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

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

December 17, 2018

Abstract

We consider estimation and inference on average treatment effects under unconfoundedness conditional on the realizations of the treatment variable and covariates. Given nonparametric smoothness and/or shape restrictions on the conditional mean of the outcome variable, we derive estimators and confidence intervals (CIs) that are optimal in finite samples when the regression errors are normal with known variance. In contrast to conventional CIs, our CIs use a larger critical value that explicitly takes into account the potential bias of the estimator. When the error distribution is unknown, feasible versions of our CIs are valid asymptotically, even when $\sqrt{n}$-inference is not possible due to lack of overlap, or low smoothness of the conditional mean. We also derive the minimum smoothness conditions on the conditional mean that are necessary for $\sqrt{n}$-inference. When the conditional mean is restricted to be Lipschitz with a large enough bound on the Lipschitz constant, the optimal estimator reduces to a matching estimator with the number of matches set to one. We illustrate our methods in an application to the National Supported Work Demonstration.

*We thank Xiaohong Chen, Jin Hahn, Guido Imbens, Pepe Montiel Olea, Christoph Rothe, Pedro Sant'Anna, Andres Santos, Azeem Shaikh, Jeff Smith, 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.

The standard approach to comparing estimators and constructing confidence intervals (CIs) in this setting is based on the theory of semiparametric efficiency bounds. If, in addition to unconfoundedness, one also assumes overlap of the covariate distributions of treated and untreated subpopulations, as well as enough smoothness of either the propensity score or the conditional mean of the outcome given the treatment and covariates, 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 CIs based on any such estimator by adding and subtracting its standard deviation times the conventional 1.96 critical value (for nominal 95% CIs).

However, in many applications, the overlap is limited, which can have drastic effects on finite-sample performance (Busso et al., 2014; Rothe, 2017) and leads to an infinite semiparametric efficiency bound (Khan and Tamer, 2010).[1] Furthermore, even under perfect overlap, the standard approach requires a large amount of smoothness: 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. Finally, as argued in, for instance, Robins and Ritov (1997), the standard approach may 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

[1] To prevent these issues, one can redefine the object of interest as a treatment effect for a subset of the population for which overlap holds. While this restores the possibility of conventional $\sqrt{n}$-asymptotics, it changes the estimand to one that is typically less relevant to the policy question at hand. For examples of this approach, see Heckman et al. (1997), Galiani et al. (2005), Bailey and Goodman-Bacon (2015) or Crump et al. (2009).

if they are asymptotically equivalent. The bias may in turn lead to undercoverage of the CIs due to incorrect centering.

In this paper, we instead treat the smoothness and/or shape restrictions on the conditional mean of the outcome given the treatment and covariates as given and determined by the researcher. We make no overlap assumptions: we do not require that the semiparametric efficiency bound be finite, or even that the average treatment effect be point identified. We view the treatment and covariates as fixed, which allows us to explicitly calculate and account for the potential finite-sample biases of estimators. In this setting, we derive estimators and CIs that are optimal or near-optimal (depending on the criterion) in finite samples when the regression errors are assumed to be normal with known variance. We show that when this assumption is dropped, feasible versions of these CIs are valid asymptotically, uniformly in the underlying distribution (i.e. they are honest in the sense of Li, 1989). Importantly, our results cover both the regular case (in which $\sqrt{n}$-inference is possible) and the irregular case (in which $\sqrt{n}$-inference may not be possible, due to lack of perfect overlap, or due to low regularity of the regression function relative to the dimension of covariates).[2] In the latter case, conventional CIs, which assume $\sqrt{n}$-convergence and do not account for bias, will have coverage converging to zero asymptotically.

We show that optimal estimators are linear in the outcomes y_i: they take the form $\sum_{i=1}^n k_i y_i$, where $\{k_i\}_{i=1}^n$ are weights that depend on the covariates and treatments. The optimal weights k_i solve a finite-sample bias-variance tradeoff problem, and we give a general characterization of them as the solution to a convex programming problem. Furthermore, optimal CIs are based on the same class of estimators. Importantly, however, in order to account for the possible bias of the estimator, the CI uses a larger critical value than the conventional 1.96 critical value. This critical value depends on the worst-case bias of the estimator, which for the optimally chosen estimator has a simple form.[3] We show that the same approach can be used to form CIs based on any estimator that is linear in the outcomes, such as kernel, series, or matching estimators. The resulting CI can then be compared to the conventional CI as a form of sensitivity analysis: if the bias-adjusted critical value is much larger than the conventional 1.96 critical value, this indicates that the finite-sample bias of

[2] Khan and Tamer (2010) use the term "irregular identification" to refer to settings in which $\sqrt{n}$-inference is impossible due to the semiparametric efficiency bound being infinite. Here, we use the term "irregular" to refer to any setting in which $\sqrt{n}$-inference is impossible.

[3] The worst-case bias, in turn, depends on the a priori smoothness restrictions on the conditional mean imposed by the researcher, including any smoothness constants. Our efficiency results in Section 2.5 imply that a priori specification of the smoothness constants is unavoidable, and we therefore recommend reporting CIs for a range of smoothness constants as a form of sensitivity analysis.

the estimator may not be negligible.

To make further progress on characterizing the weights, we focus on the case where the regression function is assumed to satisfy a Lipschitz constraint. We develop an algorithm that traces out the optimal weights as a function of the Lipschitz constant, analogous to the least angle regression algorithm for computing the LASSO solution path (Efron et al., 2004). In our empirical application, the algorithm computes the optimal estimators and CIs in a few minutes or less on a laptop computer. It follows from the form of this algorithm that the optimal estimator can be interpreted as a matching or kernel estimator with the number of matches varying between individuals and optimizing a bias-variance tradeoff. For a given sample size, when the Lipschitz constant is large enough, it becomes optimal to use a single match for each individual, and the optimal estimator reduces to a matching estimator with a single match.

The reason for asymptotic validity of our CIs is simple: because they are based on a linear estimator and account for its finite-sample bias, the CIs will be asymptotically valid—even in irregular and possibly set-identified cases—so long as the estimator is asymptotically normal, which in turn holds if the weights k_i satisfy a Lindeberg condition for the central limit theorem. However, since the weights k_i solve a bias-variance tradeoff, no single observation can have too much weight—otherwise, in large samples, a large decrease in variance could be achieved at a small cost to bias. On the other hand, asymptotic normality may fail under limited overlap for other estimators, and we show by example that this is the case for matching estimators.[4]

To formally show that conventional $\sqrt{n}$-asymptotics cannot be used when the dimension of the covariates is large relative to the smoothness of the regression function, we show that for $\sqrt{n}$-inference to be possible, one needs to bound the derivative of the conditional mean of order at least $p/2$. If one only bounds derivatives of lower order, the bias will asymptotically dominate the variance—in contrast to some nonparametric settings such as estimation of a conditional mean at a point, it is not possible to "undersmooth", and valid CIs need to take the bias into account. The smoothness condition is essentially the same 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. We then consider the asymptotic efficiency of competing

[4]These results have implications for the whether finite-sample corrections to the critical value under limited overlap, such as those proposed by Rothe (2017), are needed for asymptotic coverage in our setting. See Section 4.3.

estimators and CIs in this irregular setting, and we show that matching with a single match is asymptotically optimal under Lipschitz smoothness, so long as there is sufficient overlap between treated and untreated observations. On the other hand, we show that matching estimators may fail to be asymptotically efficient under insufficient overlap.

We illustrate the results in an application to the National Supported Work (NSW) Demonstration. We find that finite-sample optimal CIs are substantially different from those based on conventional $\sqrt{n}$-asymptotic theory, with bias determining a substantial portion of the CI width. Furthermore, our finite-sample approach allows us to investigate several questions that are moot under $\sqrt{n}$-asymptotic theory, due to the asymptotic equivalence of different estimators that achieve the semiparametric efficiency bound. For example, we examine how optimal estimators under the mean squared error (MSE) criterion differ from estimators used for optimal CIs, and we find that, in our application, the optimal CI oversmooths slightly relative to the MSE optimal estimator. We also examine alternative estimators and find that, under Lipschitz smoothness, matching estimators perform relatively well.

An important practical 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 $\sqrt{n}$-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.

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 general framework of Donoho (1994) and Cai and Low (2004) and allows us to apply sharp efficiency bounds in Armstrong and Kolesár (2018a). The form of the optimal estimator CIs follows by applying the general framework. The rest of our finite-

sample results, as well as all asymptotic results, are novel and require substantial further analysis. In particular, solving for the optimal weights k_i in general requires solving an optimization problem over the space of functions in p variables. Whereas simple strategies, such as gridding, are infeasible unless the dimension of covariates p is very small, we show that, for Lipschitz smoothness, the problem can be reduced to convex optimization in a finite number of variables and constraints, which depend only on the sample size and not on p.[5] Furthermore, our solution path algorithm uses insights from Rosset and Zhu (2007) on computation of penalized regression estimators to further speed up computation. In independent and contemporaneous work, Kallus (2017) computes optimal linear weights using a different characterization of the optimization problem.

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 bound 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. Limited overlap brings an additional layer of difficulty, and tail conditions on the distribution of the outcome variable (or the outcome variable divided by the propensity score) play a role in rates of convergence (see Khan and Tamer, 2010; Chaudhuri and Hill, 2016; Sasaki and Ura, 2017; Ma and Wang, 2018). 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, and that it is worth studying optimal estimation and inference

[5] While restricting attention to small p (say ≤ 2) in would be severely limiting in our setting, it is not restrictive in some other problems, such as regression discontinuity. For computation of optimal weights in other settings with small p, see Heckman (1988) and Imbens and Wager (2017).

in both settings, and instructive to compare the procedures. We provide such a comparison in the context of our empirical application in Section 6.4. Note also that, since our CIs are valid unconditionally, they can be used in either setting.[6]

The remainder of this paper is organized as follows. Section 2 presents the model and gives the main finite-sample results. Section 3 considers practical implementation issues. Section 4 presents asymptotic results. Section 5 discusses some possible extensions of our results. Section 6 discusses an application to the NSW data. Additional results, proofs and details of results given in the main text are given in appendices and the supplemental materials.

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

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

[6]While we define the treatment effect of interest to be one that conditions on realized covariates in the sample, our approach can be extended to construct valid CIs for the population ATE; see Section 5.1

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

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

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

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

with the (conditional on x_i and d_i) variance $\sigma^2(x_i, d_i)$ treated as known.[7]

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

$$\inf_{f\in\mathcal{F}} P_f(Lf \in \mathcal{C}) \ge 1 - \alpha, \tag{3}$$

where P_f denotes probability computed under f.

[7] 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

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

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 $\{(X_i', D_i)\}_{i=1}^n$.

2.2 Linear estimators

Consider an estimator that is linear in the outcomes y_i,

$$\hat{L}_k = \sum_{i=1}^{n} k(x_i, d_i) y_i. \tag{4}$$

This covers many estimators that are popular in practice, such as series of kernel estimators, or various matching estimators. For example, the matching estimator with M matches that matches (with replacement) on covariates 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 $\|\cdot\|_{\mathcal{X}}$), with the CATE estimate given by $L\hat{f}$. The form of $k(\cdot)$ for this estimator is given by

$$k_{\text{match},M}(x_i, d_i) = \frac{1}{n}(2d_i - 1)\left(1 + \frac{K_M(i)}{M}\right), \tag{5}$$

where $K_M(i)$ is the number of times the ith observation is matched. We begin by restricting attention to estimators that take the form (4), and to CIs based on such estimators. We then show, in Section 2.5 and Appendix A, that, provided the weights $k(\cdot)$ are optimally chosen, these 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 maximum bias

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

and variance $\text{sd}(\hat{L}_k)^2 = \sum_{i=1}^n k(x_i, d_i)^2 \sigma^2(x_i, d_i)$. By centrosymmetry of $\mathcal{F}$, $\inf_{f \in \mathcal{F}} E_f(\hat{L}_k - Lf) = -\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$, and 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 $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$ in addition to subtracting the usual normal quantile times its standard deviation—otherwise the CI will undercover for some $f \in \mathcal{F}$. A $100 \cdot (1 - \alpha)\%$ one-sided CI is therefore given by $[\hat{c}, \infty)$, where

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

and $z_{1-\alpha}$ denotes the $1 - \alpha$ quantile of a standard normal distribution.

One could form a two-sided CI centered around $\hat{L}_k$ by adding and subtracting $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) +$

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

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

Note that $\mathrm{cv}_\alpha(0) = z_{1-\alpha/2}$, 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 $\alpha \leq 0.2$, $\mathrm{cv}_\alpha(b) \approx b + z_{1-\alpha}$ up to three decimal places.[8] For large values of b, the CI is therefore approximately given by adding and subtracting $\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k) + z_{1-\alpha}\,\mathrm{sd}(\hat{L}_k)$ 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

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

The linear estimator that achieves the lowest RMSE is thus minimax optimal in the class of linear estimators (4). 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 $\sqrt{80\%} = 89.4\%$, (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

[8] The critical value $\mathrm{cv}_{1-\alpha}(b)$ be computed in statistical software as the square root of the $1-\alpha$ quantile of a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter b^2

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 (3), 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)[9]. 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.

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

For a given $\delta > 0$, let f^*_δ solve

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

and, with a slight abuse of notation, define

$$\hat{L}_\delta = \hat{L}_{k^*_\delta}, \qquad k^*_\delta(x_i, d_i) = \frac{f^*_\delta(x_i, d_i)/\sigma^2(x_i, d_i)}{\sum_{j=1}^n d_j f^*_\delta(x_j, d_j)/\sigma^2(x_j, d_j)}. \tag{9}$$

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 $k^*_{\delta_\rho}$, where δ_ρ 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), and then varying δ to find the optimal bias-variance tradeoff. Both of these optimization problems require 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, the solution to the first optimization problem can be found by solving a finite-dimensional linear program. The optimization problem (8) can be cast as a finite-dimensional convex program. Furthermore, if the program is put

into a Lagrangian form, then the solution is piecewise linear a function of the Lagrange multiplier, and one can trace the entire solution path $\{\hat{L}_\delta\}_{\delta>0}$ using an algorithm similar to the LASSO/LAR algorithm of Efron et al. (2004).

First, observe that in both optimization problems (6) and (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

$$f(x_i,d) - f(x_j,d) \le C\|x_i - x_j\|_{\mathcal{X}} \quad d \in \{0,1\},\ i,j \in \{1,\ldots,n\}. \tag{10}$$

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*

$$\sum_{i=1}^n d_i k(x_i,d_i) = 1 \quad \textit{and} \quad \sum_{i=1}^n (1-d_i)k(x_i,d_i) = -1. \tag{11}$$

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

$$\max_{f\in\mathbb{R}^{2n}} \left\{ \sum_{i=1}^n k(x_i,d_i)f(x_i,d_i) - \frac{1}{n}\sum_{i=1}^n [f(x_i,1) - f(x_i,0)] \right\}, \tag{12}$$

where the maximum is taken subject to (10) and

$$\sum_{i=1}^n f(x_i,1) = \sum_{i=1}^n f(x_i,0) = 0. \tag{13}$$

Furthermore, if $k(x_i,d_i) \ge 1/n$ if $d_i = 1$ and $k(x_i,d_i) \le -1/n$ if $d_i = 0$, it suffices to impose the following subset of the constraints in (10):

$$f(x_i,1) \le f(x_j,1) + C\|x_i - x_j\|_{\mathcal{X}}, \quad \textit{all } i,j \textit{ with } d_i = 1,\ d_j = 0 \textit{ and } k(x_i,1) > 1/n, \tag{14}$$

$$f(x_i,0) \le f(x_j,0) + C\|x_i - x_j\|_{\mathcal{X}}, \quad \textit{all } i,j \textit{ with } d_i = 1,\ d_j = 0 \textit{ and } k(x_j,1) < -1/n. \tag{15}$$

The assumption that $\hat{L}_k$ satisfies (11) is necessary to prevent the bias from becoming arbitrarily large at multiples of $f(x,d) = d$ and $f(x,d) = 1-d$. If (11) 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 (13), 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 (12).

The last part of the theorem follows by checking that the remaining constraints in (10) are automatically satisfied at the optimum (see Lemma A.2). The conditions (14) and (15) give at most $2n_0n_1$ inequalities, where n_d is the number of observations with $d_i = d$. The condition on the weights k holds, for example, for the matching estimator given in (5). Since for the matching estimator $k(x_i,d_i) = (2d_i-1)/n$ if observation i is not used as a match, the theorem says that one only needs to impose the constraint (10) for pairs of observations with opposite treatment status, and for which one of the observations is used as a match. Consequently, in settings with imperfect overlap, in which many observations are not used as a match, the number of constraints will be much lower than $2n_0n_1$.

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

Theorem 2.2. *Given $\delta > 0$, the value of the maximizer f^*_δ of (8) at $\{x_i,d_i\}_{i=1}^n$ is given by the solution to the convex program*

$$\max_{f \in \mathbb{R}^{2n}} 2Lf \quad s.t. \quad \sum_{i=1}^n \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)} \le \frac{\delta^2}{4} \quad \text{and s.t. (10).} \tag{16}$$

Furthermore, if $\sigma^2(x,d)$ doesn't depend on x, it suffices to impose the constraints (10) for $i,j \in \{1,\dots,n\}$ with $d_i = 0$ and $d_j = 1$, and the solution path $\{f^*_\delta\}_{\delta>0}$ can be computed by the piecewise linear algorithm given in Appendix A.3.

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. If the variance is homoscedastic for each treatment group, then the number of linear constraints can be reduced to $2n_0n_1$, and the entire solution path can be computed efficiently using the piecewise linear algorithm given in Appendix A.3.

As we discuss in more detail in Appendix A.3, it follows from the algorithm that the optimal estimator can be interpreted as matching (or kernel) estimator with a variable number of

matches, where the number of matches for each observation i increases with δ, and depends on the number of observations with opposite treatment status that are close to i according to a matrix of "effective distances". The "effective distance" between i and j increases in the number of times an observation j has been used as a match. Thus, observations for which there exist more good matches receive relatively more matches, since this decreases the variance of the estimator at a little cost in terms of bias. Also, since the weight $k(x_j, d_j)$ on j is increasing in the number of times it has been used as a match, using it more often as a match increases the variance of the estimator. Using the "effective distance" matrix trades off this increase in the variance against an increase in the bias that results from using a lower-quality match instead.

If the constant C is large enough, the increase in the bias from using more than a single match for each i is greater than any reduction in the variance of the estimator, and the optimal estimator takes the form of a matching estimator with a single match:

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 $\operatorname{argmin}_{j:\, d_j \neq d_i} \|x_i - x_j\|_{\mathcal{X}}$ 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 $M = 1$.*

In contemporaneous work, Kallus (2017) gives a similar result using a different method of proof. In the other direction, as $C/\delta \to 0$, the optimal estimator $\hat{L}_\delta$ converges to the difference-in-means estimator that takes the difference between the average outcome for the treated and the average outcome for the untreated units.

For the optimality result in Theorem 2.3, it is important that the metric on x used to define the matching estimator is the same as the one used to define the Lipschitz constraint. Zhao (2004) has argued that conditions on the regression function should be considered when defining the metric used for matching. Theorem 2.3 establishes a formal connection between conditions on the regression function and the optimal metric for matching. We investigate this issue further in the context of our empirical application by calculating the efficiency loss from matching with the "wrong" metric (see Section 6.5).

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 is too restrictive: 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 if $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$. This rules out 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 $\mathcal{F}$. The sharp bound follows from general results in Armstrong and Kolesár (2018a), and it gives a benchmark for the scope for improvement over the FLCI in Theorem 2.2. Theorem A.3 also gives a universal lower bound for this sharp bound.

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 71.7% 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 our empirical application in Section 6, we find that the efficiency of the FLCI is over 97% at such smooth functions g, both in our baseline specification, and for the experimental sample considered in Section 6.4. This implies 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 $f \in \mathcal{F}$. It is thus not possible to, say, 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 6. This also mirrors the common practice of reporting results for different specifications of the regression function in parametric regression problems.

The key assumption underlying these efficiency bounds is that the parameter space $\mathcal{F}$ be convex and centrosymmetric. This holds for the function class $\mathcal{F}_{\text{Lip}}(C)$, and, more generally, for parameter spaces that place bounds on derivatives of f. 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 Practical implementation

The estimators and CIs we have constructed require prior knowledge of the variance function $\sigma^2(x,d)$. Furthermore, the theoretical justification for these CIs relies on normality of the error distribution. This section discusses the implementation of our CIs in the more realistic setting where $\sigma^2(x,d)$ is unknown. We also discuss other implementation issues. In Section 4, we provide an asymptotic justification for our CIs if $\sigma^2(x,d)$ is unknown and the errors may be non-normal.

To implement feasible versions of our CIs when $\sigma^2(x,d)$ is unknown, we propose the following:[10]

1. Let $\tilde{\sigma}^2(x,d)$ be an initial (possibly incorrect) estimate or guess for $\sigma^2(x,d)$. As a default choice, we recommend taking $\tilde{\sigma}^2(x,d) = \hat{\sigma}^2$ where $\hat{\sigma}^2$ is an estimate of the variance computed under the assumption of homoskedasticity.

2. Compute the optimal weights $\{\tilde{k}^*_\delta\}_{\delta>0}$ based on the piecewise linear solution path $\{\tilde{f}^*_\delta\}_{\delta>0}$ in Appendix A.3, computed with $\tilde{\sigma}^2(x,d)$ in place of $\sigma^2(x,d)$. Let $\tilde{L}_\delta = \sum_{i=1}^n \tilde{k}^*_\delta(x_i,d_i)y_i$ denote the corresponding estimator, $\widetilde{\text{sd}}^2_\delta = \sum_{i=1}^n \tilde{k}^*_\delta(x_i,d_i)^2\tilde{\sigma}^2(x_i,d_i)$ denote its variance computed using $\tilde{\sigma}^2(x,d)$ as the variance function, and let $\overline{\text{bias}}_\delta = \overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\tilde{L}_\delta)$ denote its worst-case bias (which doesn't depend on the variance specification).

3. Compute the minimizer $\tilde{\delta}_\rho$ of $\overline{\text{bias}}^2_\delta + \widetilde{\text{sd}}^2_\delta$ and the minimizer $\tilde{\delta}_\chi$ of $\text{cv}_\alpha(\overline{\text{bias}}_\delta/\widetilde{\text{sd}}_\delta)\widetilde{\text{sd}}_\delta$.

[10] An R package implementing this procedure, including an implementation of the piecewise linear algorithm is available at https://github.com/kolesarm/ATEHonest.

Compute the standard error $\text{se}(\tilde{L}_{\tilde{\delta}_\chi})$ using the robust variance estimator

$$\text{se}(\tilde{L}_\delta)^2 = \sum_{i=1}^{n} \tilde{k}_\delta^*(x_i, d_i)^2 \hat{u}_i^2, \tag{17}$$

where $\hat{u}_i^2$ is an estimate of $\sigma^2(x_i, d_i)$. Report the estimate $\tilde{L}_{\tilde{\delta}_\rho}$, and the CI

$$\big\{\tilde{L}_{\tilde{\delta}_\chi} \pm \text{cv}_\alpha(\overline{\text{bias}}_{\tilde{\delta}_\chi} / \text{se}(\tilde{L}_{\tilde{\delta}_\chi}))\text{se}(\tilde{L}_{\tilde{\delta}_\chi})\big\}. \tag{18}$$

The values $\tilde{\delta}_\chi$ and $\tilde{\delta}_\rho$ depend on the initial variance estimate $\tilde{\sigma}^2(x, d)$, and the resulting estimator and CI will not generally be optimal if this initial estimate is incorrect. However, because the standard error estimator (17) does not use this initial estimate, the resulting CI will be asymptotically valid even if $\tilde{\sigma}^2(x, d)$ is incorrect. As an estimator of the conditional variance in (17), we can take $\hat{u}_i^2 = (y_i - \hat{f}(x_i, d_i))^2$, where $\hat{f}(x, d)$ is a consistent estimator of $f(x, d)$, or the nearest-neighbor variance estimator of Abadie and Imbens (2006) $\hat{u}_i = J/(J+1) \cdot (y_i - \hat{f}(x_i, d_i))^2$, where $\hat{f}(x_i, d_i)$ average outcome of J observations (excluding i) with treatment status d_i that are closest to i according to some distance $\|\cdot\|$. Taking the initial estimate $\tilde{\sigma}^2(x, d)$ to be constant as a default choice mirrors the practice in the linear regression model of computing ordinary least squares estimates with heteroskedasticity robust standard errors (see Section 3.3 for further discussion and Section 6 for the particular implementation in our application).

3.1 Additional practical considerations

In forming our CI, we need to choose the function class $\mathcal{F}$. While we have focused on Lipschitz classes, we still need to complete the definition of $\mathcal{F}$ by choosing the constant C and the norm on x used to define the Lipschitz condition. The results discussed in Section 2.5 imply that it is not possible to make these choices automatically in a data-driven way. Thus, we recommend that these choices be made using problem-specific knowledge wherever possible, and that CIs be reported for a range of plausible values of C as a form of sensitivity analysis. We consider these problems in more detail in the context of our application in Sections 6.1 and 6.5.

Another issue that arises in reporting the CIs and estimators in this paper is that different criteria lead to different estimators, so that the RMSE optimal estimator will, in general, differ from the estimator used to form the one- and two-sided CIs. In our empirical appli-

cation, we find that this does not matter much: in all the specifications we consider, the RMSE optimal estimator does not differ very much from the estimators used to construct CIs. However, reporting multiple estimates for different criteria can be cumbersome. To avoid recomputing estimates for different criteria, one can simply compute the CI (18) using the choice $\tilde{\delta}_\rho$ optimized for RMSE. The resulting CIs will then be based on the same estimator reported as a point estimate. While there is some efficiency loss in doing this, in our main specification in the empirical application in Section 6, we find that the resulting CI is less than 2% longer than the one that reoptimizes δ for CI length.

3.2 CIs based on other estimators

To form a feasible CI based on a linear estimator $\hat{L}_k = \sum_{i=1}^n k(x_i, d_i)y_i$, one can simply follow the same steps, using the weights $k(x_i, d_i)$ in Equation (17), and computing the worst-case bias by solving the optimization problem in Theorem 2.1. If $\hat{L}_k$ is an estimator that achieves the semiparametric efficiency bound under standard asymptotics, 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 for f.

If one applies this method to form a feasible CI based on matching estimators, one can determine the number of matches M that leads to the shortest CI (or smallest RMSE) as in Steps 2 and 3 of the procedure, with M playing the role of δ. In our application, we compare the length of the resulting CIs to those of the optimal FLCIs. Although Theorem 2.3 implies the matching estimator with a single match is suboptimal unless C is large enough, we find that, in our application, the efficiency loss is modest.

3.3 Efficiency of feasible estimators and CIs

We motivated our estimators and CIs using efficiency and coverage results under the assumption of normal errors and known variance. In what sense can the feasible versions of these procedures in this section be considered efficient and valid? In Section 4.2, we consider the asymptotic validity of the feasible versions of our CIs, when the errors may be non-normal. We find that these CIs are asymptotically valid even in "irregular" settings when $\sqrt{n}$-inference is impossible, and in cases in which the ATE is not point identified (in which case our CI has asymptotically valid coverage for points in the identified set).

Regarding finite sample optimality, note that all the arguments in Section 2 regarding bias-variance tradeoffs for linear estimators still go through so long as the variance function is correctly specified, even if the errors are not normal. Thus, if one constructs a feasible estimator using a choice of the variance function $\tilde{\sigma}^2(x, d)$ specified a priori and this guess turns out to be correct, the resulting estimator will be optimal among linear estimators even with non-normal errors. If one uses a guess $\tilde{\sigma}^2(x, d)$ that is correct up to scale (for example, if one guesses correctly that errors are homoskedastic, but one uses the wrong variance), the resulting estimate may put too much weight on bias or variance relative to the given criterion, but it will still minimize variance among all estimators with the same worst-case bias. Note also that the minimax optimality results among all estimators (including nonlinear ones) discussed in Section 2 are valid if one fixes the variance function, even if errors can be non-normal, so long as the set of possible error distributions includes normal errors (since the minimax risk of linear estimators depends on the error distribution only through its variance).

These results mirror the linear model, in which the ordinary least squares estimator is optimal in finite samples under homoskedasticity if one assumes normal errors, or if one restricts attention to linear estimators. One can then form CIs based on this estimator using heteroskedasticity robust standard errors, which leads to CIs that are asymptotically valid for heteroskedastic and non-normal errors. Similarly, our feasible procedure leads to estimators that have finite-sample optimality properties under homoskedasticity (if one uses a constant function $\tilde{\sigma}^2(x, d)$ in step 1), along with CIs that are valid under more general conditions. Alternatively, one could use a more flexible estimate $\tilde{\sigma}^2(x, d)$, mirroring the suggestion of Wooldridge (2010) and Romano and Wolf (2017) to report feasible generalized least squares estimates in the linear model, along with heteroskedasticity robust standard errors.

4 Asymptotic results

This section considers the asymptotic validity of feasible CIs with unknown error distribution, as well as bounds on the rate of convergence of estimators and CIs. In Section 4.1, we show formally that $\sqrt{n}$-inference is impossible in our setting when the number of continuous covariates in x_i is large enough relative to the order of smoothness imposed by $\mathcal{F}$. In Section 4.2, we show that our feasible CIs are asymptotically valid and centered at asymptotically normal estimators. Importantly, the conditions for this asymptotic validity result allow for irregular cases where conventional CIs suffer from asymptotic undercoverage, due

to imperfect overlap or high-dimensional covariates. Sections 4.3 and 4.4 give conditions for asymptotic validity and optimality of CIs based on matching estimators.

4.1 Impossibility of $\sqrt{n}$-inference under low smoothness

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 (2) 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 $\sqrt{n}$-consistent estimators with asymptotically negligible bias. Furthermore, Hahn (1998) shows that no regular $\sqrt{n}$-consistent estimator can have asymptotic variance lower than the linear estimator with the kernel $k_{\text{seb}}(x_i, d_i) = n^{-1}[d_i/e(x_i) - (1-d_i)/(1-e(x_i))]$. The asymptotic variance of this linear estimator is known as the semiparametric efficiency bound.[11]

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), $\sqrt{n}$-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 $\sqrt{n}$-inference should be impossible unless $\gamma_f \geq p/2$ (i.e. the conditions for impossibility of $\sqrt{n}$-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 4.1. *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 (2) holds conditional on*

[11] The results of Hahn (1998) apply to estimation of the ATE, rather than the CATE. Crump et al. (2009) give a formulation for the CATE, although they do not give a formal statement.

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.

The theorem shows that the excess length of a CI 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 $\sqrt{n}$-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).

4.2 Asymptotic validity of feasible optimal CIs

The following theorem gives sufficient conditions for the asymptotic validity of the feasible CIs given in Section 3 based on the estimator $\tilde{L}_\delta$ 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 4.2. *Consider the model* (1) *with* $1/K \leq Eu_i^2 \leq K$ *and* $E|u_i|^{2+1/K} \leq K$ *for some constant* K*. Suppose that*

$$\text{for all } \eta > 0, \min_{1\leq i\leq n} \#\{j \in \{1,\ldots,n\} : \|x_j - x_i\|_{\mathcal{X}} \leq \eta/C_n,\, d_i = d_j\} \to \infty, \tag{19}$$

and that the variance function $\sigma^2(x,d)$ is uniformly continuous in x for $d \in \{0,1\}$. Let $\mathcal{C}$ be the CI in Equation (18) based on the feasible optimal estimator $\tilde{L}_\delta$ with $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_n)$, with δ fixed and $\tilde{\sigma}^2(x,d)$ a nonrandom function bounded away from zero and infinity. Suppose the estimator $\hat{u}_i^2$ in (17) is the nearest-neighbor variance estimator based on a fixed number of nearest neighbors J, or that $\hat{u}_i^2 = (y_i - \hat{f}(x_i,d_i))^2$, where $\hat{f}(x_i,d_i)$ the Nadaraya-Watson estimator with uniform kernel and a bandwidth sequence h_n with $h_n C_n$ converging to zero slowly enough. Then $\liminf_{n\to\infty} \inf_{f\in\mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1-\alpha$.

The conditions of Theorem 4.2 are fairly weak. In particular, if $C_n = C$ does not change with n, it suffices for x_i do be drawn from a distribution with bounded support:

Lemma 4.1. *Suppose that (x_i, d_i) is drawn i.i.d. from a distribution where x_i has bounded support and $0 < P(d_i = 1) < 1$, and that $C_n = C$ is fixed. Then (19) holds almost surely.*

In particular, feasible optimal CIs are asymptotically valid in irregular settings, including high dimensional x_i (as in Theorem 4.1) or imperfect overlap (as in Khan and Tamer, 2010) including set identification due to complete lack of overlap. The "irregular" nature of the setting only shows up in the critical value cv_α, which will remain strictly larger than the conventional $z_{1-\alpha/2}$ critical value even asymptotically, reflecting the fact that the worst-case bias is asymptotically non-negligible.

4.3 Asymptotic validity of CIs based on matching estimators

We now consider asymptotic validity of feasible CIs based on matching estimators with a fixed number of matches.

Theorem 4.3. *Suppose that the conditions of Theorem 4.2 hold. Let $\mathcal{X}$ be a set with $x_i \in \mathcal{X}$ all i. Let $\overline{G} : \mathbb{R}^+ \to \mathbb{R}^+$ and $\underline{G} : \mathbb{R}^+ \to \mathbb{R}^+$ be functions with $\lim_{t\to 0} \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}] = 0$. Suppose that, for any sequence a_n with $n\underline{G}(a_n)/\log n \to \infty$, we have*

$$\underline{G}(a_n) \leq \frac{\#\{i : \|x_i - x\|_{\mathcal{X}} \leq a_n,\, d_i = d\}}{n} \leq \overline{G}(a_n) \quad \text{all } x \in \mathcal{X},\, d \in \{0,1\} \tag{20}$$

for large enough n. Let $\mathcal{C}$ be the CIs in Section 3.2 based on the matching estimator with a fixed number of matches M, and $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_n)$. Then $\liminf_{n\to\infty} \inf_{f\in\mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1-\alpha$.

Theorem 4.3 is related to results of Abadie and Imbens (2006) on asymptotic properties of matching estimators with a fixed number of matches. Abadie and Imbens (2006) note

that, when p is large enough, the bias term will dominate, so that conventional CIs based on matching estimators will not be valid. In contrast, the CIs in Theorem 4.3 are widened to take into account worst-case bias, and so they achieve coverage even when p is large. Alternatively, one can attempt to restore asymptotic coverage by subtracting an estimate of the bias based on higher-order smoothness assumptions. While this can lead to asymptotic validity when additional smoothness is available (Abadie and Imbens, 2011), it follows from Theorem 4.1 that such an approach will lead to asymptotic undercoverage under some sequence of regression functions in the Lipschitz class $\mathcal{F}_{\text{Lip}}$.

Theorem 4.3 requires the additional condition (20). By Theorem 37 of Chapter 2 of Pollard (1984), this condition will hold almost surely if (x_i, d_i) are drawn i.i.d. from a distribution where $\underline{G}(a)$ and $\overline{G}(a)$ are lower and upper bounds (up to constants) for $P(\|x_i - x\|_{\mathcal{X}} \leq a,\, d_i = d)$ for x on the support of x_i. The condition $\lim_{t \to 0} \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}] = 0$ can then be interpreted as an overlap condition on the distribution of (x_i, d_i). In particular, if x_i has a density bounded away from zero and infinity on a sufficiently regular support, then a sufficient condition is for the propensity score $e(x)$ to be bounded away from zero and one on the support of x_i.

On the other hand, if there is not sufficient overlap, then (20) will fail, and this can lead to failure of asymptotic normality for the matching estimator. As an extreme example, suppose that $p = 1$ and that $x_i < x_j$ for all observations where $d_i = 0$ and $d_j = 1$. Then each observation with $d_i = 0$ will be matched to the observation with the smallest value of x_j among observations with $d_j = 1$. Thus, the weight $k_{\text{match},M}(x_j, d_j)$ for the observation with the smallest value of x_j among observations with $d_j = 1$ will be bounded away from zero, so that the matching estimator will not be asymptotically normal. In contrast, it follows from Lemma 4.1 that the feasible estimator with optimal weights will be asymptotically normal even when there is no overlap between the distribution of x_i for treated and untreated observations.

Rothe (2017) argues that, in settings with limited overlap, estimators of the CATE may put a large amount of weight on a small number of observations. As a result, standard approaches to inference that rely on normal asymptotic approximations to the distribution of the t-statistic will be inaccurate in finite samples. Our results shed some light on when such concerns are relevant. The above example shows that such concerns may indeed persist—even in large samples—if one uses a matching estimator with a fixed number of matches. However, it follows from Theorem 4.2 that this will generally not be the case for optimal estimators and CIs, even in the case with limited overlap, at least in large samples if the

Lipschitz constant C is fixed or doesn't increase too quickly with n. In a given application, one can check this directly by examining the weights k_i. Furthermore, it follows from the proof of Theorem 4.3 that when $p > 2$, bias will dominate variance asymptotically even if one attempts to "undersmooth" by using a matching estimator with a single match. In such settings, it is important to widen the CIs to take the bias into account, in addition to accounting for the potential inaccuracy of the normal asymptotic approximation, using methods such as those proposed in Rothe (2017).[12]

4.4 Asymptotic efficiency of matching estimator with one match

We have seen that the matching estimator with $M = 1$ is efficient in the Lipschitz class when the constant C is large enough. Here, we give conditions for asymptotic optimality of this estimator.

Theorem 4.4. *Let $\{(X_i, D_i)_{i=1}^n\}$ be drawn such that the conditions of Theorem 4.1 hold, and such that the Gaussian regression model (1) and (2) holds conditional on $\{(X_i, D_i)_{i=1}^n\}$, with $\sigma^2(x, d)$ bounded away from zero and infinity. Suppose that, for some functions $\overline{G} : \mathbb{R}^+ \to \mathbb{R}^+$ and $\underline{G} : \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{t\to 0} \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}]^{2/p+1} = 0$,*

$$\underline{G}(a) \leq P(\|X_i - x\|_{\mathcal{X}} \leq a,\, D_i = d) \leq \overline{G}(a).$$

Let $R^*_{n,match,MSE}$ denote the worst-case MSE of the matching estimator with $M = 1$, and let $R^*_{n,opt,MSE}$ denote the minimax MSE among linear estimators, conditional on $\{(X_i, D_i)\}_{i=1}^n$, for the class $\mathcal{F}_{\text{Lip}}(C)$. Then $R^*_{n,match,MSE}/R^*_{n,opt,MSE} \to 1$ almost surely. The same holds with "MSE" replaced by "FLCI length" or "β quantile of excess length of a one-sided CI."

If X_i has sufficiently regular support and the conditional density of X_i given D_i is bounded away from zero on the support of X_i for both $D_i = 0$ and $D_i = 1$, then the conditions of Theorem 4.4 will hold with $\underline{G}(a)$ and $\overline{G}(a)$ both given by constants times a^p, so that $\overline{G}(\underline{G}(a))$ decreases like a as $a \to 0$. Thus, the conditions of Theorem 4.4 will hold so long as $p > 2$, which corresponds to the case in Theorem 4.1 in which $\sqrt{n}$-inference is impossible, even with a finite semiparametric efficiency bound. On the other hand, matching with $M = 1$ is suboptimal when $p = 1$, since the semiparametric efficiency bound can be achieved and, as noted by Abadie and Imbens (2006), matching with a fixed number of matches does not

[12] The CIs proposed by Rothe (2017) require perfect matches, which requires discretizing the covariates if they are continuous. This will increase the worst-case bias relative to matching on the original covariates with a single match, and so the same comment applies to the estimator based on discretized covariates.

achieve the semiparametric efficiency bound. In addition, the conditions of Theorem 4.4 will fail if there is insufficient overlap in the support of X_i when $D_i = 1$ and when $D_i = 0$, and this may lead to asymptotic inefficiency.

5 Extensions

This section discusses some possible extensions of our framework.

5.1 Population average treatment effects

Our setup obtains if we condition on observations $(x_i, d_i) = (X_i, D_i)$ where (X_i, D_i) are drawn independently from some large population, and we focus on the CATE for this sample. However, we may be interested in the average treatment effect for entire population. The population average treatment effect (PATE) is given by $E[y_i(1) - y_i(0)] = E[E[y_i(1) - y_i(0)|\{(X_i, D_i)\}_{i=1}^n]] = E[Lf]$ where $Lf = \text{CATE}(f)$ is the CATE. This suggests that one can use the same estimates for the PATE as for the CATE, with the caveat that one must take into account the additional variation of the CATE around the PATE when computing CIs and evaluating the performance of these estimates.

Our approach gives an estimate $\hat{L}$ of the CATE and upper bounds $\overline{\text{bias}}_{\mathcal{F}}(\hat{L})$ on the bias of this estimate conditional on $\{(X_i, D_i)\}_{i=1}^n$, as well as a standard error $\text{se}(\hat{L})$ for the conditional standard deviation of $\hat{L}$ given $\{(X_i, D_i)\}_{i=1}^n$. To form a one-sided CI $[\hat{c}_{\text{PATE}}, \infty)$ for the PATE, we need to subtract an upper bound on the bias of $\hat{L}$ that holds unconditionally, as well as an estimate of the unconditional standard deviation of $\hat{L}$. Since $E\,\overline{\text{bias}}_{\mathcal{F}}(\hat{L})$ is an upper bound on the unconditional bias, one can subtract $\overline{\text{bias}}_{\mathcal{F}}(\hat{L})$ as before, and simply modify the CI by subtracting $z_{1-\alpha}$ times an estimate of the unconditional standard deviation of $\hat{L}$: $\hat{c}_{\text{PATE}} = \hat{L} - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}) - z_{1-\alpha}\text{se}_{\text{PATE}}(\hat{L})$ where $\text{se}_{\text{PATE}}(\hat{L})^2 = \text{se}(\hat{L})^2 + \hat{V}_2$ is an estimate of the unconditional variance of $\hat{L}$, formed using an estimate $\hat{V}_2$ of the variance of the CATE. Estimates $\hat{V}_2$ of the variance of the CATE have been proposed by Abadie and Imbens (2006), who take this approach to forming standard errors for matching estimators. For two-sided CIs, one can take a similar approach, although, without further analysis of the behavior of the bias, one must take the slightly more conservative approach of adding and subtracting $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}) + z_{1-\alpha/2}\text{se}_{\text{PATE}}(\hat{L})$ rather than using the critical value $\text{cv}_\alpha(\overline{\text{bias}}/\text{se})$, since $\overline{\text{bias}}$ is now a random variable that depends on $\{(X_i, D_i)\}_{i=1}^n$. We conjecture that this approach will lead to valid CIs for the PATE, and that these CIs will be close to efficient when no

additional information is given on the propensity score, and we leave the question of formally verifying it for future research.

5.2 Treatment effects on subsamples and alternative populations

We have focused on the CATE, which averages the treatment effect $\tau(x) = f(x,1) - f(x,0)$ conditional on $x_i = x$ over the sample. In our application, we focus on the conditional average treatment effect on the treated (CATT). Both of these take the form of a weighted average treatment effect $\sum_{i=1}^{n} w_i[f(x_i,1) - f(x_i,0)]$ for some known weights w_i, and we state our results in Appendix A in this general framework. More generally, a mild extension of our framework covers estimation of the average treatment effect $L_F = \int [f(x,1) - f(x,0)]\, dF(x)$ weighted by any known distribution F. In particular, if we are interested in the effect of a counterfactual policy in which we apply the treatment to a new population where $x_i \sim F$, then this gives the policy relevant effect for treating this counterfactual population.

A key advantage of our approach is that we can allow the counterfactual distribution F to have completely different support from the distribution of x_i observed in the sample. In such settings, our approach uses the observed data to optimally extrapolate the treatment effect to the support of F. The counterfactual effect will then be set identified, and our approach gives confidence intervals that contain points in the identified set for the counterfactual policy effect L_F.

When a policy-relevant F requires such extrapolation, the resulting confidence interval may be wide. Of course, this reflects the inherent uncertainty in the problem of extrapolating a conditional expectation function outside of the support of the conditioning variable. One can tighten the set by making further restrictions on the regression function $f(x,d)$, and our framework will still apply so long as these restrictions are convex. Alternatively, one can settle for a less policy-relevant F that has greater overlap with the support of the observed data, thereby leading to a tighter CI. In particular, one can consider average treatment effects over a subset $\mathcal{S}$ contained in the support of x_i conditional both on $d_i = 0$ and $d_0 = 1$, as discussed, for example, in Heckman et al. (1997) and Crump et al. (2009). One can also extend our framework to find the weighting distribution F that minimizes estimation error or CI width for L_F, thereby solving the finite-sample version of a problem considered in Crump et al. (2006); if we optimize over a convex set, then this leads to another convex optimization problem.

5.3 Generalizations of the regression model

We can generalize our setup to allow the treatment d_i and covariates x_i to take values in arbitrary sets, so long as we model the conditional mean of $y_i(d)$ given $x_i = x$ as a function $f(x, d)$ that we restrict to some convex set. In particular, we can allow for multivalued or even continuously distributed treatments d_i, and consider treatment effects of moving between any two values d and d', or the average counterfactual policy outcome $\int\int f(x, d)\, dF(x, d)$ under some joint distribution $F(x, d)$ of treatments and covariates, following Stock (1989). As discussed above, a particular advantage of our approach in this setting is that it allows for and automatically incorporates extrapolation of the treatment effect outside of the observed support of (x_i, d_i).

In such settings, particularly when d_i is continuously distributed, one will want to restrict the variation of $f(x, d)$ as a function of both x and d. In addition to smoothness assumptions, one can impose separability and linearity in certain variables, as in the partly linear model $f(x, d) = d'\beta + g(x)$. These restrictions also amount to restricting the regression function to a convex set, and therefore fall into the framework used here.

We can also allow for non-Euclidean covariates x_i. For example, x_i may describe the entire the social network of an individual i, as in Auerbach (2018). Indeed, the Lipschitz assumption we use in our main analysis can be generalized directly to any metric space, and the dimension of the optimization problem used to compute the optimal estimator scales with the number of observations, rather than the dimension of x_i. Thus, there is no additional computational burden of applying our method under a Lipschitz assumption when x_i is high dimensional.

5.4 Experimental design

So far, we have taken the treatment assignments d_i as given, which is appropriate for observational data in which the researcher cannot choose treatment assignments. In the experimental design setting, a researcher observes covariates x_i and can choose the treatment assignments d_i in order to optimize the performance of estimators and CIs. If one evaluates performance after conditioning on the realized values of d_i, as suggested by Kasy (2016), this can be done by computing the minimax MSE or CI width under a given assignment $\{x_i, d_i\}_{i=1}^n$ using a guess for the variance function $\sigma^2(x, d)$, and optimizing over treatment assignments $d_1, \ldots, d_n$. While the first step can be done quickly using our methods, the second step involves a non-convex optimization problem over a discrete choice set with 2^n

elements, and so solving this problem exactly will typically not be computationally feasible. Nonetheless, one can still use our methods to optimize over a smaller set of candidate treatment assignments.

The resulting optimal treatment assignments $d_1, \ldots, d_n$ will be a deterministic function of the covariates $x_1, \ldots, x_n$. If one instead evaluates performance without conditioning on the realized values of d_i, then the optimal treatment rule will typically involve randomized treatment assignment (see Blackwell and Girshick, 1954, Section 8.7). The resulting CIs will be tighter, at the cost of covering only unconditionally. While we do not take a stance on whether conditional or unconditional coverage is appropriate (see Banerjee et al., 2017, footnote 11 for a recent discussion), we note that our methods can be used to quantify the cost in terms of CI length of requiring conditional coverage; we take up this question in our application in Section 6.4.

6 Empirical Application

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).[13] 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 Lf corresponding to $\text{CATT}(f)$ rather than the CATE (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.[14]

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

[14] 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).

6.1 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

$$\|x\|_{A,p} = \left(\sum_{i=1}^{n} \left| \left(A^{1/2} x\right)_i \right|^p \right)^{1/p}, \tag{21}$$

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, j) 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 1 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 6.5. 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.

6.2 Results

We compute the estimator $\tilde{L}_\delta$ as described in Section 3 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 $\hat{u}_i^2$ is the

nearest-neighbor estimate with $J = 3$ 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 4.2, 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 $\alpha = 0.05$ (see Appendix A). 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, 2018a), 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 1 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 $C \geq 0.2$, the point estimate is positive and remarkably stable as a function of C, ranging between 0.94 and 1.15, which suggests that the estimator and CIs are accounting for the possibility of selection bias by controlling for observables. The two-sided CIs become wider as C increases, which, as can be seen from the figure, is due to greater potential bias resulting from a less restrictive parameter space.

Figure 2 focuses on the case where $C = 1$ and plots the optimal estimator $\tilde{L}_\delta$ along with its standard deviation, worst-case bias, RMSE and CI length as a function of δ (recall that δ determines the relative weight on variance in the bias-variance tradeoff). 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. It can be seen from the figure that while the bias is increasing in δ and the variance is decreasing, the optimal resolution of the bias-variance trade-off depends on the criterion. Interestingly, the optimal δ is smallest for the RMSE criterion—it is cheaper, in terms of CI length or excess length, to use an estimator with *larger* bias and smaller variance than the RMSE-optimal estimator, and to take this bias into account by widening the CI. The resolution of the bias-variance trade-off is also different between the one-sided and two-sided CIs. The CIs are thus based on different estimators, which explains why for some values of C in Figure 1, the lower one-sided CI is below the lower endpoint of the two-sided CI.

Table 2 reports the point estimates that optimize each of the criteria plotted in Figure 2. These are simply the estimates from Figure 2 taken at the value of δ that minimizes the given criterion in the corresponding plot in the figure. In all cases, the bias is non-negligible relative to variance: this is consistent with Theorem 4.1, which implies that given the Lipschitz smoothness assumption and the dimension of the covariates, the semiparametric efficiency bound cannot be achieved here, and it is not possible to construct asymptotically unbiased estimators. Our CIs reflect this by explicitly taking the bias into account.

For comparison, Figure 3 plots the analog of Figure 2 for the matching estimator 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). Table 2 then reports the point estimates at M that optimizes each of the criteria plotted in Figure 2. As was the case for the optimal estimator, for the construction of one- and two-sided CIs, it is again optimal to oversmooth in that the optimal number of matches is greater than the RMSE-optimal number of matches.

According to Theorem 2.3, matching with $M = 1$ is efficient when C is "large enough". In our application, for $C \geq 2.9$, the efficiency of the matching estimator is at least 95% for all performance criteria, and it's at least 99.2% for RMSE.[15] Matching with $M = 1$ leads to a modest efficiency loss in our main specification, where $C = 1$: its efficiency is 89.8% for RMSE, and 85.5% for the construction of two-sided CIs.

6.3 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. In the experimental sample, the difference in means between the outcome for the

[15] In this application, matching with $M = 1$ is never 100% efficient even for large values of C since the condition that each unit has a single closest match is violated: there are multiple observations in the dataset that have the same covariate values. Consequently, $\lim_{\delta \to 0} \tilde{L}_\delta = 1.41$ is slightly different from 1.42, the matching estimate based on a single match.

treated and untreated individuals is 1.794. Treating this estimator as an estimator of the CATT, the (unconditional) robust standard error is 0.670 and non-robust standard error is 0.632.[16]

The estimates in Table 2 based on the optimal and the matching estimators are slightly lower, although the distance between the estimate and the experimental estimate is much smaller than the worst-case bias. Consequently, all of the difference between the estimates can be explained by the bias alone. The large value of the worst-case bias also suggests that the goal of recovering the experimental estimates from the NSW non-experimental data is too ambitious, unless one imposes substantially stronger smoothness assumptions. Furthermore, differences between the estimates reported here and the experimental estimate may also arise from (1) failure of the selection on observables assumption; and (2) the sampling error in the experimental and non-experimental estimates.

6.4 Optimal CIs after conditioning in experimental sample

As we argued in the introduction, the main cost of conditioning on the realized treatments and covariates is that one cannot use the knowledge of the propensity score or its smoothness. The experimental NSW sample allows us to quantify this cost in the most extreme setting, since the experimental design guarantees that the propensity score is constant. How much efficiency is lost by conditioning?

If one requires coverage conditional on realized treatments and pretreatment variables and uses our main specification for the conditional expectation function (Lipschitz with A_{main}, $p = 1$ and $C = 1$), the optimal FLCI for the CATT in the experimental sample is centered at 1.623, with worst-case bias 1.235, non-robust standard error 0.681 and robust standard error 0.715, leading to non-robust and robust CIs 1.623 ± 2.355 and 1.623 ± 2.411 respectively.

In contrast, if we do not condition on realized treatments when defining coverage, we can use the difference-in-means estimates and standard errors reported in Section 6.3, which gives the CIs 1.794 ± 1.315 and 1.794 ± 1.240, respectively. Focusing on the homoskedastic case, this implies that if we do not condition when defining coverage, we can use the knowledge that treatments were randomized to cut the CI length by 47%. On the other hand, if one requires conditional coverage, it is not optimal to assign treatment randomly, and, as discussed in Section 5.4, it is possible to reduce the cost of conditioning by optimizing the treatment

[16] If we treat the difference-in-means estimator as an estimator of the ATT (which also coincides with the ATE), the robust and non-robust standard errors are 0.671 and 0.633, respectively.

assignment. To bound the efficiency loss under optimal treatment assignment, we assign the individuals in the sample into clusters using the k-means algorithm by clustering their covariate values. Then, within each cluster, we randomly assigned a fraction π individuals to the treatment group, where π is the proportion treated in the original data. We then calculate the weights for the optimal estimator under this treatment assignment, and the length of the resulting CI (which doesn't require observing the outcome data). We then optimize the number of clusters k in the k-means algorithm. We find that $k = 210$ yields the shortest CI length, with the optimal estimator having worst-case bias 0.40 and standard deviation 0.71. The resulting CI is 32% shorter than the CI obtained under the original treatment assignment. This implies that the cost of conditioning can be reduced by at least 69% by optimizing the treatment assignment.

6.5 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 1 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 years' earnings to have a one-to-one effect, we would need to take $C = 1/\sqrt{.07^2 + .07^2} = 10.1$. For this C, the optimal FLCI is given by 1.72 ± 7.63, which is much wider than the FLCIs reported in Table 2 for A_{main} and $C = 1$.

In Theorem 2.3, we showed that the matching estimator with a single match is optimal for C large enough. For this result, it is important that the norm used to construct the matches is the same as the norm defining the Lipschitz class. To illustrate this point,

consider a matching estimator considered in Abadie and Imbens (2011), that uses $p = 2$ and $A^{1/2} = A^{1/2}_{\text{ne}}$. This yields the estimate 2.07, with homoskedastic standard error 2.20. Its worst-case bias under our main specification (A_{main}, $p = 1$ and $C = 1$) is 1.89, which implies that its efficiency is 77.5% for RMSE, and 74.6% for the construction of two-sided CIs, which is 12% and 11% lower, respectively, than the efficiencies of the matching estimator that matched on the norm defining the Lipschitz class reported in Section 6.2. Furthermore, the efficiency is never higher than 80.1%, even for large values of C.

6.6 CATE, set identification, and lack of overlap

An advantage of our finite sample approach is that our CIs apply even when average treatment effects are not identified, due to lack of overlap. In the NSW data, the covariates for the treated sample can be plausibly argued to lie on the support of the covariates for the untreated observations from the PSID, so that the ATT is point identified. However, the reverse is likely not true, so that if we are interested in the CATE rather than the CATT, the overlap conditions needed for point identification will fail. Thus, to illustrate how optimal CIs perform under set identification, we can apply our method to form CIs for the CATE in this setting.

Under our main specification (Lipschitz with A_{main}, $p = 1$ and $C = 1$), the optimal FLCI for the CATE is centered at -9.74 and has worst-case bias 10.18, non-robust standard error 2.62 and robust standard error 4.18, which gives the CIs -9.74 ± 14.49, and -9.74 ± 17.06, respectively. Thus, the point estimate for the CATE is negative, while the CIs allow for the possibility of both large positive and large negative effect of the program on average wages. The large worst-case bias and wide CIs reflect the inherent uncertainty in extrapolating the treatment effect from the individuals targeted by this intervention (who tend to be less educated and have lower pretreatment wages than the general population) to average individuals in the PSID.

Appendix A Finite-sample results: proofs and additional details

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 (2018a), 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 specializes the setup to the case with Lipschitz constraints, while allowing for possible additional monotonicity constraints, and proves Theorem 2.1. Appendix A.3 proves Theorem 2.2. Appendix A.4 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 that sum to one, $\sum_{i=1}^n w_i = 1$. Setting $w_i = 1/n$ gives the CATE, while setting $w_i = d_i/n_1$, gives the conditional average treatment effect on the treated (CATT). Here $n_d = \sum_{j=1}^n \mathbb{I}\{d_j = d\}$ gives the number of observations with treatment status equal to d. 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 (2018a), and we can specialize the general efficiency bounds and the construction of optimal affine estimators and CIs in those papers to the current setting.[17] 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, 2018a)

$$\omega(\delta) = \sup_{f,g \in \mathcal{F}} \left\{ Lg - Lf : \sum_{i=1}^n \frac{(f(x_i, d_i) - g(x_i, d_i))^2}{\sigma^2(x_i, d_i)} \leq \delta^2 \right\}, \tag{22}$$

[17] In particular, in the notation of Armstrong and Kolesár (2018a), $Y = (y_1/\sigma(x_1, d_1), \ldots, y_n/\sigma(x_n, d_n))$, $\mathcal{Y} = \mathbb{R}^n$, 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 $f + \iota_\alpha \in \mathcal{F}$[18], then by Lemma D.1 in Armstrong and Kolesár (2018a), 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 (2018a), 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 (2018a), 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 (22) 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 (22) at $\delta = \delta_\chi$. Then the shortest $1-\alpha$ FLCI among all FLCIs*

[18] 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 9, implies translation invariance.

centered at affine estimators is given by

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

(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 (22) 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 $\hat{L}_{\delta_{\text{RMSE}}}$ 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 $R^*_{\text{RMSE},A}(\mathcal{F})$ can be equivalently computed as $R^*_{\text{RMSE},A}(\mathcal{F}) = \inf_{\delta>0} \frac{1}{2}\sqrt{(\omega(\delta) - \delta\omega'(\delta))^2 + \omega'(\delta)^2} = \inf_{\delta>0} \sup_{f\in\mathcal{F}}(E(\hat{L}_\delta - Lf)^2)^{1/2}$, 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 (22) 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*

$$\frac{\inf_{\{\hat{c}\colon [\hat{c},\infty) \text{ satisfies (3)}\}} q_\beta(\hat{c}, \{g\})}{q_\beta(\hat{c}_\delta^*, \{g\})} = \frac{\omega(2\delta)}{\omega(\delta) + \delta\omega'(\delta)} \geq \frac{1}{2}.$$

(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

$$\begin{aligned}\frac{\inf_{\{\mathcal{C}\colon \mathcal{C} \text{ satisfies (3)}\}} E_g\lambda(\mathcal{C})}{\inf_{\delta>0} 2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2}\right)\omega'(\delta)} &= \frac{(1-\alpha)E\left[\omega(2(z_{1-\alpha} - Z)) \mid Z \leq z_{1-\alpha}\right]}{2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta_\chi)}{2\omega'(\delta_\chi)} - \frac{\delta_\chi}{2}\right)\cdot\omega'(\delta_\chi)} \\ &\geq \frac{z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha\Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha)}{z_{1-\alpha/2}}, \quad (23)\end{aligned}$$

where $\lambda(\mathcal{C})$ denotes the Lebesgue measure of a confidence set $\mathcal{C}$, Z is a standard normal random variable, $\Phi(z)$ and $\phi(z)$ denote the standard normal distribution and density, and $\tilde{z}_\alpha = z_{1-\alpha} - z_{1-\alpha/2}$.

Proof. Both parts of the theorem, except for the lower bound in (23), follow from Corollary 3.2 and Corollary 3.3 in Armstrong and Kolesár (2018a). The lower bound follows from Theorem C.7 in Armstrong and Kolesár (2018b). □

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 $\mathcal{F} = \mathcal{F}_{\mathrm{Lip}}(C)$, 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 (23) evaluates to 71.7%. 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 6, the efficiency of the shortest FLCI is over 97.0% at piecewise constant functions.

A.2 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 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 $f\colon \{x_1, \ldots, x_n\} \times \{0,1\} \to \mathbb{R}$, we have $f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(C)$ if and only if there exists a function $h \in \mathcal{F}_{\text{Lip}}(C)$ such that $f(x,d) = h(x,d)$ for all $(x,d) \in \{x_1, \ldots, x_n\} \times \{0,1\}$.*

The first part of Theorem 2.1 follows directly from this result. The second part follows from the following lemma and the observation that $\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_k) = C\,\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(1)}(\hat{L}_k)$.

Lemma A.2. *Suppose that w_i satisfies $w_i = w(d_i)$ for some $w(0), w(1) \geq 0$. Suppose also that the weights $k(x,d)$ satisfy $\sum_{i=1}^n d_i k(x_i,d_i) = n_0 w(0) + n_1 w(1)$, and $\sum_{i=1}^n (1 - d_i)k(x_i,d_i) = -(n_0 w(0) + n_1 w(1))$ with $k(x_i,1) \geq w(1)$ and $k(x_i,0) \leq w(0)$. Then there exists a vector $f^* = (f^*(x_1,0), \ldots, f^*(x_n,0), f^*(x_1,1), \ldots, f^*(x_n,1)) \in \mathbb{R}^{2n}$ that maximizes $\sum_{i=1}^n k(x_i,d_i) f(x_i,d_i) - Lf$ subject to*

$$f(x_i,1) \leq f(x_j,1) + \|x_i - x_j\|_{\mathcal{X}}, \quad \text{all } i,j \text{ with } d_i = 1,\ d_j = 0 \text{ and } k(x_i,1) > w(1), \tag{24}$$

$$f(x_i,0) \leq f(x_j,0) + \|x_i - x_j\|_{\mathcal{X}}, \quad \text{all } i,j \text{ with } d_i = 1,\ d_j = 0 \text{ and } k(x_j,1) < -w(0). \tag{25}$$

such that $f^*(x_i,d) \leq f^*(x_j,d) + \|x_i - x_j\|_{\mathcal{X}}$ for all $i,j \in \{1, \ldots, n\}$ and $d \in \{0,1\}$.

The condition on the weights w_i holds for the CATE (with $w(1) = w(0) = 1/n$), as well as the CATT (with $w(1) = 1/n_1$ and $w(0) = 0$). The proof is given in the supplemental materials. The implications of Lemma A.1 for the form of the optimal estimator are considered in Appendix A.3.

We now consider imposing monotonicity restrictions 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 $\mathcal{F}_{\text{Lip},\mathcal{S}\uparrow}(C) \subseteq \mathcal{F}_{\text{Lip}}(C)$ denote the subset of functions such that $f(\cdot,0)$ and $f(\cdot,1)$ are monotone for the indices in $\mathcal{S}$: for any $x, \tilde{x}$ with $x_j \geq \tilde{x}_j$ for $j \in \mathcal{S}$ and $x_j = \tilde{x}_j$ for $j \notin \mathcal{S}$, we have $f(x,d) \geq f(\tilde{x},d)$ 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 $x \in \mathbb{R}^p$, let $(x)_{\mathcal{S}+}$ denote the vector with jth element x_j for $j \notin \mathcal{S}$ and jth element $\max\{x_j, 0\}$ 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.3. *(Beliakov, 2005, Proposition 4.1) For any function $f\colon \{x_1, \ldots, x_n\} \times \{0,1\} \to \mathbb{R}$, we have $f \in \widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S}\uparrow,n}(C)$ if and only if there exists a function $h \in \mathcal{F}_{\text{Lip},\mathcal{S}\uparrow}(C)$ such that $f(x,d) = h(x,d)$ for all $(x,d) \in \{x_1, \ldots, x_n\} \times \{0,1\}$.*

Using this result, the problem of computing the maximum bias of an affine estimator $\hat{L}_{k,a}$ that satisfies (11) can again be phrased as a finite-dimensional linear program of maximizing $a + \sum_{i=1}^n k(x_i,d_i)f(x_i,d_i) - Lf$ subject to $f \in \widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S}\uparrow,n}(C)$. The optimal estimator can be computed by solving (22) with $\mathcal{F} = \widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S}\uparrow,n}(C)$, which is a finite-dimensional convex optimization problem.

A.3 Proof of Theorem 2.2

The first part follows directly from Lemma A.1. To show the second part and to give the algorithm for computing the solution path, suppose, as in Lemma A.2, that $w_i = w(d_i)$ for some $w(0), w(1) \geq 0$, and that $\sigma^2(x,d) = \sigma^2(d)$ for some $\sigma^2(0), \sigma^2(1) > 0$. The dual problem to (16) is to minimize $\sum_{i=1}^n f(x_i,d_i^2)/\sigma^2(x_i,d_i)$ subject to a lower bound on Lf/C. The Lagrangian for this problem has the form

$$\min_{f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(C)} \frac{1}{2}\sum_{i=1}^n \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)} - \mu Lf/C = \min_{f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)} \frac{C^2}{2}\sum_{i=1}^n \frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)} - \mu Lf, \tag{26}$$

where we use the observation that if $f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(C)$, then $f/C \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)$. Let g_μ^* denote the solution to the minimization problem on the right-hand side of (26). Because for each $\delta > 0$, the program (16) is strictly feasible at $f = 0$, Slater's condition holds, and the solution path $\{f_\delta^*\}_{\delta>0}$ can be identified with the solution path $\{Cg_\mu^*\}_{\mu>0}$.

Suppose, without loss of generality, that the observations are ordered, so that $d_j = 0$ for $j = 1, \ldots, n_0$, and $d_i = 1$ for $i = n_0 + 1, \ldots, n$. It will be convenient to state the algorithm using the notation $m_i = (2d_i - 1)f(x_i, d_i)$, and $r_i = (1 - 2d_i)f(x_i, 1 - d_i)$. Then $Lf = \sum_{i=1}^n w_i(m_i + r_i)$, and the constraint $f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)$ is equivalent to the constraints

$$r_j \leq m_i + \|x_i - x_j\|, \qquad d_i \neq d_j, \tag{27}$$

$$r_j \leq r_{j'} + \|x_j - x_j\|, \qquad d_j = d_{j'}, \tag{28}$$

$$m_i \leq m_{i'} + \|x_i - x_{i'}\| \qquad d_i = d_{i'}. \tag{29}$$

$$m_i \leq r_j + \|x_i - x_j\|, \qquad d_i \neq d_j. \tag{30}$$

Lemma A.4. *Consider the problem of minimizing* $\frac{1}{2}\sum_{i=1}^n m_i^2/\sigma^2(x_i, d_i) - \mu \sum_{i=1}^n w_i(m_i + r_i)$ *subject to* (27). *Then there exists a solution* $m(\mu)$ *and* $r(\mu)$ *that satisfies Equations* (28), (29) *and* (30).

If we only impose the constraints in (27), the Lagrangian for the program (26) can be written as

$$\frac{1}{2}\sum_{j=1}^{n_0} \frac{m_j^2}{\sigma^2(0)} + \frac{1}{2}\sum_{i=1}^{n_1} \frac{m_{i+n_0}^2}{\sigma^2(1)} - \mu \left(\sum_{i=1}^{n_1} w(1)(m_{i+n_0} + r_{i+n_0}) + \sum_{j=1}^{n_0} w(0)(m_j + r_j) \right)$$
$$+ \sum_{i=1}^{n_1}\sum_{j=1}^{n_0} \left[\Lambda_{ij}^0 (r_{i+n_0} - m_j - \|x_{i+n_0} - x_j\|_{\mathcal{X}}) + \Lambda_{ij}^1 (r_j - m_{i+n_0} - \|x_{i+n_0} - x_j\|_{\mathcal{X}}) \right]. \tag{31}$$

The lemma shows that the resulting first-order conditions imply the constraints Equations (28), (29) and (30) must hold at the optimum. The second part of Theorem 2.2 follows directly from Lemma A.4, the proof of which is given in the supplemental materials.

To describe the algorithm, we need additional notation. Let $m(\mu)$, $r(\mu)$, $\Lambda^0(\mu)$, and $\Lambda^1(\mu)$ denote the values of m, r, and of the Lagrange multipliers at the optimum of (31). For $d \in \{0, 1\}$, let $N^d(\mu) \in \mathbb{R}^{n_1 \times n_0}$ denote a matrix with elements $N_{ij}^d(\mu) = 1$ if the constraint associated with $\Lambda_{ij}^d(\mu)$ is active, and $N_{ij}^d(\mu) = 0$ otherwise. Let $G^0 \in \mathbb{R}^{n_0 \times n_0}$ and $G^1 \in \mathbb{R}^{n_1 \times n_1}$ denote matrices with elements $G_{jj'}^0 = \mathbb{I}\{\sum_i N_{ij}^0(\mu) N_{ij'}^0(\mu) > 0\}$, and $G_{ii'}^1 = \mathbb{I}\{\sum_j N_{ij}^1(\mu) N_{i'j}^1(\mu) > 0\}$. Then G^0 defines a graph (adjacency matrix) of a network in which j and j' are linked if the constrains associated with Λ_{ij}^0 and $\Lambda_{ij'}^0$ are both active for some i. Similarly, G^1 defines a graph of a network in which i and i' are linked if the constraints associated with $\Lambda_{i'j}^1$ and Λ_{ij}^1 are both active for some j. Let $\{\mathcal{M}_1^0, \ldots, \mathcal{M}_{K_0}^0\}$ denote a partition of $\{1, \ldots, n_0\}$ according to the connected components of G^0, so that if

$j, j' \in \mathcal{M}_k^0$ then there exists a path from j to j'. Let $\{\mathcal{R}_1^0, \ldots, \mathcal{R}_k^0\}$ be a corresponding partition of $\{1, \ldots, n_1\}$, defined by $\mathcal{R}_k^0 = \{i \in \{1, \ldots, n_1\}\colon N_{ij}^0(\mu) = 1 \text{ for some } j \in \mathcal{M}_k^0\}$. Similarly, let $\{\mathcal{M}_1^1, \ldots, \mathcal{M}_{K_1}^1\}$ denote a partition of $\{1, \ldots, n_1\}$ according to the connected components of G^1, and let $\mathcal{R}_k^1 = \{j \in \{1, \ldots, n_0\}\colon N_{ij}^1(\mu) = 1 \text{ for some } i \in \mathcal{M}_k^1\}$.

In the supplemental materials, we show that the solution path for $m(\mu)$ is piecewise linear in μ, with points of non-differentiability when either a new constraint becomes active, or else the Lagrange multiplies $\Lambda_{ij}^d(\mu)$ associated with an active constraint decreases to zero. We also derive the formulas for the slope of $m(\mu)$, $r(\mu)$, and $\Lambda^d(\mu)$ at points of differentiability. This leads to the following algorithm that is similar to the LAR algorithm in Rosset and Zhu (2007) and Efron et al. (2004) for computing the LASSO path.

1. Initialize $\mu = 0$, $m = 0$, $\Lambda^0 = 0$, and $\Lambda^1 = 0$. Let $D^0, D^1 \in \mathbb{R}^{n_1 \times n_0}$ be matrices with elements $D_{ij}^d = \|x_{i+n_0} - x_j\|_{\mathcal{X}}$, $d \in \{0, 1\}$. Let r be a vector with elements $r_j = \min_{i=1,\ldots,n_1}\{D_{ij}^1\}$, $j = 1, \ldots, n_0$ and $r_{i+n_0} = \min_{j=1,\ldots,n_0}\{D_{ij}^0\}$, $i = 1, \ldots, n_1$. Let $N^0, N^1 \in \mathbb{R}^{n_1 \times n_0}$ be matrices with elements $N_{ij}^0 = \mathbb{I}\{D_{ij}^0 = r_{i+n_0}\}$ and $N_{ij}^1 = \mathbb{I}\{D_{ij}^1 = r_j\}$.

2. While $\mu < \infty$:

   (a) Calculate the partitions $\mathcal{M}_k^d$ and $\mathcal{R}_k^d$ associated with N^d, $d \in \{0, 1\}$. Calculate directions δ for m and a direction δ_r for r as $\delta_{r,i+n_0} = \delta_j = \sigma^2(0)(w(0) + (\#\mathcal{R}_k^0/\#\mathcal{M}_k^0)w(1))$ for $i \in \mathcal{R}_k^0$ and $j \in \mathcal{M}_k^0$, and $\delta_{r,j} = \delta_{i+n_0} = \sigma^2(1)(w(1) + (\#\mathcal{R}_k^1/\#\mathcal{M}_k^1)w(0))$ for $i \in \mathcal{R}_k^0$ and $j \in \mathcal{M}_k^0$.

   (b) Calculate directions Δ^d for Λ^d by setting $\Delta_{ij}^d = 0$ if $N_{ij}^d = 0$, with the remaining elements given by a solution to the systems of n equations (i) $\sum_{i=1}^{n_0} \Delta_{ij}^1 = \delta_j/\sigma^2(0) - w(0)$, $j = 1, \ldots, n_0$ and $\sum_{j=1}^{n_0} \Delta_{ij}^0 = w(1)$, $i = 1, \ldots, n_1$ and (ii) $\sum_{j=1}^{n_0} \Delta_{ij}^1 = \delta_{i+n_0}/\sigma^2(1) - w(1)$, $i = 1, \ldots, n_1$ and $\sum_{i=1}^{n_1} \Delta_{ij}^0 = w(0)$, $j = 1, \ldots, n_0$.

   (c) Calculate step size s as $s = \min\{s_1^0, s_2^0, s_1^1, s_2^1\}$, where

   $$s_1^0 = \min\{s \geq 0\colon r_{i+n_0} + \delta_{r,i+n_0} s = \delta_j s + D_{ij}^0 \text{ some } (i,j) \text{ s.t. } N_{ij}^0 = 0,\ \delta_j > \delta_{r,i+n_0}\}$$
   $$s_1^1 = \min\{s \geq 0\colon r_j + \delta_{r,j} s = \delta_{i+n_0} s + D_{ij}^1 \text{ some } (i,j) \text{ s.t. } N_{ij}^1 = 0,\ \delta_{i+n_0} > \delta_{rj}\}$$
   $$s_2^d = \min\{s \geq 0\colon \Lambda_{ij}^d + s\Delta_{ij}^d = 0 \text{ among } (i,j) \text{ with } N_{ij}^d = 1 \text{ and } \Delta_{ij}^d < 0\}$$

   (d) Update $\mu \mapsto \mu + s$, $m \mapsto m + s\delta$, $r \mapsto r + s\delta_r$ $\Lambda^d \mapsto \Lambda^d + s\Delta^d$, $D_{ij}^0 \mapsto D_{ij}^0 + s\delta_j$, $D_{ij}^1 \mapsto D_{ij}^1 + s\delta_{i+n_0}$ If $s = s_1^d$, then update $N_{ij}^d = 1$, where (i, j) is the index defining s_1^d. If $s = s_2^d$, update $N_{ij}^d = 0$, where (i, j) is the index defining s_2^d.

Given the solution path $\{m(\mu)\}_{\mu>0}$, the optimal estimator $\hat{L}_\delta$ and its worst-case bias can then be easily computed. For simplicity, we specialize to the ATE case, $w(1) = w(0) = 1/n$. Let $\delta(\mu) = 2C\sqrt{m(\mu)'m(\mu)}$. It then follows from the formulas in Appendix A.1 and the first-order conditions associated with the Lagrangian (31) (see the supplemental materials) that the optimal estimator takes the form

$$\hat{L}_{\delta(\mu)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_\mu(x_i,1) - \hat{f}_\mu(x_i,0)),$$

where $\hat{f}_\mu(x_j,1) = \sum_{i=1}^{n_1} n\Lambda^1_{ij}(\mu)/\mu Y_{n_0+i}$ for $j = 1,\dots,n_0$; $\hat{f}_\mu(x_i,1) = Y_i$ for $i = n_0+1,\dots,n$; $\hat{f}_\mu(x_j,0) = Y_j$ for $j = 1,\dots,n_0$; and $\hat{f}_\mu(x_i,0) = \sum_{j=1}^{n_0} n\Lambda^0_{i-n_0,j}(\mu)/\mu Y_i$ for $i = n_0+1,\dots,n$. The worst-case bias of the estimator is given by $C(\sum_{i=1}^{n}(m_i(\mu)+r_i(\mu))/n - \sum_{i=1}^{n} m_i(\mu)^2/\mu)$.

For the interpretation of $\hat{L}_{\delta(\mu)}$ as a matching estimator with a variable number of matches, observe that $\sum_{i=1}^{n_1} n\Lambda^1_{ij}(\mu)/\mu = \sum_{j=1}^{n_0} n\Lambda^0_{ij}(\mu)/\mu = 1$. Also, $N^0_{ij}(\mu) = 0$ and hence $\Lambda^0_{ij}(\mu) = 0$ unless $D^0_{ij}(\mu) = \min_\ell D^0_{i\ell}(\mu)$. Similarly, $\Lambda^1_{ij}(\mu) = 0$ unless $D^1_{ij}(\mu) = \min_\ell D^1_{\ell j}(\mu)$. Thus, the counterfactual outcome for each observation i is given by a weighted average of outcomes for observations with opposite treatment status that are closest to it in terms of the "effective distance" matrices $D^0_{ik}(\mu)$ (for $i = n_0+1,\dots,n$) or $D^1_{ki}(\mu)$ (for $i = 1,\dots,n_0$). Since $D^0_{ik}(\mu) = m_k(\mu) + \|x_{n_0+i} - x_k\|_{\mathcal{X}}$ $D^1_{ki}(\mu) = m_{n_0+i}(\mu) + \|x_{n_0+i} - x_k\|_{\mathcal{X}}$, and $m_k(\mu)$ is increasing in the number of times k has been used as a match, observations that have been used more often as a match are considered to be further away according to these effective distance matrices.

A.4 Proof of Theorem 2.3

To prove Theorem 2.3, we first provide another characterization of the optimal weights given in (9). Given $\{m_i\}_{i=1}^n$, consider the optimization problem (16) 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 if and only if $|m_i - m_j| \le C\|x_i - x_j\|_{\mathcal{X}}$ 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 (16), it follows that $f^*_\delta(x_i,d_i) = (2d_i - 1)\cdot m^*_i$ where $\{m^*_i\}_{i=1}^n$

solves

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

$$\sum_{i=1}^n m_i^2/\sigma^2(x_i, d_i) \leq \delta^2/4, \tag{32}$$

$$|m_i - m_j| \leq C\|x_i - x_j\|_{\mathcal{X}} \text{ for all } i, j \text{ with } d_i = d_j, \tag{33}$$

where

$$\tilde{\omega}_i(m) = \min_{j: d_j \neq d_i} (m_j + C\|x_i - x_j\|_{\mathcal{X}}). \tag{34}$$

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 (34). 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

$$\mathcal{J}_i(m) = \{j \in \mathcal{J}_i(0) : m_j \leq m_\ell \text{ all } \ell \in \mathcal{J}_i(0)\}. \tag{35}$$

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 (33) implied by the constraint (32). Thus, specializing to the case with $w_i = 1/n$, the first order conditions are given by

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

where λ is the Lagrange multiplier on (32), 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 (35) 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 (9), 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 (5) that the optimal estimator coincides with the matching estimator based on a single match.

Appendix B Proofs for asymptotic results

This appendix proves the results given in Section 4.

B.1 Proof of Theorem 4.1

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

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

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

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

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

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n[f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,1)-f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,0)]\\ &\qquad=-\frac{1}{n}\sum_{i=1}^n\sum_{\tilde{x}\in\tilde{\mathcal{X}}_n(h,\eta)}k((X_i-\tilde{x})/h)\leq-\underline{k}\frac{\#\mathcal{I}_n(h)}{n},\quad(36)\end{aligned}$$

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,

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

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 (36), 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.

B.2 Proofs of Theorems 4.2 and 4.3

We first give a lemma that is used to prove consistency of the nearest-neighbor variance estimator. The proof is based on the arguments in Abadie and Imbens (2008) and it is deferred to the supplemental materials.

Lemma B.2. *Consider the fixed design model (1). Suppose that $1/K \leq Eu_i^2 \leq K$ and $E|u_i|^{2+1/K} \leq K$ for some constant K, and that $\sigma^2(x, d)$ is uniformly continuous in x for $d \in \{0, 1\}$. Let $\ell_j(i)$ be the jth closest unit to i, with respect to some norm $\|\cdot\|$, among units with the same value of the treatment. Let $\hat{u}_i^2 = \frac{J}{J+1}(y_i - \sum_{j=1}^J y_{\ell_j(i)}/J)^2$, and let $a_{ni} \geq 0$ be a non-random sequence such that $\max_i a_{ni} \to 0$, and that $\sum_{i=1}^n a_{ni}$ is uniformly bounded. If $\max_i C_n\|x_{\ell_J(i)} - x_i\| \to 0$, then $\sum_i a_{ni}(\hat{u}_i^2 - u_i^2)$ converges in probability to zero, uniformly over $\mathcal{F}_{\text{Lip}}(C_n)$.*

Theorems 4.2 and 4.3 follow from verifying the high level conditions of Theorem F.1 in Armstrong and Kolesár (2018a). In particular, we need to show that the weights k ($\tilde{k}^*_\delta$ for Theorem 4.2 and $k_{\text{match},M}$ for Theorem 4.3) are such that $\sum_{i=1}^n k(x_i,d_i)u_i/\operatorname{sd}_k$ converges in distribution to $N(0,1)$ (condition (S13) in Armstrong and Kolesár, 2018a) and $\sum_i \hat{u}_i^2 k(x_i,d_i)^2/\operatorname{sd}_k^2$ converges in probability to 1, uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$ (S14), where $\operatorname{sd}_k^2 = \sum_{i=1}^n \sigma^2(x_i,d_i)k(x_i,d_i)$. We claim that both (S13) and (S14) hold if the weights satisfy

$$\frac{\max_{1\le i\le n} k(x_i,d_i)^2}{\sum_{i=1}^n k(x_i,d_i)^2} \to 0. \tag{37}$$

Under the moment bounds on u_i, Equation (37) directly implies the Lindeberg condition that is needed for condition (S13) to hold. To show that it also implies (S14), note that (S14) is equivalent to the requirement that $\sum_{i=1}^n \hat{u}_i^2 a_{ni} - \sum_{i=1}^n \sigma^2(x_i,n_i)a_{ni}$ converges to zero uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$, where

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

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

$$\begin{aligned} E\left|\sum_{i=1}^n (u_i^2 - \sigma^2(x_i,d_i))a_{ni}\right|^{1+1/(2K)} &\le 2\sum_{i=1}^n a_{ni}^{1+1/(2K)} E|u_i^2 - \sigma^2(x_i,d_i)|^{1+1/(2K)} \\ &\le \max_{1\le i\le n} a_{ni}^{1/(2K)} E|u_i^2 - \sigma^2(x_i,d_i)|^{1+1/(2K)} \cdot \sum_{i=1}^n a_{ni}. \end{aligned}$$

Note that, by boundedness of $\sigma(x,d)$ away from zero and infinity, $\sum_{i=1}^n a_{ni}$ is uniformly bounded. Furthermore, it follows from (37), that $\max_{1\le i\le n} a_{ni} \to 0$. From this and the moment bounds on u_i, it follows that the above display converges to zero. It therefore suffices to show that $\sum_{i=1}^n (\hat{u}_i^2 - u_i^2)a_{ni}$ converges to zero. For the nearest-neighbor variance estimator, this follows from Lemma B.2. We therefore just need to show that this holds for the Nadaraya-Watson estimator with uniform kernel and bandwidth h_n. Denote this estimator by $\hat{u}_i^2 = (y_i - \hat{f}(x_i,d_i))^2$ where $\hat{f}(x_i,d_i) = \sum_{j\in\mathcal{N}_i} y_i/\#\mathcal{N}_i$ and $\mathcal{N}_i = \{j \in \{1,\dots,n\} : \|x_j - x_i\|_{\mathcal{X}} \le h_n,\ d_i = d_j\}$. Write

$$\sum_{i=1}^n (\hat{u}_i^2 - u_i^2)a_{ni} = \sum_{i=1}^n (2y_i - \hat{f}(x_i,d_i) - f(x_i,d_i))(f(x_i,d_i) - \hat{f}(x_i,d_i))a_{ni}$$

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

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

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

which is in turn bounded by a constant times $\max_{1\le i\le n} E_f[(f(x_i, d_i) - \hat{f}(x_i, d_i))^2]$. Since

$$E_f[(f(x_i, d_i) - \hat{f}(x_i, d_i))^2] = \frac{1}{\#\mathcal{N}_i^2}\sum_{j\in\mathcal{N}_i}^{n} E[u_j^2] + \frac{1}{\#\mathcal{N}_i^2}\left(\sum_{j\in\mathcal{N}_i}^{n}(f(x_j, d_i) - f(x_i, d_i))\right)^2$$
$$\le \max_{1\le j\le n} E[u_j^2]/\#\mathcal{N}_i + \max_{j\in\mathcal{N}_i}(f(x_j, d_i) - f(x_i, d_i))^2,$$

it follows that

$$\sup_{f\in\mathcal{F}_{\mathrm{Lip}}(C_n)} \max_{1\le i\le n} E_f[(f(x_i, d_i) - \hat{f}(x_i, d_i))^2] \le K/\min_{i=1,\ldots,n} \mathcal{N}_i + (h_n C_n)^2.$$

If condition (19) 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 $h_n C_n \to 0$ such that $\min_{1\le i\le n} \#\mathcal{N}_i \to \infty$, so that under this bandwidth sequence, the above display converges to zero.

Proof of Theorem 4.2. We need to verify that (37) holds for the weights $\tilde{k}_\delta^*$. By boundedness of $\tilde{\sigma}(x_i, d_i)$ away from zero and infinity, (37) 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 (10) 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, $|\tilde{f}_\delta^*(x, d_{i_n^*})| \geq \eta - C_n\|x - x_{i_n^*}\|$ so that, for $\|x - x_{i_n^*}\| \leq \eta/(2C_n)$, we have $|\tilde{f}_\delta^*(x, d_{i_n^*})| \geq \eta/2$. Thus, we have

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

infinitely often. This gives a contradiction so long as (19) holds. This completes the proof of Theorem 4.2. □

Proof of Theorem 4.3. We need to show that (37) holds for the weights $k_{\text{match},M}(x_i, d_i) = (1 + K_M(i))/n$. For this, it is sufficient to show that $\max_{1 \leq i \leq n} K_M(i)^2/n \to 0$. To this end, let $U_M(x, d) = \|x_j - x\|_{\mathcal{X}}$ where x_j is the Mth closest observation to x among observations i with $d_i = d$, so that $K_M(i) = \#\{j : d_j \neq d_i, \|x_j - x_i\|_{\mathcal{X}} \leq U_M(x_j, d_i)\}$. When (20) holds and n is large enough so that $n\underline{G}(a_n) \geq M$, we will have $U_M(x, d) \leq a_n$ for all $x \in \mathcal{X}$. By definition of $K_M(i)$, the upper bound in (20) then implies $K_M(i) \leq n\overline{G}(a_n)$. Thus, it suffices to show that $[n\overline{G}(a_n)]^2/n = n\overline{G}(a_n)^2 \to 0$.

Let $c_n = n\underline{G}(a_n)/\log n$ and $b(t) = \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}]$ (so that $\lim_{t \to 0} b(t) = 0$ under the conditions of Theorem 4.3). Then $a_n = \underline{G}^{-1}(c_n(\log n)/n)$ so that

$$n\overline{G}(a_n)^2 = n\overline{G}(\underline{G}^{-1}(c_n(\log n)/n))^2 = b(c_n(\log n)/n)\frac{c_n \log n}{\log n - \log c_n - \log\log n}.$$

This converges to zero so long as c_n increases slowly enough (it suffices to take c_n to be the minimum of $\log n$ and $1/\sqrt{b((\log n)^2/n)}$). □

B.3 Proof of Lemma 4.1

To prove Lemma 4.1, it suffices to show that, for i.i.d. variables w_i taking values in Euclidean space with finite support $\mathcal{W}$, we have $\inf_{w \in \mathcal{W}} \#\{i \in \{1, \ldots, n\} : \|w - w_i\| \leq \varepsilon\} \to \infty$ with probability one. To this end, for any w and r, let $B_r(w) = \{\tilde{w} : \|w - \tilde{w}\| < r\}$ denote the open ball centered at w with radius r. Given $\delta > 0$, let $\widetilde{\mathcal{W}}_\delta$ be a grid of meshwidth δ on $\mathcal{W}$. If δ is chosen to be small enough, then, for every $w \in \mathcal{W}$, there exists $\tilde{w} \in \widetilde{\mathcal{W}}_\delta$ such that $B_\delta(\tilde{w}) \subseteq B_\varepsilon(w)$. Thus, if δ is chosen small enough, the quantity of interest is bounded from below by

$$\min_{w \in \widetilde{\mathcal{W}}_\delta} \#\{i \in \{1, \ldots, n\} : \|w - w_i\| < \delta\},$$

where we note that the infimum is now a minimum over a finite set. Since each $w \in \widetilde{\mathcal{W}}_\delta$ is contained in the support of w_i, we have $\min_{w \in \widetilde{\mathcal{W}}_\delta} P(\|w - w_i\| < \delta) > 0$, so it follows from the strong law of large numbers that the quantity in the above display converges to infinity almost surely.

B.4 Proof of Theorem 4.4

Let $\text{sd}_{\delta_\rho,n}$ and $\overline{\text{bias}}_{\delta_\rho,n}$ denote the standard deviation and worst-case bias of the minimax linear estimator and let $\text{sd}_{\text{match},1}$ and $\overline{\text{bias}}_{\text{match},1}$ denote the standard deviation and worst-case bias of the estimator with a single match (conditional on $\{(X_i, D_i)_{i=1}^n\}$). Since worst-case bias is increasing in δ and variance is decreasing in δ, and since the matching estimator with $M = 1$ solves the modulus problem for small enough δ by Theorem 2.3, we have $\overline{\text{bias}}_{\delta_\rho,n} \geq \overline{\text{bias}}_{\text{match},1}$. Thus,

$$1 \leq \frac{\overline{\text{bias}}^2_{\text{match},1} + \text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta_\rho,n} + \text{sd}^2_{\delta_\rho,n}} \leq \frac{\overline{\text{bias}}^2_{\delta_\rho,n} + \text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta_\rho,n} + \text{sd}^2_{\delta_\rho,n}} \leq 1 + \frac{\text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta_\rho,n} + \text{sd}^2_{\delta_\rho,n}}.$$

By the arguments in the proof of Theorem 4.1, there exists $\varepsilon > 0$ such that $\overline{\text{bias}}_{\delta_\rho,n} \geq \varepsilon n^{-2/p}$ almost surely. In addition, by Theorem 37 in Chapter 2 of Pollard (1984), the conditions of Theorem 4.3 hold almost surely (with $\underline{G}(a)$ and $\overline{G}(a)$ multiplied by some positive constants). Arguing as in the proof of Theorem 4.3 then gives the bound $\text{sd}^2_{\text{match},1} \leq [2 \max_{1 \leq i \leq n} K_1(i)]^2/n \leq [2n\overline{G}(a_n)]^2/n$ for any sequence $a_n = \underline{G}^{-1}(c_n(\log n)/n)$ with $c_n = n\overline{G}(a_n)/\log n \to \infty$. Plugging these bounds into the above display gives a bound proportional to

$$\overline{G}(\underline{G}^{-1}(c_n(\log n)/n))^2 n^{2/p+1} = b(c_n(\log n)/n) \left[\frac{c_n(\log n)/n}{\log n - \log c_n - \log\log n}\right]^{2/p+1} n^{2/p+1},$$

where $b(t) = \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}]^{2/p+1}$. If $\lim_{t \to 0} b(t) = 0$, then this can be made to converge to zero by choosing c_n to increase slowly enough. Similar arguments apply to the FLCI and one-sided CI criteria.

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 (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74, 235–267.

——— (2008): "Estimation of the Conditional Variance in Paired Experiments," *Annales d'Économie et de Statistique*, 175–187.

——— (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 (2018a): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86, 655–683.

——— (2018b): "Sensitivity Analysis using Approximate Moment Condition Models," ArXiv: 1808.07387.

Auerbach, E. (2018): "Identification and Estimation of a Partially Linear Regression Model using Network Data," Working paper, Northwestern University.

Bailey, M. J. and A. Goodman-Bacon (2015): "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans," *American Economic Review*, 105, 1067–1104.

Banerjee, A. V., S. Chassang, and E. Snowberg (2017): "Decision Theoretic Approaches to Experiment Design and External Validity," in *Handbook of Economic Field Experiments*, ed. by A. V. Banerjee and E. Duflo, Amsterdam: North-Holland, vol. 1, chap. 4, 141–174.

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.

Blackwell, D. and M. A. Girshick (1954): *Theory of Games and Statistical Decisions*, New York: John Wiley & Sons.

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

Busso, M., J. DiNardo, and J. McCrary (2014): "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators," *The Review of Economics and Statistics*, 96, 885–897.

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

Chaudhuri, S. and J. B. Hill (2016): "Heavy tail robust estimation and inference for average treatment effects," Unpublished manuscript, University of North Carolina at Chapel Hill.

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.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2006): "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand," Working Paper 330, National Bureau of Economic Research.

——— (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

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.

Efron, B., T. Hastie, I. M. Johnstone, and R. J. Tibshirani (2004): "Least Angle Regression," *The Annals of Statistics*, 32, 407–451.

Galiani, S., P. Gertler, and E. Schargrodsky (2005): "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 113, 83–120.

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.

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching as an Econometric Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

Heckman, N. E. (1988): "Minimax Estimates in a Semiparametric Model," *Journal of the American Statistical Association*, 83, 1090–1096.

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

Imbens, G. and S. Wager (2017): "Optimized Regression Discontinuity Designs," *Review of Economics and Statistics*, forthcoming.

Kallus, N. (2017): "Generalized Optimal Matching Methods for Causal Inference," ArXiv: 1612.08321.

Kasy, M. (2016): "Why Experimenters Might Not Always Want to Randomize, and What They Could Do Instead," *Political Analysis*, 24, 324–338.

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.

Li, K.-C. (1989): "Honest confidence regions for nonparametric regression," *The Annals of Statistics*, 17, 1001–1008.

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

Ma, X. and J. Wang (2018): "Robust Inference Using Inverse Probability Weighting," ArXiv: 1810.11397.

Pollard, D. (1984): *Convergence of stochastic processes*, New York, NY: Springer.

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.

Romano, J. P. and M. Wolf (2017): "Resurrecting Weighted Least Squares," *Journal of Econometrics*, 197, 1–19.

Rosset, S. and J. Zhu (2007): "Piecewise Linear Regularized Solution Paths," *The Annals of Statistics*, 35, 1012–1030.

Rothe, C. (2017): "Robust Confidence Intervals for Average Treatment Effects Under Limited Overlap," *Econometrica*, 85, 645–660.

Sasaki, Y. and T. Ura (2017): "Inference for moments of ratios with robustness against large trimming bias and unknown convergence rate," ArXiv: 1709.00981.

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*, 125, 305–353.

Stock, J. H. (1989): "Nonparametric Policy Analysis," *Journal of the American Statistical Association*, 84, 567–575.

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.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press, second ed.

Zhao, Z. (2004): "Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence," *The Review of Economics and Statistics*, 86, 91–107.

Table 1: Diagonal elements of the weight matrix $A^{1/2}$ in definition of the norm (21) for the main specification, $A^{1/2}_{\text{main}}$, and alternative specification, $A^{1/2}_{\text{ne}}$.

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

Table 2: Results for NSW data, $p = 1$, $A = A_{\text{main}}$, $C = 1$.

					Std. error		
Criterion	δ	M	Estimate	$\overline{\text{bias}}$	homosk.	robust	$\text{cv}_{0.05}$
Optimal estimator							
RMSE	1.86		0.94	1.64	1.53	1.04	3.22
FLCI	3.30		0.94	1.81	1.40	0.96	3.52
one-sided CI	2.49		0.98	1.71	1.47	1.00	3.36
Matching estimator							
RMSE		1	1.39	1.48	2.01	1.11	2.98
FLCI		18	1.26	2.21	1.39	0.89	4.12
one-sided CI		17	1.32	2.16	1.42	0.89	4.09

Notes: 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. $\overline{\text{bias}}$ gives the worst-case bias of the estimator, and $\text{cv}_{0.05}$ is the critical value for a two-sided 95% CI that depends on the ratio of the worst-case bias to standard error.

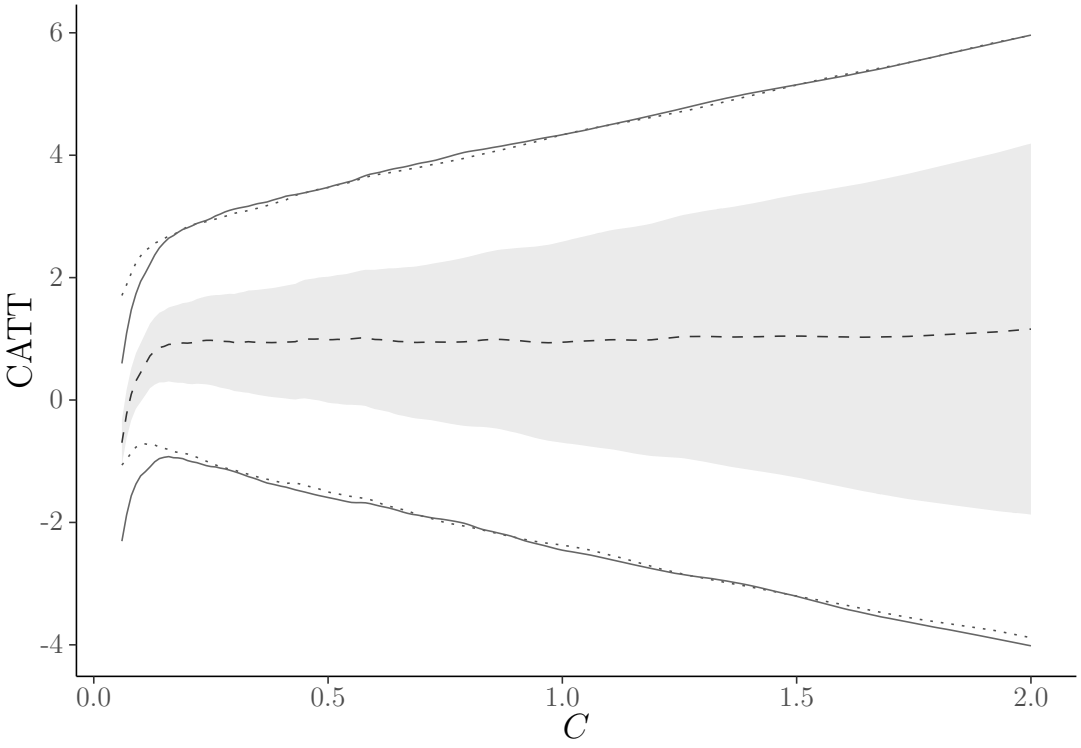

Figure 1: Optimal estimator and CIs for CATT in NSW data as a function of the Lipschitz constant C.

Notes: Dashed line corresponds to point estimate, shaded region denotes the estimate $\pm$ its worst-case bias, dotted lines give one-sided 95% heteroskedasticity-robust confidence bands, and a two-sided 95% confidence band is denoted by solid lines.

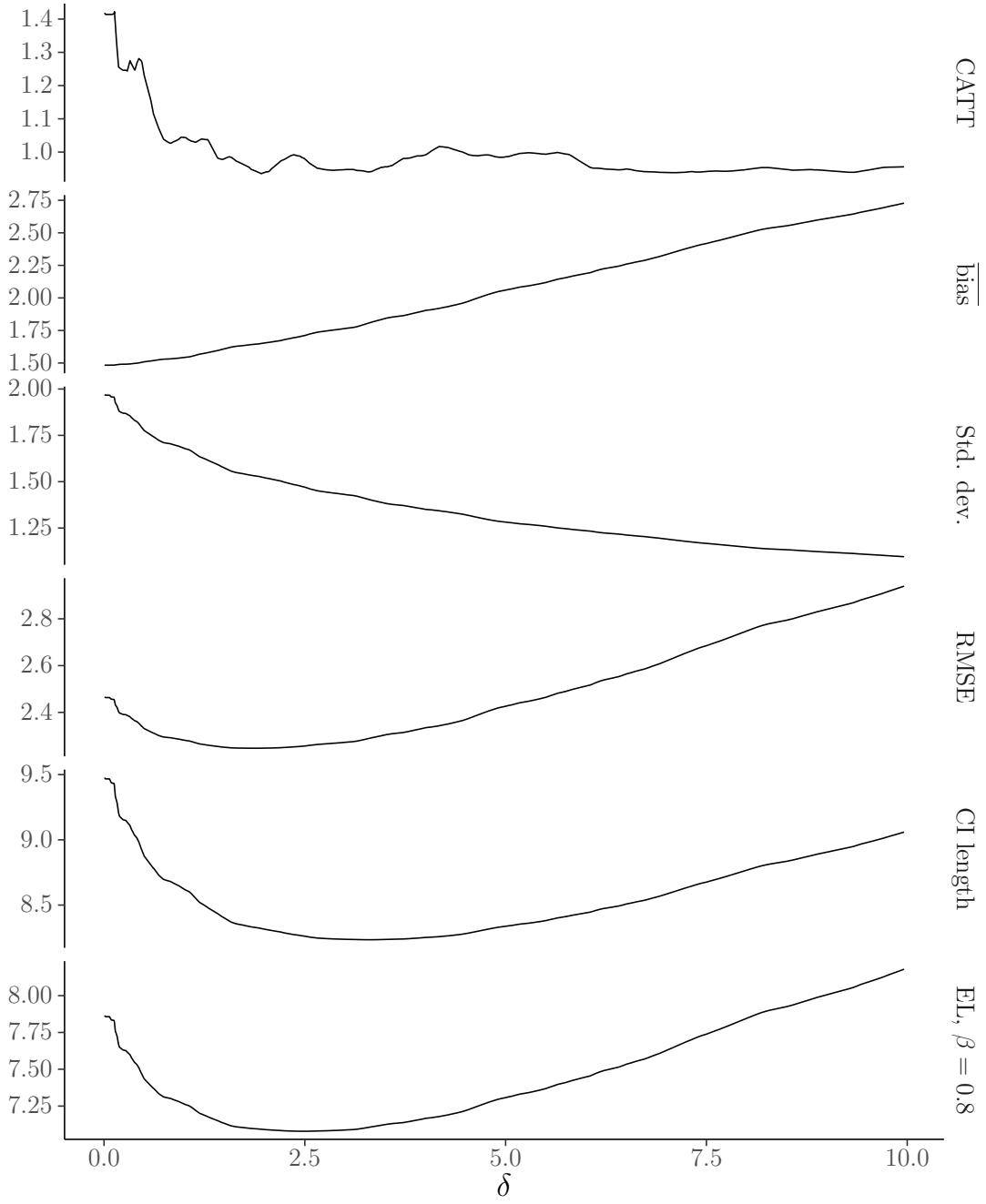

Figure 2: Performance of the optimal estimator as a function of the parameter δ.

Notes: CATT gives the value of the point estimate, $\overline{\text{bias}}$ gives the worst-case bias, and "EL, $\beta = 0.8$" corresponds to the 0.8 quantile of excess length of one-sided CI.

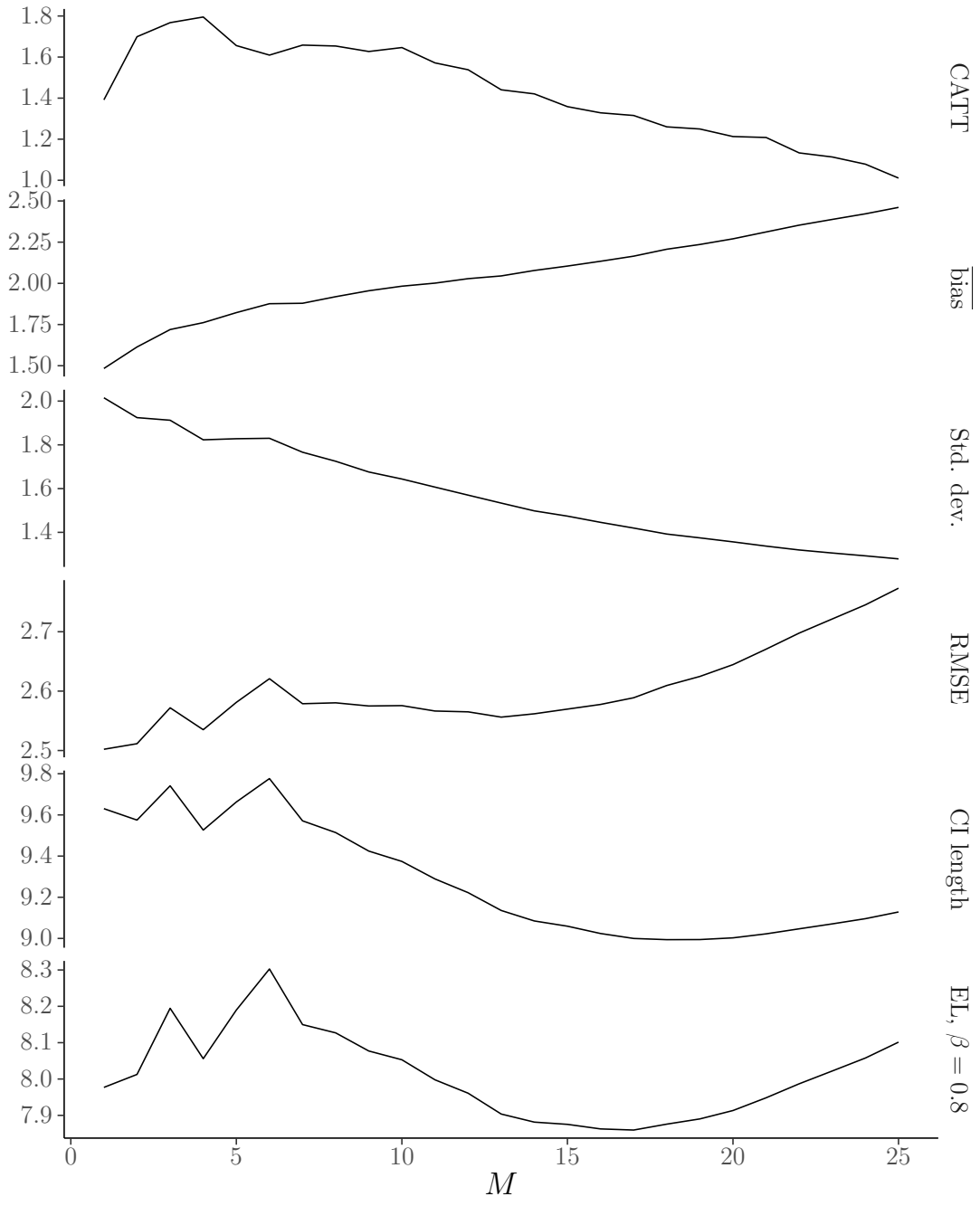

Figure 3: Performance of the matching estimator as a function of the number of matches M.

Notes: CATT gives the value of the point estimate, $\overline{\text{bias}}$ gives the worst-case bias, and "EL, $\beta = 0.8$" corresponds to the 0.8 quantile of excess length of one-sided CI.