Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers

Cowles Foundation

12-1-2017

Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness

Timothy B. Armstrong

Michal Kolesár

Follow this and additional works at: https://elischolar.library.yale.edu/cowles-discussion-paper-series

 Part of the Economics Commons

Recommended Citation

Armstrong, Timothy B. and Kolesár, Michal, "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness" (2017). *Cowles Foundation Discussion Papers*. 163.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/163

This Discussion Paper is brought to you for free and open access by the Cowles Foundation at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Cowles Foundation Discussion Papers by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

FINITE-SAMPLE OPTIMAL ESTIMATION AND INFERENCE ON AVERAGE TREATMENT EFFECTS UNDER UNCONFOUNDEDNESS

By

Timothy B. Armstrong and Michal Kolesár

December 2017

COWLES FOUNDATION DISCUSSION PAPER NO. 2115



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness*

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

December 12, 2017

Abstract

We consider estimation and inference on average treatment effects under unconfoundedness conditional on the realizations of the treatment variable and covariates. We derive finite-sample optimal estimators and confidence intervals (CIs) under the assumption of normal errors when the conditional mean of the outcome variable is constrained only by nonparametric smoothness and/or shape restrictions. When the conditional mean is restricted to be Lipschitz with a large enough bound on the Lipschitz constant, we show that the optimal estimator reduces to a matching estimator with the number of matches set to one. In contrast to conventional CIs, our CIs use a larger critical value that explicitly takes into account the potential bias of the estimator. It is needed for correct coverage in finite samples and, in certain cases, asymptotically. We give conditions under which root-n inference is impossible, and we provide versions of our CIs that are feasible and asymptotically valid with unknown error distribution, including in this non-regular case. We apply our results in a numerical illustration and in an application to the National Supported Work Demonstration.

*We thank Xiaohong Chen, Jin Hahn, Guido Imbens, Pedro Sant'Anna, Andres Santos, Azeem Shaikh, and Alex Torgovitsky for illuminating discussions. We also thank numerous seminar and conference participants for helpful comments and suggestions. All 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

To estimate the average treatment effect (ATE) of a binary treatment in observational studies, it is typically assumed that the treatment is unconfounded given a set of pretreatment covariates. This assumption implies that systematic differences in outcomes between treated and control units with the same values of the covariates are attributable to the treatment. When the covariates are continuously distributed, it is not possible to perfectly match the treated and control units based on their covariate values, and estimation of the ATE requires nonparametric regularization methods such as kernel, series or sieve estimators, or matching estimators that allow for imperfect matches.

To compare estimators, one can use the theory of semiparametric efficiency bounds. Given enough smoothness, and given overlap in the covariate distributions in the treated and control subpopulations, many regularization methods lead to estimators that are $\sqrt{n}$-consistent, asymptotically unbiased and normally distributed, with variance that achieves the semiparametric efficiency bound (see, among others, Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003; Chen et al., 2008). One can then construct confidence intervals (CIs) based on any such estimator by adding and subtracting its standard deviation times a quantile of a standard normal distribution. A common critique[1] of this approach is that it does not provide a good description of finite-sample behavior of estimators and CIs: in finite samples, regularization leads to bias, and different estimators have different finite-sample biases even if they are asymptotically equivalent. The bias may in turn lead to undercoverage of the resulting CIs due to incorrect centering. Furthermore, to achieve the semiparametric efficiency bound, regularization requires a large amount of smoothness of either the propensity score or the conditional mean of the outcome given the treatment and covariates: one typically assumes continuous differentiability of the order $p/2$ at minimum (e.g. Chen et al., 2008), and often of the order $p + 1$ or higher (e.g. Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003), where p is the dimension of the covariates. Unless p is very small, such assumptions are hard to evaluate, and may be much stronger than the researcher is willing to impose.

In this paper, we instead treat smoothness and/or shape restrictions on the conditional mean of the outcome—the regression of the outcome on the treatment and covariates—as given and determined by the researcher. To explicitly account for finite-sample biases, we consider finite-sample performance of estimators and CIs under the assumption that the

[1] See, for example, Robins and Ritov (1997).

regression errors are normal with known variance, with the treatment and covariates viewed as fixed.

We derive three main results. First, we show that if the conditional mean is assumed to satisfy a Lipschitz constraint, the minimax optimal estimator is given by a matching estimator with the number of matches set to one, so long as the Lipschitz constant is large enough. Thus, the matching estimator with a single match is finite-sample optimal when only very weak smoothness assumptions are made. More generally, we show that the optimal estimator is given by a solution to a convex programming problem. We show how the solution can be found numerically under Lipschitz smoothness.

Second, we derive minimal conditions under which the semiparametric efficiency bound can be achieved in our setting. In particular, we show that for $\sqrt{n}$-inference to be possible, one needs to bound the derivative of the conditional mean of order at least $p/2$. This is essentially the same smoothness condition as when one does not condition on treatment and covariates (Robins et al., 2009), and when no smoothness is imposed on the propensity score. Intuitively, by conditioning on the treatment and covariates, we take away any role that the propensity score may play in increasing precision of inference.

Third, we derive the form of optimal CIs. We show the optimal CI is centered around a linear estimator that is based on the same class of estimators that lead to the optimal estimator. Importantly, however, in order to account for the possible bias of the estimator, the CI uses a larger critical value than the conventional critical value based on normal quantiles. This critical value depends on the worst-case bias of the estimator, which for the optimally chosen estimator has a simple form. We show that feasible versions of the optimal CI are asymptotically valid and efficient when the distribution of errors is unknown and potentially non-normal, including in the non-regular case in which the semiparametric efficiency bound cannot be achieved. In the regular case, the large-sample bias of the estimator is negligible, and the critical value converges to the conventional critical value based on normal quantiles. However, in the non-regular case, the bias remains non-negligible even in large samples, and using this larger critical value is necessary to ensure asymptotic coverage.

We also show that by using this larger critical value, one can construct finite-sample valid CIs based on other linear estimators, such as series or kernel estimators, or matching estimators with more than a single match. This requires computing the worst-case bias of the estimator, which is a convex programming problem; we show how the solution can be found numerically under Lipschitz smoothness, in which case it reduces to a linear programming problem. One can compare this CI to the conventional CI that uses critical values based on

normal quantiles that does not take bias into account as a form of sensitivity analysis.

An important advantage of our finite sample approach is that it deals automatically with issues that normally arise with translating asymptotic results into practice. One need not worry about whether the model is point identified, "irregularly identified" (due to partial overlap as in Khan and Tamer 2010, or due to smoothness conditions being too weak to achieve root-n convergence, as in Robins et al. 2009) or set identified (due to complete lack of overlap). If the overlap in the data combined with the smoothness conditions imposed by the researcher lead to non-negligible bias, this will be incorporated into the CI. If the model is set identified due to lack of overlap, this bias term will prevent the CI from shrinking to a point, and the CI will converge to the identified set. Nor does one have to worry about whether covariates should be logically treated as having a continuous or discrete distribution. If it is optimal to do so, our estimator will regularize when covariates are discrete, and the CI will automatically incorporate the resulting finite sample bias. Thus, we avoid decisions about whether, for example, to allow for imperfect matches with a discrete covariate when an "asymptotic promise" says that, when the sample size is large enough, we will not.

We illustrate the results using a numerical example and an application to the National Supported Work (NSW) Demonstration. We find that finite-sample optimal CIs are substantially different from those based on first order asymptotic theory, with bias determining a substantial portion of the width of the CI. We also find that, under Lipschitz smoothness, matching estimators perform relatively well for a range of smoothness constants, in addition to being exactly optimal when the smoothness constant is large enough.

Our results rely on the key insight that, once one conditions on treatment assignments and pretreatment variables, the ATE is a linear functional of a regression function. This puts the problem in the framework of Donoho (1994) and Cai and Low (2004) and allows us to apply sharp efficiency bounds in Armstrong and Kolesár (2017). In contrast, if one does not condition on treatment assignments and pretreatment variables, the ATE is a nonlinear functional of two regression functions (the propensity score, and the conditional mean of the outcome variable given pretreatment variables). This makes the problem much more difficult: while upper and lower bounds have been developed that give the optimal rate (Robins et al., 2009), computing efficiency bounds that are sharp in finite samples (or even bounds on the asymptotic constant in non-regular cases) remains elusive.

Whether one should condition on treatment assignments and pretreatment covariates when evaluating estimators and CIs is itself an interesting question (see Abadie et al., 2014a,b, for a recent discussion in related settings). An argument in favor of conditioning is

that it takes into account the realized imbalance, or overlap, of covariates across treatment groups. For example, even if the treatment is assigned randomly and independently of an individual's level of education, it may happen that the realized treatments are such that the treated individuals are highly educated relative to those randomized out of treatment. Conditioning takes into account this ex-post imbalance when evaluating estimators and CIs. On the other hand, by conditioning on realized treatment assignments, one loses the ability to use knowledge of the propensity score or its smoothness to gain efficiency. We do not intend to make a blanket argument for or against the practice of conditioning on realized treatment. Rather, our view is that this choice depends on the particular empirical context, that it is worth developing efficiency bounds that are as sharp as possible in both settings, and that comparing the bounds is instructive. Since our CIs are valid unconditionally, they can be used in either setting, so long as one is willing to pay the price of not using the knowledge of the smoothness of the propensity score in the unconditional case (which would lead to tighter CIs).

The remainder of this paper is organized as follows. Section 2 presents the model and gives the main finite-sample results. Section 3 presents asymptotic results. Section 4 gives a numerical illustration of the optimal CIs. Section 5 discusses an application to the NSW data. Additional results, proofs and details of results given in the main text are given in appendices.

2 Setup and finite-sample results

This section sets up the model, and shows how to construct finite-sample optimal estimators and well as finite-sample valid and optimal CIs under general smoothness restrictions on the conditional mean of the outcome. We then specialize the results to the case with Lipschitz smoothness. Proofs and additional details are given in Appendix A.

2.1 Setup

We have a random sample of size n. Let $d_i \in \{0, 1\}$ denote the treatment indicator, and let $y_i(0)$ and $y_i(1)$ denote the potential outcomes under no treatment and under treatment, respectively, for each unit i in the sample, $i = 1 \ldots, n$. For each unit i, we observe its treatment status d_i, $y_i = y_i(1)d_i + y_i(0)(1 - d_i)$, as well as a vector of pretreatment variables $x_i \in \mathbb{R}^p$. We condition on the realized values of the treatment status and covariates, $\{x_i, d_i\}_{i=1}^n$, throughout the paper: all probability statements are taken to be with respect

to the conditional distribution of $\{y_i(0), y_i(1)\}_{i=1}^n$ conditional on $\{x_i, d_i\}_{i=1}^n$ unless stated otherwise. This leads to a fixed design regression model

$$y_i = f(x_i, d_i) + u_i, \qquad u_i \text{ are independent with } E(u_i) = 0. \tag{1}$$

Under the assumption of unconfoundedness, the conditional average treatment effect (CATE) is given by[2]

$$Lf = \frac{1}{n}\sum_{i=1}^{n}[f(x_i, 1) - f(x_i, 0)]. \tag{2}$$

In order to obtain finite-sample results, we make the further assumption that u_i is normal

$$u_i \sim N(0, \sigma^2(x_i, d_i)), \tag{3}$$

with the (conditional on x_i and d_i) variance $\sigma^2(x_i, d_i)$ treated as known.

We assume that f lies in a known function class $\mathcal{F}$, which we assume throughout the paper to be convex. We also assume that $\mathcal{F}$ is centrosymmetric in the sense that $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$. The function class $\mathcal{F}$ formalizes the "regularity" or "smoothness" that we are willing to impose. While the convexity assumption is essential for most of our results, the centrosymmetry assumption can be relaxed—see Appendix A. As a leading example, we consider classes that place Lipschitz constraints on $f(\cdot, 0)$ and $f(\cdot, 1)$:

$$\mathcal{F}_{\text{Lip}}(C) = \{f\colon |f(x, d) - f(\tilde{x}, d)| \le C\|x - \tilde{x}\|_{\mathcal{X}},\ d \in \{0, 1\}\},$$

where $\|\cdot\|_{\mathcal{X}}$ is a norm on x, and C denotes the Lipschitz constant, which for simplicity we take to be the same for both $f(\cdot, 1)$ and $f(\cdot, 0)$.

Our goal is to construct estimators and confidence sets for the CATE parameter Lf. We

[2]Formally, suppose that $\{(X_i', D_i, y_i(0), y_i(1))\}_{i=1}^n$ are i.i.d. and that the unconfoundedness assumption $y_i(1), y_i(0) \perp\!\!\!\perp D_i \mid X_i$ holds. Then

$$\frac{1}{n}\sum_{i=1}^{n} E\left[y_i(1) - y_i(0) \mid D_1, \ldots, D_n, X_1, \ldots, X_n\right] = \frac{1}{n}\sum_{i=1}^{n}(f(X_i, 1) - f(X_i, 0)),$$

where $f(x, 1) = E(y_i(1) \mid X_i = x) = E(y_i(1) \mid D_i = 1, X_i = x) = E(y_i \mid D_i = 1, X_i = x)$ and similarly for $f(x, 0)$. Furthermore, $\{y_i\}_{i=1}^n$ follows (1) conditional on $\{(X_i', D_i) = (x_i', d_i)\}_{i=1}^n$. The assumption that u_i is (conditionally) normal then follows from the assumption that each of $y_i(0)$ and $y_i(1)$ are normal (but not necessarily joint normal) conditional on $\{(X_i', D_i)\}_{i=1}^n$.

call a set $\mathcal{C}$ a $100 \cdot (1-\alpha)\%$ confidence set for Lf if it satisfies

$$\inf_{f \in \mathcal{F}} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha, \tag{4}$$

where P_f denotes probability computed under f.

2.2 Linear estimators

Consider an estimator that is linear in the outcomes y_i,

$$\hat{L}_k = \sum_{i=1}^{n} k(x_i, d_i) y_i. \tag{5}$$

This covers many estimators that are popular in practice, such as series of kernel estimators, or various matching estimators. We begin by restricting attention to estimators that take this form, and to CIs based on such estimators. We then show, in Section 2.5 and Appendix A, that such estimators and CIs are optimal or near optimal (depending on the criterion and type of CI being constructed) among all procedures, including nonlinear ones.

Since $\hat{L}_k$ is linear in $\{y_i\}_{i=1}^n$, it is normally distributed with variance

$$\mathrm{sd}(\hat{L}_k)^2 = \sum_{i=1}^{n} k(x_i, d_i)^2 \sigma^2(x_i, d_i)$$

and maximum bias

$$\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k) = \sup_{f \in \mathcal{F}} E_f(\hat{L}_k - Lf) = \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^{n} k(x_i, d_i) f(x_i, d_i) - Lf \right]. \tag{6}$$

Note that by centrosymmetry of $\mathcal{F}$, $\inf_{f \in \mathcal{F}} E_f(\hat{L}_k - Lf) = -\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)$, and that if the minimum bias obtains at f^*, then the maximum bias (6) obtains at $-f^*$.

To form a one-sided confidence interval (CI) based on $\hat{L}_k$, we must take into account its potential bias by subtracting $\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)$ in addition to subtracting the usual normal quantile times its standard deviation—otherwise the confidence interval will undercover for some $f \in \mathcal{F}$. A $100 \cdot (1-\alpha)\%$ one-sided CI is therefore given by $[\hat{c}, \infty)$, where

$$\hat{c} = \hat{L}_k - \overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k) - \mathrm{sd}(\hat{L}_k) z_{1-\alpha},$$

and $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of a $N(0,1)$ distribution.

One could form a two-sided CI centered around $\hat{L}_k$ by adding and subtracting $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) + z_{1-\alpha/2}\,\text{sd}(\hat{L}_k)$. However, this is conservative since the bias cannot be equal to $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$ and to $-\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$ at once. Instead, observe that under any $f \in \mathcal{F}$, the z-statistic $(\hat{L}_k - Lf)/\,\text{sd}(\hat{L}_k)$ is distributed $N(t,1)$ where $t = E_f(\hat{L}_k - Lf)/\,\text{sd}(\hat{L}_k)$, and that t is bounded in absolute value by $|t| \leq b$, where $b = \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)/\,\text{sd}(\hat{L}_k)$ denotes the ratio of worst-case bias to standard deviation. Thus, letting $\text{cv}_\alpha(b)$ be the $1-\alpha$ quantile of the absolute value of a $N(b,1)$ distribution, a two-sided CI can be formed as

$$\left\{\hat{L}_k \pm \text{cv}_\alpha(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)/\,\text{sd}(\hat{L}_k)) \cdot \text{sd}(\hat{L}_k)\right\}. \tag{7}$$

Note that $\text{cv}_\alpha(0) = z_{1-\alpha/2}$, so that if $\hat{L}_k$ is unbiased, the critical value reduces to the usual critical value based on standard normal quantiles. For positive values of the worst-case bias-standard deviation ratio, it will be larger: for $b \geq 1.5$ and $\alpha \leq 0.2$, $\text{cv}_\alpha(b) \approx b + z_{1-\alpha}$ up to three decimal places. For large values of b, the CI is therefore approximately given by adding and subtracting $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) + z_{1-\alpha}\,\text{sd}(\hat{L}_k)$ from $\hat{L}_k$.

Following Donoho (1994), we refer to the CI (7) as a fixed-length confidence interval (FLCI), since it takes the form $\hat{L}_k \pm \chi$ where χ is fixed in the sense that does not depend on the outcomes y_i—it only depends on the known variance function $\sigma^2(\cdot,\cdot)$ and the realized treatment and covariate values $\{x_i, d_i\}_{i=1}^n$ (in practice, the length of the feasible version of this CI will depend on the data through an estimate of the standard deviation).

2.3 Optimal estimators and CIs

To compare different linear estimators, we consider their maximum root mean squared error (RMSE), given by

$$R_{\text{RMSE},\mathcal{F}}(\hat{L}_k) = \left(\sup_{f\in\mathcal{F}} E_f(\hat{L}_k - Lf)^2\right)^{1/2} = \left(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)^2 + \text{sd}(\hat{L}_k)^2\right)^{1/2}.$$

The linear estimator that achieves the lowest RMSE is thus minimax optimal in the class of linear estimators (5). It turns out (see Theorem A.2 in Appendix A.1) that the linear minimax estimator is also highly efficient among all estimators: its efficiency is at least 89.4%, (in the sense that one cannot reduce the RMSE by more than 10.6% by considering non-linear estimators) and, in particular applications, its efficiency can be shown to be even higher. There is thus little loss of efficiency in restricting attention to linear estimators.

One-sided CIs can be compared using the maximum β-quantile of excess length, for a given β (see Appendix A). In Theorem A.1 in Appendix A.1, we show that under this optimality criterion, when the weights k are optimally chosen, a one-sided CI based on $\hat{L}_k$ is minimax among all one-sided CIs, so that, for the purposes of constructing one-sided CIs, there is no efficiency loss in focusing on linear estimators.

Fixed-length CIs are easy to compare—given two FLCIs that satisfy (4), one simply prefers the shorter one. To construct the shortest possible FLCI (in the class of FLCIs based on linear estimators), one therefore needs to choose the weight function k that minimizes the CI length

$$2\,\mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)/\,\mathrm{sd}(\hat{L}_k))\cdot\mathrm{sd}(\hat{L}_k).$$

Since the length of the CI is fixed—it doesn't depend on the data $\{y_i\}_{i=1}^n$, choosing a weighting function to minimize the length does not affect the coverage properties of the resulting CI. We discuss the efficiency of the shortest FLCI among all CIs in Section 2.5.

While in general, the optimal weight function for minimizing the length of FLCI will be different from the one that minimizes RMSE, both performance criteria depend on the weight function k only through $\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)$, and $\mathrm{sd}(\hat{L}_k)$, and they are increasing in both quantities (this is also true for one-sided CIs under the maximum β-quantile of excess length criterion; see Appendix A). Therefore, to find the optimal weights, it suffices to first find weights that minimize the worst-case bias $\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)$ subject to a bound on variance. We can then vary the bound to find the optimal bias-variance tradeoff for a given performance criterion (FLCI or RMSE). It follows from Donoho (1994) and Low (1995) that this bias-variance frontier can be traced out by solving a certain convex optimization problem indexed by δ, where δ indexes the relative weight on variance, and then vary δ.

For a simple statement of the Donoho-Low result, assume that the parameter space $\mathcal{F}$, in addition to being convex and centrosymmetric, does not restrict the value of CATE in the sense that the function $\iota_\alpha(x,d)=\alpha d$ lies in $\mathcal{F}$ for all $\alpha\in\mathbb{R}$ (see Appendix A for a general statement)[3]. Intuitively since $L\iota_\alpha=\alpha$, the set of functions $\{\iota_\alpha\}_{\alpha\in\mathbb{R}}$ is the smoothest set of functions that span the potential values of the CATE parameter Lf, so that this assumption will typically hold unless $\mathcal{F}$ places constraints on the possible values of the CATE parameter. For a given $\delta>0$, let f_δ^* solve

$$\max_{f\in\mathcal{F}} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^n \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)}\le\frac{\delta^2}{4}, \tag{8}$$

[3]We also assume the regularity condition that if $\lambda f+\iota_\alpha\in\mathcal{F}$ for all $0\le\lambda<1$, then $f+\iota_\alpha\in\mathcal{F}$.

and, with a slight abuse of notation, define

$$\hat{L}_\delta = \hat{L}_{k^*_\delta}, \qquad k^*_\delta(x_i, d_i) = \frac{f^*_\delta(x_i, d_i)/\sigma^2(x_i, d_i)}{\sum_{j=1}^n d_j f^*_\delta(x_j, d_j)/\sigma^2(x_j, d_j)}.$$

Then the maximum bias of $\hat{L}_\delta$ occurs at $-f^*_\delta$, and the minimum bias occurs at f^*_δ, so that

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) = \frac{1}{n}\sum_{i=1}^n [f^*_\delta(x_i, 1) - f^*_\delta(x_i, 0)] - \sum_{i=1}^n k^*_\delta(x_i, d_i) f^*_\delta(x_i, d_i),$$

and $\hat{L}_\delta$ minimizes the worst-case bias among all linear estimators with variance bounded by

$$\text{sd}(\hat{L}_\delta)^2 = \frac{\delta^2}{(2\sum_{j=1}^n d_j f^*_\delta(x_j, d_j)/\sigma^2(x_j, d_j))^2}.$$

Thus, the class of estimators $\{\hat{L}_\delta\}_{\delta>0}$ traces out the optimal bias-variance frontier. The variance $\text{sd}(\hat{L}_\delta)^2$ can be shown to be decreasing in δ, so that δ can be thought of as indexing the relative weight on variance.

The weights leading to the shortest possible FLCI are thus given by $k^*_{\delta_\chi}$, where δ_χ minimizes $\text{cv}_\alpha(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta)/\text{sd}(\hat{L}_\delta)) \cdot \text{sd}(\hat{L}_\delta)$ over δ. Similarly, the optimal weights for estimation are given by $k^*_{\delta_{\text{RMSE}}}$, where δ_{RMSE} minimizes $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta)^2 + \text{sd}(\hat{L}_\delta)^2$.

2.4 Estimators and CIs under Lipschitz smoothness

Computing a fixed-length CI based on a linear estimator $\hat{L}_k$ requires computing the worst-case bias (6). Computing the RMSE-optimal estimator, and the optimal FLCI requires solving the optimization problem (8). Both of these optimization problems requires optimizing over the set $\mathcal{F}$, which, in nonparametric settings, is infinite-dimensional. We now focus on the Lipschitz class $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, and show that in this case, solutions to these optimization problems can be found by solving finite-dimensional linear and convex programs, respectively.

First, observe that both in the optimization problem (6) and in (8), the objective and constraints depend on f only through its value at the points $\{(x_i, 0), (x_i, 1)\}_{i=1}^n$; the value of f at other points does not matter. Furthermore, it follows from Beliakov (2006, Theorem 4) that if the Lipschitz constraints hold at these points, then it is always possible to find a function $f \in \mathcal{F}_{\text{Lip}}(C)$ that interpolates these points (see Lemma A.1). Consequently, in solving the optimization problems (6) and (8), we identify f with the vector

$(f(x_1,0),\ldots,f(x_n,0),f(x_1,1),\ldots,f(x_n,1))' \in \mathbb{R}^{2n}$, and replace the functional constraint $f \in \mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$ with $2n(n-1)$ linear inequality constraints

$$f(x_i,d) - f(x_j,d) \leq C \left\|x_i - x_j\right\|_{\mathcal{X}} \text{ for } d \in \{0,1\}, i,j \in \{1,\ldots,n\}. \tag{9}$$

This leads to the following result:

Theorem 2.1. *Consider a linear estimator $\hat{L}_k = \sum_{i=1}^n k(x_i,d_i)y_i$ where k satisfies*

$$\sum_{i=1}^n d_i k(x_i,d_i) = 1 \text{ and } \sum_{i=1}^n (1-d_i)k(x_i,d_i) = -1. \tag{10}$$

The worst-case bias of this estimator, $\overline{\text{bias}}_{\mathcal{F}_{Lip}(C)}(\hat{L}_k)$, is given by the value of

$$\max_{f \in \mathbb{R}^{2n}} \left\{ \sum_{i=1}^n k(x_i,d_i) f(x_i,d_i) - \frac{1}{n}\sum_{i=1}^n \left[f(x_i,1) - f(x_i,0)\right] \right\}, \tag{11}$$

where the maximum is taken s.t. (9) *and s.t.*

$$\sum_{i=1}^n f(x_i,1) = \sum_{i=1}^n f(x_i,0) = 0. \tag{12}$$

The assumption that $\hat{L}_k$ satisfies (10) is necessary to prevent the bias from becoming arbitrarily large at multiples of $f(x,d) = d$ and $f(x,d) = 1-d$. If (10) holds, then the set of possible biases over $f \in \mathcal{F}_{\text{Lip}}(C)$ is the same as the set of possible biases over the restricted set of functions with the additional constraint (12), since any function in the class can be obtained by adding a function in the span of $\{(x,d) \mapsto d, (x,d) \mapsto (1-d)\}$ to such a function without affecting the bias. In particular, Theorem 2.1 implies that the formulas for one-sided CIs and two-sided FLCIs given in Section 2.2 hold with $\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_k)$ given by (11).

For RMSE-optimal estimators and optimal FLCIs, we have the following result:

Theorem 2.2. *Given $\delta > 0$, let f_δ^* solve*

$$\max_{f \in \mathbb{R}^{2n}} 2\frac{1}{n}\sum_{i=1}^n \left[f(x_i,1) - f(x_i,0)\right] \quad \text{s.t.} \quad \sqrt{\sum_{i=1}^n \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)}} \leq \frac{\delta}{2} \quad \text{and s.t. (9).} \tag{13}$$

Let $\hat{L}_\delta = \hat{L}_{k^*_\delta}$*, where*

$$k^*_\delta(x_i, d_i) = \frac{f^*_\delta(x_i, d_i)/\sigma^2(x_i, d_i)}{\sum_{j=1}^n d_j f^*_\delta(x_j, d_j)/\sigma^2(x_j, d_j)}, \tag{14}$$

and let $\overline{\text{bias}}_\delta = \frac{1}{n}\sum_{i=1}^n [f^*_\delta(x_i, 1) - f^*_\delta(x_i, 0)] - \sum_{i=1}^n k^*_\delta(x_i, d_i) f^*_\delta(x_i, d_i)$ *denote the bias of* $\hat{L}_\delta$ *at* $-f^*_\delta$*.*

Then the estimator $\hat{L}_\delta$ *attains the worst-case bias at* $-f^*$*,* $\overline{\text{bias}}_{\mathcal{F}_{Lip}(C)}(\hat{L}_\delta) = \overline{\text{bias}}_\delta$*. The estimator* $\hat{L}_{\delta_{RMSE}}$*, where* δ_{RMSE} *minimizes* $\overline{\text{bias}}^2_\delta + \text{sd}(\hat{L}_\delta)^2$ *over* δ*, minimaxes RMSE among all linear estimators. The shortest FLCI among all FLCIs centered at linear estimators is given by*

$$\left\{\hat{L}_{\delta_\chi} \pm \text{cv}_\alpha(\overline{\text{bias}}_{\delta_\chi} / \text{sd}(\hat{L}_{\delta_\chi}))\,\text{sd}(\hat{L}_{\delta_\chi})\right\},$$

where δ_χ *minimizes* $\text{cv}_\alpha(\overline{\text{bias}}_\delta / \text{sd}(\hat{L}_\delta))\,\text{sd}(\hat{L}_\delta)$ *over* δ*.*

Theorem 2.2 shows that the optimization problem (8) that involves optimization over an infinite-dimensional function space can be replaced by an optimization problem in $\mathbb{R}^{2n}$ with $2n(n-1)$ linear constraints, one quadratic constraint and a linear objective function.

While the estimators $\hat{L}_\delta$ do not, in general, have a closed form, it turns out that, when C is large enough, the optimal estimator takes the form of a matching estimator. A matching estimator with M matches (with replacement) constructs estimates $\hat{f}(x_i, d_i) = y_i$, and $\hat{f}(x_i, 1-d_i) = \hat{y}_{i,M}$, where $\hat{y}_{i,M}$ is the average outcome of the M observations closest to i (using the norm $\|\cdot\|_\mathcal{X}$) among the observations with treatment status different from i. The matching estimator of the CATE is then given by $L\hat{f}$, and can be written in the form (5), with $k(\cdot)$ given by

$$k_{\text{match},M}(x_i, d_i) = \frac{1}{n}(2d_i - 1)\left(1 + \frac{K_M(i)}{M}\right), \tag{15}$$

where $K_M(i)$ is the number of times the ith observation is matched.

Theorem 2.3. *Suppose that* $\sigma(x_i, d_i) > 0$ *for each* i*, and suppose that each unit has a single closest match, so that* $\text{argmin}_{j:\, d_j \neq d_i} \|x_i - x_j\|_\mathcal{X}$ *is a singleton for each* i*. There exists a constant* K *depending on* $\sigma^2(x_i, d_i)$ *and* $\{x_i, d_i\}_{i=1}^n$ *such that, if* $C/\delta > K$*, the optimal estimator* $\hat{L}_\delta$ *is given by the matching estimator with* $M = 1$*.*

Theorem 2.1 allows one to construct finite-sample CIs based on any linear estimator, as well as to compute the RMSE of any linear estimator. One can compare the resulting FLCI to the conventional CI that uses critical values based on normal quantiles and ignores the potential bias as a form of sensitivity analysis: if the CIs are substantively different, this indicates that conventional asymptotics may not work well for the sample at hand unless

one further restricts the parameter space $\mathcal{F}$. One can also compare the FLCI length and the RMSE to the length of the optimal FLCI and linear minimax RMSE estimator given by Theorem 2.2 to quantify the loss of efficiency by using a suboptimal estimator. In our numerical illustration and application, we conduct such comparison for matching estimators. Although Theorem 2.3 implies that matching estimators are suboptimal unless C is large enough, we find that, in our application, the efficiency loss is limited provided the number of matches is chosen optimally.

2.5 Bounds to adaptation

The results in Section 2.3 and Theorem 2.2 show how to construct the shortest FLCI based on a linear estimator. One may, however, worry that only considering fixed-length CIs based on linear estimators may be too restrictive. In particular, the length of a fixed-length CI is determined by the least-favorable function in $\mathcal{F}$ (that maximizes the potential bias), which may result in CIs that are "too long" when f turns out to be smooth. Consequently, one may prefer a variable-length CI that optimizes its expected length over a class of smoother functions $\mathcal{G} \subset \mathcal{F}$ (while maintaining coverage over the whole parameter space), especially if this leads to substantial reduction in expected length when $f \in \mathcal{G}$. When such a CI also simultaneously optimizes its length over all of $\mathcal{F}$, it is referred to as "adaptive".

A related concern is that implementing our CIs in practice requires the user to explicitly specify the parameter space $\mathcal{F}$, which typically involves specification of smoothness constants such as the Lipschitz constant C in the case of Lipschitz smoothness. This in particular rules out fully data-driven procedures that try to implicitly or explicitly estimate C from the data.

To address these concerns, in Theorem A.3 in Appendix A, we give a sharp bound on the problem of constructing a confidence set that optimizes its expected length at a smooth function of the form $g(x, d) = \alpha_0 + \alpha_1 d$, while maintaining coverage over the original parameter space $\mathcal{F}_{\text{Lip}}(C)$ for a given $C > 0$. The sharp bound follows from general results in Armstrong and Kolesár (2017), and it gives a benchmark for the scope for improvement over the FLCI in Theorem 2.2. Theorem A.3 also gives a general lower bound for this sharp bound; this result is new.

In particular, Theorem A.3 shows that the efficiency of the FLCI depends on the realized values of $\{x_i, d_i\}_{i=1}^n$ and the form of the variance function $\sigma^2(\cdot, \cdot)$, and that the efficiency can be lower-bounded by 64.6% when $1-\alpha = 0.95$. In a particular application, one can explicitly compute the sharp efficiency bound; typically it is much higher than the lower bound. For example, in the baseline specification in our empirical application in Section 5, we find that

the efficiency of the FLCI is over 99% at such smooth functions g, implying that there is very little scope for improvement over the FLCI: not only must the rate of convergence be the same even if one optimizes length g, the constant is also very tight.

Consequently, data-driven or adaptive methods for constructing CIs must either fail to meaningfully improve over the FLCI, or else undercover for some $f \in \mathcal{F}_{\text{Lip}}(C)$. It is thus not possible to estimate the Lipschitz constant C for the purposes of forming a tighter CI—it must be specified ex ante by the researcher. Because of this, by way of sensitivity analysis, we recommend reporting estimates and CIs for a range of choices of the Lipschitz constant C when implementing the FLCI in practice to see how assumptions about the parameter space affect the results. We adopt this approach in the empirical application in Section 5. This also mirrors the common practice of reporting results for different specifications of the regression function in parametric regression problems.

These efficiency results are not specific to setting $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$; the key assumption of Theorem A.3 is that $\mathcal{F}$ is convex and centrosymmetric. If additional restrictions such as monotonicity are used that break either convexity or centrosymmetry, then some degree of adaptation may be possible. While we leave the full exploration of this question for future research, we note that the approach in Section 2.3 can still be used when the centrosymmetry assumption is dropped. As an example, we show how optimal fixed-length CIs can be computed when $\mathcal{F}$ imposes Lipschitz and monotonicity constraints in Appendix A.

3 Asymptotic results

3.1 Semiparametric efficiency bound

Suppose that $\{(X_i', D_i, y_i(0), y_i(1))\}_{i=1}^n$ are drawn i.i.d., so that the Gaussian regression model given by (1) and (3) obtains conditional on the realizations $\{(X_i', D_i) = (x_i', d_i)\}_{i=1}^n$, if $y_i(0)$ and $y_i(1)$ are normal (but not necessarily joint normal) conditional on $\{(X_i', D_i)\}_{i=1}^n$. Let $e(x) = P(D_i = 1 \mid X_i = x)$ denote the propensity score. If $\mathcal{F}$ imposes sufficient smoothness, then it is possible to construct root-n consistent estimators with asymptotically negligible bias. Furthermore, Hahn (1998) shows that such estimator that is regular can have asymptotic variance lower than the linear estimator with the kernel

$$k_{\text{seb}}(x_i, d_i) = \frac{1}{n}\left[\frac{d_i}{e(x_i)} - \frac{1 - d_i}{1 - e(x_i)}\right]. \tag{16}$$

The asymptotic variance of this linear estimator is known as the semiparametric efficiency bound. We compare the kernel of the optimal estimator $\hat{L}_{\delta_{\text{RMSE}}}$ to k_{seb} in our numerical illustration in Section 4.

The semiparametric efficiency bound gives only a lower bound for the asymptotic variance: it cannot be achieved unless $\mathcal{F}$ imposes sufficient smoothness relative to the dimension of x_i. Let $\Sigma(\gamma, C)$ denote the set of ℓ-times differentiable functions f such that, for all integers $k_1, k_2, \ldots, k_p$ with $\sum_{j=1}^{p} k_j = \ell$, $\left| \frac{d^\ell}{dx_1^{k_1} \cdots dx_p^{k_p}} f(x) - \frac{d^\ell}{dx_1^{k_1} \cdots dx_p^{k_p}} f(x') \right| \leq C \|x - x'\|_{\mathcal{X}}^{\gamma - \ell}$, where ℓ is the greatest integer strictly less than γ and $\|\cdot\|_{\mathcal{X}}$ denotes the Euclidean norm on $\mathbb{R}^p$. Note that $f \in \mathcal{F}_{\text{Lip}}(C)$ is equivalent to $f(\cdot, 1), f(\cdot, 0) \in \Sigma(1, C)$. Robins et al. (2009) consider minimax rates of testing and estimation when (X_i, D_i) are not conditioned on, and $f(\cdot, 0), f(\cdot, 1) \in \Sigma(\gamma_f, C)$ and $e \in \Sigma(\gamma_e, C)$. Their results imply that if one requires unconditional coverage of Lf (rather than conditional coverage conditional on the realizations of covariates and treatment), root-n inference is impossible unless $\gamma_e + \gamma_f \geq p/2$ where p is the dimension of the (continuously distributed) covariates.

Since conditioning on the realizations $\{x_i, d_i\}_{i=1}^n$ essentially takes away the role of smoothness of $e(\cdot)$, this suggests that conditional root-n inference should be impossible unless $\gamma_f \geq p/2$ (i.e. the conditions for impossibility of root-n inference in our setting with fixed x_i and d_i should correspond to the conditions derived by Robins et al. 2009 in the case where no smoothness is imposed on $e(\cdot)$). This intuition turns out to be essentially correct:

Theorem 3.1. *Let $f(\cdot, 0), f(\cdot, 1) \in \Sigma(\gamma, C)$, and let $\{X_i, D_i\}$ be i.i.d. with $X_i \in \mathbb{R}^p$ and $D_i \in \{0, 1\}$. Suppose that the Gaussian regression model (1) and (3) holds conditional on the realizations of the treatment and covariates. Suppose that the marginal probability that $D_i = 1$ is not equal to zero or one and that X_i has a bounded density conditional on D_i. Let $[\hat{c}_n, \infty)$ be a sequence of CIs with asymptotic coverage at least $1 - \alpha$ for the CATE conditional on $\{X_i, D_i\}_{i=1}^n$:*

$$\liminf_{n \to \infty} \inf_{f(\cdot,0), f(\cdot,1) \in \Sigma(C,\gamma)} P_f \left(\frac{1}{n} \sum_{i=1}^{n} [f(X_i, 1) - f(X_i, 0)] \in [\hat{c}_n, \infty) \;\middle|\; \{X_i, D_i\}_{i=1}^n \right) \geq 1 - \alpha$$

almost surely. Then, under the zero function $f(x, d) = 0$, $\hat{c}_n$ cannot converge to the CATE (which is 0 in this case) more quickly than $n^{-\gamma/p}$: there exists $\eta > 0$ such that

$$\liminf_{n} P_0 \left(\hat{c}_n \leq -\eta n^{-\gamma/p} | \{X_i, D_i\}_{i=1}^n \right) \geq 1 - \alpha$$

almost surely.

Proof of this result is given in Appendix B. The theorem shows that the excess length of a confidence interval with conditional coverage in the class with $f(\cdot,0), f(\cdot,1) \in \Sigma(\gamma, C)$ must be of order at least $n^{-\gamma/p}$, even at the "smooth" function $f(x,d) = 0$. The Lipschitz case we consider throughout most of this paper corresponds to $\gamma = 1$, so that root-n inference is possible only when $p \leq 2$.

On the other hand, when $\gamma/p > 1/2$, Chen et al. (2008) show that the semiparametric efficiency bound can be achieved (for example, using series estimators) without smoothness assumptions on the propensity score (while Chen et al. 2008 do not condition on treatments and pretreatment variables, their arguments appear to extend to the conditional case).

3.2 Unknown error distribution

In practice, the error distribution is typically unknown, which makes estimators and CIs that depend on $\sigma^2(x,d)$ infeasible. To implement feasible versions of the CIs proposed in this paper, we propose the following. Let $\tilde{\sigma}^2(x,d)$ be a (possibly incorrect) guess or estimate of the conditional variance function. Let $\tilde{L}_\delta$, $\tilde{k}^*_\delta$ and $\widetilde{\text{bias}}_\delta$ denote the estimator, weights and worst-case bias computed using $\tilde{\sigma}^2(x,d)$ as the conditional variance. The worst-case bias calculations do not depend on the correct specification of the variance, so $\widetilde{\text{bias}}_\delta$ still gives the worst-case bias of $\tilde{L}_\delta$. We then form the standard error using an estimate that does not impose correct specification of the conditional variance:

$$\text{se}(\tilde{L}_\delta) = \sqrt{\sum_{i=1}^n \tilde{k}^*_\delta(x_i,d_i)^2 \hat{u}_i^2}$$

where $\hat{u}_i = y_i - \hat{f}(x_i,d_i)$ and $\hat{f}(x,d)$ is an estimate of $f(x,d)$. The FLCI is then given by

$$\left\{\tilde{L}_\delta \pm \text{cv}_\alpha(\widetilde{\text{bias}}_\delta/\text{se}(\tilde{L}_\delta))\text{se}(\tilde{L}_\delta)\right\}$$

and the one-sided CI is given by

$$[\tilde{L}_\delta - \widetilde{\text{bias}}_\delta - \text{se}(\tilde{L}_\delta)z_{1-\alpha}, \infty).$$

The following theorem gives sufficient conditions for the asymptotic validity of this CI for the case where $\mathcal{F}$ is the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$. To allow for the possibility that the researcher may want to choose a more conservative parameter space when the sample size is

large, we allow for the possibility that $C = C_n \to \infty$ as the sample size n increases.

Theorem 3.2. *Consider the fixed design model* (1) *with* u_i *distributed independently (but not identically distributed) with* $Eu_i = 0$ *and* $1/K \leq Eu_i^2 \leq K$ *and* $E|u_i|^{2+1/K} \leq K$ *for some* K*. Let* $\mathcal{C}$ *be one of the CIs described above with* $\mathcal{F} = \mathcal{F}_{Lip}(C_n)$*, with* $\tilde{\sigma}^2(x,d)$ *a nonrandom function bounded away from zero and infinity. Suppose that*

$$\text{for all } \eta > 0, \min_{1 \leq i \leq n} \#\{j \in \{1,\ldots,n\} : \|x_j - x_i\| \leq \eta/C_n,\, d_i = d_j\} \to \infty. \tag{17}$$

Then, if the estimator $\hat{f}(x_i, d_i)$ *used to construct the variance estimate satisfies*

$$\max_{1 \leq i \leq n} \sup_{f \in \mathcal{F}} E_f[(f(x_i,d_i) - \hat{f}(x_i,d_i))^2] \to 0,$$

we will have $\liminf_{n\to\infty} \inf_{f \in \mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha$*. In particular, this holds when* $\hat{f}$ *is the Nadaraya-Watson estimator with uniform kernel and a bandwidth sequence* h_n *with* $h_n C_n$ *converging to zero slowly enough.*

Proof is given in Appendix C. Importantly, Theorem 3.2 allows for non-regular cases such as cases discussed in Section 3.1 in which the semiparametric efficiency bound cannot be achieved. It also follows from the proof of this theorem that the estimator $\tilde{L}_\delta$ is asymptotically normal, including in non-regular cases. In such cases, the worst-case bias can form a non-negligible portion of CI length, even asymptotically.

4 Numerical illustration

To get a sense of what the optimal kernels look like, we generate $\{x_i, d_i\}_{i=1}^n$ i.i.d. with $x_i \sim \text{unif}(0,1)$ and $P(d_i = 1 | x_i = x) = e(x) = 2(x - 1/2)^2 + 1/4$ for a range of sample sizes n. We then compute the optimal kernel k_δ^* with $\sigma^2(x_i, d_i) = 1$ and Lipschitz constant $C = 1$ and $\delta = 2z_{.95}$ so that a minimax test with level .05 has power .95. For comparison, we compute the kernel associated with the matching estimator with M matches for a range of values of M, which is given by (15). We also compare the optimal weights to the weights corresponding to the semiparametric efficiency bound, given in (16).

Figures 1, 2 and 3 plot the minimax optimal weight function k_δ^* and $k_{\text{match},M}$, with $M = 5$, along with k_{seb} for a single draw of the data for $n = 100$, $n = 250$ and $n = 500$ (each of the weight functions are scaled by n to make them comparable across sample sizes). For this

draw of the dgp with $n = 100$, the estimator based on k_b^* has worst-case bias 0.0201 and standard deviation 0.2053. The worst-case bias for the matching estimator with $M = 5$ is 0.0202, and its standard deviation is 0.2081. For $n = 250$, the estimator based on k_b^* has worst-case bias 0.0087 and standard deviation 0.1331. The worst-case bias for the matching estimator with $M = 5$ is 0.0079, and its standard deviation is 0.1353. For $n = 500$, the worst-case bias for the minimax estimator is 0.0057, and the standard deviation is 0.0963, while the $M = 5$ matching estimator has worst-case bias 0.0048 and standard deviation 0.0983. Overall, the matching estimators seem to be close to optimal.

5 Application to National Supported Work demonstration

We now consider an application to the National Supported Work (NSW) demonstration. The dataset that we use is the same as the one analyzed by Dehejia and Wahba (1999) and Abadie and Imbens (2011).[4] The sample with $d_i = 1$ corresponds to the experimental sample of 185 men who received job training in a randomized evaluation of the NSW program. The sample with $d_i = 0$ is a non-experimental sample of 2490 men taken from the PSID. We are interested in the conditional average treatment effect on the treated (assuming unconfoundedness):

$$\text{CATT}(f) = \frac{\sum_{i=1}^n \left[f(x_i, 1) - f(x_i, 0)\right] d_i}{\sum_{i=1}^n d_i}.$$

The analysis in Section 2 goes through essentially unchanged, with CATT(f) replacing CATE(f) throughout (see Appendix A).

In this data, y_i denotes earnings in 1978 (after the training program) in thousands of dollars. The variable x_i contains the following variables (in the same order): age, education, indicators for Black and Hispanic, indicator for marriage, earnings in 1974, earnings in 1975 (before the training program), and employment indicators for 1974 and 1975.[5]

[4] Taken from Rajeev Dehejia's website, `http://users.nber.org/~rdehejia/nswdata2.html`.

[5] Following Abadie and Imbens (2011), the no-degree indicator variable is dropped, and the employment indicators are defined as an indicator for nonzero earnings (Abadie and Imbens, 2011, do not give details of how they constructed the employment variables, but these definitions match their summary statistics).

5.1 Choice of norm for Lipschitz class

The choice of the norm on $\mathbb{R}^p$ used in the definition of the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$ and in determining matches is important both for minimax estimators and for matching estimators. For a positive definite symmetric $p \times p$ matrix A, define the norm

$$\|x\|_{A,p} = \left(\sum_{i=1}^{n} \left| \left(A^{1/2}x\right)_i \right|^p \right)^{1/p} \tag{18}$$

where $\left(A^{1/2}x\right)_i$ denotes the ith element of Ax. Ideally, the parameter space $\mathcal{F}_{\text{Lip}}(C)$ should reflect the a priori restrictions the researcher is willing to place on the conditional mean of the outcome variable under treatment and control. If we take A to be a diagonal matrix, then, when $C = 1$, the j, jth element gives the a priori bound on the derivative of the regression function with respect to x_j.

We use $A = A_{\text{main}}$ given in Table 2 in defining the distance in our main specification. To make the distance more interpretable, we use $p = 1$ in defining the distance, so that the Lipschitz condition places a bound on the cumulative effect of all the variables. We discuss other choices of the weights A in Section 5.4. The elements of A_{main} are chosen to give restrictions on $f(x, d)$ that are plausible when $C = 1$, and we report results for a range of choices of C as a form of sensitivity analysis. It is perhaps easiest to interpret the bounds in terms of percentage increase in expected earnings. As a benchmark, consider deviations from expected earnings when $f(x_i, d_i) = 10$, that is \$10,000. Since the average earnings of for the $d_i = 1$ sample is 6.4 thousand dollars, with 78% of the treated sample reporting income below 10 thousand dollars, the implied percentage bounds for most people in the treated sample will be even more conservative. When $C = 1$, and $A = A_{\text{main}}$, the implied bounds for the effect of age and education on expected earnings at 10 thousand dollars are 1.5% and 6%, respectively, which is in line with the 1980 census data. Similarly, the wage gap implied by the black, Hispanic, and married indicators is bounded at 25%. The A_{main} coefficients on 1974 and 1975 earnings imply that their cumulative effect on 1978 earnings is at most a one-to-one increase. Including the employment indicators allows for a small discontinuous jump in addition for people with zero previous years' earnings.

5.2 Results

We compute the estimator $\hat{L}_\delta$ as described in Section 3.2 with the initial guess for the variance function given by the constant function $\tilde{\sigma}^2(x, d) = \hat{\sigma}^2$, where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2$ and

$\hat{u}_i = y_i - \hat{f}(x_i, d_i)$ where $\hat{f}(x_i, d_i)$ is the nearest-neighbor estimate with 30 neighbors, with the nearest neighbors defined using Mahalanobis distance (using the metric $\|\cdot\|_{A_{\text{main}},1}$, as in the definition of the Lipschitz class leads to very similar results). The robust standard deviation estimate follows the formula in Section 3.2, while the non robust estimate is computed under the assumption that the variance is constant and equal to $\hat{\sigma}^2$. For one-sided CIs, we calibrate δ so that the test is optimal for worst-case 0.8 quantile with $\alpha = 0.05$. Since the problem is translation invariant, the minimax one-sided CI inverts minimax tests with size 0.05 and power 0.8 (see Armstrong and Kolesár, 2017), which is a common benchmark in the literature on statistical power analysis (Cohen, 1988). For two-sided CIs, δ is calibrated to minimize the width of the resulting CI, and for estimation, it is calibrated to minimize the worst-case RMSE.

Figure 4 plots the optimal one-sided CIs in both directions along with the optimal affine FLCI and RMSE optimal affine estimator as a function of C. For very small values of C—smaller than 0.1—the Lipschitz assumption implies that selection on pretreatment variables does not lead to substantial bias, and the optimal estimator and CIs incorporates this by tending toward the raw difference in means between treated and untreated individuals, which in this data set is negative. For $C \geq 0.1$, the point estimate is positive and remarkably stable as a function of C, which suggests that the estimator and CIs are accounting for the possibility of selection bias by controlling for observables. Note also that the two-sided FLCIs become wider as C increases, reflecting greater potential bias resulting from a less restrictive parameter space.

Interestingly, the upper one-sided CI is above the upper endpoint of the two-sided CI for some values of C. This occurs because the one-sided CI criterion resolves the bias-variance tradeoff differently than the two-sided FLCI: the FLCI and one-sided CI are based on the estimator $\hat{L}_\delta$ with different choices of δ (recall that $\hat{L}_\delta$ minimizes the variance subject to a bound on worst-case bias subject, with δ determining the relative weights given to bias and variance). In particular, the one-sided CI uses a smaller value of δ for a given C when applied to this data set—they "undersmooth" relative to two-sided CIs, which leads to the one-sided CI being based on a different point estimate than the two-sided FLCI.

To examine this more closely, Figure 5 focuses on the case where $C = 1$ and plots the optimal estimator along with its standard deviation, worst-case bias, RMSE and CI length as a function of δ. For this figure, the standard deviation is computed under the assumption of homoskedasticity, so that the standard deviation, RMSE and CI length are identical to those optimized by the estimator. For comparison, we also plot the same quantities for matching

estimators as a function of M, the number of matches, using the linear programming problem described in Section 2.4 to compute worst-case bias (the distance used to define matches is the same as the one used for the Lipschitz condition). For the matching estimator, M plays the role of a tuning parameter that trades off bias and variance, just as δ does for the class of optimal estimators: larger values of M tend to lower the variance and increase the bias (although the relationship is not always monotonic). As required by Theorem 2.3, $\hat{L}_\delta$ approaches the matching estimator with $M = 1$ as δ gets small enough.

Table 1 reports the point estimates that optimize each of the criteria plotted in Figure 5 along with worst-case bias, standard errors, and the value of the tuning parameter (δ or M) that optimizes the given criterion. These are simply the estimates from Figure 5 taken at the value of δ or M where the given criterion takes the minimum in the corresponding plot in the figure. Note that, in all cases, the bias is non-negligible relative to variance: unlike CIs based on conventional asymptotics, the CIs computed here reflect the "nonparametric" nature of the problem by explicitly taking bias into account. One can see that the bias-variance trade-off for both the matching estimator and the optimal estimator is resolved differently for different optimality criteria, with two-sided CIs employing the most smoothing.

5.3 Comparison with experimental estimates

The present analysis follows LaLonde (1986), Dehejia and Wahba (1999), Smith and Todd (2001), Smith and Todd (2005) and Abadie and Imbens (2011) (among others) in using a non-experimental sample to estimate treatment effects of the NSW program. A major question in this literature has been whether a non-experimental sample can be used to obtain the same results (or, at least, results that are the same up to sampling error) as estimates based on the original experimental sample of individuals who were randomized out of the NSW program. Taking the difference in means between the outcome for the treated and untreated individuals in the subset of the experimental sample that corresponds to the data used here gives an estimate of the average treatment effect on the treated (ATT) of 1.794 with a standard error of 0.633 (see Dehejia and Wahba, 1999, Table 3).

As can be seen from Table 1, the matching estimator (when RMSE is the optimality criterion) gets remarkably close to this experimental benchmark. For constructing one-and two-sided CIs, it is optimal to use a larger number of matches, which leads to a lower point estimate; the decrease in the point estimate is accompanied by a corresponding increase in the worst-case bias. This is consistent with the hypothesis that the extra smoothing increases the bias in the estimator. The estimate based on the optimal estimator is lower,

although the distance between the estimate and the experimental estimate is much smaller than the worst-case bias. This suggest that bias alone can explain the difference between the estimates. However, differences between the estimates reported here and the experimental estimate can also arise from (1) differences between the CATT for our sample and the ATT (2) failure of the selection on observables assumption; and (3) variance in estimating the CATT, as well as sampling error in the experimental estimates of the ATT.

5.4 Other choices of distance

A disadvantage of the distance based on $A = A_{\text{main}}$ is that it requires prior knowledge of the relative importance of different pretreatment variables in explaining the outcome variable. An alternative is to specify the distance using moments of the pretreatment variables in a way that ensures invariance to scale transformations. For example, Abadie and Imbens (2011) form matching estimators using $p = 2$ and $A^{1/2} = A_{\text{ne}}^{1/2} \equiv \text{diag}(1/\text{std}(x_1), \ldots, 1/\text{std}(x_p))$, where std denotes sample standard deviation. Table 2 shows the diagonal elements of A_{ne}, which are simply the inverses of the standard deviations of each control variable. From this table, it can be seen that this distance is most likely not the best way of encoding a researcher's prior beliefs about Lipschitz constraints. For example, the bound on the difference in average earnings between Blacks and non-Black non-Hispanics is substantially smaller than the bound on the difference in average earnings between Hispanics and non-Black non-Hispanics.

If the constant C is to be chosen conservatively, the derivative of $f(x, d)$ with respect to each of these variables must be bounded by C times the corresponding element in this table. If one allows for somewhat persistent earnings, this would suggest that C should be chosen in the range of 10 or above: to allow previous year's earnings to have a one-to-one effect, we would need to take $C = 1/\sqrt{.07^2 + .07^2} = 10.11$. For $C = 10$, the FLCI is $1.7179 \pm 7.9901 = [-6.2722, 9.7080]$, which is much wider than the FLCIs reported for A_{main} when $C = 1$.

Appendix A Proofs and additional derivations

This appendix contains proofs and derivations in Section 2, as well as additional results. Appendix A.1 maps a generalization of the setup in Section 2.1 to the framework of Donoho (1994) and Armstrong and Kolesár (2017), and specializes their general efficiency bounds and optimal estimator and CI construction to the current setting. This gives the formulas

for optimal estimators and CIs given in Section 2.3, and the efficiency bounds discussed in Section 2.5. Appendix A.2 proves Theorem 2.2, as well as a generalization of the theorem to Lipschitz classes with monotonicity. The proof of Theorem 2.1 follows from the arguments in the main text and it is omitted. Appendix A.3 proves Theorem 2.3.

A.1 General setup and results

We consider a generalization of the setup in Section 2.1 by letting the parameter of interest be a general weighted conditional average treatment effect of the form

$$Lf = \sum_{i=1}^{n} w_i[f(x_i, 1) - f(x_i, 0)]$$

where $\{w_i\}_{i=1}^n$ is a set of known weights with $\sum_{i=1}^n w_i = 1$. Setting $w_i = 1/n$ gives the CATE, while setting $w_i = d_i / \left(\sum_{j=1}^n d_j\right)$ gives the conditional average treatment effect on the treated (CATT). We retain the assumption that $\mathcal{F}$ is convex, but drop the centrosymmetry assumption. We also slightly generalize the class of estimators we consider by allowing for a recentering by some constant a. This leads to affine estimators of the form

$$\hat{L}_{k,a} = a + \sum_{i=1}^{n} k(x_i, d_i) y_i,$$

with the notational convention $\hat{L}_k = \hat{L}_{k,0}$. Define maximum and minimum bias

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a}) = \sup_{f \in \mathcal{F}} E_f(\hat{L}_{k,a} - Lf), \qquad \underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a}) = \inf_{f \in \mathcal{F}} E_f(\hat{L}_{k,a} - Lf).$$

A fixed-length CI around $\hat{L}_{k,a}$ can be formed as

$$\left\{ \hat{L}_{k,a} \pm \text{cv}_\alpha(b/\,\text{sd}(\hat{L}_{k,a})) \cdot \text{sd}(\hat{L}_{k,a}) \right\}, \quad \text{where} \quad b = \max\left\{ |\,\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a})|, |\,\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a})| \right\}.$$

The RMSE of $\hat{L}_{k,a}$ is given by

$$R_{\text{RMSE},\mathcal{F}}(\hat{L}_{k,a}) = \sqrt{b^2 + \text{sd}(\hat{L}_{k,a})^2}, \quad \text{where} \quad b = \max\left\{ |\,\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a})|, |\,\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a})| \right\}.$$

For comparisons of one-sided CIs $[\hat{c}, \infty)$, we focus on quantiles of excess length. Given a subset $\mathcal{G} \subseteq \mathcal{F}$, define the worst-case βth quantile of excess length over $\mathcal{G}$:

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

where $q_{g,\beta}(\cdot)$ denotes the βth quantile under the function g, and $Lg - \hat{c}$ is the excess length of the CI $[\hat{c}, \infty)$. Taking $\mathcal{G} = \mathcal{F}$, a CI that optimizes $q_\beta(\hat{c}, \mathcal{F})$ is minimax. Taking $\mathcal{G}$ to correspond to a smaller set of smoother functions amounts to "directing power" at such smooth functions. For a one-sided CI $[\hat{c}, \infty)$ with $\hat{c} = \hat{L}_{k,a} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a}) - z_{1-\alpha}\,\text{sd}(\hat{L}_{k,a})$, we have

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

This follows from the fact that the worst-case βth quantile of excess length over $\mathcal{G}$ is taken at the function $g \in \mathcal{G}$ that achieves $\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_{k,a})$ (i.e. when the estimate is biased downward as much as possible).

Note that if the performance criterion is RMSE or length of FLCI, it is optimal to set the centering constant a such that $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a}) = -\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a})$ (which yields $a = 0$ as the optimal choice under centrosymmetry), while the centering constant does not matter for constructing one-sided CIs. If the performance criterion is RMSE, length of FLCI, or $q_\beta(\cdot, \mathcal{F})$, and the centering constant chosen in this way, then the weight function k matters only through $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a})$ and $\text{sd}(\hat{L}_{k,a})$, and the criterion is increasing in both quantities, as stated in Section 2.3.

For constructing optimal estimators and CIs, observe that our setting is a fixed design regression model with normal errors and known variance, with the parameter of interest given by a linear functional of the regression function. Therefore, our setting falls into the framework of Donoho (1994) and Armstrong and Kolesár (2017), and we can specialize the general efficiency bounds and the construction of optimal affine estimators and CIs in those papers to the current setting.[6] To state these results, define the (single-class) modulus of continuity of L (see p. 244 in Donoho, 1994, and Section 3.2 in Armstrong and Kolesár, 2017)

$$\omega(\delta) = \sup_{f,g \in \mathcal{F}} \left\{ Lg - Lf \colon \sum_{i=1}^{n} \frac{(f(x_i, d_i) - g(x_i, d_i))^2}{\sigma^2(x_i, d_i)} \leq \delta^2 \right\}, \tag{19}$$

[6] In particular, in the notation of Armstrong and Kolesár (2017), $Y = (y_1/\sigma(x_1, d_1), \ldots, y_n/\sigma(x_n, d_n))$, $\mathcal{Y} = \mathbb{R}^n$, and $Kf = (f(x_1, d_1), \sigma(x_1, d_1), \ldots, f(x_n, d_n)/\sigma(x_n, d_n))$. Donoho (1994) denotes the outcome vector Y by $\mathbf{y}$, and uses $\mathbf{x}$ and $\mathbf{X}$ in place of f and $\mathcal{F}$.

and let f^*_δ and g^*_δ a pair of functions that attain the supremum (assuming the supremum is attained). When $\mathcal{F}$ is centrosymmetric, then $f^*_\delta = -g^*_\delta$, and the modulus problem reduces to the optimization problem (8) in the main text (in the main text, the notation f^*_δ is used for the function denoted g^*_δ in this appendix). Let $\omega'(\delta)$ denote an (arbitrary) element of the superdifferential at δ (the superdifferential is non-empty since the modulus can be shown to be concave). Typically, $\omega(\cdot)$ is differentiable, and $\omega'(\delta)$ corresponds uniquely to the derivative at δ. Define $\hat{L}_\delta = \hat{L}_{k^*_\delta, a^*_\delta}$, where

$$k^*_\delta(x_i, d_i) = \frac{\omega'(\delta)}{\delta} \frac{g^*(x_i, d_i) - f^*(x_i, d_i)}{\sigma^2(x_i, d_i)},$$

and

$$a^*_\delta = \frac{1}{2}\left[L(f^*_\delta + g^*_\delta) - \sum_{i=1}^n k^*_\delta(x_i, d_i)(f^*_\delta(x_i, d_i) + g^*_\delta(x_i, d_i)) \right].$$

If the class $\mathcal{F}$ is translation invariant in the sense that $f \in \mathcal{F}$ implies $f + \iota_\alpha \in \mathcal{F}$[7], then by Lemma D.1 in Armstrong and Kolesár (2017), the modulus is differentiable, with $\omega'(\delta)/\delta = 1/\sum_{i=1}^n d_i(g^*_\delta(x_i, d_i) - f^*_\delta(x_i, d_i))/\sigma^2(x_i, d_i)$. The formula for $\hat{L}_\delta$ in the main text follows from this result combined with fact that, under centrosymmetry, $f^*_\delta = -g^*_\delta$. By Lemma A.1 in Armstrong and Kolesár (2017), the maximum and minimum bias of $\hat{L}_\delta$ is attained at g^*_δ and f^*_δ, respectively, which yields

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) = -\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) = \frac{1}{2}(\omega(\delta) - \delta\omega'(\delta)).$$

Note that $\text{sd}(\hat{L}_\delta) = \omega'(\delta)$.

Corollary 3.1 in Armstrong and Kolesár (2017), and the results in Donoho (1994) then yield the following result:

Theorem A.1. *Let $\mathcal{F}$ be convex, and fix $\alpha > 0$. (i) Suppose that f^*_δ and g^*_δ attain the supremum in* (19) *with $\sum_{i=1}^n \frac{(f(x_i,d_i)-g(x_i,d_i))^2}{\sigma^2(x_i,d_i)} = \delta^2$, and let $\hat{c}^*_\delta = \hat{L}_\delta - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) - z_{1-\alpha}\,\text{sd}(\hat{L}_\delta)$. Then $[\hat{c}^*_\delta, \infty)$ is a $1-\alpha$ CI over $\mathcal{F}$, and it minimaxes the βth quantile of excess length among all $1-\alpha$ CIs for Lf, where $\beta = \Phi(\delta - z_{1-\alpha})$, and Φ denotes the standard normal cdf. (ii) Let δ_χ be the minimizer of $\text{cv}_\alpha\left(\omega(\delta)/2\omega'(\delta) - \delta/2\right)\omega'(\delta)$ over δ, and suppose that $f^*_{\delta_\chi}$ and $g^*_{\delta_\chi}$ attain the supremum in* (19) *at $\delta = \delta_\chi$. Then the shortest $1-\alpha$ FLCI among all FLCIs*

[7]In the main text, we assume that $\{\iota_\alpha\}_{\alpha\in\mathbb{R}} \subset \mathcal{F}$. By convexity, for any $\lambda < 1$, $\lambda f + (1-\lambda)\iota_\alpha = \lambda f + \iota_{(1-\lambda)\alpha} \in \mathcal{F}$, which implies that for all $\lambda < 1$ and $\alpha \in \mathbb{R}$, $\lambda f + \iota_\alpha \in \mathcal{F}$. This, under the assumption in footnote 3, implies translation invariance.

centered at affine estimators is given by

$$\left\{\hat{L}_{\delta_\chi} \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\delta_\chi} / \mathrm{sd}(\hat{L}_{\delta_\chi}))\,\mathrm{sd}(\hat{L}_{\delta_\chi})\right\}.$$

(iii) Let δ_{RMSE} minimize $\frac{1}{4}(\omega(\delta) - \delta\omega'(\delta))^2 + \omega'(\delta)^2$ over δ, and suppose that $f^*_{\delta_\chi}$ and $g^*_{\delta_\chi}$ attain the supremum in (19) at $\delta = \delta_{RMSE}$. Then the estimator $\hat{L}_{\delta_{RMSE}}$ minimaxes RMSE among all affine estimators.

The theorem shows that a one-sided CI based on $\hat{L}_\delta$ is minimax optimal for β-quantile of excess length if $\delta = z_\beta + z_{1-\alpha}$. Therefore, restricting attention to affine estimators does not result in any loss of efficiency if the criterion is $q_\beta(\cdot, \mathcal{F})$.

If the criterion is RMSE Theorem A.1 only gives minimax optimality in the class of affine estimators. However, Donoho (1994) shows that one cannot substantially reduce the maximum risk by considering non-linear estimators. To state the result, let $\rho_A(\tau) = \tau/\sqrt{1+\tau}$ denote the minimax RMSE among affine estimators of θ in the bounded normal mean model in which we observe a single draw from the $N(\theta, 1)$ distribution, and $\theta \in [-\tau, \tau]$, and let $\rho_N(\tau)$ denote the minimax RMSE among all estimators (affine or non-linear). Donoho et al. (1990) give bounds on $\rho_N(\tau)$, and show that $\sup_{\tau>0} \rho_A(\tau)/\rho_N(\tau) \leq \sqrt{5/4}$, which is known as the Ibragimov-Hasminskii constant.

Theorem A.2 (Donoho, 1994)**.** *Let $\mathcal{F}$ be convex. The minimax RMSE among affine estimators risk equals $R^*_{RMSE,A}(\mathcal{F}) = \sup_{\delta>0} \frac{\omega(\delta)}{\delta}\rho_A(\delta/2)$. The minimax RMSE among all estimators is bounded below by $\sup_{\delta>0} \frac{\omega(\delta)}{\delta}\rho_N(\delta/2) \geq \sqrt{4/5}\sup_{\delta>0} \frac{\omega(\delta)}{\delta}\rho_A(\delta/2) = \sqrt{4/5}R^*_{RMSE,A}(\mathcal{F})$.*

The theorem shows that the minimax efficiency of $\hat{L}_{\delta_{\mathrm{RMSE}}}$ among all estimators is at least $\sqrt{4/5} = 89.4\%$. In particular applications, the efficiency can be shown to be even higher by lower bounding $\sup_{\delta>0} \frac{\omega(\delta)}{\delta}\rho_N(\delta/2)$ directly, rather than using the Ibragimov-Hasminskii constant. The arguments in Donoho (1994) also imply $R^*_{\mathrm{RMSE},A}(\mathcal{F})$ can be equivalently computed as $R^*_{\mathrm{RMSE},A}(\mathcal{F}) = \inf_{\delta>0} \frac{1}{2}\sqrt{(\omega(\delta) - \delta\omega'(\delta))^2 + \omega'(\delta)^2} = \inf_{\delta>0}\sup_{f\in\mathcal{F}}(E(\hat{L}_\delta - Lf)^2)^{1/2}$, as implied by Theorem A.1.

The one-dimensional subfamily argument used in Donoho (1994) to derive Theorem A.2 could also be used to obtain the minimax efficiency of the fixed-length CI based on $\hat{L}_{\delta_\chi}$ among all CIs when the criterion is expected length. However, when the parameter space $\mathcal{F}$ is centrosymmetric, we can obtain a stronger result that gives sharp bounds for the scope of adaptation to smooth functions:

Theorem A.3. *Let $\mathcal{F}$ be convex and centrosymmetric, and fix $g \in \mathcal{F}$ such that $f - g \in \mathcal{F}$ for all $f \in \mathcal{F}$. (i) Suppose $-f_\delta^*$ and f_δ^* attain the supremum in (19) with $\sum_{i=1}^n \frac{(f(x_i,d_i)-g(x_i,d_i))^2}{\sigma^2(x_i,d_i)} = \delta^2$, with $\delta = z_\beta + z_{1-\alpha}$, and define $\hat{c}_\delta^*$ as in Theorem A.1. Then the efficiency of $\hat{c}_\delta^*$ under the criterion $q_\beta(\cdot, \{g\})$ is given by*

$$\frac{\inf_{\{\hat{c}\colon [\hat{c},\infty) \text{ satisfies (4)}\}} q_\beta(\hat{c}, \{g\})}{q_\beta(\hat{c}_\delta^*, \{g\})} = \frac{\omega(2\delta)}{\omega(\delta) + \delta\omega'(\delta)} \geq \frac{1}{2}.$$

(ii) Suppose the minimizer f_{L_0} of $\sum_{i=1}^n \frac{(f(x_i,d_i)-g(x_i,d_i))^2}{\sigma^2(x_i,d_i)}$ subject to $Lf = L_0$ and $f \in \mathcal{F}$ exists for all $L_0 \in \mathbb{R}$. Then the efficiency of the fixed-length CI around $\hat{L}_{\delta_\chi}$ at g relative to all confidence sets is

$$\begin{aligned}\frac{\inf_{\{\mathcal{C}\colon \mathcal{C} \text{ satisfies (4)}\}} E_g\lambda(\mathcal{C})}{\inf_{\delta>0} 2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2}\right)\omega'(\delta)} &= \frac{(1-\alpha)E\left[\omega(2(z_{1-\alpha} - Z)) \mid Z \leq z_{1-\alpha}\right]}{2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta_\chi)}{2\omega'(\delta_\chi)} - \frac{\delta_\chi}{2}\right)\cdot\omega'(\delta_\chi)} \\ &\geq \frac{((1-\alpha)z_{1-\alpha} + \phi(z_{1-\alpha}) - \phi(0))}{z_{1-\alpha/2}}, \quad (20)\end{aligned}$$

where $\lambda(\mathcal{C})$ denotes the Lebesgue measure of a confidence set $\mathcal{C}$, Z is a standard normal random variable, and $\phi(z)$ denotes standard normal density.

Proof. Both parts of the theorem, except for the lower bound in (20), follow from Corollary 3.2 and Corollary 3.3 in Armstrong and Kolesár (2017). It therefore suffices to prove the lower bound in (20). Since the shortest FLCI is shorter than the FLCI around $\hat{L}_{2z_{1-\alpha/2}}$, and since $\mathrm{cv}_\alpha(b) \leq b + z_{1-\alpha/2}$, the denominator in (20) can be bounded by

$$\begin{aligned}&2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta_\chi)}{2\omega'(\delta_\chi)} - \frac{\delta_\chi}{2}\right)\cdot\omega'(\delta_\chi) \\ &\qquad \leq 2\,\mathrm{cv}_\alpha\left(\frac{\omega(2z_{1-\alpha/2})}{2\omega'(2z_{1-\alpha/2})} - z_{1-\alpha/2}\right)\cdot\omega'(2z_{1-\alpha/2}) \leq \omega(2z_{1-\alpha/2}) \leq \frac{z_{1-\alpha/2}}{z_{1-\alpha}}\omega(2z_{1-\alpha}).\end{aligned}$$

On the other hand, using the fact that the modulus is concave and non-decreasing, the

numerator in (20) can be lower-bounded by

$$\begin{aligned}
&(1-\alpha)E\left[\omega(2(z_{1-\alpha}-Z)) \mid Z \leq z_{1-\alpha}\right] \\
&\quad = E\left[\omega(2(z_{1-\alpha}-Z))\mathbb{I}\{0 \leq Z \leq z_{1-\alpha}\}\right] + E\left[\omega(2(z_{1-\alpha}-Z))\mathbb{I}\{Z \leq 0\}\right] \\
&\quad \geq \omega\left(2z_{1-\alpha}\right) E\left[\frac{z_{1-\alpha}-Z}{z_{1-\alpha}}\mathbb{I}\{0 \leq Z \leq z_{1-\alpha}\}\right] + \omega(2z_{1-\alpha})P(Z \leq 0) \\
&\quad = \omega\left(2z_{1-\alpha}\right)\left(1/2-\alpha+\frac{\phi(z_{1-\alpha})-\phi(0)}{z_{1-\alpha}}+1/2\right).
\end{aligned}$$

Taking the ratio of the bounds in the two preceding displays then yields the result. □

The theorem gives sharp efficiency bounds for one-sided CIs as well as fixed-length CIs relative to CIs that direct all power at a particular function g. The condition on g is satisfied if g is smooth enough relative to $\mathcal{F}$. For example, if $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, it holds if g is piecewise constant, $g(x,d) = \alpha_0 + d\alpha_1$ for some $\alpha_0, \alpha_1 \in \mathbb{R}$. The theorem also gives lower bounds for these efficiencies—for one-sided CIs, the theorem implies that the β-quantile excess of length of the CI $[\hat{c}^*_\delta, \infty)$ at g cannot be reduced by more than 50%. For 95% fixed-length CIs, the efficiency lower bound in (20) evaluates to 64.6%. In a particular application, sharp lower bounds can be computed directly by computing the modulus; typically this gives much higher efficiencies—for example in the baseline specification in the empirical application in Section 4, the efficiency of the shortest FLCI is over 99% at piecewise constant functions.

A.2 Optimal estimators and CIs under Lipschitz smoothness

We now specialize the results from Appendix A.1 to the case with Lipschitz smoothness, $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, as well as versions of these classes that impose monotonicity conditions.

To that end, let $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$ denote the set of functions $f : \{x_1, \ldots, x_n\} \times \{0,1\} \to \mathbb{R}$ such that $|f(x,d) - f(\tilde{x},d)| \leq C\|x - \tilde{x}\|_{\mathcal{X}}$ for all $x, \tilde{x} \in \{x_1, \ldots, x_n\}$ and each $d \in \{0,1\}$. That is, $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$ denotes the class of functions with domain $\{x_1, \ldots, x_n\} \times \{0,1\}$ that satisfy the Lipschitz condition on this domain. If we take the restriction of any function $f \in \mathcal{F}_{\text{Lip}}(C)$ to the domain $\{x_1, \ldots, x_n\} \times \{0,1\}$, then the resulting function will clearly be in $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$. The following result, from Beliakov (2006), shows that, given a function in $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$, one can always interpolate the points $x_1, \ldots, x_n$ to obtain a function in $\mathcal{F}_{\text{Lip}}(C)$.

Lemma A.1. *(Beliakov, 2006, Theorem 4) For any function $f : \{x_1, \ldots, x_n\} \times \{0,1\} \to \mathbb{R}$, we have $f \in \widetilde{\mathcal{F}}_{Lip,n}(C)$ iff. there exists a function $h \in \mathcal{F}_{Lip}(C)$ such that $f(x,d) = h(x,d)$ for all $(x,d) \in \{x_1, \ldots, x_n\} \times \{0,1\}$.*

We also consider the case where monotonicity restrictions are imposed in addition to the Lipschitz restriction. Let $\mathcal{S} \subseteq \{1, \ldots, p\}$ denote the subset of indices of x_i for which monotonicity is imposed, and normalize the variables so that the monotonicity condition states that $f(\cdot, d)$ is nondecreasing in each of these variables (by taking the negative of variables for which $f(\cdot, d)$ is non-increasing). Let $\mathcal{F}_{\text{Lip},\mathcal{S}\uparrow}(C)$ denote the set of functions in $\mathcal{F}_{\text{Lip}}(C)$ such that $f(\cdot, 0)$ and $f(\cdot, 1)$ are monotone for the indices in $\mathcal{S}$: for any $t, \tilde{t}$ with $t_j \geq \tilde{t}_j$ for $j \in \mathcal{S}$ and $t_j = \tilde{t}_j$ for $j \notin \mathcal{S}$, we have $f(t, d) \geq f(\tilde{t}, d)$ for each $d \in \{0, 1\}$ (that is, increasing the elements in $\mathcal{S}$ and holding others fixed weakly increases the function).

We use a result on necessary and sufficient conditions for interpolation by monotonic Lipschitz functions given by Beliakov (2005). For a vector $t \in \mathbb{R}^p$, let $(t)_{\mathcal{S}+}$ denote the vector with jth element t_j for $j \notin \mathcal{S}$ and jth element $\max\{t_j, 0\}$ for $j \in \mathcal{S}$. Let $\widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S}\uparrow,n}(C)$ denote the set of functions $f : \{x_1, \ldots, x_n\} \times \{0, 1\} \to \mathbb{R}$ such that, for all $i, j \in \{1, \ldots, n\}$ and $d \in \{0, 1\}$

$$f(x_i, d) - f(x_j, d) \leq C\|(x_i - x_j)_{\mathcal{S}+}\|_{\mathcal{X}}.$$

Lemma A.2. *(Beliakov, 2005, Proposition 4.1) For any function $f : \{x_1, \ldots, x_n\} \times \{0, 1\} \to \mathbb{R}$, we have $f \in \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C)$ iff. there exists a function $h \in \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C)$ such that $f(x, d) = h(x, d)$ for all $(x, d) \in \{x_1, \ldots, x_n\} \times \{0, 1\}$.*

Using these results and the fact that $\mathcal{F}_{\text{Lip}}(C)$ is centrosymmetric, we can phrase the problem of computing the modulus, optimal weights and worst-case biases as a finite dimensional convex optimization problem.

Lemma A.3. *The modulus of continuity $\omega(\delta)$ with $\mathcal{F} = \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C)$ is given by the value of (19) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C)$. Furthermore, the functions $f^*_\delta, g^*_\delta \in \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C)$ are solutions to the modulus problem (19) with $\mathcal{F} = \mathcal{F}_{Lip,\mathcal{S}\uparrow,n}(C)$ if and only if there exist $\tilde{f}^*_\delta$ and $\tilde{g}^*_\delta$ that maximize (19) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C)$ such that $\tilde{f}^*_\delta(x, d) = f^*_\delta(x, d)$ and $\tilde{g}^*_\delta(x, d) = g^*_\delta(x, d)$ for $(x, d) \in \{x_1, \ldots, x_n\} \times \{0, 1\}$. In particular, the corresponding estimator and CIs can be computed using $\tilde{f}^*_\delta$ and $\tilde{g}^*_\delta$ in place of f^*_δ and g^*_δ.*

Similarly, the modulus of continuity $\omega(\delta)$ with $\mathcal{F} = \mathcal{F}_{Lip}(C)$ is given by the value of (8) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,n}(C)$. The function $f^*_\delta \in \mathcal{F}_{Lip}(C)$ is a solution to the modulus problem (8) with $\mathcal{F} = \mathcal{F}_{Lip}(C)$ if and only if there exists $\tilde{f}^*_\delta$ that maximizes (8) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C)$ such that $\tilde{f}^*_\delta(x, d) = f^*_\delta(x, d)$ for $(x, d) \in \{x_1, \ldots, x_n\} \times \{0, 1\}$. In particular, the corresponding estimator and CIs can be computed using $\tilde{f}^*$ and $\tilde{g}^*_\delta = -\tilde{f}^*_\delta$ in place of f^*_δ and g^*_δ.

Theorem 2.2 now follows immediately from Lemma A.3 and Theorem A.1.

A.3 Proof of Theorem 2.3

To prove Theorem 2.3, we first provide another characterization of the optimal weights given in (14). Given $\{m_i\}_{i=1}^n$, consider the optimization problem (13) with the additional constraint that $f(x_i, d_i) = m_i$ for $d_i = 1$ and $f(x_i, d_i) = -m_i$ for $d_i = 0$. It follows from Beliakov (2006) that there exists a function $f \in \mathcal{F}_{\text{Lip}}(C)$ satisfying these constraints iff. $|m_i - m_j| \leq C\|x_i - x_j\|_{\mathcal{X}}$ for all i, j with $d_i = d_j$. Furthermore, when this condition holds, $f(x, 1)$ is maximized simultaneously for all x subject to the constraint that $f(x_i, d_i) = m_i$ for all i by taking $f(x, 1) = \min_{i:d_i=1}(m_i + C\|x - x_i\|_{\mathcal{X}})$. Similarly, $f(x, 0)$ is minimized simultaneously for all x by taking $f(x, 0) = -\min_{i:d_i=0}(m_i + C\|x - x_i\|_{\mathcal{X}})$ (see Beliakov, 2006, p. 25). Plugging this into (13), it follows that $f_\delta^*(x_i, d_i) = (2d_i - 1) \cdot m_i^*$ where $\{m_i^*\}_{i=1}^n$ solves

$$\max_m 2\sum_i w_i(m_i + \tilde{\omega}_i(m)) \quad \text{s.t.}$$

$$\sum_{i=1}^n m_i^2/\sigma^2(x_i, d_i) \leq \delta^2/4, \tag{21}$$

$$|m_i - m_j| \leq C\|x_i - x_j\|_{\mathcal{X}} \text{ for all } i, j \text{ with } d_i = d_j, \tag{22}$$

where

$$\tilde{\omega}_i(m) = \min_{j:d_j \neq d_i}(m_j + C\|x_i - x_j\|_{\mathcal{X}}). \tag{23}$$

This is a convex optimization problem and constraint qualification holds since $m = 0$ satisfies Slater's condition (see Boyd and Vandenberghe, 2004, p. 226). Thus, the solution (or set of solutions) is the same as the solution to the Lagrangian.

To characterize the solution, let $\mathcal{J}_i(m)$ denote the set of indices that achieve the minimum in (23). Note that $\mathcal{J}_i(0)$ is the set of the nearest neighbors to i (i.e. the set of indices j of observations such that $\|x_j - x_i\|_{\mathcal{X}}$ is minimized). Furthermore, if $\|m\|$ is smaller than some constant that depends only on the design points, we will have

$$\mathcal{J}_i(m) = \{j \in \mathcal{J}_i(0) : m_j \leq m_\ell \text{ all } \ell \in \mathcal{J}_i(0)\}. \tag{24}$$

The superdifferential $\partial\tilde{\omega}_i(m)$ of $\tilde{\omega}_i(m)$ is given by the convex hull of $\cup_{j\in\mathcal{J}_i(m)}\{e_j\}$. For δ/C small enough, if the values of x_i and x_j for $d_i = d_j$ are distinct (which is implied by the assumption that each observation has a unique closest match), the constraints (22) implied by the constraint (21). Thus, specializing to the case with $w_i = 1/n$, the first order

conditions are given by

$$\iota - \lambda n \Sigma^{-1} m \in -\sum_{i=1}^{n} \partial \tilde{\omega}_i(m)$$
$$= \left\{ \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} e_j \colon b_{ij} = 0 \text{ all } j \notin \mathcal{J}_i(m),\ b_{ij} \geq 0,\ \text{all } i, j \text{ and } \sum_{j=1}^{n} b_{ij} = 1 \text{ all } i \right\}.$$

where λ is the Lagrange multiplier on (21), ι is a vector of ones, and Σ is a diagonal matrix with (i,i) element given by $\sigma(x_i, d_i)^2$. Let $\|m\|$ be small enough so that (24) holds, and suppose that each observation has a unique closest match. Then $\mathcal{J}_i(m) = \mathcal{J}_i(0)$ for small enough m and $\mathcal{J}_i(0)$ is a singleton for each i, so that m_j^* is proportional to $\sigma^2(x_i, d_i)(1 + \#\{i : j \in \mathcal{J}_i(m)\}) = \sigma^2(x_i, d_i)(1 + K_1(i))$, so that by (14), the optimal weights are given by

$$k_\delta^*(x_i, d_i) = \frac{(2d_i - 1)(1 + K_1(i))}{\sum_i d_i (2d_i - 1)(1 + K_1(i))} = \frac{(2d_i - 1)(1 + K_1(i))}{n},$$

where the second equality follows from $\sum_i \sum_i d_i (2d_i - 1)(1 + K_1(i)) = \sum_i d_i (1 + K_1(i)) = \sum_i d_i + \sum_i (1 - d_i) = n$. It then follows from (15) that the optimal estimator coincides with the matching estimator based on a single match.

Appendix B Asymptotic efficiency bound

This section proves Theorem 3.1. The fact that X_i has a bounded density conditional on D_i means that there exists some $a < b$ such that X_i has a density bounded away from zero and infinity on $[a,b]^p$ conditional on $D_i = 1$. Let $\mathcal{N}_{d,n} = \{i \colon D_i = d, i \in \{1, \ldots, n\}\}$ and let

$$\mathcal{I}_n(h) = \{i \in \mathcal{N}_{1,n} \colon X_i \in [a,b]^p \text{ and for all } j \in \mathcal{N}_{0,n}, \|X_i - X_j\|_{\mathcal{X}} > 2h\}.$$

Let $\mathcal{E}$ denote the σ-algebra generated by $\{D_i\}_{i=1}^{\infty}$ and $\{X_i \colon D_i = 0, i \in \mathbb{N}\}$. Note that, conditional on $\mathcal{E}$, the observations $\{X_i \colon i \in \mathcal{N}_{1,n}\}$ are i.i.d. with density bounded away from zero and infinity on $[a,b]^p$.

Lemma B.1. *There exists $\eta > 0$ such that, if $\limsup_n h_n n^{1/p} \leq \eta$, then almost surely, $\liminf_n \#\mathcal{I}_n(h_n)/n \geq \eta$.*

Proof. Let $A_n = \{x \in [a,b]^p | \text{there exists } j \text{ such that } D_j = 0 \text{ and } \|x - X_j\|_{\mathcal{X}} \leq 2h\}$. Then $\#\mathcal{I}_n(h) = \sum_{i \in \mathcal{N}_{1,n}} [\mathbb{I}\{X_i \in [a,b]^p\} - \mathbb{I}\{X_i \in A_n\}]$. Note that, conditional on $\mathcal{E}$, the random

variables $\mathbb{I}\{X_i \in A_n\}$ with $i \in \mathcal{N}_{1,n}$ are i.i.d. Bernoulli(ν_n) with $\nu_n = P(X_i \in A_n|\mathcal{E}) = \int \mathbb{I}\{x \in A_n\} f_{X|D}(x|1)\,dx \leq K\lambda(A_n)$ where $f_{X|D}(x|1)$ is the conditional density of X_i given $D_i = 1$, λ is the Lebesgue measure and K is an upper bound on this density. Under the assumption that $\limsup_n h_n n^{1/p} \leq \eta$, we have $\lambda(A_n) \leq (4h_n)^p n \leq 8^p \eta^p$ where the last inequality holds for large enough n. Thus, letting $\overline{\nu} = 8^p \eta^p K$, we can construct random variables Z_i for each $i \in \mathcal{N}_{1,n}$ that are i.i.d. Bernoulli($\overline{\nu}$) conditional on $\mathcal{E}$ such that $\mathbb{I}\{X_i \in A_n\} \leq Z_i$. Applying the strong law of large numbers, it follows that

$$\begin{aligned}\liminf_n \#\mathcal{I}_n(h)/n &\geq \liminf_n \frac{\#\mathcal{N}_{1,n}}{n} \frac{1}{\#\mathcal{N}_{1,n}} \sum_{i \in \mathcal{N}_{1,n}} (\mathbb{I}\{X_i \in [a,b]^p\} - Z_i) \\ &\geq P(D_i = 1)(P(X_i \in [a,b]^p | D_i = 1) - 8^p \eta^p K)\end{aligned}$$

almost surely. This will be greater than η for η small enough. □

Let $\tilde{\mathcal{X}}_n(h,\eta)$ be the set of elements $\tilde{x}$ in the grid

$$\{a + jh\eta \colon j = (j_1, \ldots, j_p) \in \{1, \ldots, \lfloor h^{-1} \rfloor (b-a)\}^p\}$$

such that there exists $i \in \mathcal{I}_n(h)$ with $\max_{1 \leq k \leq p} |\tilde{x}_k - X_{i,k}| \leq h\eta$. Note that, for any $\tilde{x} \in \tilde{\mathcal{X}}_n(h,\eta)$, the closest element X_i with $i \in \mathcal{I}_n(h)$ satisfies $\|\tilde{x} - X_i\|_{\mathcal{X}} \leq ph\eta$. Thus, for any X_j with $D_j = 0$, we have

$$\|\tilde{x} - X_j\|_{\mathcal{X}} \geq \|X_j - X_i\|_{\mathcal{X}} - \|\tilde{x} - X_i\|_{\mathcal{X}} \geq 2h - p\eta h > h$$

for η small enough, where the first inequality follows from rearranging the triangle inequality. Let $k \in \Sigma(1,\gamma)$ be a nonnegative function with support contained in $\{x \colon \|x\|_{\mathcal{X}} \leq 1\}$, with $k(x) \geq \underline{k}$ on $\{x \colon \max_{1 \leq k \leq p} |x_k| \leq \eta\}$ for some $\underline{k} > 0$. By the above display, the function

$$f_n(x,d) = f_{n,\{X_i,D_i\}_{i=1}^n}(x,d) = \sum_{\tilde{x} \in \tilde{\mathcal{X}}_n(h,\eta)} (1-d)k((x-\tilde{x})/h)$$

is equal to zero for $(x,d) = (X_i, D_i)$ for all $i = 1, \ldots, n$. Thus, it is observationally equivalent to the zero function conditional on $\{X_i, D_i\}_{i=1}^n$: $P_{f_{n,\{X_i,D_i\}_{i=1}^n}}(\cdot|\{X_i,D_i\}_{i=1}^n) = P_0(\cdot|\{X_i,D_i\}_{i=1}^n)$. Furthermore, we have

$$\frac{1}{n}\sum_{i=1}^n [f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,1) - f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,0)]$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\sum_{\tilde{x}\in\tilde{\mathcal{X}}_n(h,\eta)} k((X_i-\tilde{x})/h) \leq -\underline{k}\frac{\#\mathcal{I}_n(h)}{n} \tag{25}$$

where the last step follows since, for each $i \in \mathcal{I}_n(h)$, there is a $\tilde{x} \in \tilde{\mathcal{X}}_n(h,\eta)$ such that $\max_{1\leq k\leq p}|\tilde{x}_k - X_{i,k}|/h \leq \eta$.

Now let us consider the Hölder condition on $f_{n,\{X_i,D_i\}_{i=1}^n}$. Let ℓ be the greatest integer strictly less than γ and let D^r denote the derivative with respect to the multi-index $r = r_1,\ldots,r_p$ for some r with $\sum_{i=1}^p r_i = \ell$. Let $x, x' \in \mathbb{R}^p$. Let $\mathcal{A}(x,x') \subseteq \tilde{\mathcal{X}}_n(h,\eta)$ denote the set of $\tilde{x} \in \tilde{\mathcal{X}}_n(h,\eta)$ such that $\max\{k((x-\tilde{x})/h), k((x'-\tilde{x})/h)\} > 0$. By the support conditions on k, there exists a constant K depending only on p such that $\#\mathcal{A}(x,x') \leq K/\eta^p$. Thus,

$$\begin{aligned}
&\left|D^r f_{n,\{X_i,D_i\}_{i=1}^n}(x,d) - D^r f_{n,\{X_i,D_i\}_{i=1}^n}(x',d)\right| \\
&\leq h^{-\ell}(K/\eta^p)\sup_{\tilde{x}\in\mathcal{A}(x,x')}|D^r k((x-\tilde{x})/h) - D^r k((x'-\tilde{x})/h)| \\
&\leq h^{-\ell}(K/\eta^p)\|(x-x')/h\|_{\mathcal{X}}^{\gamma-\ell} = h^{-\gamma}(K/\eta^p)\|x-x'\|_{\mathcal{X}}^{\gamma},
\end{aligned}$$

which implies that $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n} \in \Sigma(C,\gamma)$ where $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}(x,d) = \frac{h^\gamma C}{K/\eta^p} f_{n,\{X_i,D_i\}_{i=1}^n}(x,d)$. By (25), the CATE under $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}$ is bounded from above by $-\underline{k}\frac{h^\gamma C}{K/\eta^p}\frac{\#\mathcal{I}_n(h)}{n}$, which, by Lemma B.1, is bounded from above by a constant times h_n^γ for large enough n on a probability one event for h_n a small enough multiple of $n^{-1/p}$. Thus, there exists $\varepsilon > 0$ such that the CATE under $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}$ is bounded from above by $-\varepsilon n^{-1/p}$ for large enough n with probability one. On this probability one event,

$$\begin{aligned}
&\liminf_n P_0\left(\hat{c}_n \leq -\varepsilon n^{-\gamma}|\{X_i,D_i\}_{i=1}^n\right) = \liminf_n P_{\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}}\left(\hat{c}_n \leq \varepsilon n^{-\gamma}|\{X_i,D_i\}_{i=1}^n\right) \\
&\geq \liminf_n \inf_{f(\cdot,0),f(\cdot,1)\in\Sigma(C,\gamma)} P_f\left(\frac{1}{n}\sum_{i=1}^n [f(X_i,1)-f(X_i,0)] \in [\hat{c}_n,\infty)\middle|\{X_i,D_i\}_{i=1}^n\right) \geq 1-\alpha,
\end{aligned}$$

which gives the result.

Appendix C Asymptotic validity with unknown error distribution

This section proves Theorem 3.2. To prove this theorem, we verify the high level conditions (S13) and (S14) in Armstrong and Kolesár (2017). For the central limit theorem condition

(S13), it suffices to verify that the weights $\tilde{k}^*_\delta(x_i, d_i)$ satisfy

$$\frac{\max_{1\le i\le n} \tilde{k}^*_\delta(x_i, d_i)^2}{\sum_{i=1}^n \tilde{k}^*_\delta(x_i, d_i)^2} \to 0, \tag{26}$$

since this implies the Lindeberg condition under the moment bounds on u_i. To this end, we follow arguments similar to the proof of Theorem F.3 in Armstrong and Kolesár (2017), with the additional complication that the parameter space $\mathcal{F}_{\text{Lip}}(C_n)$ changes with n.

By boundedness of $\tilde{\sigma}(x_i, d_i)$ away from zero and infinity, (26) is equivalent to showing that

$$\frac{\max_{1\le i\le n} \tilde{f}^*_\delta(x_i, d_i)^2}{\sum_{i=1}^n \tilde{f}^*_\delta(x_i, d_i)^2} \to 0$$

where $\tilde{f}^*_\delta$ is the solution to the optimization problem defined by (8) and (9) with $\tilde{\sigma}(x, d)$ in place of $\sigma(x, d)$. Since the constraint on $\sum_{i=1}^n \frac{\tilde{f}^*_\delta(x_i,d_i)^2}{\tilde{\sigma}^2(x_i,d_i)}$ in (8) binds, the denominator is bounded from above and below by constants that depend only on δ and the upper and lower bounds on $\tilde{\sigma}^2(x_i, d_i)$. Thus, it suffices to show that

$$\max_{1\le i\le n} \tilde{f}^*_\delta(x_i, d_i)^2 \to 0.$$

To get a contradiction, suppose that there exists $\eta > 0$ and a sequence i^*_n such that $\tilde{f}^*_\delta(x_{i^*_n}, d_{i^*_n})^2 > \eta^2$ infinitely often. Then, by the Lipschitz condition, $|\tilde{f}^*_\delta(x, d_{i^*_n})| \ge \eta - C_n\|x - x_{i^*_n}\|$ so that, for $\|x - x_{i^*_n}\| \le \eta/(2C_n)$, we have $|\tilde{f}^*_\delta(x, d_{i^*_n})| \ge \eta/2$. Thus, we have

$$\sum_{i=1}^n \tilde{f}^*_\delta(x_i, d_i)^2 \ge \sum_{i:d_i=d_{i^*_n}} \tilde{f}^*_\delta(x_i, d_i)^2 \ge (\eta/2)^2 \#\{i : \|x_i - x_{i^*_n}\| \le \eta/(2C_n), d_i = d_{i^*_n}\}$$

infinitely often. This gives a contradiction so long as (17) holds.

Now consider the variance estimator $\sum_{i=1}^n (y_i - \hat{f}(x_i, d_i))^2 k(x_i, d_i)^2$ based on an estimate $\hat{f}(x_i, d_i)$. To show that this converges to one when divided by the true variance (condition (S14) in Armstrong and Kolesár, 2017), we need to show that $\sum_{i=1}^n (y_i - \hat{f}(x_i, d_i))^2 a_{n,i} - \sum_{i=1}^n \sigma^2(x_i, n_i) a_{n,i}$ converges to zero uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$, where

$$a_{n,i} = k(x_i, d_i)^2 / \sum_{j=1}^n [\sigma^2(x_j, d_j) k(x_j, d_j)^2].$$

By an inequality of von Bahr and Esseen (1965),

$$E\left|\sum_{i=1}^{n} u_i^2 a_{n,i} - \sum_{i=1}^{n} \sigma^2(x_i, d_i) a_{n,i}\right|^{1+1/(2K)}$$

is bounded by a constant times

$$\sum_{i=1}^{n} a_{n,i}^{1+1/(2K)} E|u_i^2 - \sigma^2(x_i, d_i)|^{1+1/(2K)} \leq \left[\max_{1\leq i\leq n} a_{n,i}^{1/(2K)} E|u_i^2 - \sigma^2(x_i, d_i)|^{1+1/(2K)}\right] \sum_{i=1}^{n} a_{n,i}$$

Note that, by boundedness of $\sigma(x, d)$ away from zero and infinity, $\sum_{i=1}^{n} a_{n,i}$ is uniformly bounded. Furthermore, it follows from (26), that $\max_{1\leq i\leq n} a_{n,i} \to 0$. From this and the moment bounds on u_i, it follows that the above display converges to zero.

It therefore suffices to bound

$$\begin{aligned}
&\sum_{i=1}^{n} (y_i - \hat{f}(x_i, d_i))^2 a_{n,i} - \sum_{i=1}^{n} (y_i - f(x_i, d_i))^2 a_{n,i} \\
&= \sum_{i=1}^{n} (2y_i - \hat{f}(x_i, d_i) - f(x_i, d_i))(f(x_i, d_i) - \hat{f}(x_i, d_i)) a_{n,i} \\
&= \sum_{i=1}^{n} (2u_i + f(x_i, d_i) - \hat{f}(x_i, d_i))(f(x_i, d_i) - \hat{f}(x_i, d_i)) a_{n,i}.
\end{aligned}$$

The expectation of the absolute value of this display is bounded by

$$\sum_{i=1}^{n} E_f[(f(x_i, d_i) - \hat{f}(x_i, d_i))^2] a_{n,i} + 2 \sum_{i=1}^{n} E_f[|u_i||f(x_i, d_i) - \hat{f}(x_i, d_i)|] a_{n,i}.$$

The above display is bounded by a constant times

$$\max_{1\leq i\leq n} E_f[(f(x_i, d_i) - \hat{f}(x_i, d_i))^2].$$

If condition (17) holds for all $\eta > 0$, then the same condition also holds with η replaced by a sequence η_n converging to zero. It follows that, under this condition, there exists a bandwidth sequence h_n with $h_n C_n \to 0$ and $\min_{1\leq i\leq n} \#\{j \in \{1, \ldots, n\} : \|x_j - x_i\| \leq h_n,\, d_i = d_j\} \to \infty$. Let

$$\hat{f}(x_i, d_i) = \frac{\sum_{j=1}^{n} y_j \mathbb{I}\{\|x_j - x_i\| \leq h_n,\, d_i = d_j\}}{\#\{j \in \{1, \ldots, n\} : \|x_j - x_i\| \leq h_n,\, d_i = d_j\}}$$

be the Nadaraya-Watson estimator with uniform kernel and bandwidth h_n. Then

$$\sup_{f\in\mathcal{F}_{\text{Lip}}(C_n)} \max_{1\leq i\leq n} E_f[(f(x_i,d_i)-\hat{f}(x_i,d_i))^2]$$
$$\leq \max_{1\leq i\leq n} Eu_i^2/\#\{j\in\{1,\ldots,n\} : \|x_j-x_i\|\leq h_n,\, d_i=d_j\} + (h_nC_n)^2 \to 0.$$

This proves the result.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014a): "Finite Population Causal Standard Errors," Tech. Rep. 20325, National Bureau of Economic Research.

Abadie, A. and G. W. Imbens (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11.

Abadie, A., G. W. Imbens, and F. Zheng (2014b): "Inference for Misspecified Models With Fixed Regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Armstrong, T. B. and M. Kolesár (2017): "Optimal inference in a class of regression models," ArXiv:1511.06028v4.

Beliakov, G. (2005): "Monotonicity Preserving Approximation of Multivariate Scattered Data," *BIT Numerical Mathematics*, 45, 653–677.

——— (2006): "Interpolation of Lipschitz functions," *Journal of Computational and Applied Mathematics*, 196, 20–44.

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

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

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *The Annals of Statistics*, 36, 808–843.

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

Dehejia, R. H. and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053–1062.

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

Donoho, D. L., R. C. Liu, and B. MacGibbon (1990): "Minimax Risk Over Hyperrectangles, and Implications," *The Annals of Statistics*, 18, 1416–1437.

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Heckman, J. J., H. Ichimura, and P. Todd (1998): "Matching As An Econometric Evaluation Estimator," *The Review of Economic Studies*, 65, 261–294.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Khan, S. and E. Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *The American Economic Review*, 76, 604–620.

Low, M. G. (1995): "Bias-Variance Tradeoffs in Functional Estimation Problems," *The Annals of Statistics*, 23, 824–835.

Robins, J., E. T. Tchetgen, L. Li, and A. van der Vaart (2009): "Semiparametric minimax rates," *Electronic Journal of Statistics*, 3, 1305–1321.

Robins, J. M. and Y. Ritov (1997): "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models." *Statistics in medicine*, 16, 285.

Smith, J. A. and P. E. Todd (2001): "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," *The American Economic Review*, 91, 112–118.

——— (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics*, 305–353.

VON BAHR, B. AND C.-G. ESSEEN (1965): "Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$," *The Annals of Mathematical Statistics*, 36, 299–303.

Criterion	δ	M	Estimate	Worst-case bias	Std. error	
					homosk.	robust
Optimal estimator						
one-sided CI	2.49		0.951	1.737	1.553	1.095
FLCI	3.30		0.956	1.846	1.478	1.034
RMSE	1.95		0.950	1.662	1.620	1.142
Matching estimator						
one-sided CI		17	1.315	2.164	1.519	0.936
FLCI		19	1.250	2.236	1.471	0.927
RMSE		2	1.699	1.614	2.059	1.282

Table 1: Results for NSW data, $p = 1$, $A = A_{\text{main}}$, $C = 1$. The tuning parameters δ (for the optimal estimator) and M (the number of matches for the matching estimator) are chosen to optimize a given optimality criterion.

	Age	Educ.	Black	Hispanic	Married	Earnings		Employed	
						1974	1975	1974	1975
$A_{\text{main}}^{1/2}$	0.15	0.60	2.50	2.50	2.50	0.50	0.50	0.10	0.10
$A_{\text{ne}}^{1/2}$	0.10	0.33	2.20	5.49	2.60	0.07	0.07	2.98	2.93

Table 2: Diagonal elements of the weight matrix $A^{1/2}$ in definition of the norm (18) for the main specification, $A_{\text{main}}^{1/2}$, and alternative specification, $A_{\text{ne}}^{1/2}$.

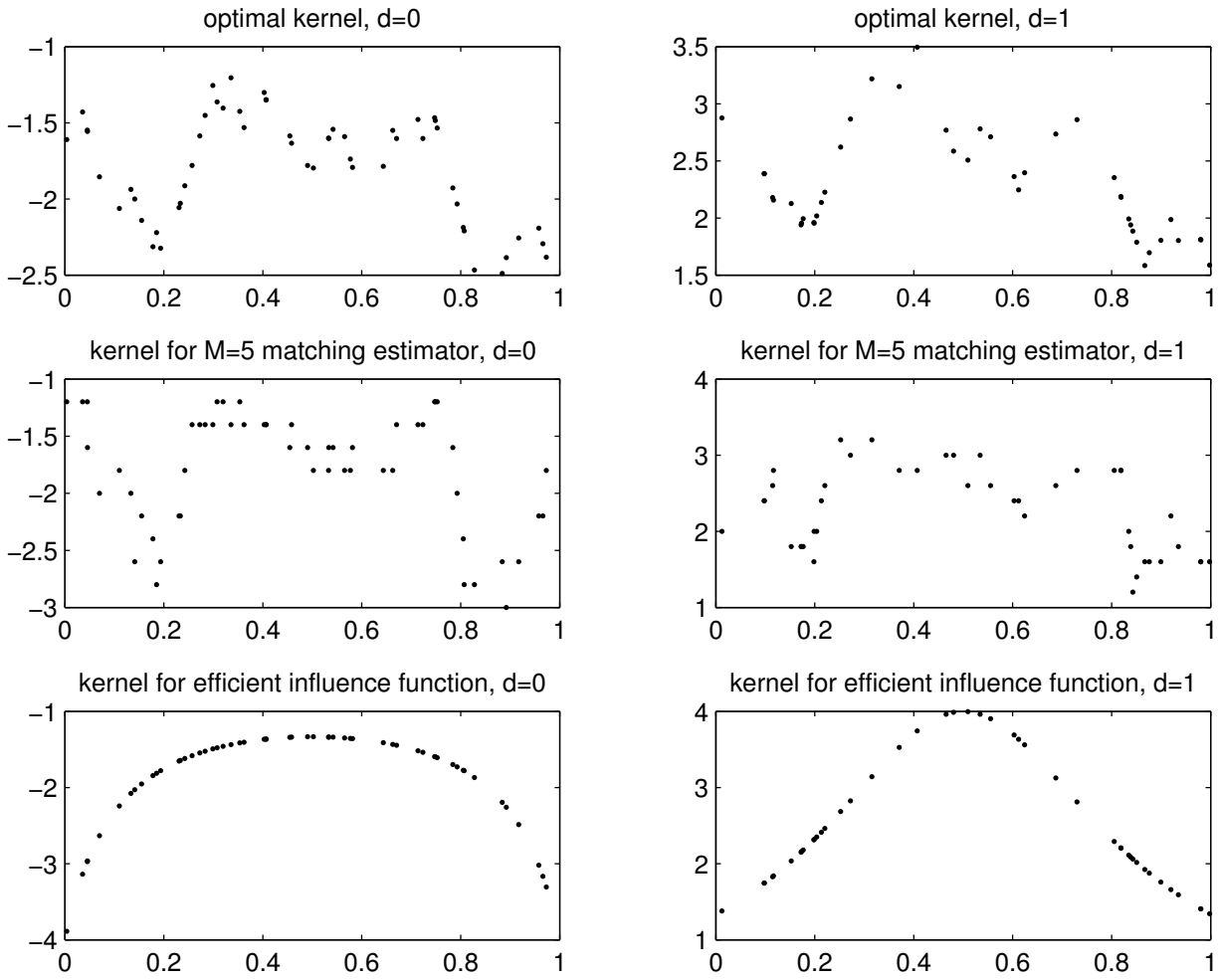


Figure 1: Estimator weights for $n = 100$

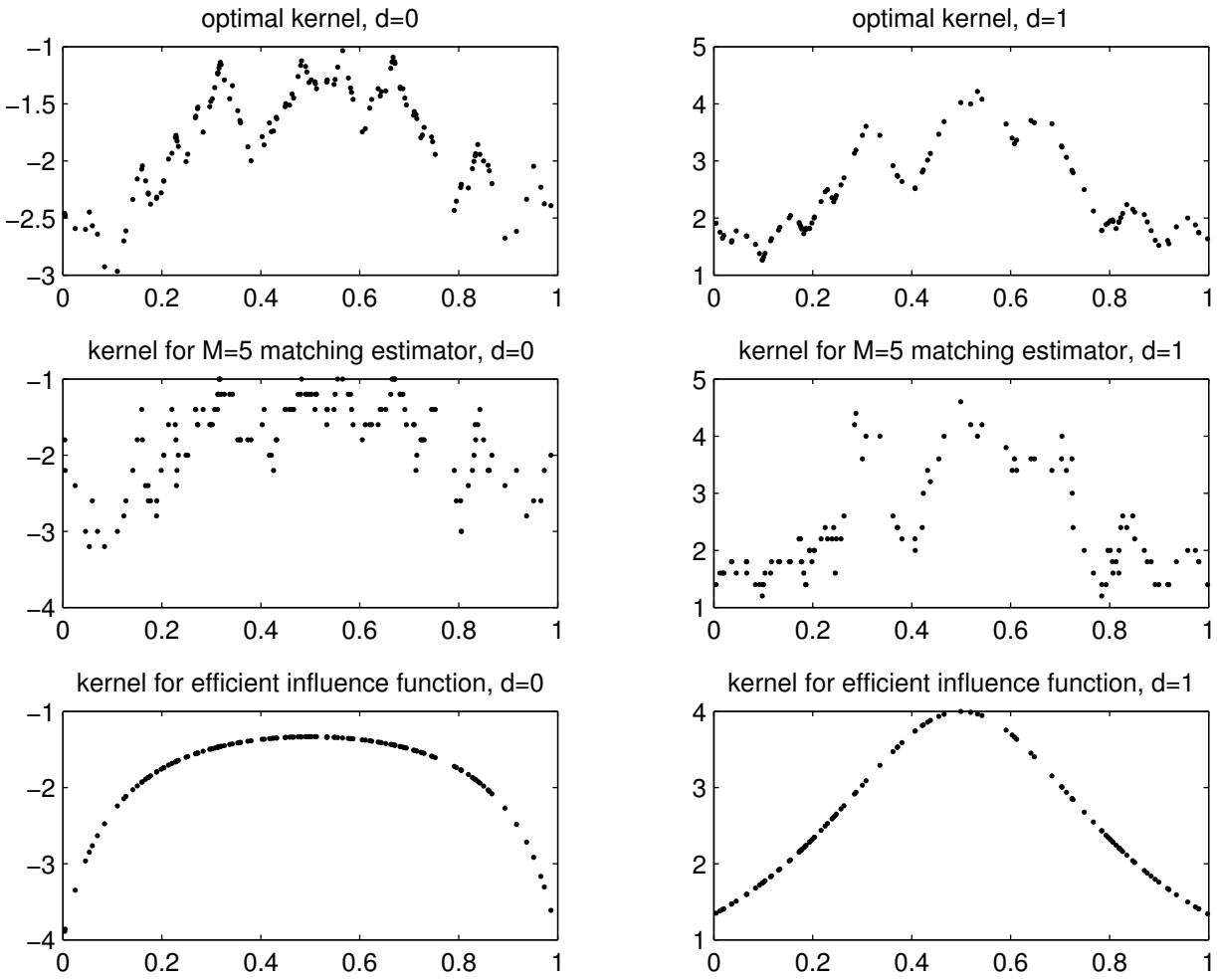


Figure 2: Estimator weights for $n = 250$

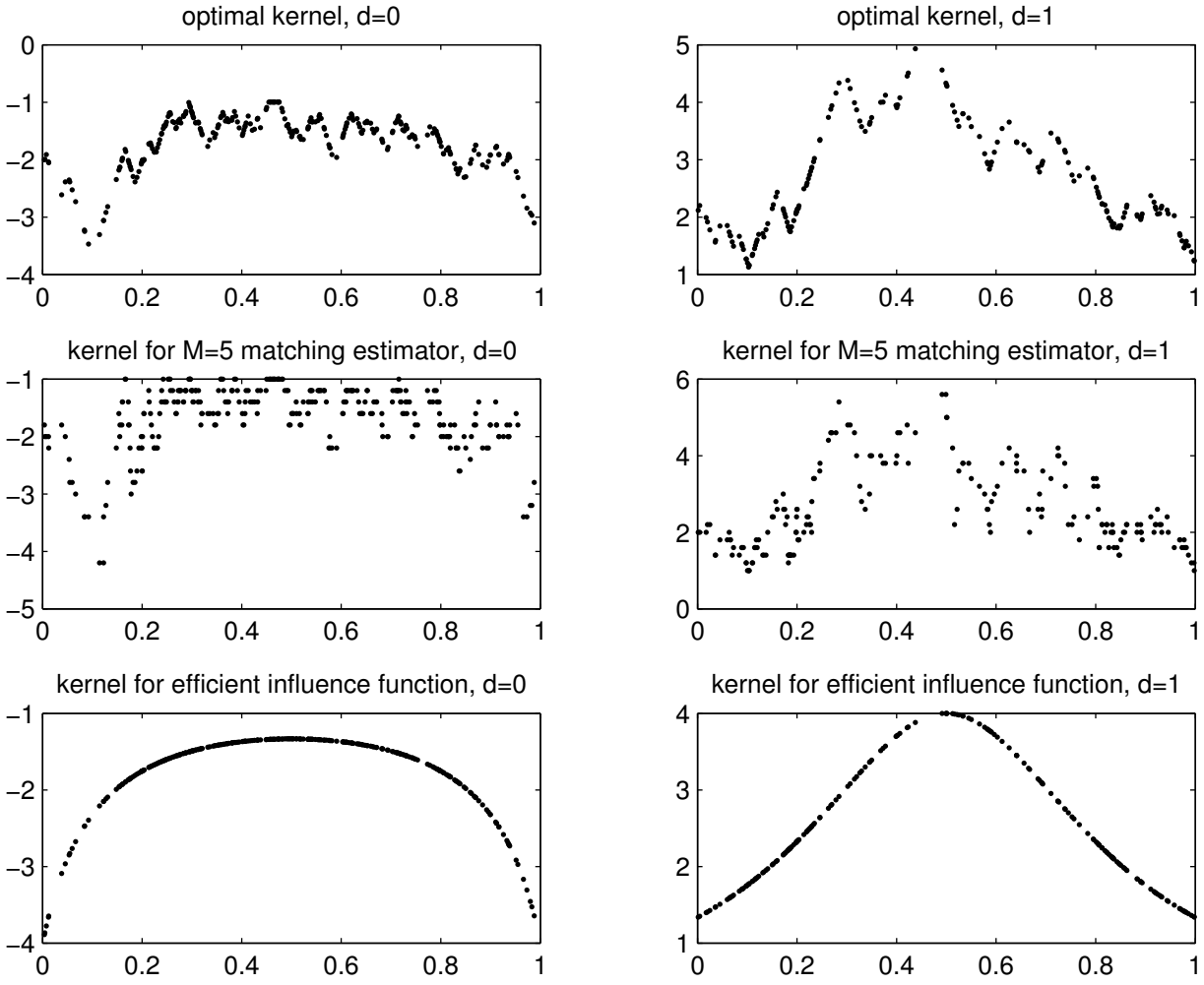


Figure 3: Estimator weights for $n = 500$

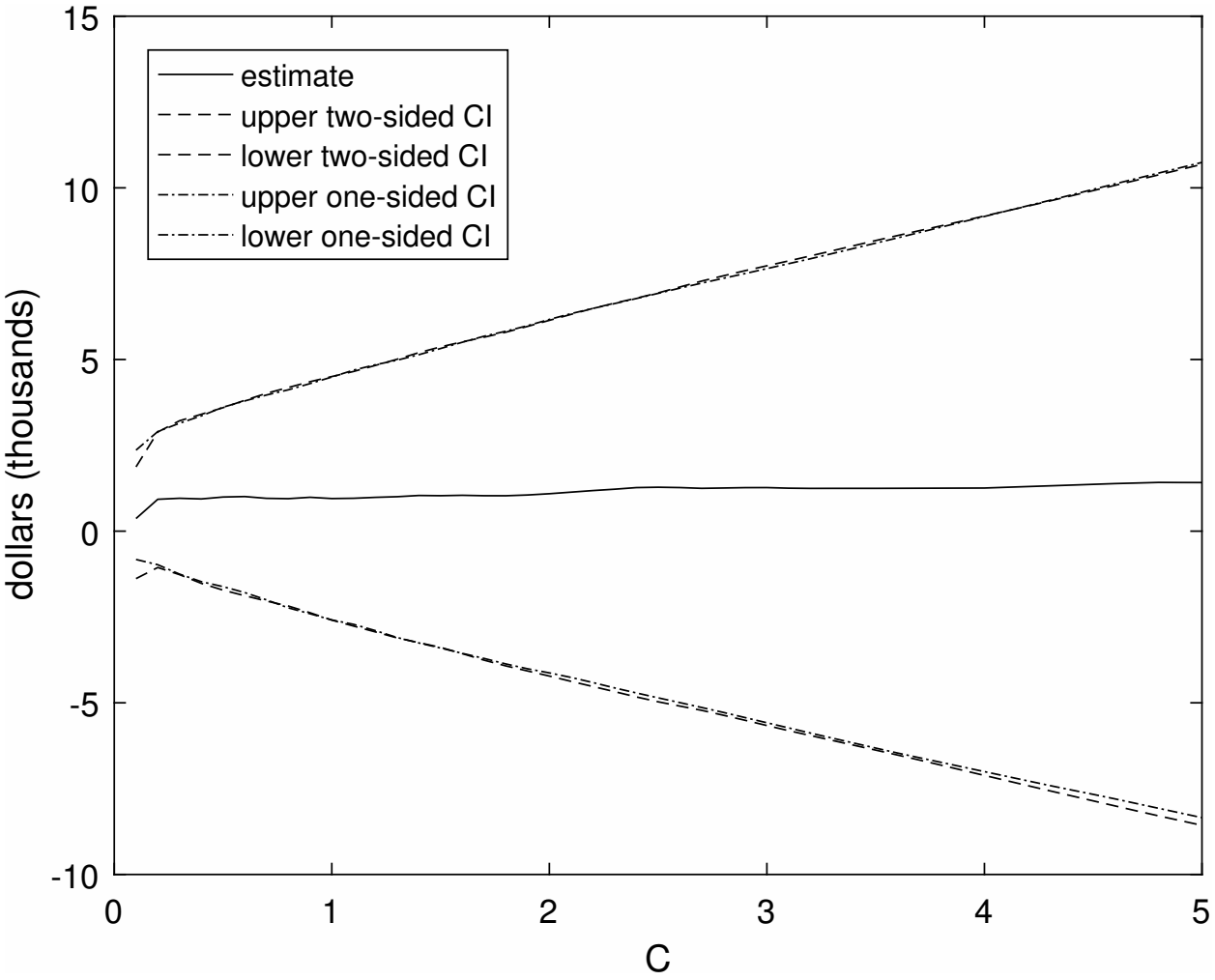


Figure 4: Optimal estimator and CIs for CATT in NSW data

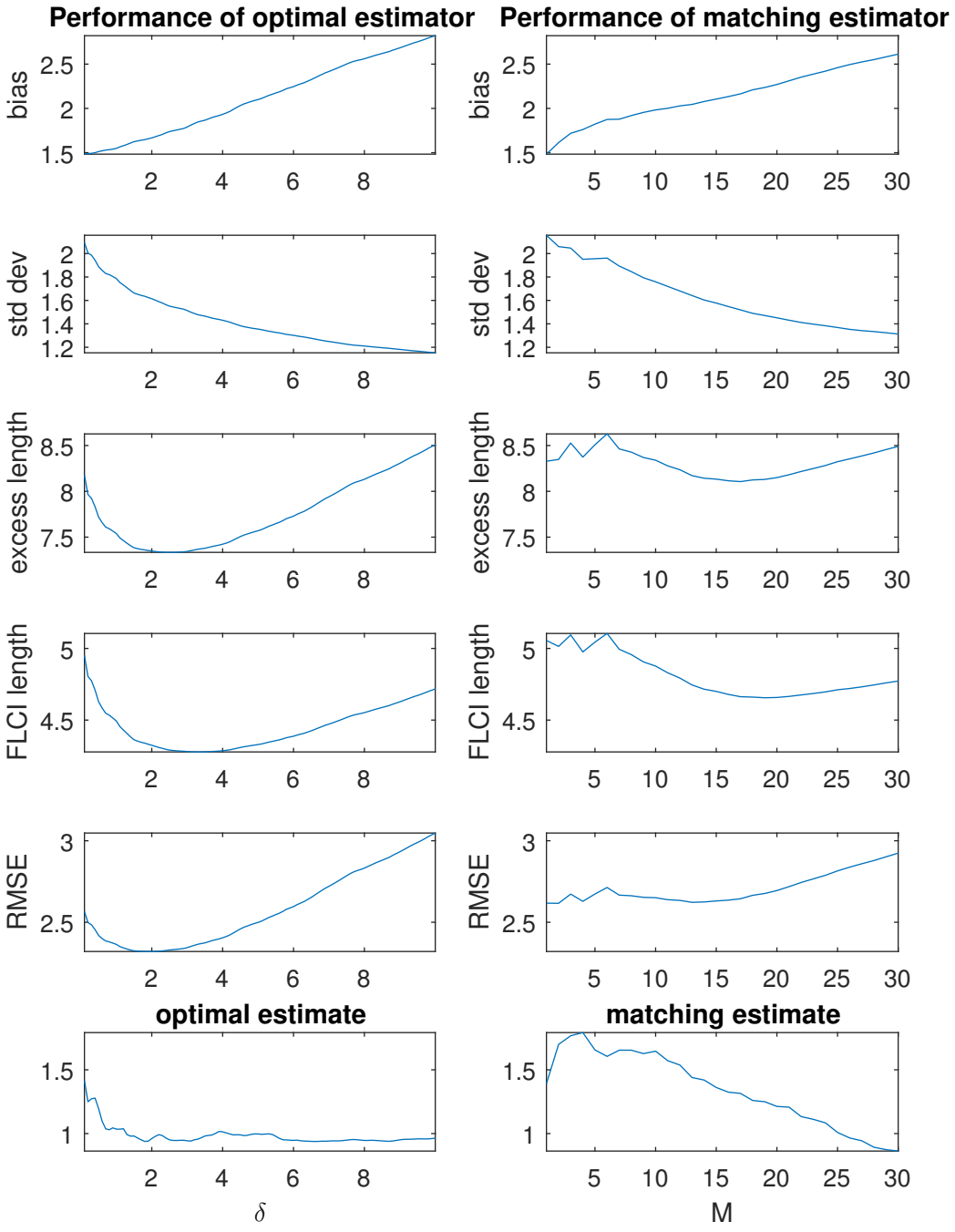


Figure 5: Performance of optimal and matching estimators