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

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

September 19, 2017

Preliminary and incomplete, please do not circulate

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 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 different 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 in the case of 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 in the case in which one does not condition on treatment and covariates (Robins et al., 2009), but where no regularity 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 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 a more than a single match. This requires computing the worst-case bias of the estimator, which reduces to a convex programming problem; we show how the solution can be found numerically under Lipschitz smoothness. 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 the overlap in the data combined with the smoothness conditions imposed by the researcher lead to nonnegligible 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 often substantially different than 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 (2016). 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 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 main results. Section 3 gives a numerical illustration of the optimal CIs. Section 4 presents the results of 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 main results

We consider the following setting. For observations $i = 1, \ldots, n$, we observe y_i where $y_i = y_i(1)d_i + y_i(0)(1 - d_i)$ and $y_i(1)$ and $y_i(0)$ are potential outcomes, along with pretreatment variables $x_i \in \mathbb{R}^p$ and treatment indicator $d_i \in \{0, 1\}$. We condition on the realized values of $\{x_i, d_i\}_{i=1}^n$ 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

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

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

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

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 d_i, x_i) variance $\sigma^2(x_i, d_i)$ treated as known.

We assume that f is in a known function class $\mathcal{F}$, which we assume throughout the paper to be convex. The function class $\mathcal{F}$ formalizes the "regularity" or "smoothness" that we are willing to impose. For many of the results in this paper, we focus on classes that place Lipschitz constraints on $f(\cdot, 0)$ and $f(\cdot, 1)$:

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

where $\|\cdot\|_{\mathcal{X}}$ is a norm on x, although our general results hold for any convex function class.

For a given level α, a $100 \cdot (1-\alpha)$ CI $\mathcal{C}$ for a parameter Lf (e.g. for the CATE, we take $Lf = \text{CATE}(f)$ as defined in (2)) must satisfy

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

where P_f denotes probability computed under f. Subject to (4), we derive CIs that optimize minimax performance over $\mathcal{F}$ or over a smaller subset of the parameter space. For a one-sided CI $[\hat{c}, \infty)$ for a parameter Lf, we focus on quantiles of excess length. Given a subset

[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

$$y_i(1), y_i(0) \perp\!\!\!\perp D_i | X_i$$

holds. Then

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

where $f(x,1) = E(y_i(1)|X_i = x) = E(y_i(1)|D_i = 1, X_i = x) = E(y_i|D_i = 1, X_i = x)$ and similarly for $f(x,0)$, and $\{y_i\}_{i=1}^n$ 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 $\{(X_i', D_i)\}_{i=1}^n$.

$\mathcal{G} \subseteq \mathcal{F}$, 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 $q_{g,\beta}(Lg - \hat{c})$ denotes the βth quantile of excess length $Lg - \hat{c}$ for the CI $[\hat{c}, \infty)$ under the function g. Taking $\mathcal{G} = \mathcal{F}$, a CI that optimizes $q_\beta(\hat{c}, \mathcal{F})$ is called minimax, which is the case we consider throughout most of this paper.

For two-sided CIs, we focus on fixed length CIs, which take the form $\hat{L} \pm \chi$ for an estimator $\hat{L}$ and a constant χ (since we take the variance function of u_i as known, χ can depend on the variance of the errors). Since χ is constant, optimal fixed length CIs simply minimize the half-length χ. As we discuss below, fixed length CIs can be shown to have nearly optimal expected length in our setting among all CIs.

2.1 Linear Estimators

To derive optimal CIs, we note that our problem falls into the general framework of Donoho (1994). Thus, we can use results from Donoho (1994), Cai and Low (2004) and Armstrong and Kolesár (2016) to find estimators and CIs that are optimal or "close to" optimal, with "close to" defined using tight finite-sample bounds. In particular, optimal CIs are based on linear estimators, which in our setting take the form

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

Since $\hat{L}_{k(\cdot)}$ is linear in $\{y_i\}_{i=1}^n$, it is normal with variance

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

and bias bounded from above by

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

and from below by

$$\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k(\cdot)}) = \inf_{f\in\mathcal{F}} E_f(\hat{L}_{k(\cdot)} - Lf) = \inf_{f\in\mathcal{F}} \left[\sum_{i=1}^{n} k(x_i, d_i)f(x_i, d_i) - Lf\right].$$

To form a one-sided CI based on $\hat{L}_{k(\cdot)}$, we must take into account bias by subtracting $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k(\cdot)})$ in addition to the usual normal quantile: a $100 \cdot (1-\alpha)\%$ CI is given by $[\hat{c}, \infty)$ where

$$\hat{c} = \hat{L}_{k(\cdot)} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k(\cdot)}) - \text{sd}(\hat{L}_{k(\cdot)})z_{1-\alpha}$$

where $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of a $N(0,1)$ distribution. To form a two-sided CI, note that, under any $f \in \mathcal{F}$, the z-statistic $(\hat{L}_{k(\cdot)} - Lf)/\text{sd}(\hat{L}_{k(\cdot)})$ is distributed $N(t,1)$ for some t with $|t| \leq \max\left\{|\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k(\cdot)})|, |\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k(\cdot)})|\right\}/\text{sd}(\hat{L}_{k(\cdot)})$. Thus, letting $\text{cv}_\alpha(t)$ be the $1-\alpha$ quantile of the absolute value of a $N(0,1)$ random variable, a two-sided CI can be formed as

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

Following Donoho (1994), we refer to this as a fixed-length CI (FLCI), since it takes the form $\hat{L} \pm \chi$ where χ is constant (in practice, the length of the feasible version of this CI will depend on the data through an estimate of the standard deviation)[3]. For the one-sided CI, the worst-case βth quantile of excess length over $\mathcal{G}$ is taken at the function $g \in \mathcal{G}$ that achieves $\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_{k(\cdot)})$ (i.e. when the estimate is biased downward as much as possible). This gives

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

Finally, we consider estimation as well as inference. For estimation, we consider minimax root mean squared error (RMSE), given by

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

The optimal weighting $k(\cdot)$ follows from results in Donoho (1994) and Armstrong and

[3]For general convex classes $\mathcal{F}$, the optimal FLCI is centered at $\hat{L}_{k(\cdot)} + a$ for a nonrandom constant a. However, most if this paper focuses on settings where $\mathcal{F}$ is centrosymmetric, which leads to $a = 0$. See Appendix A

Kolesár (2016). To get some intuition for this, note that the width of the FLCI given above is increasing in both variance and worst-case bias. Thus, computing the optimal $k(\cdot)$ amounts to tracing out the minimum worst-case bias subject to a bound on variance, and varying this bound on variance to find the optimal bias/standard deviation ratio. Optimizing the worst-case bias is a minimax problem (minimizing the maximum bias), and the results in Donoho (1994) reduce it to a single convex optimization problem. We provide details in Appendix A.

In general, computing the optimal CI requires optimizing over the set $\mathcal{F}$, which, in nonparametric settings, is infinite dimensional. We now focus on the Lipschitz class $\mathcal{F}_{\text{Lip}}(C)$ and provide a finite dimensional convex optimization problem that characterizes the solution.

2.2 Optimal CIs Under Lipschitz Smoothness

Given $\delta > 0$, let f^*_δ solve

$$\max_f 2\frac{1}{n}\sum_{i=1}^n [f(x_i,1) - f(x_i,0)] \quad \text{s.t.} \quad \sqrt{\sum_{i=1}^n \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)}} \leq \frac{\delta}{2} \quad \text{and}$$
$$|f(x_i,d) - f(x_j,d)| \leq C\left\|x_i - x_j\right\|_{\mathcal{X}} \text{ for } d \in \{0,1\}, i,j \in \{1,\ldots,n\},$$

and let $\omega(\delta)$ denote the value of this problem. The constraints in the second line of the above display are equivalent to imposing that there exists a function $f \in \mathcal{F}_{\text{Lip}}(C)$ that extrapolates these points (see Beliakov, 2006, Theorem 4). In solving this and other maximization problems over f where the objective and constraints depend only on f evaluated at points in $\{(x_i,0),(x_i,1)\}_{i=1}^n$, 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 optimize over the constrained subset of $\mathbb{R}^{2n}$: the value of f at other points does not matter for our purposes. Note that this is a convex optimization problem in $\mathbb{R}^{2n}$ with $2n(n-1)$ linear constraints, one quadratic constraint and a linear objective.

Let

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

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 the corresponding estimator at $-f_\delta^*$.

Theorem 2.1. *The estimator $\hat{L}_\delta = \hat{L}_{k_\delta^*(\cdot)}$ has worst case bias*

$$\overline{\text{bias}}_{\mathcal{F}_{Lip}(C)}(\hat{L}_\delta) = -\underline{\text{bias}}_{\mathcal{F}_{Lip}(C)}(\hat{L}_\delta) = \overline{\text{bias}}_\delta,$$

where $\overline{\text{bias}}_\delta$ is given above. Let

$$\hat{c}_{\alpha,\delta} = \hat{L}_\delta - \overline{\text{bias}}_\delta - \text{sd}(\hat{L}_\delta) z_{1-\alpha}.$$

Then $[\hat{c}_{\alpha,\delta}, \infty)$ is a $1-\alpha$ CI over $\mathcal{F}_{Lip}(C)$, and it minimizes $q_\beta(\hat{c}, \mathcal{F}_{Lip}(C))$ over all $1-\alpha$ CIs where $\beta = \Phi(\delta - z_{1-\alpha})$ and Φ denotes the standard normal cdf. The optimal fixed length CI centered at an affine estimator 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 2.1 shows that the CI that is minimax for βth quantile excess length is based on $\hat{L}_\delta$ where $\delta = z_{1-\alpha} + z_\beta$. The result follows from an application of results in Armstrong and Kolesár (2016) and Donoho (1994) to the present setting. We provide details in Appendix A. The FLCI in Theorem 2.1 is exactly optimal only among affine FLCIs. However, bounds in Donoho (1994) and Armstrong and Kolesár (2016) can be used to show that little is lost by restricting attention to affine FLCIs.

The one-sided CI in Theorem 2.1 can be computed by solving a single convex optimization problem. For the two-sided CI, one can solve the optimization problem for each δ and then perform a grid search over δ to find δ_χ. Alternatively, one can use the characterization of δ_χ given in Donoho (1994) using least favorable one-dimensional subfamilies.

While the optimal CIs do not, in general, have a closed form, it turns out that, when C is large enough, the optimal CI takes the form of a matching estimator. For a positive

integer M, the matching estimator takes the form in (5) with $k(\cdot)$ given by

$$k_{\text{match},M}(x_i, d_i) = \frac{1}{n}(2d_i - 1)\left(1 + \frac{K_M(i)}{M}\right) \tag{6}$$

where $K_M(i)$ is the number of times the ith observation is matched (see equation (3) in Abadie and Imbens 2006).

Theorem 2.2. *Consider the case where $\sigma^2(x_i, d_i) = \sigma$ is constant and the distances $\|x_i - x_j\|_{\mathcal{X}}$ take on unique values as i and j vary. There exists a constant K depending on σ 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 $M = 1$.*

2.3 Computing CIs Based on Suboptimal Estimators

The analysis in Section 2.1 allows one to construct a finite-sample CI based on any estimator that is linear in the y_i's. One can then compute the length of the FLCI or worst-case βth quantile of excess length, and compare it to the performance of the optimal CI. The only difficulty is in computing the worst-case bias. We now show that this reduces to a linear programming problem for $\mathcal{F}_{\text{Lip}}(C)$. In our numerical illustration and application, we compare the optimal CI to CIs based on matching estimators (which are suboptimal unless C is large and $M = 1$).

Consider a (possibly suboptimal) linear estimator $\hat{L}_{k(\cdot)}$ as defined in (5). We will assume that

$$\sum_{i=1}^n d_i k(x_i, d_i) = 1 \text{ and } \sum_{i=1}^n (1 - d_i) k(x_i, d_i) = -1, \tag{7}$$

since otherwise the bias would be arbitrarily large at multiples of $f(x, d) = d$ and $f(x, d) = 1 - d$. If this 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 $\sum_{i=1}^n f(x_i, 1) = \sum_{i=1}^n f(x_i, 0) = 0$ (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, which will not

change the bias). Thus, the worst-case bias of $\hat{L}_{k(\cdot)}$ is given by the maximized value of

$$\begin{aligned} &\max \sum_{i=1}^{n} k(x_i, d_i) f(x_i, d_i) - \frac{1}{n} \sum_{i=1}^{n} [f(x_i, 1) - f(x_i, 0)] \\ &\text{s.t. } |f(x_i, d) - f(x_j, d)| \leq C \left\| x_i - x_j \right\|_{\mathcal{X}} \text{ for } d \in \{0, 1\}, i, j \in \{1, \ldots, n\} \\ &\text{and } \sum_{i=1}^{n} f(x_i, 1) = \sum_{i=1}^{n} f(x_i, 0) = 0. \end{aligned} \tag{8}$$

where we again use Beliakov (2006, Theorem 4). This is a linear programming problem with $2n(n-1)$ inequality constraints and two equality constraints. We record this in the following theorem.

Theorem 2.3. *Consider the estimator $\hat{L}_{k(\cdot)} = \sum_{i=1}^{n} k(x_i, d_i) y_i$ where k satisfies (7). The worst-case bias of this estimator $\overline{\text{bias}}_{\mathcal{F}_{Lip}(C)}(\hat{L}_{k(\cdot)}) = -\underline{\text{bias}}_{\mathcal{F}_{Lip}(C)}(\hat{L}_{k(\cdot)})$ is given by the maximized value of (8).*

In particular, Theorem 2.3 shows that the formulas for CIs given in Section 2.1 hold with $\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_{k(\cdot)}) = -\underline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_{k(\cdot)})$ given by the maximized value of (8), along with the formula for minimax excess length $q_\beta(\hat{c}, \mathcal{F}_{\text{Lip}}(C))$ of the one-sided CI.

2.4 Bounds to Adaptation

In our numerical illustration and empirical example, we focus on computing minimax CIs. In the one-sided case, this corresponds to optimizing $g_\beta(\hat{c}, \mathcal{F})$. An alternative is to optimize $g_\beta(\hat{c}, \mathcal{G})$ for some smaller class $\mathcal{G} \subsetneq \mathcal{F}$, or to try to do this simultaneously for multiple classes $\mathcal{G}$. Such CIs are refereed to as adaptive. Unfortunately, in the case where $\mathcal{F}$ is centrosymmetric ($f \in \mathcal{F} \implies -f \in \mathcal{F}$), it can be shown that there is little scope for improving upon minimax CIs (see Armstrong and Kolesár, 2016). In particular, this is the case for the Lipschitz classes used in much of this paper, which means that one cannot estimate the Lipschitz constant C for the purposes of forming a CI. Because of this, we recommend reporting estimates and CIs for a range of choices of the Lipschitz constant C when implementing these estimators in practice.

Alternatively, if additional restrictions such as monotonicity are used, then some degree of adaptation may be possible. While we leave the full exploration of this question for future research, we note that our approach can be used to bound the potential gains from adaptation. For example, one can define $\mathcal{F}$ to be the class of functions such that $f(\cdot, d)$ is

monotone in certain variables and Lipschitz with constant C, and let $\mathcal{G}$ be the same class, but with C replaced by a smaller constant C'. One can then use our approach to compute the optimal excess length over $\mathcal{G}$ subject to coverage over $\mathcal{F}$. We show how optimal CIs can be computed when $\mathcal{F}$ and $\mathcal{G}$ impose Lipschitz and monotonicity constraints in Appendix A.

2.5 Semiparametric Efficiency Bound

Consider a setting where x_i, d_i and y_i are random with $p(x) = P(d_i = 1|x_i = x)$ denoting the propensity score. If $\mathcal{F}$ imposes sufficient smoothness, optimal estimators will be root-n consistent with asymptotically negligible bias, and will be asymptotically equivalent to a linear estimator with kernel

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

(see Hahn, 1998). We compare the optimal kernel to k_{seb} in our numerical illustration in Section 3.

On the other hand, the semiparametric efficiency bound gives only an upper bound, and it cannot be achieved unless $\mathcal{F}$ imposes sufficient smoothness relative to the dimension of x_i. With random x_i and d_i, Robins et al. (2009) derive optimal rates under a bound on the γ_fth derivative of $f(\cdot, 0)$ and $f(\cdot, 1)$ along with a bound on the γ_pth derivative of $p(\cdot)$. They find that root-n inference is impossible unless $\gamma_p + \gamma_f \geq p/2$ where p is the dimension x_i when x_i is continuously distributed.

Since conditioning on x_i and d_i essentially takes away the role of smoothness of $p(\cdot)$, this suggests that root-n inference should be impossible in our setting when $\gamma_f \geq p/2$ (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 $p(\cdot)$). This intuition turns out to be essentially correct. Since this question does not appear to have been addressed in the existing literature, we provide a formal result in Appendix B. In particular, the Lipschitz case we consider throughout most of this paper corresponds to $\gamma_f = 1$, so that root-n inference is possible only when $p \leq 2$.

2.6 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

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

We prove the asymptotic validity of this approach using results from Armstrong and Kolesár (2016) in Appendix C. In particular, our CIs are asymptotically valid even in cases discussed in Section 2.5 where the semiparametric efficiency bound cannot be achieved. Given that achieving the semiparametric efficiency bound requires bounding high order smoothness when p is moderate, this includes many practically relevant cases.

3 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 $P(d_i = 1|x_i = x) = p(x) = 2(x-1/2)^2 + 1/4$ 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 (6). We also compare the optimal weights to the weights corresponding to the semiparametric efficiency bound, given in (2.5).

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.

4 Application to National Supported Work Demonstration

We now consider an application to the National Supported Work (NSW) demonstration. The sample with $d_i = 1$ is given by a sample of people who received job training in this program. The sample with $d_i = 0$ is taken from the PSID. The data is the same as the data used by Dehejia and Wahba (1999) and Abadie and Imbens (2011).[4] Following these papers, we are interested in the sample 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, 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).

4.1 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

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

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. With this in mind, we use

$$A^{1/2} = A^{1/2}_{\text{main}} \equiv \text{diag}(0, 1, 20, 20, 0, 1, 0, 0, 0)$$

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 of the variables. We discuss other choices of the A in Section 4.4.

This choice of distance assumes that it suffices to control for education, previous year's earnings and the Black/Hispanic indicators when making the selection-on-observables assumption. 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. When $C = 1$, the bound on the Lipschitz constants for earnings state that earnings from the previous year do not have a greater than one-to-one effect on current earnings, and that last year's earnings are sufficient to control for employment and earnings in previous years. The bound on the Lipschitz constant for education (earnings cannot increase by more than \$1000 per year of education) may be somewhat strong, although it is large in percentage terms for many people in the sample.

4.2 Results

We compute the estimator $\hat{L}_\delta$ as described in Section 2.6 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

nearest neighbors defined using the same norm to define distance as for the Lipschitz class. The robust standard deviation estimate follows the formula in Section 2.6, 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 .8 quantile with $\alpha = .05$. Since the problem is translation invariant, the minimax one-sided CI inverts minimax tests with size .05 and power .8 (see Armstrong and Kolesár, 2016), which is a common benchmark in the literature on statistical power analysis (Cohen, 1988).

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 small values of C, 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 larger values of C, the point estimate is larger and becomes positive, 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 uncertainty 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 in a different way 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, which leads to the one-sided CI being based on a larger point estimate than the two-sided FLCI. On the other hand, the point estimate for the FLCI is never very far from the RMSE optimal estimate, reflecting the fact that the FLCI and RMSE criteria resolve the bias-variance tradeoff in a similar way.

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 plot the same quantities for matching estimators as a function of M, the number of matches, using the linear programming problem described in Section 2.3 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 2.2, $\hat{L}_\delta$ 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.

4.3 Comparison to 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).

Differences between the estimates reported here and the experimental estimate can arise from (1) differences between the CATT for our sample and the ATT (2) failure of the selection on observables assumption (which may lead to $\frac{\sum_{i=1}^n [f(x_i,1)-f(x_i,0)]d_i}{\sum_{i=1}^n d_i}$ not giving the actual CATT) and (3) bias and variance in estimating $\frac{\sum_{i=1}^n [f(x_i,1)-f(x_i,0)]d_i}{\sum_{i=1}^n d_i}$ as well as sampling error in the experimental estimates of the ATT. Since our CIs contain the experimental estimate of the ATT except for when the Lipschitz constant is very small, our results do not lead one to reject the null hypothesis that all of the difference between our estimates and the experimental estimate comes from (3). Indeed, our results show that a substantial portion of the difference between experimental estimates and the estimates based on non-experimental data reported here can be explained by bias in estimating $\frac{\sum_{i=1}^n [f(x_i,1)-f(x_i,0)]d_i}{\sum_{i=1}^n d_i}$: for $C = 1$, the optimal FLCI is centered at 0.5906 and has a worst-case bias of 0.5060 for estimating

$\frac{\sum_{i=1}^n [f(x_i,1)-f(x_i,0)]d_i}{\sum_{i=1}^n d_i}$, which is almost half the difference between the center of the FLCI and the experimental estimate of 1.794.

4.4 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 $C = 1/.0729 = 13.7174$. For $C = 10$, the optimal 95% affine FLCI is 1.7176 ± 7.6797, which is much wider than the FLCIs reported for A_{main} when $C = 1$ (which corresponds to a greater bound on the derivative of the conditional mean with respect to last year's earnings).

4.5 Monotonicity Restrictions

To tighten the bounds further, one can impose monotonicity restrictions. The optimal estimators can be obtained by solving an optimization problem similar to the one in the case where monotonicity is not imposed (see Appendix A). As an example to show how this leads to a tighter CI, if we assume that, for each $d = 0, 1$, $f(\cdot, d)$ is weakly increasing in age, education and both income and employment variables, and weakly decreasing in the Black and Hispanic indicator variables, the minimax affine FLCI with $C = 1$ and $A = A_{\text{main}}$ is 0.3912 ± 1.8762 (compared to $0.5906 \pm \text{cv}_{.05}(0.5060/0.9009) \cdot 0.9009 = 0.5906 \pm 2.0111$ when

monotonicity is not imposed).

Once monotonicity is imposed, the class of functions is no longer centrosymmetric, so the argument for using minimax fixed length confidence intervals is less clear. One may want to "direct power" at smooth alternatives, or attempt to adapt to different levels of smoothness. The problem of adaptive inference on average treatment effect parameters under unconfoundedness when the conditional mean satisfies shape restrictions is an interesting question that we leave for future research.

Appendix A Proofs and additional derivations

This appendix contains proofs and derivations used in the main text. Section A.1 proves Theorem 2.1 and derives the formulas for optimal estimators and CIs given in Section 2.2 as well as the generalization to Lipschitz classes with monotonicity discussed in Section 4.5. Section A.2 proves Theorem 2.2.

A.1 Derivation of Optimal CIs

We first note that our setting is a fixed design regression model with normal errors and known variance, and therefore falls into the framework used in Armstrong and Kolesár (2016) with (in the notation of that paper) $Y = (y_1/\sigma(x_1, d_1), \ldots, y_n/\sigma(x_n, d_n))$, $\mathcal{Y} = \mathbb{R}^n$, $Kf = (f(x_1, d_1), \sigma(x_1, d_1), \ldots, f(x_n, d_n)/\sigma(x_n, d_n))$. The functional of interest (L in the notation of Armstrong and Kolesár (2016)) is given by the CATE or CATT. To accommodate both of these cases, we consider a general weighted sample average treatment effect of the form

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

where $\{a_i\}_{i=1}^n$ are a set of known weights with $\sum_{i=1}^n a_i = 1$. Setting $a_i = 1/n$ gives the CATE, while setting $a_i = d_i / \left(\sum_{j=1}^n d_j\right)$ gives the CATT.

The ordered class modulus of continuity $\omega(\delta; \mathcal{F}, \mathcal{G})$ for classes $\mathcal{F}$ and $\mathcal{G}$ is given by the

maximized value of

$$\sup_{f,g} \sum_{i=1}^{n} a_i \left\{ [g(x_i, 1) - g(x_i, 0)] - [f(x_i, 1) - f(x_i, 0)] \right\} \tag{9}$$
$$\text{s.t.} \sqrt{\sum_{i=1}^{n} \frac{[f(x_i, d_i) - g(x_i, d_i)]^2}{\sigma^2(x_i, d_i)}} \leq \delta,\ f \in \mathcal{F},\ g \in \mathcal{G}$$

(see Armstrong and Kolesár, 2016, p. 14). Let f^*_δ, g^*_δ denote a pair of functions that achieves the maximum. For general convex classes $\mathcal{F}$ and $\mathcal{G}$, optimal CIs and efficiency bounds can be derived by solving this problem. We specialize to the Lipschitz classes $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$ and $\mathcal{G} = \mathcal{F}_{\text{Lip}}(C')$ as well as versions of these classes that impose monotonicity conditions and show that, in these cases, the constraints on the infinite dimensional objects f and g can be phrased as a finite set of linear constraints.

For the function classes we consider here, the problem is translation invariant (as defined by Armstrong and Kolesár, 2016, p. 15) with ι given by the function $f(x, d) = d$. This gives the optimal weights as

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

The corresponding estimator $\hat{L}_\delta$ is then given by

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

where

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

and the worst-case biases are taken at f^*_δ and g^*_δ, and are given by

$$\begin{aligned} \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) &= a^*_\delta + \sum_{i=1}^{n} k^*_\delta(x_i, d_i) f^*_\delta(x_i, d_i) - L f^*_\delta \\ &= L g^*_\delta - a^*_\delta - \sum_{i=1}^{n} k^*_\delta(x_i, d_i) g^*_\delta(x_i, d_i) = -\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_\delta) \end{aligned}$$

(see Armstrong and Kolesár, 2016, p. 15).

The Lipschitz class (without monotonicity imposed) is centrosymmetric, so, if we take $\mathcal{F} = \mathcal{G} = \mathcal{F}_{\text{Lip}}(C)$, the solutions to the modulus problem for $\omega(\delta; \mathcal{F}) = \omega(\delta; \mathcal{F}, \mathcal{F})$ will be given by f_δ^* and $g_\delta^* = -f_\delta^*$ where f_δ^* solves

$$\sup_f 2\sum_{i=1}^n a_i \left[f(x_i, 1) - f(x_i, 0)\right] \text{ s.t. } \sqrt{\sum_{i=1}^n \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)}} \leq \frac{\delta}{2},\ f \in \mathcal{F}. \tag{10}$$

The formula for k_δ^* and the worst-case biases given above then hold with $a_\delta^* = 0$ and $g_\delta^* = -f_\delta^*$, which gives the formulas in the main text (in the main text, the notation f_δ^* is used for the function denoted g_δ^* in this appendix).

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 that $f(\cdot, 0)$ and $f(\cdot, 1)$ are monotonic 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, we can phrase the problem of computing the modulus, optimal weights and worst-case biases as a finite dimensional convex optimization problem.

Theorem A.1. *The modulus of continuity $\omega(\delta; \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C), \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C'))$ is given by the value of (9) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C)$ and $\mathcal{G} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C')$. Furthermore, the functions $f^* \in \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C)$ and $g^* \in \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C')$ are solutions to the modulus problem (9) with $\mathcal{F} = \mathcal{F}_{Lip,\mathcal{S}\uparrow,n}(C)$ and $\mathcal{G} = \mathcal{F}_{Lip,\mathcal{S}\uparrow,n}(C')$ iff. there exist $\tilde{f}^*$ and $\tilde{g}^*$ that maximize (9) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C)$ and $\mathcal{G} = \widetilde{\mathcal{F}}_{Lip,\mathcal{S}\uparrow,n}(C')$ such that $\tilde{f}^*(x,d) = f^*(x,d)$ and $\tilde{g}^*(x,d) = g^*(x,d)$ 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 $\tilde{g}^*$ in place of f^*_δ and g^*_δ.*

Similarly, the modulus of continuity $\omega(\delta; \mathcal{F}_{Lip}(C), \mathcal{F}_{Lip}(C))$ is given by the value of (10) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,n}(C)$. The function $f^* \in \mathcal{F}_{Lip}(C)$ is a solution to the modulus problem (10) with $\mathcal{F} = \mathcal{F}_{Lip}(C)$ iff. there exists $\tilde{f}^* \in \widetilde{\mathcal{F}}_{Lip,n}(C)$ that maximizes (10) with $\mathcal{F} = \widetilde{\mathcal{F}}_{Lip,n}(C)$ such that $\tilde{f}^*(x,d) = f^*(x,d)$ 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 $\tilde{g}^* = -\tilde{f}^*$ in place of f^*_δ and g^*_δ.

Theorem 2.1 now follows immediately from Theorem A.1 along with Corollary 3.1 in Armstrong and Kolesár (2016) for the one-sided CI and Theorem 1 and calculations in Donoho (1994) for the two-sided case.

A.2 Proof of Theorem 2.2

To prove Theorem 2.2, we first provide another characterization of the optimal weights given in (10). Given $\{m_i\}_{i=1}^n$, consider the optimization problem (10) 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 $f(x,1) = \min_{i:d_i=1}(m_i + \|x - x_i\|_{\mathcal{X}})$. Similarly, $f(x,0)$ is minimized simultaneously for all x by taking $f(x,0) = -\min_{i:d_i=0}(m_i + \|x - x_i\|_{\mathcal{X}})$ (see Beliakov, 2006, p. 25). Plugging this into (10), it follows that $f^*_\delta(x_i, d_i) = (2d_i - 1) \cdot m^*_i$ where $\{m^*_i\}_{i=1}^n$ solves

$$\max_m \sum_{i=1}^n a_i \left[m_i + \min_{j:d_j \neq d_i} (m_j + \|x_i - x_j\|_{\mathcal{X}}) \right]$$
$$\text{s.t. } \sum_{i=1}^n m_i^2/\sigma^2(x_i, d_i) \leq \delta^2/4, \quad |m_i - m_j| \leq C\|x_i - x_j\|_{\mathcal{X}} \text{ for all } i, j \text{ with } d_i = d_j.$$

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 $\tilde{\omega}_i(m) = \min_{j:d_j \neq d_i}(m_j + \|x_i - x_j\|_{\mathcal{X}})$ and let $\mathcal{J}_i(m)$ denote the set of indices such that this minimum is achieved. Note that $\mathcal{J}_i(0)$ is the set of 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

$$\mathcal{J}_i(m) = \{j \in \mathcal{J}_i(0) : m_j \leq m_\ell \text{ all } \ell \in \mathcal{J}_i(0)\}. \tag{11}$$

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, the constraints $|m_i - m_j| \leq C\|x_i - x_j\|_{\mathcal{X}}$ are implied by the constraint on $\sum_{i=1}^n m_i^2/\sigma^2(x_i, d_i)$. Thus, specializing to the case in Theorem 2.2 where $\sigma^2(x_i, d_i) = \sigma^2$ is constant and $a_i = 1/n$, the first order conditions can be written

$$\iota - 2(\lambda n/\sigma^2)m \in -\sum_{i=1}^n \partial\tilde{\omega}_i(m)$$
$$= \left\{ \sum_{i=1}^n \sum_{j=1}^n b_{ij} e_j | b_{ij} = 0 \text{ all } j \notin \mathcal{J}_i(m),\ b_{ij} \geq 0,\ \text{all } i, j \text{ and } \sum_{j=1}^n b_{ij} = 1 \text{ all } i \right\}$$

where ι is a vector of ones. Let $\|m\|$ be small enough so that (11) holds. Then $m_j = m_\ell$ for $j, \ell \in \mathcal{J}_i(m)$. Thus, the nonzero b_{ij}'s in the superdifferential must take the form $b_{ij} = \frac{1}{\#\mathcal{J}_i(m)}$, which gives

$$2(\lambda n/\sigma^2)m_j = 1 + \sum_{i:j \in \mathcal{J}_i(m)} \frac{1}{\#\mathcal{J}_i(m)}.$$

In the case where the values of $\|x_i - x_j\|_{\mathcal{X}}$ are unique, we have $\mathcal{J}_i(m) = \mathcal{J}_i(0)$ for small enough m and $\mathcal{J}_i(0)$ is a singleton for each i, so that m_j is proportional to $1 + \#\{i : j \in \mathcal{J}_i(m)\}$, which gives the matching estimator with a single match.

Appendix B Asymptotic efficiency bound in non-regular case

In this appendix, we derive conditions under which root-n inference on the CATE conditional on treatments and outcomes is impossible. We consider Hölder classes of functions, which place bounds on (possibly higher order) derivatives of $f(\cdot, 0)$ and $f(\cdot, 0)$. Let $\Sigma(\gamma, C)$ denote the set of 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 $\mathbb{R}^p$. We consider the class $\mathcal{F}$ given by functions $f(x, d)$ such that $f(\cdot, 0)$ and $f(\cdot, 1)$ are both in $\Sigma(\gamma, C)$. Here, $\gamma = 1$ corresponds to the Lipschitz class used in most of the paper.

We consider a setup where i.i.d. random variables $\{X_i, D_i\}_{i=1}^n$ are drawn, and the Gaussian regression model defined in (1) and (3) is considered with $\{x_i, d_i\}_{i=1}^n = \{X_i, D_i\}_{i=1}^n$ treated as fixed. Under regularity conditions on the distribution generating these covariates and treatment indicators, we show 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 $f(x, d) = 0$.

Theorem B.1. *Let $\{X_i, D_i\}$ be i.i.d. with X_i a random variable on $\mathbb{R}^p$ and D_i taking values in $\{0, 1\}$. 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 $\{X_i, D_i\}_{i=1}^n$:*

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

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.

The condition $\gamma/p < 1/2$ corresponds to conditions under which equivalence with a Brownian sheet fails, as noted by Brown and Zhang (1998). The CATE has a similar form to the parameter used in a counterexample in the $p = 1$ case by Brown and Low (1996). As discussed in Section 2.5, the condition $\gamma/p < 1/2$ corresponds to the case where root-n inference is impossible when one does not condition on treatments and pretreatment variables in the case where no smoothness is imposed on the propensity score. However, when smoothness is imposed on the propensity score, weakening the coverage requirement to require only marginal coverage allows one to obtain inference at a root-n rate under weaker conditions on $f(x, d)$. 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).

We now prove Theorem B.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 $\liminf_n \#\mathcal{I}_n(h_n)/n \geq \eta$ almost surely.*

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 $\#\mathcal{I}_n(h) = \sum_{i \in \mathcal{N}_{1,n}} [I(X_i \in [a, b]^p) - I(X_i \in A_n)]$. Note that, conditional on $\mathcal{E}$, the random variables $I(X_i \in A_n)$ with $i \in \mathcal{N}_{1,n}$ are i.i.d. Bernoulli(ν_n) with $\nu_n = P(X_i \in A_n | \mathcal{E}) = \int I(x \in A_n) f_{X|D}(x|1)\, dx \leq K\lambda(A_n)$ 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 $I(X_i \in A_n) \leq Z_i$. Applying

the strong law of large numbers, it follows that

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

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

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n[f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,1)-f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,0)]\\ &=-\frac{1}{n}\sum_{i=1}^n\sum_{\tilde{x}\in\tilde{\mathcal{X}}_n(h,\eta)}k((X_i-\tilde{x})/h)\leq-\underline{k}\frac{\#\mathcal{I}_n(h)}{n}\end{aligned}\tag{12}$$

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 (12), 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 appendix gives conditions for asymptotic coverage of the feasible CI with unknown error distribution described in Section 2.6. The result follows by verifying the conditions in Theorem E.2 in Armstrong and Kolesár (2016).

Theorem C.1. *Consider the fixed design model with u_i distributed independently (but not identically distributed) with $Eu_i = 0$ and $1/K \leq \mathrm{var}(u_i) \leq K$ and $E|u_i|^{2+\eta} \leq K$ for some $\eta > 0$ and some K. Suppose that, for all $\eta > 0$ and $d \in \{0, 1\}$, $\min_{1 \leq i \leq n} \sum_{j=1}^n I(\|x_i - x_j\| \leq \eta, d_i = d) \to \infty$, and that $\max_{1 \leq i \leq n} |\hat{f}(x_i, d_i) - f(x_i, d_i)| \xrightarrow{p} 0$ uniformly over $f \in \mathcal{F}_{Lip}(C)$. Let $\mathcal{C}$ denote either of the CIs described in Section 2.6 with $\tilde{\sigma}(x, d)$ taken to be nonrandom and bounded away from zero and infinity. Then $\liminf_{n\to\infty} \inf_{f \in \mathcal{F}} P(Lf \in \mathcal{C}) \geq 1 - \alpha$, where*

$\mathcal{F} = \mathcal{F}_{Lip}(C)$ *or* $\mathcal{F} = \mathcal{F}_{Lip,\mathcal{S}\uparrow}(C)$ *denotes the parameter space used to compute the CI.*

In addition to asymptotic coverage of the CIs, it follows from Theorem E.2 in Armstrong and Kolesár (2016) that $\hat{L}_\delta$ is asymptotically normal (conditional on the x_i's and d_i's). In particular, it is interesting to note that the optimal estimator is asymptotically normal even in non-regular cases, albeit with a potentially non-negligible asymptotic bias and non-root-n rate of convergence.

References

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2014a): "Finite Population Causal Standard Errors," Tech. rep., nBER Working Paper No. 20325.

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

——— (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 (2016): "Optimal inference in a class of regression models," ArXiv:1511.06028v2.

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.

Brown, L. D. and M. G. Low (1996): "Asymptotic equivalence of nonparametric regression and white noise," *Annals of Statistics*, 24, 2384–2398.

Brown, L. D. and C.-H. Zhang (1998): "Asymptotic Nonequivalence of Nonparametric Experiments When the Smoothness Index is 1/2," *The Annals of Statistics*, 26, 279–287.

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*, Routledge.

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.

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.

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.

Optimal estimator

criterion	optimal δ	estimate	worst-case bias	non robust std estimate	robust std estimate
one-sided CI	2.4865	0.9652	0.2376	1.4806	0.9752
FLCI	5.5988	0.5906	0.5060	1.3399	0.9009
RMSE	5.4672	0.6019	0.4920	1.3449	0.9033

Matching estimator

criterion	optimal M	estimate	worst-case bias	non robust std estimate	robust std estimate
one-sided CI	1	1.4353	0.1137	1.6496	0.9950
FLCI	12	0.8409	0.6739	1.4393	0.8540
RMSE	11	0.9609	0.6132	1.4641	0.8541

Table 1: Results for NSW data, $p = 1$, $A = A_{\text{main}}$, $C = 1$

age	educ.	Black	Hispanic	married	earnings in 1974	earnings in 1975	emp. in 1974	emp. in 1975
0.0952	0.3275	2.1998	5.4864	2.5993	0.0729	0.0721	2.9793	2.9297

Table 2: Diagonal elements of $A_{\text{ne}}^{1/2}$

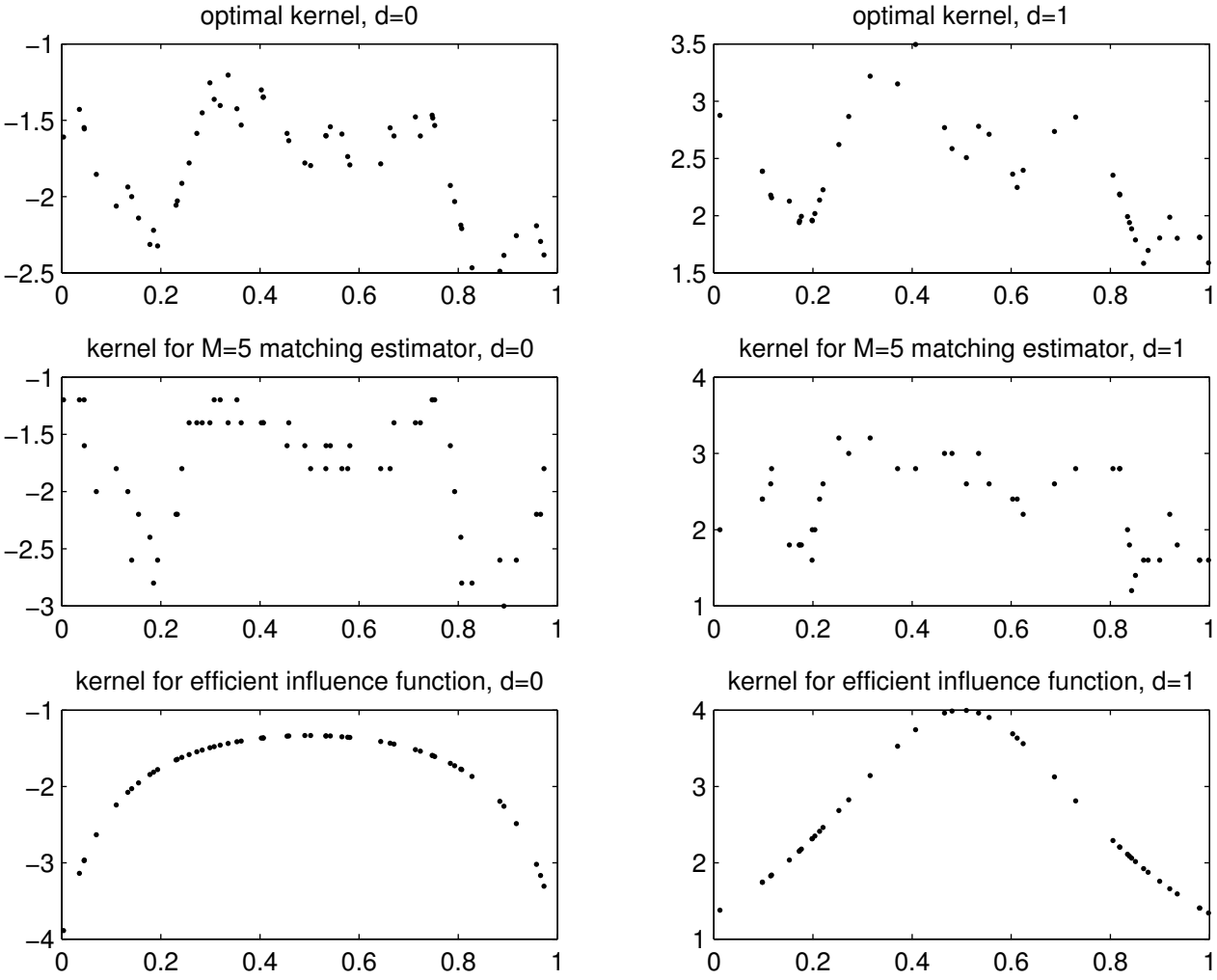

Figure 1: Estimator weights for $n = 100$

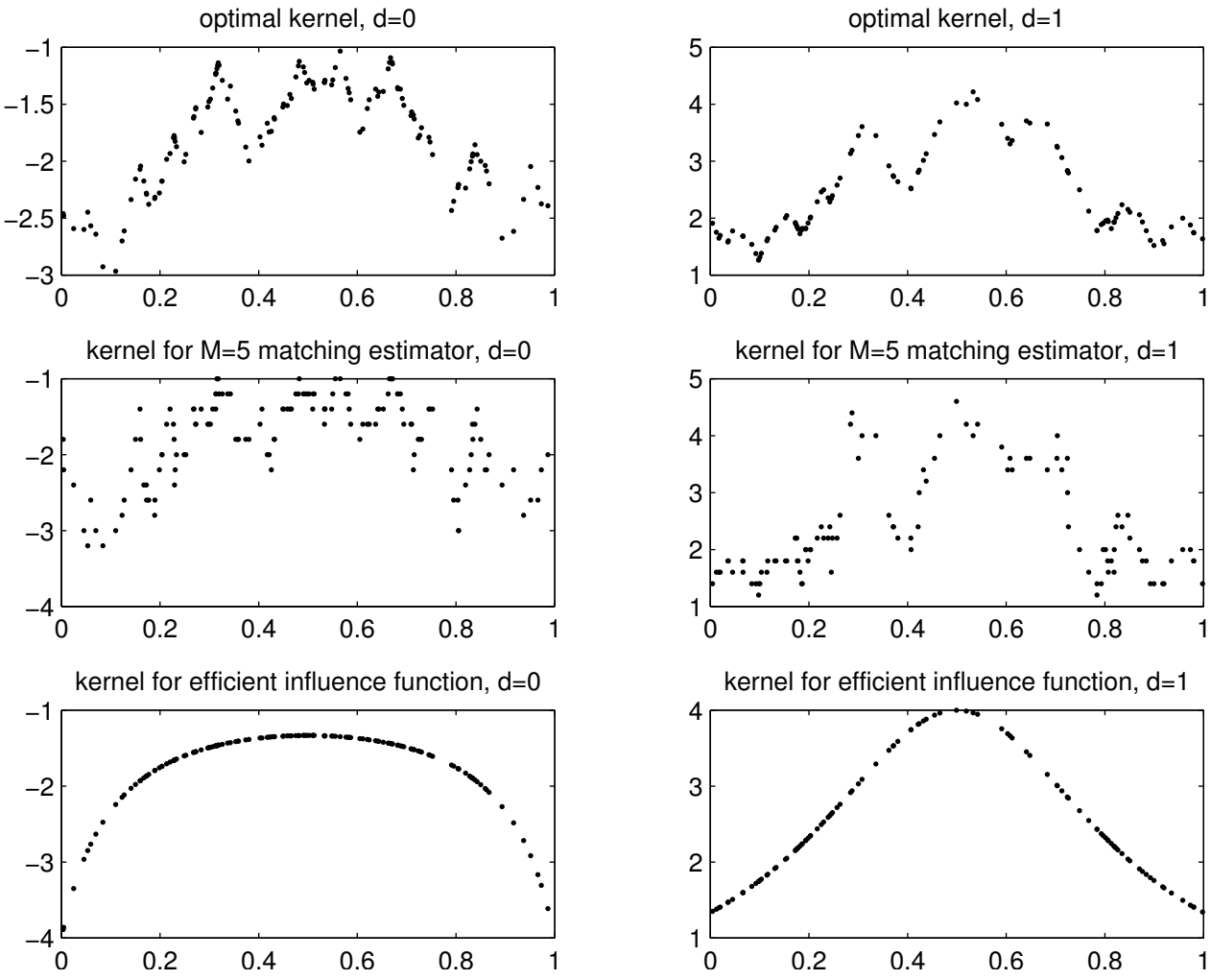

Figure 2: Estimator weights for $n = 250$

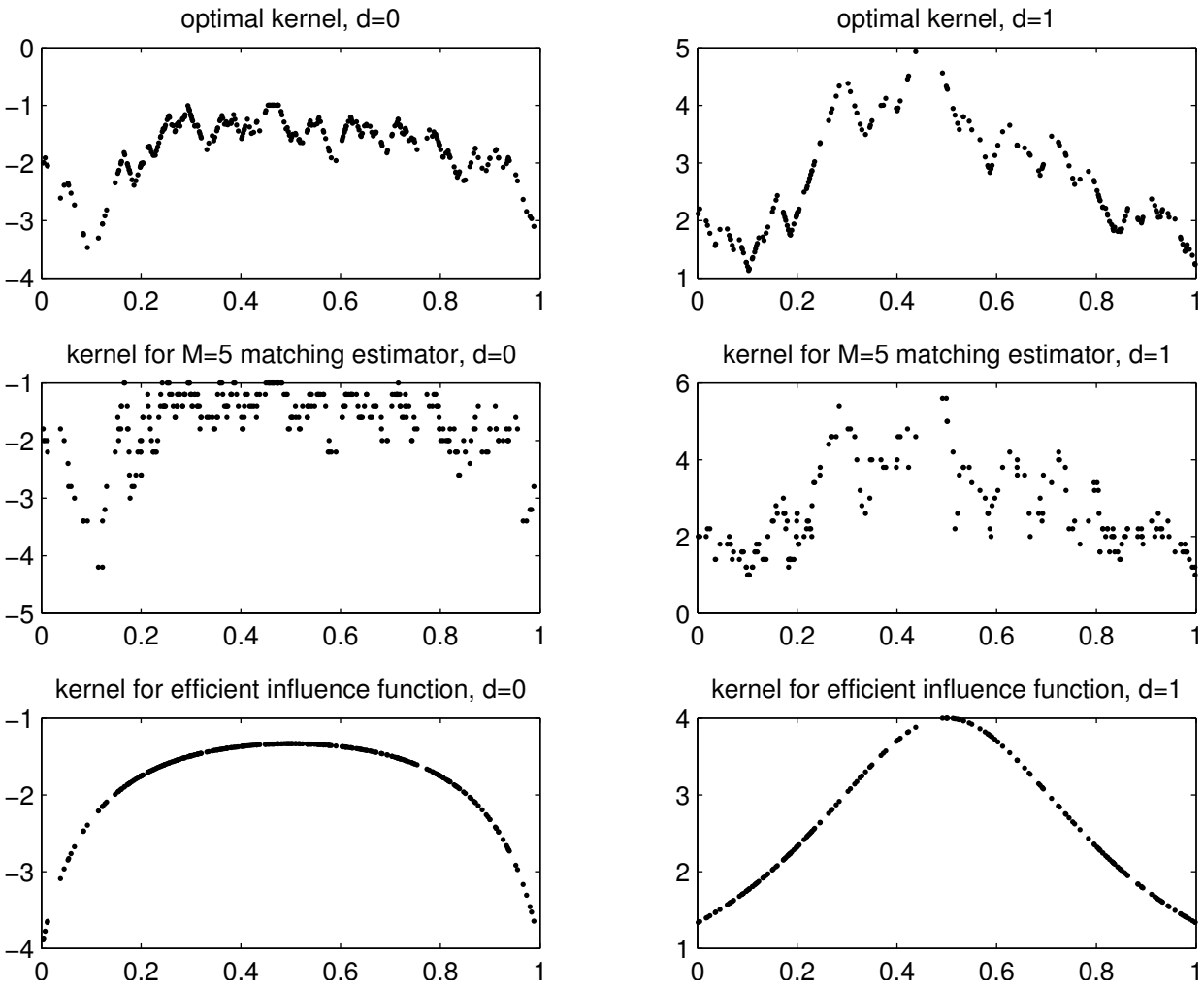

Figure 3: Estimator weights for $n = 500$

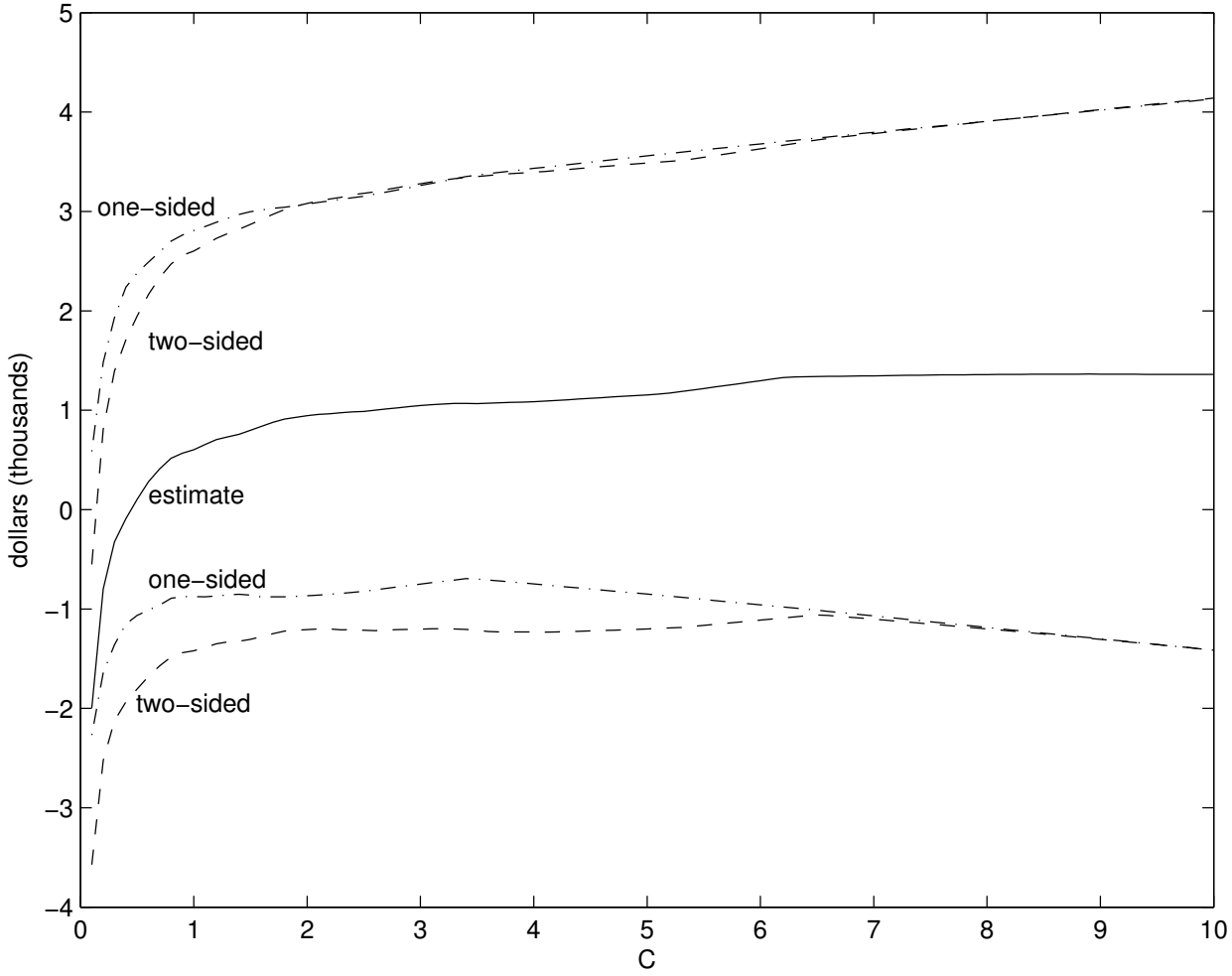

Figure 4: Optimal estimator and CIs for CATT in NSW data

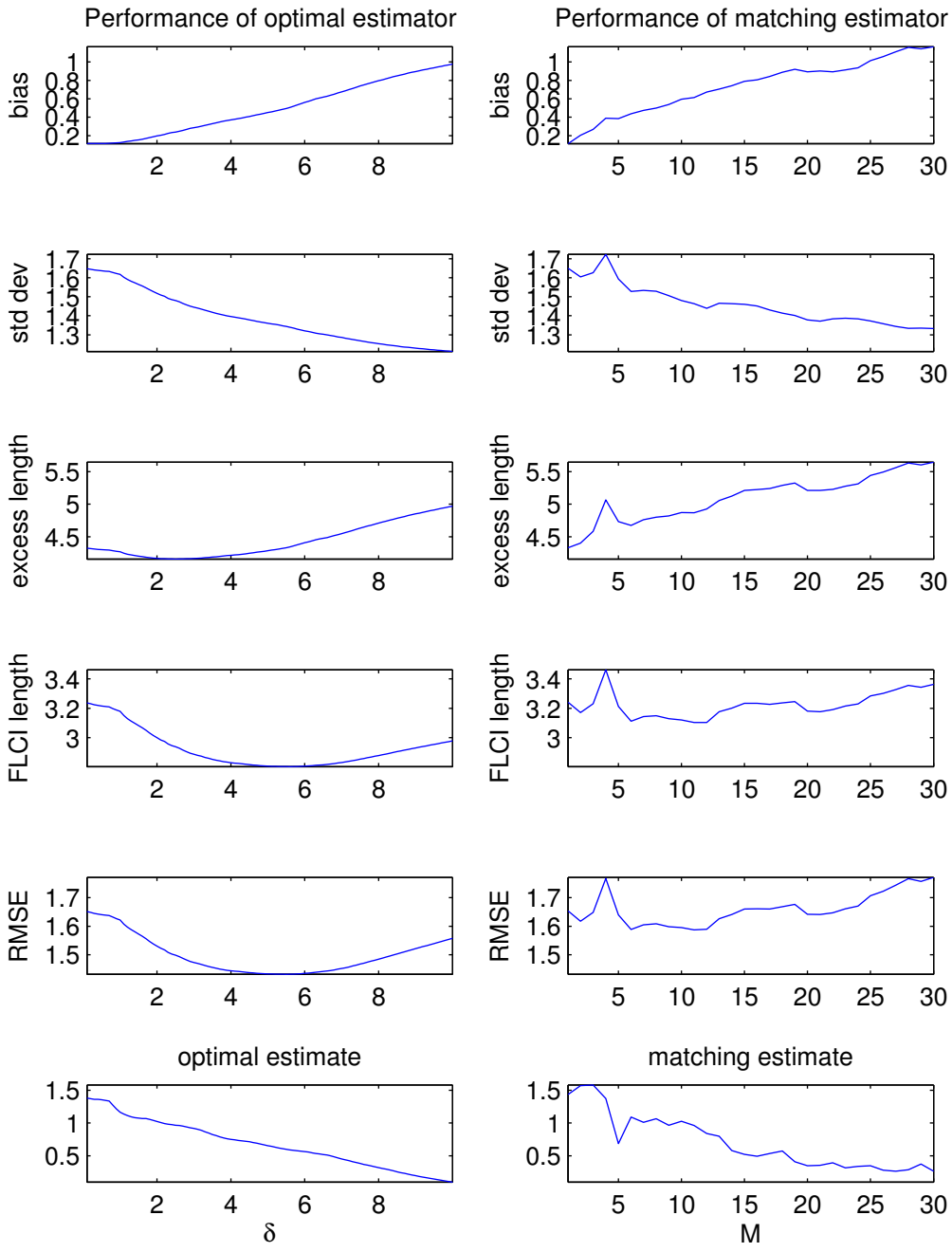

Figure 5: Performance of optimal and matching estimators