delta + (1-\lambda)\tilde{\delta}$ so that $\omega(\lambda\delta + (1-\lambda)\tilde{\delta}) \geq Lg_\lambda - Lf_\lambda = \lambda\omega(\delta) + (1-\lambda)\omega(\tilde{\delta})$.

The definition of $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ in (23) depends on the choice of $\omega'(\delta; \mathcal{F}, \mathcal{G}) \in \partial\omega(\delta; \mathcal{F}, \mathcal{G})$ and $f_{\delta,\mathcal{F},\mathcal{G}}^*, g_{\delta,\mathcal{F},\mathcal{G}}^*$. As we explain in Supplemental Appendix D.2, Theorem 3.1 holds for any choice of $\omega'(\delta; \mathcal{F}, \mathcal{G})$ so long as the same element is used in the definition of the estimator and worst-case bias formula. Regarding the choice of the particular solution $f_{\delta,\mathcal{F},\mathcal{G}}^*, g_{\delta,\mathcal{F},\mathcal{G}}^*$ used to construct the estimator and CIs, it turns out that, under the conditions of Theorem 3.1, the choice does not affect the definition of $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ or the CIs based on it, as we now explain. If (f_0^*, g_0^*) and (f_1^*, g_1^*) solve the modulus problem with $K(g_0^* - f_0^*) \neq K(g_1^* - f_1^*)$, a strict convex combination (f_λ, g_λ) will satisfy $\|K(f_\lambda - g_\lambda)\| \leq \delta - \eta$ for some $\eta > 0$, which implies $\omega(\delta - \eta; \mathcal{F}, \mathcal{G}) = L(g_\lambda - f_\lambda) = \omega(\delta; \mathcal{F}, \mathcal{G})$. Since the modulus is nondecreasing, this implies that it is constant in a neighborhood of δ, so that $\partial\omega(\delta; \mathcal{F}, \mathcal{G}) = \{0\}$. Thus, either $K(g_\delta^* - f_\delta^*)$ is defined uniquely or $\partial\omega(\delta; \mathcal{F}, \mathcal{G}) = \{0\}$. In either case, $\omega'(\delta; \mathcal{F}, \mathcal{G}) \cdot K(f_\delta^* - g_\delta^*)$ is defined uniquely up to the choice of $\omega'(\delta; \mathcal{F}, \mathcal{G})$, which means that, for any two estimators $\hat{L}_0$ and $\hat{L}_1$ that satisfy the definition of $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ with the same choice of $\omega'(\delta; \mathcal{F}, \mathcal{G})$, we must have $\hat{L}_1 = \hat{L}_0 + a$ for some constant a. The bias formula (24), which follows from Lemma A.1 below, then implies that $a = 0$. Similarly, the CIs $[\hat{c}_{\alpha,\mathcal{F},\mathcal{G}}, \infty)$ and $\hat{L}_{\delta,\mathcal{F},\mathcal{G}} \pm \chi_\alpha(\hat{L}_{\delta,\mathcal{F},\mathcal{G}})$ are defined uniquely up to the choice of $\omega'(\delta; \mathcal{F}, \mathcal{G})$.

A.1 Auxiliary lemmas

The following lemma extends Lemma 4 in Donoho (1994) to the two class modulus (see also Theorem 2 in Cai and Low, 2004b, for a similar result in the Gaussian white noise model). The proof is essentially the same as for the single class case.

Lemma A.1. *Let $\mathcal{F}$ and $\mathcal{G}$ be convex sets and let f^* and g^* solve the optimization problem for $\omega(\delta_0; \mathcal{F}, \mathcal{G})$ with $\|K(f^* - g^*)\| = \delta_0$, and let $d \in \partial\omega(\delta_0; \mathcal{F}, \mathcal{G})$. Then, for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$Lg - Lg^* \leq d\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} \text{ and } Lf - Lf^* \geq d\frac{\langle K(g^* - f^*), K(f - f^*)\rangle}{\|K(g^* - f^*)\|}. \quad (31)$$

In particular, $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ achieves maximum bias over $\mathcal{F}$ at f^* and minimum bias over $\mathcal{G}$ at g^*.

Proof. Denote the ordered modulus $\omega(\delta; \mathcal{F}, \mathcal{G})$ by $\omega(\delta)$. Suppose that the first inequality in (31) does not hold for some g. Then, for some $\varepsilon > 0$,

$$Lg - Lg^* > (d + \varepsilon)\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|}. \quad (32)$$

Let $g_\lambda = (1 - \lambda)g^* + \lambda g$. Since $g_\lambda - g^* = \lambda(g - g^*)$, we have $\lambda L(g - g^*) = Lg_\lambda - Lf^* - L(g^* - f^*) = Lg_\lambda - Lf^* - \omega(\delta_0)$. Furthermore, since $g_\lambda \in \mathcal{G}$ by convexity, $Lg_\lambda - Lf^* \leq \omega(\|K(g_\lambda - f^*)\|)$ so multiplying (32) by λ gives

$$\omega(\|K(g_\lambda - f^*)\|) - \omega(\delta_0) \geq \lambda L(g - g^*) > \lambda(d + \varepsilon)\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|}. \quad (33)$$

Note that

$$\frac{d}{d\lambda}\|K(g_\lambda - f^*)\|\Big|_{\lambda=0} = \frac{1}{2}\frac{\frac{d}{d\lambda}\|K(g_\lambda - f^*)\|^2\Big|_{\lambda=0}}{\|K(g^* - f^*)\|} = \frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} \quad (34)$$

so that $\|K(g_\lambda - f^*)\| = \delta_0 + \lambda\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} + o(\lambda)$. Combining this with (33), we have

$$\omega\left(\delta_0 + \lambda\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} + o(\lambda)\right) - \omega(\delta_0) > \lambda(d + \varepsilon)\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|},$$

which is a contradiction unless $\langle K(g^* - f^*), K(g - g^*)\rangle = 0$.

If $\langle K(g^* - f^*), K(g - g^*)\rangle = 0$, then (32) gives $Lg - Lg^* > 0$, which, by the first

inequality in (33) implies $\omega(\|K(g_\lambda - f^*)\|) - \omega(\delta_0) \geq \lambda c$ where $c = Lg - Lg^* > 0$. But in this case (34) implies $\|K(g_\lambda - f^*)\| = \delta_0 + o(\lambda)$, again giving a contradiction. This proves the first inequality, and a symmetric argument applies to the inequality involving $Lf - Lf^*$, thereby giving the first result.

Now consider the test statistic $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$. Under $g \in \mathcal{G}$, the bias of this statistic is equal to a constant that does not depend on g plus

$$d\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} - (Lg - Lg^*).$$

It follows from (31) that this is minimized over $g \in \mathcal{G}$ by taking $g = g^*$. Similarly, the maximum bias over $\mathcal{F}$ is taken at f^*. □

The next lemma is used in the proof of Theorem 3.2.

Lemma A.2. *Let $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$ be convex sets, and suppose that f^* and g^* minimize $\|K(f-g)\|$ over $f \in \tilde{\mathcal{F}}$ and $g \in \tilde{\mathcal{G}}$. Then, for any level α, the minimax test of $H_0 : \tilde{\mathcal{F}}$ vs $H_1 : \tilde{\mathcal{G}}$ is given by the Neyman-Pearson test of f^* vs g^*. It rejects when $\langle K(f^* - g^*), Y\rangle$ is greater than its $1-\alpha$ quantile under f^*. The minimum power of this test over $\tilde{\mathcal{G}}$ is taken at g^*.*

Proof. The result is immediate from results stated in Section 2.4.3 in Ingster and Suslina (2003), since the sets $\{Kf\colon f \in \tilde{\mathcal{F}}\}$ and $\{Kg\colon g \in \tilde{\mathcal{G}}\}$ are convex. □

A.2 Proof of Theorem 3.1

For ease of notation in this proof, let $f^* = f^*_\delta$ and $g^* = g^*_\delta$ denote the functions that solve the modulus problem with $\|K(f^* - g^*)\| = \delta$, and let $d = \omega'(\delta; \mathcal{F}, \mathcal{G}) \in \partial\omega(\delta; \mathcal{F}, \mathcal{G})$ so that, plugging the worst-case bias formula (24) into the definition of $\hat{c}_\alpha$, we have

$$\hat{c}_\alpha = \hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = Lf^* + d\frac{\langle K(g^* - f^*), Y\rangle}{\|K(g^* - f^*)\|} - d\frac{\langle K(g^* - f^*), Kf^*\rangle}{\|K(g^* - f^*)\|} - z_{1-\alpha}\sigma d.$$

Note that $\hat{c}_\alpha = \hat{L}_{\delta,\mathcal{F},\mathcal{G}} + a$ for a chosen so that the $1-\alpha$ quantile of $\hat{c}_\alpha - Lf^*$ under f^* is zero. Thus, it follows from Lemma A.1 that $[\hat{c}_\alpha, \infty)$ is a valid $1-\alpha$ CI for Lf over $\mathcal{F}$, and that all quantiles of excess coverage $Lg - \hat{c}_\alpha$ are maximized over $\mathcal{G}$ at g^*. In particular, $q_\beta(\hat{c}_\alpha; \mathcal{G}) = q_{g^*,\beta}(Lg^* - \hat{c}_\alpha)$. To calculate this quantile, note that, under g^*, $Lg^* - \hat{c}_\alpha$ is normal with variance $d^2\sigma^2$ and mean

$$Lg^* - Lf^* - d\frac{\langle K(g^* - f^*), K(g^* - f^*)\rangle}{\|K(g^* - f^*)\|} + z_{1-\alpha}\sigma d = \omega(\delta; \mathcal{F}, \mathcal{G}) + d(z_{1-\alpha}\sigma - \delta).$$

The probability that this normal variable is less than or equal to $\omega(\delta; \mathcal{F}, \mathcal{G})$ is given by the probability that a normal variable with mean $d(z_{1-\alpha}\sigma - \delta)$ and variance $d^2\sigma^2$ is less than or equal to zero, which is $\Phi(\delta/\sigma - z_{1-\alpha}) = \beta$. Thus, $q_\beta(\hat{c}_\alpha; \mathcal{G}) = \omega(\delta; \mathcal{F}, \mathcal{G})$ as claimed.

It remains to show that no other $1-\alpha$ CI can strictly improve on this. Suppose that some other $1-\alpha$ CI $[\tilde{c}, \infty)$ obtained $q_\beta(\tilde{c}; \mathcal{G}) < q_\beta(\hat{c}_\alpha; \mathcal{G}) = \omega(\delta; \mathcal{F}, \mathcal{G})$. Then the β quantile of excess length at g^* would be strictly less than $\omega(\delta; \mathcal{F}, \mathcal{G})$, so that, for some $\eta > 0$,

$$P_{g^*}(Lg^* - \tilde{c} \leq \omega(\delta; \mathcal{F}, \mathcal{G}) - \eta) \geq \beta.$$

Let $\tilde{f}$ be given by a convex combination between g^* and f^* such that $Lg^* - L\tilde{f} = \omega(\delta; \mathcal{F}; \mathcal{G}) - \eta/2$. Then the above display gives

$$P_{g^*}(\tilde{c} > L\tilde{f}) \geq P_{g^*}(\tilde{c} \geq L\tilde{f} + \eta/2) = P_{g^*}(Lg^* - \tilde{c} \leq Lg^* - L\tilde{f} - \eta/2) \geq \beta.$$

But this would imply that the test that rejects when $\tilde{c} > L\tilde{f}$ is level α for $H_0 : \tilde{f}$ and has power β at g^*. This can be seen to be impossible by calculating the power of the Neyman-Pearson test of $\tilde{f}$ vs g^*, since β is the power of the Neyman-Pearson test of f^* vs g^*, and $\tilde{f}$ is a strict convex combination of these functions.

A.3 Proof of Corollary 3.2

Under (27), if $f^*_{\delta,\mathcal{F},\mathcal{G}}$ and $g^*_{\delta,\mathcal{F},\mathcal{G}}$ solve the modulus problem $\omega(\delta, \mathcal{F}, \mathcal{G})$, then $f^*_{\delta,\mathcal{F},\mathcal{G}} - g^*_{\delta,\mathcal{F},\mathcal{G}}$ and 0 (the zero function) solve $\omega(\delta; \mathcal{F}, \{0\})$. Thus, $\omega(\delta; \mathcal{F}, \mathcal{G}) = \omega(\delta; \mathcal{F}, \{0\})$, and the estimators $\hat{L}_{\delta,\mathcal{F},\mathcal{G}}$ and $\hat{L}_{\delta,\mathcal{F},\{0\}}$ and the corresponding CIs are equal up to the choice of the element in the superdifferential. It therefore suffices to prove the result for $\mathcal{G} = \{0\}$.

We have

$$\omega(\delta; \mathcal{F}, \{0\}) = \sup\{-Lf \colon \|Kf\| \leq \delta, f \in \mathcal{F}\} = \frac{1}{2}\omega(2\delta; \mathcal{F}),$$

where the last equality obtains because under centrosymmetry, maximizing $-Lf = L(-f)$ and maximizing Lf are equivalent, so that the maximization problem is equivalent to (25). Furthermore, we can take $g^*_{2\delta,\mathcal{F}}, f^*_{2\delta,\mathcal{F}}$ to satisfy $g^*_{2\delta,\mathcal{F}} = -f^*_{2\delta,\mathcal{F}}$ with $f^*_{2\delta,\mathcal{F}}$ solving the above optimization problem, so that $g^*_{\delta,\mathcal{F},\{0\}} - f^*_{\delta,\mathcal{F},\{0\}} = -f^*_{\delta,\mathcal{F},\{0\}} = -f^*_{2\delta,\mathcal{F}} = \frac{1}{2}(g^*_{2\delta,\mathcal{F}} - f^*_{2\delta,\mathcal{F}})$. Thus, $\hat{L}_{\delta,\mathcal{F},\{0\}}$ and $\hat{L}_{2\delta,\mathcal{F}}$ are equal up to a constant, which implies $\hat{c}_{\alpha,\delta,\mathcal{F},\{0\}} = \hat{c}_{\alpha,2\delta,\mathcal{F}}$. This proves the first part of the corollary. The second part of the corollary follows by plugging $\underline{\text{bias}}_{\{0\}}(\hat{L}_{\delta_\beta,\mathcal{F}}) = 0$ and the formulas for $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{\delta_\beta,\mathcal{F}})$ and $\text{sd}(\hat{L}_{\delta_\beta,\mathcal{F}})$ given in Section 3.2 into the expression (21) to obtain $q_\beta(\hat{c}_{\alpha,\delta_\beta,\mathcal{F}}, \{0\}) = (\omega(\delta_\beta; \mathcal{F}) + \delta_\beta\omega'(\delta_\beta; \mathcal{F}))/2$.

A.4 Proof of Theorem 3.2

Following Pratt (1961), note that, for any confidence set $\mathcal{C}$ for $\vartheta = Lf$, we have

$$E_g\lambda(\mathcal{C}) = E_g \int (1 - \phi_{\mathcal{C}}(\vartheta))\, d\vartheta = \int E_g(1 - \phi_{\mathcal{C}}(\vartheta))\, d\vartheta$$

by Fubini's theorem, where $\phi_{\mathcal{C}}(\vartheta) = 1(\vartheta \notin \mathcal{C})$. Thus, the CI that minimizes this inverts the family of most powerful tests of $H_0 \colon Lf = \vartheta, f \in \mathcal{F}$ against $H_1 \colon f = g$. By Lemma A.2 since the sets $\{f \colon Lf = \vartheta, f \in \mathcal{F}\}$ and $\{g\}$ are convex, the least favorable function f_ϑ minimize $\|K(g - f)\|$ subject to $Lf = \vartheta$, which gives the first part of the theorem.

To derive the expression for expected length, note that if $Lg \leq \vartheta$, then the minimization problem is equivalent to solving the inverse ordered modulus problem $\omega^{-1}(\vartheta - Lg; \{g\}, \mathcal{F})$, and if $Lg \geq \vartheta$, it is equivalent to solving $\omega^{-1}(Lg - \vartheta; \mathcal{F}, \{g\})$. This follows because if the ordered modulus $\omega(\delta; \mathcal{F}, \{g\})$ is attained at some f_δ^* and g, then the inequality $\|K(f - g)\| \leq \delta$ must be binding: otherwise a convex combination of $\tilde{f}$ and f_δ^*, where $\tilde{f}$ is such that $L(g - f_\delta^*) < L(g - \tilde{f})$ would achieve a strictly larger value, and similarly for $\omega(\delta; \{g\}, \mathcal{F})$. Such $\tilde{f}$ always exists since by the assumption that f_ϑ exists for all ϑ. The above argument assumes that $\vartheta - Lg \geq \omega(0; \{g\}, \mathcal{F})$ so that $\vartheta - Lg$ is in the range of the modulus; if $0 \leq \vartheta - Lg \leq \omega(0; \{g\}, \mathcal{F})$, then $\|K(f_\vartheta - g)\| = 0$ so the minimization problem is still equivalent to the inverse modulus if we define the inverse to be 0 in this case (and similarly for $0 \leq Lg - \vartheta \leq \omega(0; \mathcal{F}, \{g\})$).

Next, it follows from the proof of Theorem 3.1 that the power of the test ϕ_ϑ at g is given by $\Phi(\delta_\vartheta/\sigma - z_{1-\alpha})$, where $\delta_\vartheta = \|f_\vartheta - g\|$. Therefore,

$$E_g[\lambda(\mathcal{C}_g(Y))] = \int \Phi\left(z_{1-\alpha} - \frac{\delta_\vartheta}{\sigma}\right) \mathrm{d}\vartheta = \iint 1(\delta_\vartheta \leq \sigma(z_{1-\alpha} - z))\, \mathrm{d}\vartheta\, \mathrm{d}\Phi(z),$$

where the second equality swaps the order of integration. Splitting the inner integral, using fact that $\delta_\vartheta = \omega^{-1}(Lg - \vartheta; \mathcal{F}, \{g\})$ for $\vartheta \leq Lg$ and $\delta_\vartheta = \omega^{-1}(\vartheta - Lg; \{g\}, \mathcal{F})$ for $\vartheta \geq Lg$, and taking a modulus on both sides of the inequality of the integrand then yields

$$\begin{aligned} E_g[\lambda(\mathcal{C}_g(Y))] &= \iint_{\vartheta \leq Lg} 1(Lg - \vartheta \leq \omega\left(\sigma(z_{1-\alpha} - z); \mathcal{F}, \{g\}\right)) 1(z \leq z_{1-\alpha})\, \mathrm{d}\vartheta\, \mathrm{d}\Phi(z) \\ &\quad + \iint_{\vartheta > Lg} 1(\vartheta - Lg \leq \omega\left(\sigma(z_{1-\alpha} - z); \{g\}, \mathcal{F}\right)) 1(z \leq z_{1-\alpha})\, \mathrm{d}\vartheta\, \mathrm{d}\Phi(z) \\ &= (1 - \alpha) E\left[\left(\omega(\sigma(z_{1-\alpha} - Z); \mathcal{F}, \{g\}) + \omega(\sigma(z_{1-\alpha} - Z); \{g\}, \mathcal{F})\right) \mid Z \leq z_{1-\alpha}\right], \end{aligned}$$

where Z is standard normal, which yields the result.

Appendix B Extension to RD with covariates

This section discusses extensions to the RD setup when we have available a set of covariates z_i that are independent of the treatment. If the object of interest is still the average treatment effect at $x = 0$, then ignoring the additional covariates will still lead to a valid CI. However, one may want to use the information that z_i is independent of treatment to gain precision. We discuss this in Appendix B.1. Alternatively, one may want to estimate the treatment effect at $x = 0$ conditional on different values of z, which leads to a different approach, discussed in Appendix B.2.

B.1 Using covariates to improve precision

As argued by Calonico et al. (2017), if z_i is independent of treatment, the conditional mean of z_i given the running variable x_i should be smooth near the cutoff. We can fit this into our setup using the model

$$\begin{aligned} y_i &= h_y(x_i) + u_i, \\ z_i &= h_z(x_i) + v_i, \end{aligned} \qquad \begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim \mathcal{N}\left(0, \Sigma(x_i)\right),\ h_y \in \mathcal{H}_y,\ h_z \in \mathcal{H}_z,$$

where $\mathcal{H}_y$ and $\mathcal{H}_z$ are convex smoothness classes, and we treat $\Sigma(\cdot)$ as known. We incorporate the constraint that z_i is independent of treatment by choosing a class $\mathcal{H}_z$ such that $\lim_{x\downarrow 0} h_z(x) - \lim_{x\uparrow 0} h_z(x) = 0$ for all $h_z \in \mathcal{H}_z$. For example, we can take $\mathcal{H}_y = \mathcal{F}_{RDT,p}(C_y)$ and $\mathcal{H}_z = \mathcal{F}_{RDT,p}(C_z) \cap \{h\colon \lim_{x\downarrow 0} h_z(x) - \lim_{x\uparrow 0} h_z(x) = 0\}$ for some constants C_y and C_z.

Using our general results, one can compute optimal CIs and bounds for adaptation. For example, our adaptation bounds show that, when $\mathcal{H}_y$ and $\mathcal{H}_z$ are centrosymmetric, there are severe limitations to adapting to the smoothness constant for either class. Thus, CIs that take into account the covariates z_i will have to depend explicitly on the smoothness constant that h_z is assumed to satisfy.

In the remainder of this section, we consider a particular smoothness class, and we construct CIs that are optimal or near-optimal when $\Sigma(x)$ is constant as well as feasible versions of these CIs that are valid when $\Sigma(x)$ is unknown and may not be constant. Given Σ, let Σ_{22} denote the bottom-right $d_z \times d_z$ submatrix of Σ and let Σ_{21} denote the bottom-left

$d_z \times d_1$ submatrix of Σ, where d_z is the dimension of z_i. Let $\tilde{y}_i = y_i - z_i'\Sigma_{22}^{-1}\Sigma_{21}$ so that

$$\tilde{y}_i = h_y(x_i) - h_z(y_i)'\Sigma_{22}^{-1}\Sigma_{21} + u_i - v_i'\Sigma_{22}^{-1}\Sigma_{21} = \tilde{h}_y(x_i) + \tilde{u}_i$$

where $\tilde{h}_y(x_i) = h_y(x_i) - h_z(y_i)'\Sigma_{22}^{-1}\Sigma_{21}$ and $\tilde{u}_i = u_i - v_i'\Sigma_{22}^{-1}\Sigma_{21}$. Note also that $\lim_{x\downarrow 0}\tilde{h}_y(x) - \lim_{x\uparrow 0}\tilde{h}_y(x) = \lim_{x\downarrow 0} h_y(x) - \lim_{x\uparrow 0} h_y(x)$, so that the RD parameter for $\tilde{h}_y$ is the same as the RD parameter for h_y. Suppose that we model the smoothness of $\tilde{h}_y$ directly, and take the parameter space for $(\tilde{h}_y, h_z)$ to be $\mathcal{F}_{RDT,p}(\tilde{C}) \times \mathcal{H}_z$. Since $\tilde{u}_i$ is independent of v_i and the RD parameter depends only on $\tilde{h}_y$, it can be seen that minimax optimal estimators and CIs can be formed by ignoring the z_i's after this transformation is made. Thus, one can proceed as in Section 2.2 with $\tilde{y}_i$ in place of y_i.[8]

To make this procedure feasible, we need an estimate of $\Sigma_{22}^{-1}\Sigma_{21}$. We propose the estimates $\hat{\Sigma}_{22} = \frac{1}{nh}\sum_{i=1}^n \hat{v}_i\hat{v}_i' k(x_i/h)$ and $\hat{\Sigma}_{21} = \frac{1}{nh}\sum_{i=1}^n \hat{v}_i y_i k(x_i/h)$ where $\hat{v}_i$ is the residual from the local polynomial regression of z_i on a pth order polynomial of x_i and its interaction with $1(x_i > 0)$, with weight $k(x_i/h)$. To form CIs, one proceeds as in Section 2.2 with $\tilde{y}_i = y_i - z_i'\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$ in place of y_i and $\tilde{C}$ playing the role of C. A simple calculation shows that, if one uses the local polynomial weights (14), with the same kernel and bandwidth used to estimate Σ, the resulting CIs will be centered at a local polynomial estimate where z_i is included as a regressor in the local polynomial regression. This corresponds exactly to an estimator proposed by Calonico et al. (2017). Thus, our relative efficiency results can be used to show that this estimator is close to optimal under these assumptions.

B.2 Estimating the treatment effect conditional on $z_i = z$

If one is interested in how the treatment effect at $x = 0$ varies with z, one can use the model $y_i = f(x_i, z_i) + u_i$ where f is placed in a smoothness class and the object of interest is $L_z f = \lim_{x\downarrow 0} f(x, z) - \lim_{x\uparrow 0} f(x, z)$ for different values of z. This fits into our general framework once one fixes the point z at which $L_z f$ is evaluated, and one can use our results to obtain CIs for different values of z. A natural smoothness class is to place a bound on the pth order multivariate Taylor approximation of $f(x, z)1(x > 0)$ and $f(x, z)1(x < 0)$ at $x = 0$ and z equal to the value of interest. The analysis of optimal and near optimal estimators

[8] If one places smoothness assumptions on h_y rather than $\tilde{h}_y$ by taking $\mathcal{H}_y = \mathcal{F}_{RDT,p}(C_y)$ and $\mathcal{H}_z = \mathcal{F}_{RDT,p}(C_z) \cap \{h\colon \lim_{x\downarrow 0} h_z(x) - \lim_{x\uparrow 0} h_z(x) = 0\}$, then $\tilde{h}_y \in \mathcal{F}_{RDT,p}(C_y + C_z\iota'\Sigma_{22}^{-1}\Sigma_{21})$ where ι is a vector of ones. It follows that the CIs discussed here will be valid for $\tilde{C} \geq C_y + C_z\iota'\Sigma_{22}^{-1}\Sigma_{21}$. However, the resulting parameter space for $(\tilde{h}_y, h_z)$ will be different (in particular, it will not take the form $\mathcal{H}_y \times \mathcal{H}_z$), so that optimal estimators will be different for this class.

then follows from a generalization of the results described in Section 2.2. In particular, one can use multivariate local polynomial estimators (with worst-case bias computed using a generalization of the calculations in Supplemental Appendix E.1), or optimal weights can be computed by generalizing the calculations in Supplemental Appendix E.2.

Estimating the treatment effect conditional on different values of z can be a useful way of exploring treatment effect heterogeneity. However, unless one places some additional parametric structure on $f(x, z)$, the resulting estimates will suffer from imprecision when the dimension of z is moderate due to the curse of dimensionality.

References

Abadie, A. and G. W. Imbens (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74, 235–267.

Andrews, D. W. K. and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77, 721–762.

Armstrong, T. B. (2015): "Adaptive testing on a regression function at a point," *The Annals of Statistics*, 43, 2086–2101.

Armstrong, T. B. and M. Kolesár (2016a): "Optimal inference in a class of regression models," ArXiv:1511.06028v2, `https://arxiv.org/abs/1511.06028v2`.

——— (2016b): "Simple and honest confidence intervals in nonparametric regression," ArXiv:1606.01200, `https://arxiv.org/abs/1606.01200`.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650.

Cai, T. T. and M. G. Low (2004a): "An adaptation theory for nonparametric confidence intervals," *Annals of Statistics*, 32, 1805–1840.

——— (2004b): "Minimax estimation of linear functionals over nonconvex parameter spaces," *Annals of Statistics*, 32, 552–576.

Cai, T. T., M. G. Low, and Z. Ma (2014): "Adaptive Confidence Bands for Nonparametric Regression Functions," *Journal of the American Statistical Association*, 109, 1054–1070.

Cai, T. T., M. G. Low, and Y. Xia (2013): "Adaptive confidence intervals for regression functions under shape constraints," *The Annals of Statistics*, 41, 722–750.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2017): "Regression Discontinuity Designs Using Covariates," Unpublished Manuscript, University of Michigan.

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

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On automatic boundary corrections," *The Annals of Statistics*, 25, 1691–1708.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014): "Anti-concentration and honest, adaptive confidence bands," *The Annals of Statistics*, 42, 1787–1818.

Cohen, J. (1988): *Statistical Power Analysis for the Behavioral Sciences*, Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Donoho, D. L. and M. G. Low (1992): "Renormalization Exponents and Optimal Pointwise Rates of Convergence," *The Annals of Statistics*, 20, 944–970.

Fan, J. (1993): "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Giné, E. and R. Nickl (2010): "Confidence bands in density estimation," *The Annals of Statistics*, 38, 1122–1170.

Hall, P. and J. Horowitz (2013): "A simple bootstrap method for constructing nonparametric confidence bands for functions," *The Annals of Statistics*, 41, 1892–1921.

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

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

Ingster, Y. I. and I. A. Suslina (2003): *Nonparametric goodness-of-fit testing under Gaussian models*, New York: Springer.

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

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

Lehmann, E. L. and J. P. Romano (2005): *Testing statistical hypotheses*, New York: Springer, third ed.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25, 2547–2554.

McCloskey, A. (2017): "Bonferroni-Based Size-Correction for Nonstandard Testing Problems," *Journal of Econometrics*, 200, 17–35.

Pratt, J. W. (1961): "Length of confidence intervals," *Journal of the American Statistical Association*, 56, 549–567.

Sacks, J. and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *The Annals of Statistics*, 6, 1122–1137.

Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360.

	α		
b	0.01	0.05	0.1
0.0	2.576	1.960	1.645
0.1	2.589	1.970	1.653
0.2	2.626	1.999	1.677
0.3	2.683	2.045	1.717
0.4	2.757	2.107	1.772
0.5	2.842	2.181	1.839
0.6	2.934	2.265	1.916
0.7	3.030	2.356	2.001
0.8	3.128	2.450	2.093
0.9	3.227	2.548	2.187
1.0	3.327	2.646	2.284
1.5	3.826	3.145	2.782
2.0	4.326	3.645	3.282

Table 1: Critical values $\mathrm{cv}_\alpha(b)$ for selected confidence levels and values of maximum absolute bias b. For $b \geq 2$, $\mathrm{cv}_\alpha(b) \approx b + z_{1-\alpha}$ up to 3 decimal places for these values of α.

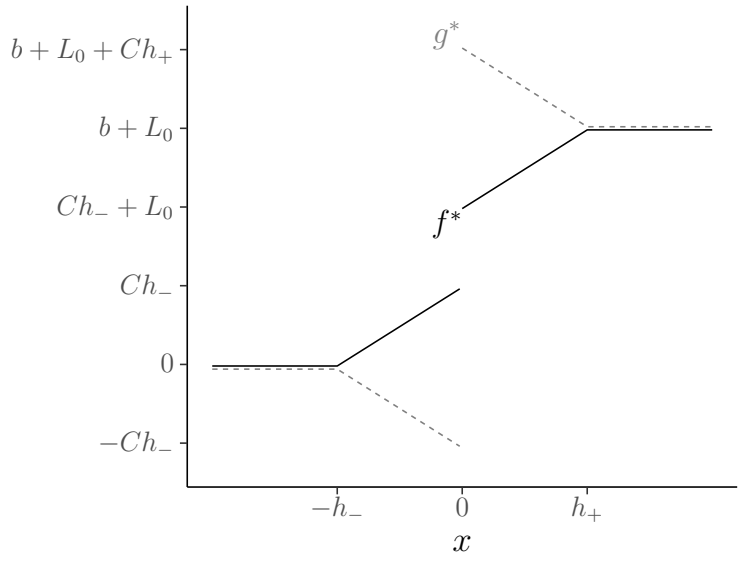

Figure 1: The least favorable null and alternative functions f^* and g^* from Equation (3) in Section 2.1.

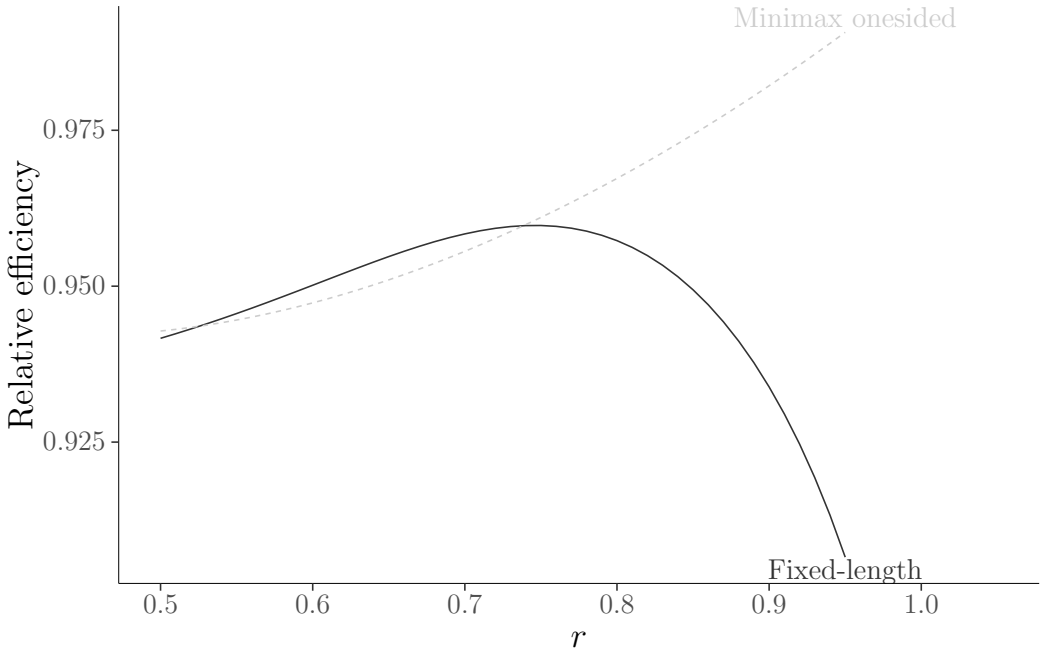

Figure 2: Asymptotic efficiency bounds for one-sided and fixed-length CIs as function of the optimal rate of convergence r under centrosymmetry. Minimax one-sided refers to ratio of β-quantile of excess length of CIs that direct power at smooth functions relative to minimax one-sided CIs given in (28). Shortest fixed-length refers the ratio of expected length of CIs that direct power at a given smooth function relative to shortest fixed-length affine CIs given in Theorem E.1.

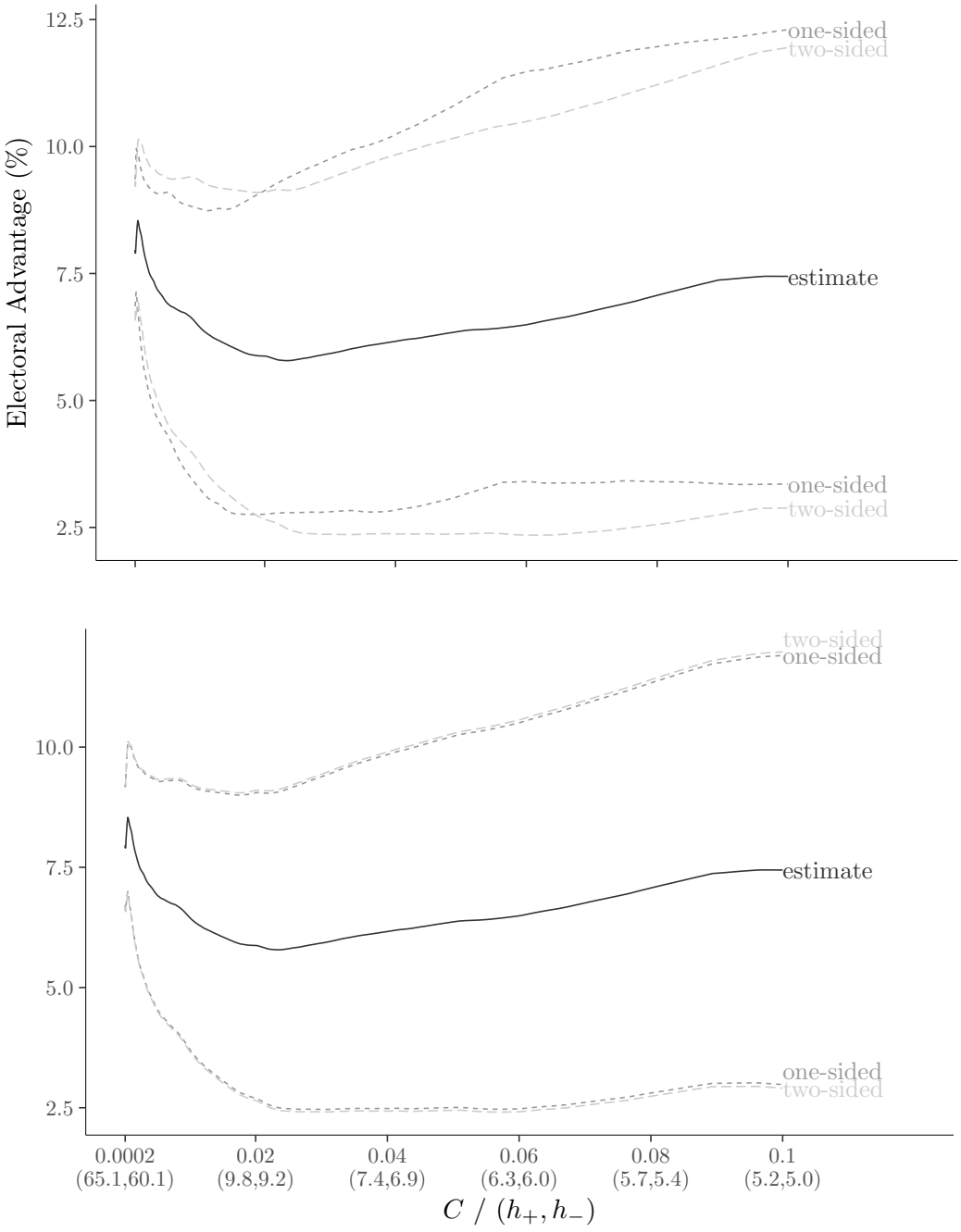

Figure 3: Lee (2008) RD example. Top panel displays minimax MSE estimator (estimator), and lower and upper limits of minimax one-sided confidence intervals for 0.8 quantile (one-sided), and fixed-length CIs (two-sided) as function of smoothness C. Bottom panel displays one-and two-sided CIs around the minimax MSE estimator. h_+, h_- correspond to the optimal smoothness parameters for the minimax MSE estimator.

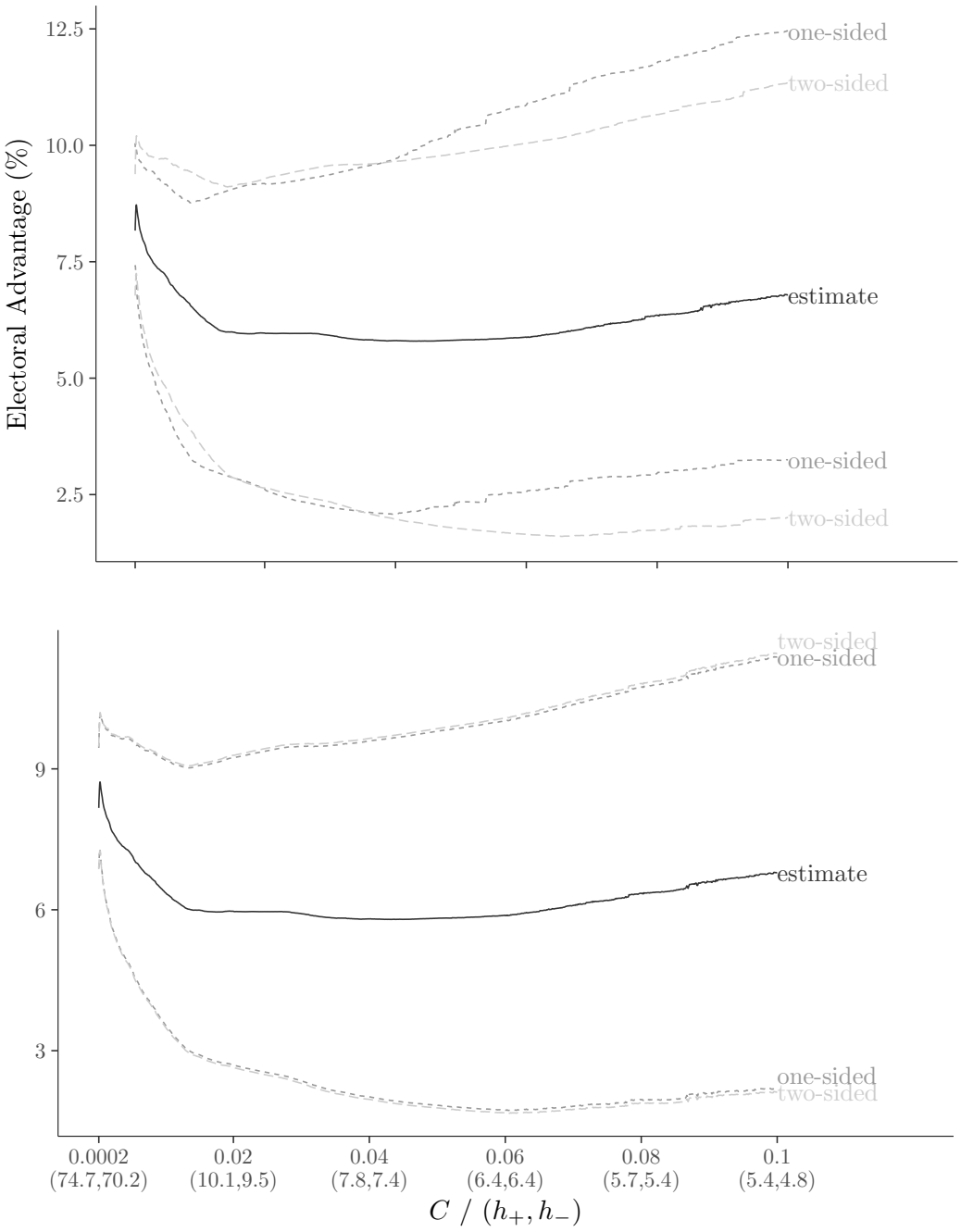

Figure 4: Lee (2008) RD example: local linear regression with triangular kernel. Top panel displays estimator based on minimax MSE bandwidths (estimator), lower and upper limits of one-sided CIs with bandwidths that are minimax for 0.8 quantile of excess length (one-sided), and shortest fixed-length CIs (two-sided) as function of smoothness C. Bottom panel displays one-and two-sided CIs around and estimator based on minimax MSE bandwidths. h_+, h_- correspond to the minimax MSE bandwidths.