

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

Timothy B. Armstrong[†]

Michal Kolesár[‡]

Yale University

Princeton University

November 11, 2020

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, Luke Miratrix, 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. Armstrong acknowledges support by the National Science Foundation Grant SES-1628939. Kolesár acknowledges support by the Sloan Research Fellowship, and by the 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 often 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.

Many of these estimators are \sqrt{n} -consistent, asymptotically unbiased and normally distributed, provided that, 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 (see, among others, [Hahn, 1998](#); [Heckman et al., 1998](#); [Hirano et al., 2003](#); [Chen et al., 2008](#)). The standard approach to constructing confidence intervals (CIs), which we refer to as \sqrt{n} -inference, is to take any such estimator, and add and subtract its standard deviation times the conventional 1.96 critical value (for nominal 95% CIs).

However, in many applications, this approach may not perform well for two reasons. First, often the overlap is limited, which may lead to failure of asymptotic normality and slower convergence rates (e.g. [Khan and Tamer, 2010](#); [Khan and Nekipelov, 2013](#); [Busso et al., 2014](#); [Rothe, 2017](#)).¹ Second, even under perfect overlap, asymptotic unbiasedness requires a lot of smoothness: one typically needs 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. Furthermore, even if the asymptotic bias is negligible, the actual finite-sample bias may affect coverage of standard CIs (e.g. [Robins and Ritov, 1997](#)).

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. Given these restrictions, we show how to construct finite-sample valid CIs based on any estimator $\sum_{i=1}^n k_i Y_i$ that is linear in the outcomes Y_i , where the weights $\{k_i\}_{i=1}^n$ depend on the covariates and treatment.² To do so, we assume that the regression errors are

¹To deal with limited overlap, one can redefine the estimand to a treatment effect for a subset of the population for which overlap holds, as in [Heckman et al. \(1997\)](#), [Galiani et al. \(2005\)](#), [Bailey and Goodman-Bacon \(2015\)](#) or [Crump et al. \(2009\)](#). However, this estimand is typically less policy relevant.

²If the restrictions are asymmetric, we allow for the estimators to be affine rather than linear, taking the

normal with known variance, and we view the treatment and covariates as fixed. The latter allows us to explicitly calculate the worst-case finite-sample bias of the estimator under the maintained restrictions on the conditional mean. Our CIs are constructed by simply adding and subtracting the estimator’s standard error times a critical value that is larger than the usual 1.96 critical value, and takes into account the potential bias of the estimator.³ We then show how to choose the weights $\{k_i\}_{i=1}^n$ optimally, in order to minimize the root mean squared error (RMSE) of the estimator, or the length of the resulting CI: one needs to solve a finite-sample bias-variance tradeoff problem, which can be cast as a convex programming problem. Furthermore, we show that, once the weights are optimized, such estimators and CIs are highly efficient among all procedures.

To make further progress on characterizing the optimal weights, we focus on the case where the conditional mean is assumed to satisfy a Lipschitz constraint. We show that, for a given sample size, when the Lipschitz constant is large enough, the optimal estimator reduces to a matching estimator with a single match. While the optimal estimator does not admit a closed form in general, we develop a computationally fast 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).

When the assumption of normal errors and known variance is dropped, feasible versions of our CIs are valid asymptotically, uniformly in the underlying distribution (i.e. they are honest in the sense of Li, 1989). Importantly, this result obtains whether \sqrt{n} -inference is possible or whether it is impossible, be it due to low regularity of the regression function⁴, limited overlap, or complete lack of overlap; we do not even require that the ATE be point identified. As we further discuss in Remark 3.2, our efficiency theory is analogous to the classic theory in the linear regression model, where finite-sample optimality results rely on the strong assumption of normal homoskedastic errors, but asymptotic validity of CIs based on heteroskedasticity robust standard errors obtains under weak assumptions.

The key condition underlying the asymptotic validity of our CIs is that the estimator doesn’t put too much weight k_i on any individual observation: this ensures that the estimator is asymptotically normal when normalized by its standard deviation. However, since the

form $a + \sum_{i=1}^n k_i Y_i$ where the intercept a may also depend on the covariates and treatment. See Appendix A.

³The worst-case bias calculations require the researcher to fully specify the restrictions on the conditional mean, including any smoothness constants. The 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.

⁴We show that for \sqrt{n} -inference to be possible, one needs to assume a bound on 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, say, estimation of a conditional mean at a point, it is not possible to “undersmooth”.

weights solve a bias-variance tradeoff, no single observation can receive too much weight—otherwise, in large samples, regardless of the amount of overlap, 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: we show that asymptotic normality for matching estimators generally requires strong overlap.

We illustrate the results in an application to the National Supported Work (NSW) Demonstration. We find that our CIs are substantially different from those based on conventional \sqrt{n} -asymptotic theory, with bias determining a substantial portion of the CI width. In line with the theoretical results above, we also find evidence that, in contrast to the estimator we propose, the weights for the matching estimator are too large for a normal approximation to its sampling distribution to be reliable.

Our results rely on the key insight that, if we condition on the treatment and covariates, 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 the sharp efficiency bounds in [Armstrong and Kolesár \(2018b\)](#). The form of the optimal estimator and 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, which makes simple strategies, such as gridding, infeasible unless the dimension of covariates p is very small. We show that under Lipschitz smoothness, the problem can be recast so that the computational complexity depends only on n and not on p , and our solution algorithm uses insights from [Rosset and Zhu \(2007\)](#) to further speed up the computation. In independent and contemporaneous work, [Kallus \(2020\)](#) computes optimal linear weights using a different characterization of the optimization problem.

In contrast, without conditioning on the treatment and covariates, the problem is more difficult: while upper and lower bounds for the rate of convergence have been developed ([Robins et al., 2009](#)), efficiency bounds that are sharp in finite samples remain elusive. Whether one should condition on the treatment and covariates when evaluating estimators and CIs is itself an interesting question. A previous version of this paper ([Armstrong and Kolesár, 2018a](#), Section 6.4) considers this question in the context of our empirical application, and [Abadie et al. \(2014, 2020\)](#) give a discussion in related settings. Since our CIs are valid unconditionally, they can be used in either setting.⁵

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.

⁵While we focus on a treatment effect that conditions on realized covariates in the sample, our approach can also be used to construct CIs for the population ATE; see Section 3.2.

Section 4 presents asymptotic results. Section 5 discusses an application to the NSW data. Additional results, details, and proofs are collected in the 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 n units. Each unit $i = 1, \dots, n$ is characterized by a pair of potential outcomes $Y_i(0)$ and $Y_i(1)$ under no treatment and treatment, respectively, a covariate vector $X_i \in \mathbb{R}^p$, and a treatment indicator $D_i \in \{0, 1\}$. Unless stated otherwise, we condition on the realized values $\{x_i, d_i\}_{i=1}^n$ of the covariates and treatment $\{X_i, D_i\}_{i=1}^n$, so that probability statements are with respect to the conditional distribution of $\{Y_i(0), Y_i(1)\}_{i=1}^n$. The realized outcome is given by $Y_i = Y_i(1)d_i + Y_i(0)(1 - d_i)$. Letting $f(x_i, d_i)$ denote the conditional mean of Y_i , we obtain a fixed design regression model

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

We are interested in the conditional average treatment effect (CATE)⁶, which, under the assumption of unconfoundedness, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$, is given by

$$Lf = \frac{1}{n} \sum_{i=1}^n [f(x_i, 1) - f(x_i, 0)] = \frac{1}{n} \sum_{i=1}^n E[Y_i(1) - Y_i(0) \mid X_i = x_i]. \quad (2)$$

To obtain finite-sample results, we further assume that u_i is normal,

$$u_i \sim N(0, \sigma^2(x_i, d_i)), \quad (3)$$

and that the variance function $\sigma^2(x_i, d_i)$ is known. We relax both assumptions in Section 3.

We assume that f lies in a known function class \mathcal{F} , which formalizes the “regularity” or “smoothness” that we are willing to impose. We require that \mathcal{F} be convex and centrosym-

⁶We note that the terminology varies in the literature. Some papers call this object the sample average treatment effect (SATE); other papers use the terms CATE and SATE for different objects entirely.

metric, i.e. that $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$. While convexity is essential for most of our results, centrosymmetry can be relaxed—see Appendix A. As a leading example, we place Lipschitz constraints on $f(\cdot, 0)$ and $f(\cdot, 1)$:

$$\mathcal{F}_{\text{Lip}}(C) = \{f: |f(x, d) - f(\tilde{x}, d)| \leq C\|x - \tilde{x}\|_{\mathcal{X}}, d \in \{0, 1\}\}, \quad (4)$$

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 $f(\cdot, 0)$. As we discuss in Section 2.5, the function class \mathcal{F} needs to be specified ex ante: for the Lipschitz case, for instance, one cannot use data-driven procedures to estimate C , or pick the norm $\|\cdot\|_{\mathcal{X}}$. We discuss how to make these choices in practice in the context of our empirical application in Section 5.

Our goal is to construct estimators and confidence sets for the CATE parameter Lf . Letting P_f denote the probability computed under f , a set \mathcal{C} is a $100 \cdot (1 - \alpha)\%$ confidence set for Lf if

$$\inf_{f \in \mathcal{F}} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha. \quad (5)$$

2.2 Linear estimators

We start by showing how to construct CIs based on estimators that are linear in the outcomes,

$$\hat{L}_k = \sum_{i=1}^n k(x_i, d_i) Y_i. \quad (6)$$

For now, we treat the weights k as given—in Section 2.3, we will show how to choose them optimally. In Section 2.5 and Appendix A, we show that, if we choose the weights optimally, the resulting estimator and CIs are optimal or near optimal among all procedures, including nonlinear ones. The class of linear estimators covers many estimators that are popular in practice, such as series or kernel estimators, or various matching estimators.⁷ For example, the matching estimator with M matches that matches (with replacement) on covariates takes the form $L\hat{f}_M$, where $\hat{f}_M(x_i, d_i) = Y_i$, and $\hat{f}_M(x_i, 1 - d_i) = \sum_{j=1}^n W_{M,ij} Y_j$. Here $W_{M,ij} = 1/M$ if j is among the M observations with treatment status $d_j = 1 - d_i$ that are the closest to i (using the norm $\|\cdot\|_{\mathcal{X}}$), and zero otherwise. For this estimator, the weights take the form

$$k_{\text{match}, M}(x_i, d_i) = \frac{1}{n}(2d_i - 1) \left(1 + \frac{K_M(i)}{M}\right), \quad (7)$$

⁷Nonlinear estimators include those based on regression trees and artificial neural networks, as well as those using nonlinear thresholding to perform variable selection; see Donoho and Johnstone (1998) for a discussion of cases where, in contrast to the present setting, nonlinear estimators outperform linear estimators.

where $K_M(i) = M \sum_{j=1}^n W_{M,ji}$ is the number of times observation i is matched.

The estimator \hat{L}_k is normally distributed with variance $\text{sd}(\hat{L}_k)^2 = \sum_{i=1}^n k(x_i, d_i)^2 \sigma^2(x_i, d_i)$ and maximum bias

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) = \sup_{f \in \mathcal{F}} E_f(\hat{L}_k - Lf) = \sup_{f \in \mathcal{F}} \left[\sum_{i=1}^n k(x_i, d_i) f(x_i, d_i) - Lf \right]. \quad (8)$$

By centrosymmetry of \mathcal{F} , the minimum bias is given by $-\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$.

We form one-sided $100 \cdot (1 - \alpha)\%$ CIs based on \hat{L}_k , as

$$[\hat{c}, \infty), \quad \hat{c} = \hat{L}_k - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) - z_{1-\alpha} \text{sd}(\hat{L}_k), \quad (9)$$

where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of a standard normal distribution. Subtracting the maximum bias, in addition to subtracting $z_{1-\alpha} \text{sd}(\hat{L}_k)$, is necessary to prevent undercoverage.

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

$$\left\{ \hat{L}_k \pm \text{cv}_{\alpha}(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) / \text{sd}(\hat{L}_k)) \cdot \text{sd}(\hat{L}_k) \right\}. \quad (10)$$

If \hat{L}_k is unbiased, the critical value reduces to the conventional critical value: $\text{cv}_{\alpha}(0) = z_{1-\alpha/2}$. If $b > 0$, it will be larger: for $b \geq 1.5$ and $\alpha \leq 0.2$, $\text{cv}_{\alpha}(b) \approx b + z_{1-\alpha}$ up to three decimal places,⁸ Following [Donoho \(1994\)](#), we refer to the CI in (10) as a fixed-length confidence interval (FLCI), since its length does not depend on the realized outcomes, only on the known variance function $\sigma^2(\cdot, \cdot)$ and on $\{x_i, d_i\}_{i=1}^n$.

2.3 Optimal linear estimators and CIs

We now show how to choose the weights k optimally. To that end, we need to define the criteria that we wish to optimize. To evaluate estimators, we consider their maximum root mean squared error (RMSE),

$$R_{\text{RMSE}, \mathcal{F}}(\hat{L}_k) = (\sup_{f \in \mathcal{F}} E_f(\hat{L}_k - Lf)^2)^{1/2} = (\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)^2 + \text{sd}(\hat{L}_k)^2)^{1/2}. \quad (11)$$

⁸The critical value $\text{cv}_{1-\alpha}(b)$ can be computed 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 .

One-sided CIs can be compared using quantiles of excess length (see Appendix A). Finally, to evaluate FLCIs, we simply consider their length, $2 \text{cv}_\alpha(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) / \text{sd}(\hat{L}_k)) \cdot \text{sd}(\hat{L}_k)$. Since the length is fixed—it doesn't depend on the data $\{Y_i\}_{i=1}^n$ —choosing the weights k to minimize the length does not affect the coverage properties of the resulting CI.

Both performance criteria—FLCI and RMSE—depend on k only through $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$, and $\text{sd}(\hat{L}_k)$, and they are increasing in both quantities (this is also true for performance of one-sided CIs; see Appendix A). Therefore, to find the optimal weights, it suffices to first find weights that minimize the worst-case bias $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$ 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 convex optimization indexed by a parameter δ that plays a role analogous to that of a bandwidth; it can be thought of as indexing the relative weight on the variance. Varying δ then traces out the optimal bias-variance frontier.

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 $\iota_\kappa(x, d) = \kappa d$ lies in \mathcal{F} for all $\kappa \in \mathbb{R}$ (see Appendix A for a general statement)⁹. Given $\delta > 0$, let f_δ^* solve

$$\max_{f \in \mathcal{F}} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^n \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4}. \quad (12)$$

With a slight abuse of notation, define

$$\hat{L}_\delta = \hat{L}_{k_\delta^*}, \quad k_\delta^*(x_i, d_i) = \frac{f_\delta^*(x_i, d_i) / \sigma^2(x_i, d_i)}{\sum_{j=1}^n d_j f_\delta^*(x_j, d_j) / \sigma^2(x_j, d_j)}. \quad (13)$$

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

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

Also, \hat{L}_δ minimizes the worst-case bias among all linear estimators with variance bounded by $\text{sd}(\hat{L}_\delta)^2 = \delta^2 / (2 \sum_{j=1}^n d_j f_\delta^*(x_j, d_j) / \sigma^2(x_j, d_j))^2$. Thus, the estimators $\{\hat{L}_\delta\}_{\delta > 0}$ trace out the optimal bias-variance frontier.

The weights leading to the shortest possible FLCI are given by $k_{\delta_{\text{FLCI}}}^*$, where δ_{FLCI} min-

⁹We also assume the regularity condition that if $\lambda f + \iota_\kappa \in \mathcal{F}$ for all $0 \leq \lambda < 1$, then $f + \iota_\kappa \in \mathcal{F}$. Since $L\iota_\kappa = \kappa$, $\{\iota_\kappa\}_{\kappa \in \mathbb{R}}$ is the smoothest set of functions that span the potential values of the CATE parameter, so this assumption will typically hold if the possible values of Lf are unrestricted.

minimizes $\text{cv}_\alpha(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta)/\text{sd}(\hat{L}_\delta)) \cdot \text{sd}(\hat{L}_\delta)$ over δ . Similarly, the optimal weights for estimation are given by $k_{\delta_{\text{RMSE}}}^*$, where δ_{RMSE} minimizes $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta)^2 + \text{sd}(\hat{L}_\delta)^2$.

2.4 Estimators and CIs under Lipschitz smoothness

Computing a FLCI based on a linear estimator \hat{L}_k with a given set of weights k requires computing the worst-case bias (8). Computing the RMSE-optimal estimator, and the optimal FLCI requires solving the optimization problem (12), and then varying δ to find the optimal bias-variance tradeoff. Both 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 (8) can be found by solving a finite-dimensional linear program. The optimization problem (12) can be cast as a finite-dimensional convex program. Furthermore, if the program is put into a Lagrangian form, then the solution is a piecewise linear 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).

We leverage the fact that in both optimization problems, we can identify f with the vector $(f(x_1, 0), \dots, f(x_n, 0), f(x_1, 1), \dots, f(x_n, 1))' \in \mathbb{R}^{2n}$, and replace the functional constraint $f \in \mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$ with $2n(n-1)$ linear inequality constraints

$$f(x_i, d) - f(x_j, d) \leq C\|x_i - x_j\|_{\mathcal{X}} \quad d \in \{0, 1\}, \quad i, j \in \{1, \dots, n\}. \quad (14)$$

This fact follows from the observation that both in (8) and in (12), the objective and constraints depend on f only through its value at these $2n$ points, and from the result that if the Lipschitz constraints hold at these points, it is always possible to find a function $f \in \mathcal{F}_{\text{Lip}}(C)$ that interpolates these points (Beliakov, 2006, Theorem 4, which we restate in Lemma A.1).

Theorem 2.1. *Consider a linear estimator (6) with weights k that satisfy*

$$\sum_{i=1}^n d_i k(x_i, d_i) = 1, \quad \text{and} \quad \sum_{i=1}^n (1 - d_i) k(x_i, d_i) = -1. \quad (15)$$

Then

$$\overline{\text{bias}}_{\mathcal{F}_{\text{Lip}}(C)}(\hat{L}_k) = \max_{f \in \mathbb{R}^{2n}} \left\{ \sum_{i=1}^n k(x_i, d_i) f(x_i, d_i) - \frac{1}{n} \sum_{i=1}^n [f(x_i, 1) - f(x_i, 0)] \right\}, \quad \text{s.t. (14)}. \quad (16)$$

Furthermore, if $k(x_i, d_i) \geq 1/n$ if $d_i = 1$ and $k(x_i, d_i) \leq -1/n$ if $d_i = 0$, it suffices to impose (14) for $i, j \in \{1, \dots, n\}$ with $d_i = 1, d_j = 0$ such that either $k(x_i, 1) > 1/n$, or

$$k(x_j, 0) < -1/n.$$

The assumption that \hat{L}_k satisfies (15) is necessary to prevent the bias from becoming arbitrarily large at multiples of $f(x, d) = d$ and $f(x, d) = 1 - d$. 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 in (16).

The last part of the theorem says that it suffices to impose at most $2n_0n_1$ of the constraints in (14), 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 (7). 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 (14) for pairs of observations with opposite treatment status for which one of the observations is used as a match.

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 (12) under $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$ is given by the solution to the convex program*

$$\max_{f \in \mathbb{R}^{2n}} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^n \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4} \quad \text{and s.t. (14)}. \quad (17)$$

Furthermore, if $\sigma^2(x, d)$ doesn't depend on x , it suffices to impose the constraints (14) 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 reduces the infinite-dimensional program (12) to a quadratic optimization problem in \mathbb{R}^{2n} with one quadratic and $2n(n - 1)$ linear constraints, and a linear objective function. The computational difficulty can be shown to be polynomial in n , and it does not depend on the covariate dimension p . If the variance is homoskedastic 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 a result, implementing the estimator is quite fast: the main specification in the empirical application in Section 5 takes less than a minute on a laptop computer.

As we discuss in more detail in Appendix A.3, it follows from the algorithm that, similarly to the matching estimator (see eq. (7)), the optimal estimator takes the form $\hat{L}_\delta = L\hat{f}_\delta$, where $\hat{f}_\delta(x_i, d_i) = Y_i$, and $\hat{f}_\delta(x_i, 1 - d_i) = \sum_{j=1}^n W_{\delta, ij} Y_j$, and the weights $W_{\delta, ij}$ correspond to the Lagrange multipliers associated with the constraints (14) for $d = d_i$, scaled to sum to one, $\sum_{j=1}^n W_{\delta, ij} = 1$. The weights are zero unless $d_j = 1 - d_i$ and j is close to i according to a matrix of “effective distances.” The “effective distance” between i and j increases with the

total weight $\sum_{i=1}^n W_{\delta,ij}$ that we already put on j . Thus, we may interpret observations j with non-zero weight $W_{\delta,ij}$ as being “matched” to i . The number of matches varies across observations i , increases with δ , and depends on the number of units with opposite treatment status that are close to i according to the matrix of effective distances. Observations for which there are more good matches receive relatively more matches, since this decreases the variance of the estimator at little cost in terms of bias. On the other hand, the weight $k_{\delta}^*(x_j, d_j) = \frac{1}{n}(1 - 2d_j)(1 + \sum_{i=1}^n W_{\delta,ij})$ on j increases with the number of times it has been used as a match, which 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}_{δ} is given by the matching estimator with $M = 1$.*

In contemporaneous work, [Kallus \(2020\)](#) gives a similar result using a different method of proof. In the other direction, as $C/\delta \rightarrow 0$, \hat{L}_{δ} converges to the difference-in-means estimator that compares the average outcomes for the treated and untreated units.

Theorem 2.3 does not imply that one should choose a large C simply to justify matching with a single match as an optimal estimator: the chosen value of C should instead represent a priori bounds on the smoothness of f formulated by the researcher. Nonetheless, if it is difficult to formulate such bounds, a conservative choice of C may be appropriate. If the choice is conservative enough, Theorem 2.3 will be relevant.

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 that 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. In the supplemental materials, we illustrate the efficiency loss from matching with the “wrong” metric in the context of our empirical application presented in Section 5.

A disadvantage of imposing the Lipschitz condition directly on f is that it rules out the simple linear model for f , unless we impose a priori bounds on the magnitude of the regression coefficients. In Appendix A.2, we give analogs of Theorems 2.1, 2.2 and 2.3 under

an alternative specification for \mathcal{F} that imposes the Lipschitz condition on f after partialling out the best linear predictor (and thus allows for unrestricted linear response). We show that in this case, if the constant C is large enough, the optimal estimator takes the form of a regression-adjusted matching estimator with a single match.

2.5 Adaptation bounds and optimality among nonlinear procedures

The results in Section 2.3 and Theorem 2.2 show how to construct RMSE-optimal linear estimators, and the shortest FLCI based on a linear estimator. Are these results useful, or do they overly restrict the class of procedures?

For estimation under the RMSE criterion, Theorem A.2 in Appendix A.1 shows that the estimator $\hat{L}_{\delta_{\text{RMSE}}}$ achieving the lowest RMSE in the class of linear estimators (6) is also highly efficient among all estimators: one cannot reduce the RMSE by more than 10.6% by considering non-linear estimators in general, and, in particular applications, its efficiency can be shown to be even higher.

For FLCIs, an even stronger result obtains, addressing two concerns. First, their length 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 all of \mathcal{F}), especially if this leads to substantial reduction in expected length when $f \in \mathcal{G}$. When such a CI also simultaneously achieves near-optimal length over all of \mathcal{F} , it is referred to as “adaptive.” A related second concern is that implementing our CIs in practice requires the user to explicitly specify the parameter space \mathcal{F} , which in the case $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$ includes the Lipschitz constant C and the norm $\|\cdot\|_{\mathcal{X}}$. This rules out data-driven procedures that try to implicitly or explicitly pick C or the norm using the data.

To address these concerns, Theorem A.3 in Appendix A.1 gives a sharp bound on the length of a confidence set that optimizes its expected length at a smooth function of the form $g(x, d) = \kappa_0 + \kappa_1 d$, while maintaining coverage over the original parameter space \mathcal{F} . The sharp bound follows from general results in Armstrong and Kolesár (2018b), and it gives a benchmark for the scope for improvement over the FLCI centered at $\hat{L}_{\delta_{\text{FLCI}}}$ (the theorem also gives an analogous result for one-sided CIs). The theorem also gives a universal lower bound for this sharp bound, which evaluates to 71.7% when $1 - \alpha = 0.95$. The sharp bound depends on the realized values of $\{x_i, d_i\}_{i=1}^n$ and the variance function $\sigma^2(\cdot, \cdot)$, and can be explicitly computed in a given application. We find that it typically is much higher than this

lower bound. For example, in our empirical application in Section 5, the FLCI efficiency is over 97% at such smooth functions g in our baseline specification. This implies that there is very little scope for improvement over the FLCI.

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 order of differentiability of f , or to estimate the Lipschitz constant C for the purposes of forming a tighter CI—the parameter space \mathcal{F} , including any smoothness constants, must be specified ex ante by the researcher. Because of this, by way of sensitivity analysis, since the CI may undercover if C is chosen too small in the sense that $\mathcal{F}_{\text{Lip}}(C)$ excludes the true f , we recommend reporting estimates and CIs for a range of choices of C to see how assumptions about the parameter space affect the results (the estimates and CIs vary with C because C affects both the optimal tuning parameters δ_{RMSE} and δ_{FLCI} , and the critical value, via the worst-case bias). We adopt this approach in the empirical application in Section 5. This mirrors the common practice of reporting results for different specifications of the regression function in parametric regression problems.

The key assumption needed for these efficiency bounds is that \mathcal{F} be convex and centrosymmetric. This holds for $\mathcal{F}_{\text{Lip}}(C)$, 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, a previous version of this paper (Armstrong and Kolesár, 2018a, Section 6.4) shows to implement our approach when \mathcal{F} imposes Lipschitz and monotonicity constraints.

3 Practical implementation

We now discuss implementation of feasible versions of the estimators and CIs given in Section 2 when the variance function $\sigma^2(x, d)$ is unknown and the errors u_i may be non-normal. We discuss the optimality and validity of these feasible procedures. In Section 3.2, we show how to use our approach to form CIs for the population average treatment effect (PATE).

3.1 Baseline implementation

As a baseline, we propose the following implementation of our procedure:¹⁰

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

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 $\{\tilde{k}_\delta^*\}_{\delta>0}$, using $\tilde{\sigma}^2(x, d)$ in place of $\sigma^2(x, d)$. When $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, this can be done using the piecewise linear solution path $\{\tilde{f}_\delta^*\}_{\delta>0}$ in Appendix A.3. Let $\tilde{L}_\delta = \sum_{i=1}^n \tilde{k}_\delta^*(x_i, d_i) Y_i$ denote the corresponding estimator, $\tilde{\text{sd}}_\delta^2 = \sum_{i=1}^n \tilde{k}_\delta^*(x_i, d_i)^2 \tilde{\sigma}^2(x_i, d_i)$ denote its variance computed using $\tilde{\sigma}^2(x, d)$ as the variance function, and let $\overline{\text{bias}}_\delta = \overline{\text{bias}}_{\mathcal{F}}(\tilde{L}_\delta)$ denote its worst-case bias (which doesn't depend on the variance specification).
3. Compute the minimizer $\tilde{\delta}_{\text{RMSE}}$ of $\overline{\text{bias}}_\delta^2 + \tilde{\text{sd}}_\delta^2$. Compute the standard error $\text{se}(\tilde{L}_{\tilde{\delta}_{\text{RMSE}}})$ using the robust variance estimator

$$\text{se}(\tilde{L}_\delta)^2 = \sum_{i=1}^n \tilde{k}_\delta^*(x_i, d_i)^2 \hat{u}_i^2, \quad (18)$$

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

$$\{\tilde{L}_\delta \pm \text{cv}_\alpha(\overline{\text{bias}}_\delta / \text{se}(\tilde{L}_\delta)) \text{se}(\tilde{L}_\delta)\}, \quad (19)$$

at $\delta = \tilde{\delta}_{\text{FLCI}}$, the minimizer of $\text{cv}_\alpha(\overline{\text{bias}}_\delta / \tilde{\text{sd}}_\delta) \cdot \tilde{\text{sd}}_\delta$.

The conditional variance, can, for instance, be estimated with 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 $\hat{f}(x_i, d_i)$ is the average outcome of J observations (excluding i) with treatment status d_i that are closest to i according to some distance $\|\cdot\|$. Note that since the length of the feasible CI in eq. (19) depends on the variance estimates, it is no longer fixed, in contrast to the infeasible FLCI.

In general, the point estimate $\tilde{L}_{\tilde{\delta}_{\text{RMSE}}}$ will differ from the estimate $\tilde{L}_{\tilde{\delta}_{\text{FLCI}}}$ used to form the CI. Since reporting multiple estimates can be cumbersome, one can simply compute the CI (19) at $\delta = \tilde{\delta}_{\text{RMSE}}$. The CI will then be based on the same estimator reported as a point estimate. While this leads to some efficiency loss, in our main specification in the empirical application in Section 5, we find that the resulting CI is only 1.2% longer than the one that reoptimizes δ for CI length.

Remark 3.1 (Specification of \mathcal{F}). In forming the CI, we need to choose the function class \mathcal{F} . In case of the Lipschitz class (4), we need to complete the specification by choosing the constant C and the norm $\|\cdot\|_{\mathcal{X}}$ on x . The results discussed in Section 2.5 imply that it is not possible to do this 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 illustrate this approach in our application in Section 5.1. Note that conducting this sensitivity analysis comes at essentially no added computational cost, since the solution path $\{\tilde{f}_\delta^*\}_{\delta>0}$ only needs to be computed once. This is because multiplying both δ and C by any constant scales the constraints in (17), so that the solution simply scales with the given constant. In particular, letting $\tilde{f}_{\delta,C}^*$ denote the solution under a given δ and C , we have $\tilde{f}_{\delta,C}^* = C\tilde{f}_{\delta/1}^*$.

Remark 3.2 (Efficiency and validity). How do the finite-sample optimality and validity of the infeasible estimators and CIs discussed in Section 2 map into optimality and validity properties of the feasible procedures?

The values $\tilde{\delta}_{\text{RMSE}}$ and $\tilde{\delta}_{\text{FLCI}}$ depend on the initial guess $\tilde{\sigma}^2(x, d)$. Thus, the resulting CI in eq. (19) will not be optimal if this guess is incorrect. However, because the standard error estimator (18) does not use this initial estimate, the CI remains asymptotically valid even if $\tilde{\sigma}^2(x, d)$ is incorrect. Furthermore, we show in Section 4.2 that the CI is asymptotically valid even in “irregular” settings when \sqrt{n} -inference is impossible, including when the CATE is set identified (in which case the CI is asymptotically valid for points in the identified set).

Second, the worst-case bias calculations do not depend on the error distribution. Although its coverage guarantees are only asymptotic, the feasible CI reflects the finite-sample impact of the covariate and treatment realizations (including the degree of overlap in the data) on the bias of the estimator through the critical value. Similarly, the bias-variance tradeoffs discussed in Section 2.3 still go through even if the errors are not normal, since only the variance of the error distribution affects the underlying calculations. Our recommendation to assume homoskedasticity when setting the initial variance estimates is motivated by the fact that under homoskedasticity, using a constant initial variance function yields an estimator that has the finite-sample optimality property of minimizing variance among estimators with the same worst-case bias. Also, if the initial variance estimate is correct, then the feasible estimator $\tilde{L}_{\tilde{\delta}_{\text{RMSE}}}$ will be optimal (in the sense discussed in Section 2.3) even when the errors are non-normal.¹¹ We can take advantage of this fact in large samples under homoskedasticity, when the initial variance estimator is consistent. At the same time, the CIs retain asymptotic validity under weak conditions. These optimality and validity properties mirror those of the ordinary least squares (OLS) estimator along with heteroskedasticity robust standard errors in a linear regression model: the estimator is optimal under homoskedasticity if one assumes normal errors, or if one restricts attention to linear estimators,

¹¹The result on finite-sample optimality among non-linear procedures discussed in Section 2.5 likewise goes through under non-normal errors, so long as the set of possible distributions for u_i includes normal errors, and its second moment is bounded.

while the CIs are asymptotically valid under heteroskedastic and non-normal errors.

Remark 3.3 (CIs based on other estimators). Feasible CIs based on linear estimators \hat{L}_k (6) can be formed as in the baseline implementation, using the weights $k(x_i, d_i)$ in eq. (18), and computing the worst-case bias by solving the optimization problem in Theorem 2.1.

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.

3.2 CI for the population average treatment effect

We now show how our approach can be adapted to construct CIs for the PATE based on linear estimators \hat{L}_k of the form (6). To do so, we treat the covariates and treatment as random. Under random sampling, the PATE is given by $\tau = E[Y_i(1) - Y_i(0)] = E[Lf]$, where Lf is the CATE as in eq. (2).

If we view \hat{L}_k as an estimator of τ , the quantity $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$ now represents its worst-case bias conditional on $\{X_i, D_i\}_{i=1}^n$, and is therefore random under i.i.d. sampling. We thus cannot use the arguments underlying the construction in eq. (10). Instead, we simply add and subtract $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k)$, in addition to adding and subtracting the usual critical value times a standard error based on the marginal, rather than conditional, variance of \hat{L}_k ,

$$\{\hat{L}_k \pm (\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) + z_{1-\alpha/2} \text{se}_{\tau}(\hat{L}_k))\}. \quad (20)$$

Here $\text{se}_{\tau}(\hat{L}_k)^2 = \text{se}(\hat{L}_k)^2 + \text{se}(Lf)^2$, where $\text{se}(\hat{L}_k)^2$ is the conditional variance of the estimator using the weights $k(\cdot)$ in eq. (18), and $\text{se}(Lf)^2$ is a consistent estimator of the variance of Lf , $\frac{1}{n}E[(f(X_i, 1) - f(X_i, 0) - \tau)^2]$. For the latter, if we use the nearest neighbor variance estimator \hat{u}_i^2 in (18), and the linear estimator takes the form $\hat{L}_k = L\hat{f}$, where $\hat{f}(x_i, d_i) = Y_i$ and $\hat{f}(x_i, 1 - d_i) = \sum_{j=1}^n W_{ij}Y_j$, where $\sum_{j=1}^n W_{ij} = 1$ and $k_j = (1 - 2d_j)(1 + \sum_{i=1}^n W_{ij})/n$ (as discussed in Section 2.2 and following Theorem 2.2, this includes matching estimators, as well as the optimal estimator), one may use the nearest neighbor estimator suggested by [Abadie and Imbens \(2006, Theorem 7\)](#), $\frac{1}{n^2} \sum_{i=1}^n (\hat{f}(x_i, 1) - \hat{f}(x_i, 0) - \hat{L}_k)^2 - \frac{1}{n^2} \sum_{i=1}^n (1 + \sum_{j=1}^n W_{ji}^2) \hat{u}_i^2$. We note that the optimality results discussed in Remark 3.2 do not apply to the CI in (20). In Appendix B.1, we provide formal asymptotic coverage results for this CI and its one-sided analog.

4 Asymptotic results

In Section 4.1, we show that \sqrt{n} -inference is impossible when the dimension of covariates is high enough relative to the order of smoothness imposed by \mathcal{F} . Section 4.2 gives conditions for asymptotic validity of the feasible CIs when the variance function is unknown and the errors u_i may be non-normal. Section 4.3 discusses conditions for asymptotic validity and optimality of CIs based on matching estimators.

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

As discussed in the introduction, the standard approach to inference, which we refer to as \sqrt{n} -inference, is based on estimators that are \sqrt{n} -consistent and asymptotically normal, with asymptotically negligible bias. We now show that if the dimension of the (continuously distributed) covariates p is high enough relative to the smoothness of \mathcal{F} , this approach is infeasible.

To state the result, let $\Sigma(\gamma, C)$ denote the set of ℓ -times differentiable functions f such that, for all integers k_1, k_2, \dots, k_p with $\sum_{j=1}^p k_j = \ell$, $\left| \frac{\partial^\ell f(x)}{\partial x_1^{k_1} \dots \partial x_p^{k_p}} - \frac{\partial^\ell f(x')}{\partial x_1^{k_1} \dots \partial x_p^{k_p}} \right| \leq C \|x - x'\|_{\mathcal{X}}^{\gamma-\ell}$, where ℓ is the greatest integer strictly less than γ and $\|x\|_{\mathcal{X}}^2 = \sum_{j=1}^p x_j^2$. Note that $f \in \mathcal{F}_{\text{Lip}}(C)$ is equivalent to $f(\cdot, 1), f(\cdot, 0) \in \Sigma(1, C)$.

Theorem 4.1. *Let $\{X_i, D_i\}$ be i.i.d. with $X_i \in \mathbb{R}^p$ and $D_i \in \{0, 1\}$. Suppose that the Gaussian regression model (1) and (3) holds conditional on the realizations of the treatment and covariates. Suppose that the marginal probability that $D_i = 1$ is not equal to zero or one and that X_i has a bounded density conditional on D_i . Given γ, C , let $[\hat{c}_n, \infty)$ be a sequence of CIs for Lf with asymptotic coverage at least $1 - \alpha$ under $\Sigma(\gamma, C)$ conditional on $\{X_i, D_i\}_{i=1}^n$:*

$$\liminf_{n \rightarrow \infty} \inf_{f(\cdot, 0), f(\cdot, 1) \in \Sigma(\gamma, C)} P_f(Lf \in [\hat{c}_n, \infty) \mid \{X_i, D_i\}_{i=1}^n) \geq 1 - \alpha$$

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

$$\liminf_{n \rightarrow \infty} P_0(\hat{c}_n \leq -\eta n^{-\gamma/p} \mid \{X_i, D_i\}_{i=1}^n) \geq 1 - \alpha \quad \text{almost surely.}$$

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 corresponds to $\gamma = 1$, so that \sqrt{n} -inference is possible only when $p \leq 2$. While Theorem 4.1 considers a setting with normal errors, the same bound applies if the normality assumption is dropped (so long as the class of possible distributions

for u_i includes the normal distribution), since including other distributions only makes the problem more difficult. Theorem 4.1 requires coverage conditional on the realizations of the covariates and treatment. For unconditional coverage, the results of Robins et al. (2009) imply that if $e \in \Sigma(\gamma_e, C)$, where $e(x) = P(D_i = 1 \mid X_i = x)$ denotes the propensity score, and if $f(\cdot, 0), f(\cdot, 1) \in \Sigma(\gamma, C)$, then \sqrt{n} -inference is impossible unless $\gamma_e + \gamma \geq p/2$. Thus, conditioning effectively takes away any role of smoothness of the propensity score.

4.2 Asymptotic validity of feasible CIs

The following theorem gives sufficient conditions for the asymptotic validity of the feasible CIs given in Section 3.1 based on the estimator \tilde{L}_δ when $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$. To allow us to capture cases in which the finite-sample bias of the estimator is non-negligible even though the asymptotic bias is negligible under standard asymptotics where the parameter space is fixed with the sample size, we allow $C = C_n \rightarrow \infty$ as $n \rightarrow \infty$. For concreteness, we restrict attention to standard errors based on nearest neighbor estimates, and for simplicity, we focus on the case where the preliminary variance estimate $\tilde{\sigma}^2(x, d)$ is non-random, leaving the extension to random $\tilde{\sigma}(x, d)$ to future research.

Theorem 4.2. *Consider the model (1). Suppose that (a) $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_n)$, $1/K \leq Eu_i^2 \leq K$ and $E|u_i|^{2+1/K} \leq K$ for some constant K , and that the variance function $\sigma^2(x, d)$ is uniformly continuous in x for $d \in \{0, 1\}$; and (b)*

$$\text{for all } \eta > 0 \quad \min_{1 \leq i \leq n} \#\{j \in \{1, \dots, n\} : \|x_j - x_i\|_X \leq \eta/C_n, d_i = d_j\} \rightarrow \infty. \quad (21)$$

Let \mathcal{C} be the CI in eq. (19) based on the feasible estimator \tilde{L}_δ , 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 (18) is the nearest-neighbor variance estimator based on a fixed number of nearest neighbors J . Then $\liminf_{n \rightarrow \infty} \inf_{f \in \mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha$.

The main regularity condition, eq. (21), only requires that the covariate distribution for the treated and control units is well-behaved: we do not require overlap between these two distributions. If $C_n = C$ does not change with n , a bounded support condition suffices:

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 (21) holds almost surely.*

Thus, by Theorem 4.2, feasible CIs are asymptotically valid in settings in which \sqrt{n} -inference is impossible, including in settings in which the covariate dimension p is high enough so that Theorem 4.1 applies, and settings with imperfect overlap (as in Khan and

(Tamer, 2010) including set identification due to complete lack of overlap. The estimator \tilde{L}_δ remains asymptotically normal, when normalized by its standard deviation: the “irregular” nature of the setting only shows up through non-negligible asymptotic bias, which is captured by the critical value cv_α when constructing the CI, and through a slower rate of convergence of the estimator (so the impossibility result of Theorem 4.1 is not contradicted).

Remark 4.1 (Lindeberg weights). The key to establishing Theorem 4.2 is showing that the estimator, when normalized by its standard deviation, converges to a normal distribution. This follows from a central limit theorem (CLT), provided that the Lindeberg condition holds. This in turn requires that the estimator doesn’t put too much weight $\tilde{k}_\delta^*(x_j, d_j)$ on any individual observation in the sense that for $k = \tilde{k}_\delta^*$, as $n \rightarrow \infty$,

$$\text{Lind}(k) = \frac{\max_{1 \leq j \leq n} k(x_j, d_j)^2}{\sum_{i=1}^n k(x_i, d_i)^2} \rightarrow 0. \quad (22)$$

To give intuition for why (22) indeed holds, recall that, as discussed below Theorem 2.2, putting more weight on any individual observation j (by using it often as a match) increases the variance of the estimator. Under condition (21), there are other observations that are almost as good of a match as j (since there are within distance η/C_n to j), so it can’t be optimal to place too much weight on j : using these other observations as a match instead of j would lower the variance of the estimator at little cost to bias.

One may nonetheless be concerned that in finite-samples, the CLT approximation is not accurate. This can be assessed directly by computing $\text{Lind}(\tilde{k}_\delta^*)$ and checking whether it is close to 0. We do this in our application in Section 5.¹²

4.3 Asymptotic properties of CIs based on matching estimators

For feasible CIs based on matching estimators, we obtain the following result:

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

$$\underline{G}(a_n) \leq \frac{\#\{i: \|x_i - x\|_{\mathcal{X}} \leq a_n, d_i = d\}}{n} \leq \bar{G}(a_n) \quad \text{all } x \in \mathcal{X}, d \in \{0, 1\} \quad (23)$$

¹²One can also directly ensure that $\text{Lind}(\tilde{k}_\delta^*)$ is small by only optimizing the RMSE or CI length in step 3 of the baseline implementation in Section 3.1 over values of δ large enough so that $\text{Lind}(\tilde{k}_\delta^*)$ is below a pre-specified cutoff. This is analogous to the suggestion by Noack and Rothe (2020) to only consider large enough bandwidth values in a regression discontinuity setting so that the resulting Lindeberg weights are small.

for large enough n . Let \mathcal{C} be the CIs in Remark 3.3 based on the matching estimator with a fixed number of matches M , and $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_n)$. Then $\liminf_{n \rightarrow \infty} \inf_{f \in \mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha$.

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 remain valid even when p is large, since they are widened to take into account the potential bias of the estimator. 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 $\mathcal{F}_{\text{Lip}}(C)$.

Relative to Theorem 4.2, Theorem 4.3 requires the additional condition (23). This condition holds 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 (Pollard, 1984, Theorem 37, p. 34). The condition $\lim_{t \rightarrow 0} \overline{G}(\underline{G}^{-1}(t))^2 / [t / \log t^{-1}] = 0$ can thus be interpreted as an overlap condition. In particular, if the density of X_i is bounded away from zero and infinity on a sufficiently regular support, then the condition holds if the propensity score $P(D_i = 1 | X_i)$ is bounded away from zero and one.

In the supplemental materials, we strengthen the conclusions of Theorem 4.3 and give an asymptotic analog of Theorem 2.3 showing that, if there is sufficient overlap and $p > 2$, matching with $M = 1$ matches is asymptotically efficient under the RMSE and FLCI criteria. Thus, while it may at first appear that, as argued in Abadie and Imbens (2006), the matching estimator is inefficient due to its slower than \sqrt{n} rate of convergence, by Theorem 4.1, the \sqrt{n} -rate is not feasible in this setting: the matching estimator in fact achieves the fastest possible rate, and, when $M = 1$, the constant is also asymptotically optimal.

On the other hand, if there is not sufficient overlap, asymptotic normality may fail for the matching estimator. As an extreme example, suppose that $p = 1$ and that $x_j < x_i$ for all observations where $d_j = 0$ and $d_i = 1$. Then each untreated observation will be matched to the same treated observation, the one with the smallest value of x_i among treated observations. Consequently, the Lindeberg weight defined in eq. (22) will be bounded away from zero for this observation, and the CLT will fail. In contrast, by Lemma 4.1, the estimator \tilde{L}_δ will be asymptotically normal when scaled by its standard deviation 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 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. Similarly to the discussion in Remark 4.1, in finite samples, one can assess these concerns directly by computing the maximum Lindeberg weight $\text{Lind}(k_{\text{match},M})$. 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.

5 Empirical application

We now illustrate our methods with an application to the National Supported Work (NSW) demonstration. We use the same dataset as Dehejia and Wahba (1999) and Abadie and Imbens (2011).¹³ In particular, the treated sample corresponds to 185 men in the NSW experimental sample with non-missing prior earnings data who were randomly assigned to receive job training after December 1975, and completed it by January 1978; the sample with $d_i = 0$ is a non-experimental sample of 2,490 men taken from the PSID. The outcome Y_i corresponds to earnings in 1978 in thousands of dollars, and the covariate vector x_i contains the variables: age, education, indicators for black and Hispanic, indicator for marriage, earnings in 1974, earnings in 1975, and employment indicators for 1974 and 1975.¹⁴ We assume that the unconfoundedness assumption holds given this covariate vector.

We are interested in the conditional average treatment effect on the treated (CATT),

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

The analysis in Section 2 goes through essentially unchanged with this definition of Lf (see Appendix A).

¹³We use the data from Rajeev Dehejia’s website, <http://users.nber.org/~rdehejia/nswdata2.html>.

¹⁴Following Abadie and Imbens (2011), the no-degree indicator variable is dropped, and the employment indicators are defined as an indicator for nonzero earnings.

Table 1: Diagonal elements of the weight matrix A in definition of the norm (24) for the main specification, A_{main} .

	Age	Educ.	Black	Hispanic	Married	Earnings		Employed	
						1974	1975	1974	1975
A_{main}	0.15	0.60	2.50	2.50	2.50	0.50	0.50	0.10	0.10

5.1 Implementation details

To construct feasible versions of our estimators and CIs, we follow the baseline implementation in Section 3.1. Implementing the procedure requires us to fix the norm $\|\cdot\|_{\mathcal{X}}$ and the smoothness constant C in the definition (4). We consider weighted ℓ_q norms of the form

$$\|x\|_{A,q} = \left(\sum_{j=1}^p |A_{jj}x_j|^q \right)^{1/q}, \quad (24)$$

where A is a diagonal matrix. Let us now describe our main specification; we discuss other specifications in the supplemental materials. To make the restrictions on f implied by the choice of norm and C interpretable, in our main specification, we set $C = 1$, and $A = A_{\text{main}}$, with the diagonal elements of A_{main} given in Table 1. The j th diagonal element A_{jj} then gives the a priori bound on the derivative of the regression function with respect to x_j (i.e. the partial effect of increasing x_j by one unit). We set $q = 1$, so that the cumulative effect of changing multiple elements of x_j by one unit is bounded by the sum of the corresponding elements A_{jj} .

The elements of A_{main} are chosen to give restrictions on the conditional mean f that are plausible when $C = 1$; we report results for a range of choices of C as a form of sensitivity analysis. While the outcome is measured in levels, it is easier to interpret the bounds in terms of percentage increase in expected earnings. As a benchmark, consider deviations from expected earnings when the expected earnings are \$10,000, that is $f(x_i, d_i) = 10$. Since the average earnings for the $d_i = 1$ sample are \$6,400, 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 than this benchmark. We set the coefficients on black, Hispanic, and married to 2.5, implying that the wage gap due to race and marriage status is at most 25% at this benchmark. We set the coefficients on 1974 and 1975 earnings so that increasing earnings in each of these years by x units leads to at most $(0.5 + 0.5)x$ increase in 1978 earnings, that 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. The implied bounds for the effect of age and education on expected earnings

Table 2: Results for NSW application, main specification with $q = 1$, $A = A_{\text{main}}$, and $C = 1$.

Criterion	Feasible estimator \tilde{L}_δ		Matching estimator	
	(1) RMSE	(2) CI length	(3) RMSE	(4) CI length
Panel A: Inference on the CATT				
δ	1.82	3.30		
M			1	18
Estimate	0.96	0.94	1.39	1.26
Worst-case bias	1.63	1.78	1.48	2.21
Rob. std. error	1.01	0.94	1.09	0.89
Critical value ($cv_{0.05}$)	3.27	3.55	3.01	4.13
95% conf. interval	(−2.33, 4.25)	(−2.38, 4.27)	(−1.88, 4.66)	(2.41, 4.93)
Panel B: Inference on the PATT				
Rob. marginal std. error	1.06	1.00	1.14	0.94
95% conf. interval	(−2.75, 4.67)	(−2.80, 4.69)	(−2.32, 5.11)	(−2.78, 5.30)
Lind(k)	0.073	0.062	0.192	0.062

Notes: In each column, the results in both panels are based on an estimator with the smoothing parameter (δ or M) chosen to optimize the criterion listed under “Criterion”. Lind(k) corresponds to the maximum Lindeberg weight given in eq. (22).

at 10 thousand dollars are 1.5% and 6%, respectively, in line with the 1980 census data.

With this choice of norm, $\|x\|_{\mathcal{X}} = \|x\|_{A_{\text{main}},1} = \sum_{j=1}^p |A_{\text{main},jj}x_j|$, we follow the baseline implementation in Section 3.1. In step 1, we use the homoskedastic estimate $\tilde{\sigma}^2(x, d) = \hat{\sigma}^2$. Here $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2$, where \hat{u}_i^2 is the nearest-neighbor variance estimator with $J = 2$ neighbors, using Mahalanobis distance (using the metric $\|\cdot\|_{A_{\text{main}},1}$ leads to very similar results). We compare our estimators and CIs to those based on matching estimators: here we again use the norm $\|\cdot\|_{A_{\text{main}},1}$ to define distance. While all reported confidence intervals and standard errors use the heteroskedasticity-robust formula (18) (using the nearest-neighbor estimate \hat{u}_i^2 above), when making efficiency comparisons, we use homoskedastic standard errors, so that RMSE and CI length are the same as those used to optimize the choice of the smoothing parameter δ , or the number of matches M .

5.2 Results

Table 2 reports the point estimates and CIs at values of δ and M that optimize the RMSE and CI length criteria. There are three aspects of the results worth highlighting. First, in all cases,

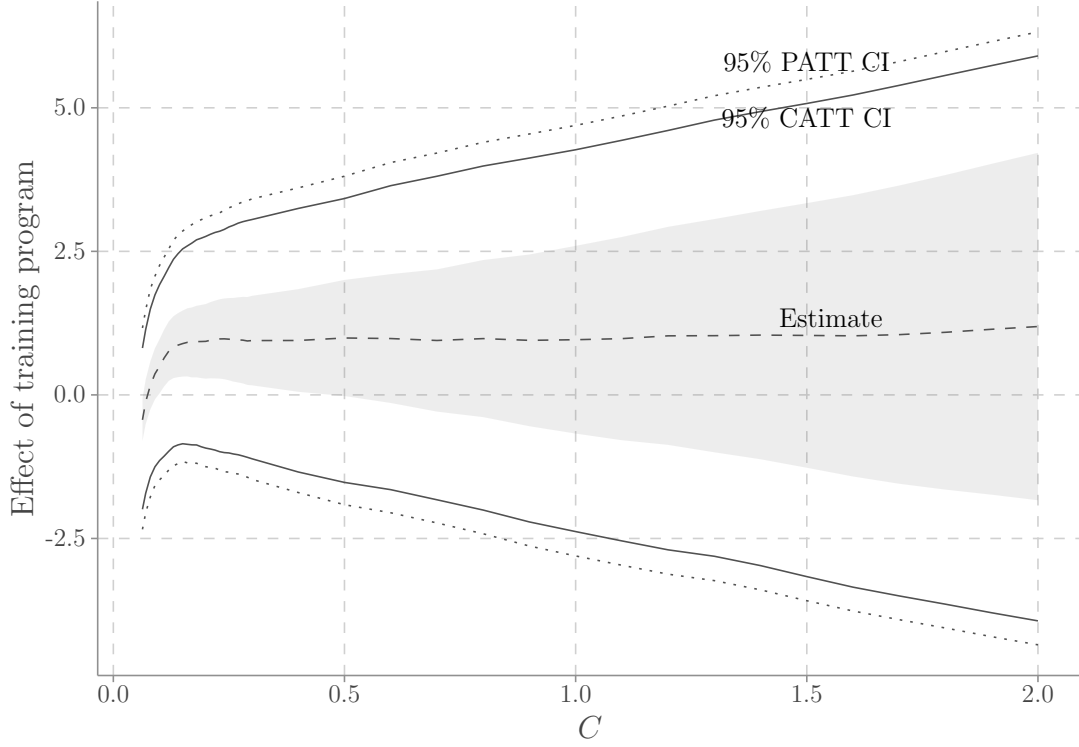


Figure 1: Feasible estimator and CIs for CATT and PATT in NSW application as a function of the Lipschitz constant C . Dashed line corresponds to point estimator $\tilde{L}_{\delta_{\text{RMSE}}}$, shaded region denotes the estimator \pm its worst-case bias. Solid lines give 95% robust CIs for the CATT based on the estimator $\tilde{L}_{\delta_{\text{FLCI}}}$, and dotted lines give CIs for the PATT.

the worst-case bias is non-negligible relative to the standard error. This is consistent with Theorem 4.1 on impossibility of \sqrt{n} -inference: under the Lipschitz smoothness assumption it is not possible to construct asymptotically unbiased estimators in this application given that the dimension of continuously distributed covariates is 4. Our CIs reflect this by explicitly taking the bias into account. Second, in line with the predictions of Theorem 4.2, the maximum Lindeberg weights (22) are low for the feasible estimators \tilde{L}_{δ} in columns (1) and (2). However, due to imperfect overlap, for the matching estimator with $M = 1$ match in column (3), it is considerably higher, and above the 0.075 cutoff suggested in Noack and Rothe (2020). Thus, its distribution may not be well-approximated by a normal distribution, in line with the discussion in Section 4.3. Third, in Panel B, the table also reports CIs for the population average treatment effect on the treated (PATT), constructed using the formula (20), using the nearest neighbor variance estimator to estimate the marginal variance. These CIs are only slightly longer than those for the CATT: depending on the column, the increase in length is 10.3–13.6%.

To examine sensitivity of the results to the specification of the parameter space \mathcal{F} , Fig-

Figure 1 plots the estimator \tilde{L}_δ at the RMSE-optimal choice of δ , as well as CI for the CATT and PATT when the smoothing parameter δ is chosen to optimize CI length. 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 estimator and CIs incorporate 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 1.14, 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.

According to Theorem 2.3, matching with $M = 1$ is efficient when C is “large enough”. In our application, for $C \geq 2.8$, the efficiency of the matching estimator is at least 95% for both RMSE and CI length. Matching with $M = 1$ leads to a modest efficiency loss in our main specification, where $C = 1$: its efficiency is 90.4% for RMSE, and 86.0% for the construction of two-sided CIs. However, inference results based on the matching estimator should be taken with a grain of caution due to the concerns with the accuracy of the CLT approximation discussed above.

5.3 Comparison with experimental estimates

The present analysis follows, among other, LaLonde (1986), Dehejia and Wahba (1999), Smith and Todd (2001), Smith and Todd (2005) and Abadie and Imbens (2011) in using a non-experimental sample to estimate treatment effects of the NSW program. A major question in this literature has been whether the non-experimental sample can be used to obtain results that are in line with the 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 treated and untreated individuals is 1.79. Treating this estimator as an estimator of the CATT, the (unconditional) robust standard error is 0.64; treating it as an estimator of the PATT (which also coincides with the PATE), it is 0.67.

The estimates in columns (1) and (2) of Table 2 are slightly lower, although the difference between them and the experimental estimate is much smaller than the worst-case bias. Consequently, all 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 using the current non-experimental dataset is too ambitious, unless one is willing to impose 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.

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 (2018b). Specializing their general results to the current setting yields the formulas for optimal estimators and CIs given in Section 2.3, and the efficiency bounds discussed in Section 2.5. Appendix A.2 considers the case with Lipschitz constraints, and when the Lipschitz constraints are imposed after partialling out the best linear predictor (BLP). It also gives proofs for Theorem 2.1, Theorem 2.3, and the first part of Theorem 2.2. Appendix A.3 proves the second part of Theorem 2.2.

A.1 General setup and results

We generalize the setup in Section 2.1 by letting the parameter of interest be a weighted CATE 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. To handle non-centrosymmetric cases, we allow for affine estimators that recenter by some constant a ,

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

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

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

While the centering constant has no effect on one-sided CIs, centering by $a^* = -(\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,0}) + \underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,0}))/2$, so that $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a^*}) + \underline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a^*}) = 0$ reduces the estimator's RMSE and

the length of the resulting FLCI (note this yields $a^* = 0$ under centrosymmetry). To simplify the results below, we assume that the estimator is recentered in this way.

One-sided CIs and FLCIs based on \hat{L}_{k,a^*} can be formed as in eqs. (9) and (10), with \hat{L}_{k,a^*} in place of \hat{L}_k , with its RMSE given by eq. (11). For comparisons of one-sided CIs, we focus on quantiles of excess length. Given a subset $\mathcal{G} \subseteq \mathcal{F}$, define the worst-case β th quantile of excess length over \mathcal{G} of a CI $[c, \infty)$, $q_\beta(\hat{c}, \mathcal{G}) = \sup_{g \in \mathcal{G}} q_{g,\beta}(Lg - \hat{c})$ where $Lg - \hat{c}$ is the excess length of the CI $[\hat{c}, \infty)$, and $q_{g,\beta}(\cdot)$ denotes the β th quantile under the function g . For a one-sided CI based on \hat{L}_{k,a^*} ,

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

This follows from the fact that the worst-case β th quantile of excess length over \mathcal{G} is attained when the estimate is biased downward as much as possible. 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 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 \(2018b\)](#), and we can specialize their general efficiency bounds and the construction of optimal affine estimators and CIs to the current setting.¹⁵ 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, 2018b](#))

$$\omega(\delta) = \sup_{f,g \in \mathcal{F}} \left\{ Lg - Lf : \sum_{i=1}^n \frac{(f(x_i, d_i) - g(x_i, d_i))^2}{\sigma^2(x_i, d_i)} \leq \delta^2 \right\}, \quad (25)$$

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 (12) 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). Define $\hat{L}_\delta = \hat{L}_{k_\delta^*, a_\delta^*}$, where

$$k_\delta^*(x_i, d_i) = \frac{\omega'(\delta)}{\delta} \frac{g_{\delta,i}^* - f_{\delta,i}^*}{\sigma^2(x_i, d_i)}, \quad a_\delta^* = \frac{L(f_\delta^* + g_\delta^*) - \sum_{i=1}^n k_\delta^*(x_i, d_i)(f_{\delta,i}^* + g_{\delta,i}^*)}{2},$$

¹⁵In particular, in the notation of [Armstrong and Kolesár \(2018b\)](#), $Y = (Y_1/\sigma(x_1, d_1), \dots, Y_n/\sigma(x_n, d_n))$, $\mathcal{Y} = \mathbb{R}^n$, and $Kf = (f(x_1, d_1)/\sigma(x_1, d_1), \dots, f(x_n, d_n)/\sigma(x_n, d_n))$. [Donoho \(1994\)](#) denotes the outcome vector Y by \mathbf{y} , and uses \mathbf{x} and \mathbf{X} in place of f and \mathcal{F} .

with $g_{\delta,i}^* = g_{\delta}^*(x_i, d_i)$ and $f_{\delta,i}^* = f_{\delta}^*(x_i, d_i)$. If the class \mathcal{F} is translation invariant in the sense that $f \in \mathcal{F}$ implies $f + \iota_{\kappa} \in \mathcal{F}$ ¹⁶, then by Lemma D.1 in [Armstrong and Kolesár \(2018b\)](#), the modulus is differentiable, with the derivative $\omega'(\delta) = \delta / \sum_{i=1}^n d_i (g_{\delta,i}^* - f_{\delta,i}^*) / \sigma^2(x_i, d_i)$. The formula for \hat{L}_{δ} 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 \(2018b\)](#), the maximum and minimum bias of \hat{L}_{δ} 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 \(2018b\)](#), 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 (25) 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 β th 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_{FLCI} be the minimizer of $\text{cv}_{\alpha}(\omega(\delta)/2\omega'(\delta) - \delta/2)\omega'(\delta)$ over δ , and suppose that $f_{\delta_{FLCI}}^*$ and $g_{\delta_{FLCI}}^*$ attain the supremum in (25) at $\delta = \delta_{FLCI}$. Then the shortest $1 - \alpha$ FLCI among all FLCIs centered at affine estimators is given by*

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

(iii) Let δ_{RMSE} minimize $\frac{1}{4}(\omega(\delta) - \delta\omega'(\delta))^2 + \omega'(\delta)^2$ over δ , and suppose that $f_{\delta_{FLCI}}^*$ and $g_{\delta_{FLCI}}^*$ attain the supremum in (25) 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}_{δ} 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.

¹⁶In the main text, we assume that $\{\iota_{\kappa}\}_{\kappa \in \mathbb{R}} \subset \mathcal{F}$. By convexity, for any $\lambda < 1$, $\lambda f + (1 - \lambda)\iota_{\kappa} = \lambda f + \iota_{(1-\lambda)\kappa} \in \mathcal{F}$, which implies that for all $\lambda < 1$ and $\kappa \in \mathbb{R}$, $\lambda f + \iota_{\kappa} \in \mathcal{F}$. This, under the assumption in footnote 9, implies translation invariance.

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

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 FLCI based on $\hat{L}_{\delta_{\text{FLCI}}}$ 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 (25) 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 $\delta = z_\beta + z_{1-\alpha}$. Define \hat{c}_δ^* as in Theorem A.1. Then the efficiency of \hat{c}_δ^* under the criterion $q_\beta(\cdot, \{g\})$, is given by*

$$\frac{\inf_{\{\hat{c}: [\hat{c}, \infty) \text{ satisfies (5)}\}} q_\beta(\hat{c}, \{g\})}{q_\beta(\hat{c}_\delta^*, \{g\})} = \frac{\omega(2\delta)}{\omega(\delta) + \delta\omega'(\delta)} \geq \frac{1}{2}.$$

(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 FLCI around $\hat{L}_{\delta_{\text{FLCI}}}$ at g relative to all confidence sets, $\frac{\inf_{\{\mathcal{C}: \mathcal{C} \text{ satisfies (5)}\}} E_g \lambda(\mathcal{C})}{\text{cv}_\alpha(\text{bias}_{\delta_{\text{FLCI}}} / \text{sd}(\hat{L}_{\delta_{\text{FLCI}}})) \text{sd}(\hat{L}_{\delta_{\text{FLCI}}})}$, is given by

$$\frac{(1 - \alpha) E[\omega(2(z_{1-\alpha} - Z)) \mid Z \leq z_{1-\alpha}]}{2 \text{cv}_\alpha \left(\frac{\omega(\delta_{\text{FLCI}})}{2\omega'(\delta_{\text{FLCI}})} - \frac{\delta_{\text{FLCI}}}{2} \right) \cdot \omega'(\delta_{\text{FLCI}})} \geq \frac{z_{1-\alpha}(1 - \alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha)}{z_{1-\alpha/2}}, \quad (26)$$

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 (26), follow from Corollary 3.2 and Corollary 3.3 in Armstrong and Kolesár (2018b). The lower bound follows from Theorem C.7 in Armstrong and Kolesár (2020). \square

The theorem gives sharp efficiency bounds for one-sided CIs as well as FLCIs relative to CIs that direct all power at a particular function g . The condition on g is satisfied if g is smooth enough relative to \mathcal{F} . For example, if $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$, it holds if g is piecewise constant, $g(x, d) = \kappa_0 + \kappa_1 d$ for some $\kappa_0, \kappa_1 \in \mathbb{R}$. 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% FLCIs, the efficiency lower bound in (26) evaluates to 71.7%. In a particular application, sharp lower bounds can be computed directly by computing the modulus; as we show in Section 5, this typically yields much higher efficiencies.

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, and versions that impose the Lipschitz smoothness after partialling out the best linear predictor (BLP). We focus on the CATT and CATE estimands by requiring the weights w_i in the definition of Lf to depend only on d_i : we assume that $w_i = w(d_i)$, with $w(1), w(0) \geq 0$ and $w(1)n_1 + w(0)n_0 = 1$. For CATT, $w(1) = 1/n_1$ while $w(0) = 0$, and for CATE, $w(1) = w(0) = 1/n$.

We begin by defining a version of the Lipschitz class that imposes the Lipschitz condition after partialling out the BLP. Let z be a subset of the covariates x that includes the intercept. The regression coefficients β_0, β_1 in a weighted least squares regression of Y_i onto $(1 - d_i)z_i$ and onto $d_i z_i$, weighed by the precision weights $\sigma^{-2}(x_i, d_i)$, are given by¹⁷ $\beta_d = \text{argmin}_\beta \sum_{i=1}^n \sigma^{-2}(x_i, d_i) \mathbb{I}\{d_i = d\} (E[Y_i(d) \mid Z_i = z_i] - z_i' \beta)^2$. The part of f left over after partialling out the BLP is given by $g(x, d) = f(x, d) - (1 - d)z\beta_0 - dz\beta_1$, and it satisfies the orthogonality restriction

$$\sum_{i=1}^n \mathbb{I}\{d_i = d\} \frac{z_i g(x_i, d)}{\sigma^2(x_i, d_i)} = 0, \quad d \in \{0, 1\}. \quad (27)$$

Imposing a Lipschitz smoothness condition on g then leads to the class

$$\mathcal{F}_{z, \text{Lip}}(C) = \{z' \beta_d + g(x, d) : g \in \mathcal{F}_{\text{Lip}}(C), g \text{ satisfies (27)}, \beta_0, \beta_1 \in \mathbb{R}^p\}. \quad (28)$$

If z comprises just the intercept, then the constraint (27) simply normalizes $f(\cdot, d)$ to have empirical mean equal to zero, $g(x, d) = f(x, d) - \sum_i \{d_i = d\} f(x_i, d) / \sigma^2(x_i, d_i)$. Since this is

¹⁷See Abadie et al. (2014) for a comparison of this conditional definition of the BLP with the unconditional BLP, $\beta_d = \text{argmin}_\beta E \sigma^{-2}(X_i, D_i) (Y_i(d) - Z_i' \beta)^2$

just a normalization, $\mathcal{F}_{1,\text{Lip}}(C) = \mathcal{F}_{\text{Lip}}(C)$. Results for the class $\mathcal{F}_{z,\text{Lip}}(C)$ thus directly imply results for the class $\mathcal{F}_{\text{Lip}}(C)$ that the main text focuses on. On the other hand, if $z = x$, we can think of the space $\mathcal{F}_{x,\text{Lip}}(C)$ as formalizing the notion that f is “approximately linear” by requiring that the residual is small in the sense that it lies in $\mathcal{F}_{\text{Lip}}(C)$, with $\mathcal{F}_{x,\text{Lip}}(0)$ corresponding to the linear model. In practice, when the variance function is unknown, one may wish to impose this class under the assumption of homoskedasticity, so that we partial out the OLS, rather than the weighted least squares estimand.

To see how we can replace the functional constraint $g \in \mathcal{F}_{\text{Lip}}(C)$ with inequality constraints, let $\tilde{\mathcal{F}}_{\text{Lip},n}(C)$ denote the set of functions $f: \{x_1, \dots, x_n\} \times \{0, 1\} \rightarrow \mathbb{R}$ such that $|f(x, d) - f(\tilde{x}, d)| \leq C\|x - \tilde{x}\|_{\mathcal{X}}$ for all $x, \tilde{x} \in \{x_1, \dots, x_n\}$ and each $d \in \{0, 1\}$. That is, $\tilde{\mathcal{F}}_{\text{Lip},n}(C)$ denotes the class of functions that satisfy the Lipschitz conditions when restricted to the domain $\{x_1, \dots, x_n\} \times \{0, 1\}$. Clearly, $f \in \mathcal{F}_{\text{Lip}}(C)$ implies $f \in \tilde{\mathcal{F}}_{\text{Lip},n}(C)$. In the other direction, we have the following result:

Lemma A.1. (*Beliakov, 2006, Theorem 4*) *For any function $f: \{x_1, \dots, x_n\} \times \{0, 1\} \rightarrow \mathbb{R}$, we have $f \in \tilde{\mathcal{F}}_{\text{Lip},n}(C)$ 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 $(x, d) \in \{x_1, \dots, x_n\} \times \{0, 1\}$.*

As a direct consequence, we obtain:

Lemma A.2. *Consider a linear estimator $\hat{L}_k = \sum_i k_i Y_i$ with weights k that satisfy $\sum_i \mathbb{I}\{d_i = d\} z_i k_i = (2d - 1)[w(d) \sum_i \mathbb{I}\{d_i = d\} z_i + w(1 - d) \sum_i \mathbb{I}\{d_i = 1 - d\} z_i]$. The worst-case bias of this estimator over $\mathcal{F}_{z,\text{Lip}}(C)$ is given by the value of*

$$\max_{g \in \mathbb{R}^{2n}} \left\{ \sum_{i=1}^n k_i g(x_i, d_i) - \sum_{i=1}^n w(d_i) [g(x_i, 1) - g(x_i, 0)] \right\},$$

where the maximum is taken subject to (27) and subject to

$$g(x_i, d) - g(x_j, d) \leq C\|x_i - x_j\|_{\mathcal{X}}, \quad d \in \{0, 1\} \quad i, j \in \{1, \dots, n\}. \quad (29)$$

The constraints on the weights imply that the values of β_0 and β_1 do not affect the bias. Specializing to the case with $z = 1$ yields the first part of Theorem 2.1 (using the observation above that in this case the constraint (27) is just a normalization that under the conditions on the weights does not affect the bias, so not imposing it doesn’t affect the worst-case bias).

To show the second part of Theorem 2.1, we use the following lemma, the proof of which is deferred to the supplemental materials.

Lemma A.3. *Fix $d \in \{0, 1\}$, and consider a vector $(g(x_1, d), \dots, g(x_n, d))'$. (i) Suppose (29) holds for all i, j with $d_i = d_j = d$, and also for all i, j with $d_j = 1 - d_i = d$. Suppose further*

that for each i with $d_i = 1 - d$, there exists a j with $d_j = d$ such that (29) holds with equality. Then (29) holds for all i, j . (ii) Suppose (29) holds for all i, j with $d_j = 1 - d_i = d$. Suppose further that for each i with $d_i = 1 - d$, there exists a j with $d_j = d$ such that (29) holds with equality. Suppose also that for each j with $d_j = d$, there exists an i with $d_i = 1 - d$ such that (29) holds with equality. Then (29) holds for all i, j .

We now show that if (14) holds for all i, j with $d_i = 1$, $d_j = 0$ and $k_i = k(x_i, 1) > w(1)$, then it holds for $d = 1$ and all i, j . The argument for $d = 0$ is analogous and omitted. Observe that if $k_i = w(1)$ for some i with $d_i = 1$, we can set $f(x_i, 1) = \min_{j: d_j=0} \{f(x_j, 1) + C\|x_i - x_j\|_{\mathcal{X}}\}$ without affecting the bias, so that we may suppose that (14) holds for all i, j with $d_i = 1 - d_j = 1$. Theorem 2.1 then follows by observing that $g = -(f(x_1, 1), \dots, f(x_n, 1))$ must satisfy the assumptions of part (ii) of Lemma A.3, otherwise we could increase the bias by increasing $f(x_i, 1)$ (if $d_i = 1$) or decreasing $f(x_j, 1)$ (if $d_j = 0$).

For the optimal estimator, Lemma A.1 implies the following result:

Theorem A.4. *Given $\delta > 0$, the value of the maximizer $f_\delta^*(x_i, d_i) = z_i' \beta_{d_i, \delta}^* + g_\delta^*(x_i, d_i)$ of eq. (12) is given by the solution to the convex program*

$$\max_{g \in \mathbb{R}^{2n}, \beta_0, \beta_1 \in \mathbb{R}^{\dim(z_i)}} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^n \frac{g(x_i, d_i)^2 + (z_i' \beta_{d_i})^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4} \quad \text{and s.t. (29) and (27)}. \quad (30)$$

Specializing to the case with $z = 1$ gives the first part of Theorem 2.2. The proof for the second part is deferred to Appendix A.3.

Next, we consider the form of the optimal estimator when δ/C is small. To that end, let us first define the regression-adjusted matching estimator. The linear regression estimator imputes the counterfactual outcome $Y_i(1 - d_i)$ as $z_i' \hat{\beta}_{1-d_i}$, where $\hat{\beta}_d = (Z_d' \Sigma_d^{-1} Z_d)^{-1} Z_d' \Sigma_d^{-1} Y_d$. Here Y_d and Z_d are the subsets of the outcome vector and design matrix of the covariates z_i corresponding to units with $d_i = d$, and Σ_d is a diagonal matrix of dimension $n_d \times n_d$ with elements $\sigma^2(x_i, d_i)$ corresponding to observations with $d_i = d$ on the diagonal. Recall from Section 2.2 that in contrast, the matching estimator with M matches uses the imputation $\sum_{j=1}^n W_{M,ij} Y_j$, where $W_{M,ij} = 1/M$ if j is among the M observations with treatment status $d_j = 1 - d_i$ that are closest to i , and zero otherwise. The regression-adjusted matching estimator (Rubin, 1979) combines these approaches, imputing the counterfactual outcome as $\sum_{j=1}^n W_{M,ij} (Y_j - z_j' \hat{\beta}_{1-d_i}) + z_i' \hat{\beta}_{1-d_i} = \sum_{j=1}^n W_{M,ij} (Y_j + z_i' \hat{\beta}_{1-d_i} - z_j' \hat{\beta}_{1-d_i})$. Relative to the matching estimator, we adjust the matching imputation for the difference in covariate values. The estimator thus takes the form

$$\hat{L}_k = \sum_i (2d_i - 1)w(d_i) \left[Y_i - z_i' \hat{\beta}_{1-d_i} - \sum_j W_{M,ij} (Y_j - z_j' \hat{\beta}_{d_i}) \right].$$

Theorem 2.3 follows directly the next result by setting $z = 1$.

Theorem A.5. *Suppose that $\sigma(x_i, d_i) > 0$ for each i , and suppose that each unit has a single closest match, so that $\operatorname{argmin}_{j: d_i \neq d_j} \|x_i - x_j\|_{\mathcal{X}}$ 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}_δ is given by the regression adjusted matching estimator with $M = 1$.*

Proof of Theorem A.5. For notational convenience, let m and r denote the vectors of length n with elements $m_i = (2d_i - 1)g(x_i, d_i)$, and let $r_i = (1 - 2d_i)g(x_i, 1 - d_i)$. Abusing notation, let m_d and r_d denote the subvectors of m and r corresponding to units with treatment status d . Finally, let ι_d denote the vector of ones with length n_d . With this notation,

$$Lf = w(1)\iota'_1(m_1 + r_1) + w(0)\iota'_0(r_0 + m_0) + w(1)\iota'_1 Z_1(\beta_1 - \beta_0) + w(0)\iota'_0 Z_0(\beta_1 - \beta_0),$$

and (30) can be written as

$$\max_{\beta_0, \beta_1, m \in \mathbb{R}^n, r \in \mathbb{R}^n} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^n \frac{m_i^2 + (z'_i \beta_{d_i})^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4} \quad (31)$$

subject to

$$Z'_d \Sigma_d^{-1} m_d = 0, \quad \text{for } d \in \{0, 1\} \quad (32)$$

and subject to (29). Now, note that if $w(d) = 0$, setting $r_d = 0$ and $m_d = 0$ is optimal and the constraints (29) hold trivially. If $w(d) > 0$, observe that

$$r_i - m_j \leq C \|x_i - x_j\|_{\mathcal{X}} \quad \text{for all } i, j \text{ with } d_i = 1 - d_j = d, \quad (33)$$

holds with equality for each i for at least one j , otherwise we could increase the value of the objective function. Therefore, by Lemma A.3, to maximize (30), we can replace the constraints in (29) with (33) and

$$m_i - m_j \leq C \|x_i - x_j\|_{\mathcal{X}} \quad \text{for all } i, j \text{ with } d_i = d_j = d. \quad (34)$$

Next, the assumption that each observation has a unique closest match implies that the values of x_i and x_j for $d_i = d_j$ are distinct, so that, for δ/C small enough and hence $\|m_i\|$ small enough, the constraint (31) implies (34). For δ/C small enough, it thus suffices to maximize (31) subject to (32) and (33). This is a convex optimization problem and constraint qualification holds since $m = 0$ satisfies Slater's condition. Thus, the solution (or set of solutions) is the same as the solution to the Lagrangian,

$$\begin{aligned}
& -Lf + \frac{\lambda}{2} \left(\sum_{d=0}^1 (m'_d \Sigma_d^{-1} m_d + \beta'_d Z'_d \Sigma_d^{-1} Z_d \beta_d) - \frac{\delta^2}{4} \right) + \nu'_0 Z'_0 \Sigma_0^{-1} m_0 + \nu_1 Z'_1 \Sigma_1^{-1} m_1 + \\
& \sum_{i,j: d_i=1-d_j=1} [\Lambda_{ij}^0 (r_i - m_j - C \|x_i - x_j\|_{\mathcal{X}}) + \Lambda_{ij}^1 (r_j - m_i - C \|x_i - x_j\|_{\mathcal{X}})],
\end{aligned}$$

where ν_0, ν_1 are vectors of Lagrange multipliers associated with the constraints (32), and Λ^0, Λ^1 are matrices of Lagrange multipliers with dimension $n_1 \times n_0$ associated with the constraints (33). The first-order conditions are given by

$$\Sigma_d^{-1} m_d = \frac{1}{\lambda} (w(d) \iota_d + \Lambda^{d'} \iota_{1-d} - \Sigma_d^{-1} Z_d \nu_d) \quad (35)$$

$$w(d) \iota_d = \Lambda^{1-d} \iota_{1-d} \quad (36)$$

$$\beta_d = \frac{2d-1}{\lambda} (Z_d \Sigma_d^{-1} Z_d)^{-1} (w(1) Z'_1 \iota_1 + w(0) Z'_0 \iota_0), \quad (37)$$

for $d \in \{0, 1\}$. Furthermore, the constraint $Z'_d \Sigma_d^{-1} m_d = 0$, combined with (35) implies that

$$\nu_d = w(d) (Z'_d \Sigma_d^{-1} Z_d)^{-1} Z'_d \iota_d + (Z'_d \Sigma_d^{-1} Z_d)^{-1} Z'_d \Lambda^{d'} \iota_{1-d}$$

so that

$$\Sigma_d^{-1} m_d = \frac{1}{\lambda} (I - \Sigma_d^{-1} Z_d (Z'_d \Sigma_d^{-1} Z_d)^{-1} Z'_d) (w(d) \iota_d + \Lambda^{d'} \iota_{1-d}).$$

Now, since by assumption, $(w(1)n_1 + w(0)n_0) = 1$, we have $\iota'_1 \Sigma_1^{-1} m_1 + \iota'_1 \Sigma_1^{-1} Z_1 \beta_1 = \frac{1}{\lambda}$. Hence, by eq. (13), the optimal weights for the treated and untreated units, respectively, take the form

$$\begin{aligned}
k_d &= (2d-1) \lambda \Sigma_d^{-1} m_d + \lambda \Sigma_d^{-1} Z_d \beta_d = (2d-1) [w(d) \iota_d \\
&+ (I - \Sigma_d^{-1} Z_d (Z'_d \Sigma_d^{-1} Z_d)^{-1} Z'_d) \Lambda^{d'} \iota_{1-d} + \Sigma_d^{-1} Z_d (Z_d \Sigma_d^{-1} Z_d)^{-1} w(1-d) Z'_{1-d} \iota_{1-d}],
\end{aligned}$$

with the optimal estimator given by $\hat{L}_\delta = \sum_{i=1}^n (2d_i - 1) [w(d_i) (Y_i - z'_i \hat{\beta}_{1-d_i}) - \sum_j \Lambda_{ij}^{1-d_i} (Y_j - z'_j \hat{\beta}_{d_i})]$. The result then follows from the fact that if $w(d) = 0$, $\Lambda^{1-d} = 0$ at optimum by eq. (36) since $\Lambda^{1-d} \geq 0$ by the complementary slackness constraint. Otherwise, if $w(d) > 0$, for each i , Λ_{ij}^{1-d} is non-zero for the set corresponding to the argmin of eq. (33). However, for δ/C and hence $\|m\|$ small enough, since each observation has a unique closest match, the set of minimizers is a singleton given by the closest match of i . By eq. (36), for this closest match j , $\Lambda_{ij}^{1-d} = w(d)$, which yields the result. \square

A.3 Proof of Theorem 2.2

The dual problem to (17) 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 . When $\sigma^2(x, d) = \sigma^2(d)$, the Lagrangian for this problem has the form

$$\min_{f \in \tilde{\mathcal{F}}_{\text{Lip},n}(C)} \frac{1}{2} \sum_{i=1}^n \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} - \mu Lf/C = \min_{f \in \tilde{\mathcal{F}}_{\text{Lip},n}(1)} \frac{C^2}{2} \sum_{i=1}^n \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} - \mu Lf, \quad (38)$$

where we use the observation that if $f \in \tilde{\mathcal{F}}_{\text{Lip},n}(C)$, then $f/C \in \tilde{\mathcal{F}}_{\text{Lip},n}(1)$. Let g_μ^* denote the solution to the minimization problem on the right-hand side of (38). Because for each $\delta > 0$, the program (17) 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 $\{Cg_\mu^*\}_{\mu>0}$.

It will be convenient to state the algorithm using the notation $m_i = (2d_i - 1)g(x_i, d_i)$, and $r_i = (1 - 2d_i)g(x_i, 1 - d_i)$, in analogy to the notation in the proof of Theorem A.5. Then $Lf = \sum_{i=1}^n w_i(m_i + r_i)$. Next, we claim that the constraint $g \in \tilde{\mathcal{F}}_{\text{Lip},n}(1)$ can be replaced with the constraint

$$r_j \leq m_i + \|x_i - x_j\|_{\mathcal{X}}, \quad d_i \neq d_j, \quad (39)$$

This follows by observing that at optimum, if $w(d) = 0$, then $m_i = \mu w(d)\sigma^2(d)$ and $r_j = \mu w(d)\sigma^2(d)$ achieves the optimum, so the constraint holds trivially. If $w(d) > 0$, at least one of the constraints in eq. (39) must bind, for each i , otherwise increasing r_j would increase the value of the objective function. Thus, by part (i) Lemma A.3, we can replace the constraint $g \in \tilde{\mathcal{F}}_{\text{Lip},n}(1)$ with eq. (39) and

$$m_i \leq m_{i'} + \|x_i - x_{i'}\|_{\mathcal{X}}, \quad d_i = d_{i'}. \quad (40)$$

If we only impose the constraints in (39), the Lagrangian for the program (38) can be written as

$$\begin{aligned} & \frac{1}{2} \sum_{i=1}^n \frac{m_i^2}{\sigma^2(d_i)} - \mu \left(\sum_{i=1}^n w(d_i)(m_i + r_i) \right) \\ & + \sum_{i,j: d_i=1-d_j=1} [\Lambda_{ij}^0(r_i - m_j - \|x_i - x_j\|_{\mathcal{X}}) + \Lambda_{ij}^1(r_j - m_i - \|x_i - x_j\|_{\mathcal{X}})]. \end{aligned} \quad (41)$$

This Lagrangian implies that (40) must hold automatically at the optimum. Otherwise, if for some i, i' with $d_i = d_{i'} = 1$, $m_i > m_{i'} + \|x_i - x_{i'}\|_{\mathcal{X}}$, then for all j with $d_i = 0$,

$$r_j \leq m_{i'} + \|x_{i'} - x_j\|_{\mathcal{X}} \leq m_{i'} + \|x_{i'} - x_i\|_{\mathcal{X}} + \|x_i - x_j\|_{\mathcal{X}} < m_i + \|x_i - x_j\|_{\mathcal{X}},$$

The complementary slackness condition $\Lambda_{ij}^1(r_j - m_i - \|x_i - x_j\|_{\mathcal{X}}) = 0$ then implies that $\sum_j \Lambda_{ij}^1 = 0$, and it follows from the first-order condition that $m_i/\sigma^2(1) = \mu w(1) \leq m_{i'}/\sigma^2(1)$, which contradicts the assertion that $m_i > m_{i'} + \|x_i - x_{i'}\|_{\mathcal{X}}$.

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 (41). 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 constraints 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 Λ_{ij}^1 and $\Lambda_{i'j}^1$ are both active for some j . Let $\{\mathcal{M}_1^0, \dots, \mathcal{M}_{K_0}^0\}$ denote a partition of $\{1, \dots, n_0\}$ 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 $\{\mathcal{R}_1^0, \dots, \mathcal{R}_k^0\}$ be a corresponding partition of $\{1, \dots, n_1\}$, defined by $\mathcal{R}_k^0 = \{i \in \{1, \dots, n_1\} : N_{ij}^0(\mu) = 1 \text{ for some } j \in \mathcal{M}_k^0\}$. Similarly, let $\{\mathcal{M}_1^1, \dots, \mathcal{M}_{K_1}^1\}$ denote a partition of $\{1, \dots, n_1\}$ according to the connected components of G^1 , and let $\mathcal{R}_k^1 = \{j \in \{1, \dots, n_0\} : N_{ij}^1(\mu) = 1 \text{ for some } i \in \mathcal{M}_k^1\}$.

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 multipliers $\Lambda_{ij}^d(\mu)$ 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 $D_{ij}^d = \|x_i - x_j\|_{\mathcal{X}}$, $d \in \{0, 1\}$, $d_i = 1 - d_j = 1$. Let r be a vector with elements $r_j = \min_{i=1, \dots, n_1} \{D_{ij}^1\}$, if $d_j = 0$, and $r_i = \min_{j=1, \dots, n_0} \{D_{ij}^0\}$, if $d_i = 1$. Let $N^0, N^1 \in \mathbb{R}^{n_1 \times n_0}$ be matrices with elements $N_{ij}^0 = \mathbb{I}\{D_{ij}^0 = r_i\}$ 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 $\delta_{r,i} = \delta_j = \sigma^2(0)(w(0) + (\#\mathcal{R}_k^0/\#\mathcal{M}_k^0)w(1))$ for $i \in \mathcal{R}_k^0$ and $j \in \mathcal{M}_k^0$, and $\delta_{r,j} = \delta_i = \sigma^2(1)(w(1) + (\#\mathcal{R}_k^1/\#\mathcal{M}_k^1)w(0))$ 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) $\sum_{i=1}^{n_0} \Delta_{ij}^1 =$

$\delta_j/\sigma^2(0) - w(0)$, $j = 1, \dots, n_0$ and $\sum_{j=1}^{n_0} \Delta_{ij}^0 = w(1)$, $i = 1, \dots, n_1$ and (ii) $\sum_{j=1}^{n_0} \Delta_{ij}^1 = \delta_i/\sigma^2(1) - w(1)$, $i = 1, \dots, n_1$ and $\sum_{i=1}^{n_1} \Delta_{ij}^0 = w(0)$, $j = 1, \dots, n_0$.

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

$$\begin{aligned} s_1^0 &= \min\{s \geq 0: r_i + \delta_{r,i}s = \delta_js + D_{ij}^0 \text{ some } (i, j) \text{ s.t. } N_{ij}^0 = 0, \delta_j > \delta_{r,i}\} \\ s_1^1 &= \min\{s \geq 0: r_j + \delta_{r,j}s = \delta_is + D_{ij}^1 \text{ some } (i, j) \text{ s.t. } N_{ij}^1 = 0, \delta_i > \delta_{r,j}\} \\ s_2^d &= \min\{s \geq 0: \Lambda_{ij}^d + s\Delta_{ij}^d = 0 \text{ among } (i, j) \text{ with } N_{ij}^d = 1 \text{ and } \Delta_{ij}^d < 0\} \end{aligned}$$

(d) Update $\mu \mapsto \mu + s$, $m \mapsto m + s\delta$, $r \mapsto r + s\delta_r$ $\Lambda^d \mapsto \Lambda^d + s\Delta^d$, $D_{ij}^0 \mapsto D_{ij}^0 + s\delta_j$, $D_{ij}^1 \mapsto D_{ij}^1 + s\delta_i$. 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}_δ and its worst-case bias can then be easily computed. For simplicity, we specialize to the CATE 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 (41) (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 $\hat{f}_\mu(x_j, 1) = \sum_i n\Lambda_{ij}^1(\mu)/\mu Y_i$ if $d_j = 0$; $\hat{f}_\mu(x_i, 1) = Y_i$ if $d_i = 1$; $\hat{f}_\mu(x_j, 0) = Y_j$ if $d_j = 0$; and $\hat{f}_\mu(x_i, 0) = \sum_j n\Lambda_{ij}^0(\mu)/\mu Y_j$ if $d_i = 1$. 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}_δ as a matching estimator with a variable number of matches, observe that $\sum_i n\Lambda_{ij}^1(\mu)/\mu = \sum_j n\Lambda_{ij}^0(\mu)/\mu = 1$. Also, $N_{ij}^0(\mu) = 0$ and hence $\Lambda_{ij}^0(\mu) = 0$ unless $D_{ij}^0(\mu) = \min_\ell D_{i\ell}^0(\mu)$. Similarly, $\Lambda_{ij}^1(\mu) = 0$ unless $D_{ij}^1(\mu) = \min_\ell D_{\ell j}^1(\mu)$. 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 $D_{ik}^0(\mu)$ (if $d_i = 1$) or $D_{ki}^1(\mu)$ (if $d_i = 0$). Since $D_{ik}^0(\mu) = m_k(\mu) + \|x_i - x_k\|_\mathcal{X}$ $D_{ki}^1(\mu) = m_i(\mu) + \|x_i - x_k\|_\mathcal{X}$, 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.

B Proofs for asymptotic results

This appendix gives additional details and proofs for the results in Sections 3.2 and 4.

B.1 Proofs and details for Section 3.2

We now analyze the asymptotic coverage properties of the CI in (20) and its one-sided analog $[\hat{c}, \infty)$, where

$$\hat{c} = \hat{L}_k - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) - z_{1-\alpha} \text{se}_{\tau}(\hat{L}_k). \quad (42)$$

We give coverage results that are uniform over a certain class \mathcal{P} of underlying distributions. To this end, we index the population from which the data is drawn by P , and use P and E_P to denote probability statements and expectations taken under n i.i.d. draws from the population. We make the dependence of conditional distributions on P explicit by writing $f_P(x, d) = E_P[Y_i \mid X_i = x, D_i = d]$ and $\sigma_P^2(x, d) = \text{var}_P(Y_i \mid X_i = x, D_i = d)$, with Lf_P and $\tau(P) = E_P Lf_P$ denoting the CATE and PATE, respectively. The asymptotic coverage requirement for a sequence of CIs \mathcal{C} based on data $\{Y_i, X_i, D_i\}_{i=1}^n$ is

$$\liminf_n \inf_{P \in \mathcal{P}} P(\tau(P) \in \mathcal{C}) \geq 1 - \alpha. \quad (43)$$

To construct \mathcal{P} , we assume that $f_P \in \mathcal{F}$ for all $P \in \mathcal{P}$, and that \mathcal{F} is convex and centrosymmetric. We allow the distribution of $u_i = Y_i - f_P(X_i)$ and X_i to vary over a class that places uniform integrability or moment conditions on u_i and conditions on the support of X_i . As before, we require \mathcal{F} to be fully specified. However, we do not need to know the constants governing the regularity of u_i and X_i . The weights k in (6) may depend on the entire set $\{X_i, D_i\}_{i=1}^n$, but not on the outcomes Y_i . Let $V_{2,n}(P) = E_P((f_P(X_i, 1) - f_P(X_i, 0) - \tau(P))^2)/n$ denote the variance of the CATE.

We now give asymptotic coverage results for the CIs in eqs. (20) and (42) under the following high-level assumption.

Assumption B.1. *For some sequence of constants $V_{1,n}(P)$, (i) $\frac{\sum_{i=1}^n k(X_i, D_i)^2 \sigma_P^2(X_i, D_i)}{V_{1,n}(P)} \xrightarrow{P} 1$ uniformly over $P \in \mathcal{P}$ for all $\varepsilon > 0$ and (ii) uniformly over $P \in \mathcal{P}$,*

$$\frac{\sum_{i=1}^n E_P[k(X_i, D_i)^2 u_i^2 \mathbb{I}\{k(X_i, D_i)^2 u_i^2 > \varepsilon V_{1,n}(P)\}]}{V_{1,n}(P)} \rightarrow 0.$$

The supplemental materials verify Assumption B.1 for the matching estimator, and discuss consistency of standard errors.

Theorem B.1. *Suppose that Assumption B.1 holds, $[V_{1,n}(P) + V_{2,n}(P)]/\text{se}_{\tau}(\hat{L}_k)^2 \xrightarrow{P} 1$ uniformly over $P \in \mathcal{P}$, and that for some $\eta > 0$ and $d \in \{0, 1\}$ $E_P|f_P(X_i, d)|^{2+\eta} < 1/\eta$ for all $P \in \mathcal{P}$. Then the confidence sets in eqs. (20) and (42) satisfy (43).*

Proof. Given a sequence $P_n \in \mathcal{P}$, we need to show that $\liminf_n P_n(\tau(P_n) \in [\hat{c}, \infty)) \geq 1 - \alpha$.

Since the CI in (20) is given by the intersection of two one-sided CIs, its asymptotic validity then follows immediately from Bonferroni's inequality.

To this end, use a martingale representation, analogous to the martingale representation for matching estimators noted by [Abadie and Imbens \(2012\)](#). For $i = 1, \dots, n$, let $\xi_{n,i} = \frac{1}{n}(f_{P_n}(X_i, 1) - f_{P_n}(X_i, 0) - \tau(P_n))$ and let $\mathcal{H}_{n,i}$ be the σ -algebra generated by D_1, \dots, D_n and X_1, \dots, X_i . For $i = n+1, \dots, 2n$, let $\xi_{n,i} = k(X_{i-n}, D_{i-n})u_{i-n}$, and let $\mathcal{H}_{n,i}$ be the σ -algebra generated by $\mathcal{H}_{n,n}$ and u_1, \dots, u_{i-n} . Then $\sum_{i=1}^j \xi_{n,i}$ is a martingale with respect to the filtration $\mathcal{H}_{n,j}$. Let $s_n(P_n)^2 = \sum_{i=1}^{2n} E_{P_n}(\xi_{n,i}^2 \mid \mathcal{H}_{n,i-1}) = V_{2,n}(P_n) + \sum_{i=1}^n k(X_i, D_i)^2 \sigma_P^2(X_i, D_i)$ and let $\tilde{s}_n(P_n)^2 = V_{2,n}(P_n) + V_{1,n}(P_n)$. We apply the martingale central limit theorem, Theorem 35.12 in [Billingsley \(1995\)](#), to the martingale $\sum_{i=1}^j \tilde{\xi}_{n,i}$ where $\tilde{\xi}_{n,i} = \xi_{n,i}/\tilde{s}_n(P_n)$. By Assumption B.1, $s_n(P_n)/\tilde{s}_n(P_n)$ converges in probability to one under P_n . Thus, $\sum_{i=1}^{2n} E_{P_n}(\tilde{\xi}_{n,i}^2 \mid \mathcal{H}_{n,i-1}) = s_n(P_n)^2/\tilde{s}_n(P_n)^2$ converges to one under P_n , which gives condition (35.35) in [Billingsley \(1995\)](#). To verify the Lindeberg condition (35.36) in [Billingsley \(1995\)](#), note that, for $i = 1, \dots, n$, $\tilde{\xi}_{n,i}^2 = (f_{P_n}(X_i, 1) - f_{P_n}(X_i, 0) - \tau(P_n))^2/(n^2 \tilde{s}_n(P_n)^2) = W_i^2/[nE_{P_n}(W_i^2) + n^2 V_{1,n}(P_n)] \leq W_i^2/[nE_{P_n}(W_i^2)] = \tilde{W}_i^2/n$ where $W_i = f_{P_n}(X_i, 1) - f_{P_n}(X_i, 0) - \tau(P_n)$ and $\tilde{W}_i = W_i/\sqrt{E_{P_n}(W_i^2)}$. Thus, for any $\varepsilon > 0$, $\sum_{i=1}^n \tilde{\xi}_{n,i}^2 \mathbb{I}\{\tilde{\xi}_{n,i}^2 > \varepsilon\} \leq E_{P_n} \tilde{W}_i^2 \mathbb{I}\{\tilde{W}_i^2 > n\varepsilon\}$ which converges to zero by the uniform $2 + \eta$ moment bounds. For the remaining terms, we have

$$\sum_{i=n+1}^{2n} \tilde{\xi}_{n,i}^2 \mathbb{I}\{\tilde{\xi}_{n,i}^2 > \varepsilon\} = \frac{\sum_{i=1}^n E_{P_n}[k(X_i, D_i)^2 u_i^2 \mathbb{I}\{k(X_i, D_i)^2 u_i^2 > \varepsilon(V_{2,n}(P_n) + V_{1,n}(P_n))\}]}{V_{2,n}(P_n) + V_{1,n}(P_n)}$$

which converges to 0 by Assumption B.1. Thus, $\sum_{i=1}^{2n} \tilde{\xi}_{n,i} \xrightarrow{d} N(0, 1)$ under P_n .

To complete the proof, note that

$$\begin{aligned} P_n(\tau(P_n) \notin [\hat{c}_\alpha, \infty)) &= P_n\left(\tau(P_n) < \hat{L}_k - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) - z_{1-\alpha} \text{se}_\tau(\hat{L}_k)\right) \\ &\leq P_n\left(\tau(P_n) < \sum_{i=1}^n k(X_i, D_i)u_i + Lf_{P_n} - z_{1-\alpha} \text{se}_\tau(\hat{L}_k)\right) = P_n\left(z_{1-\alpha} < \frac{\tilde{s}_n(P_n)}{\text{se}_\tau(\hat{L}_k)} \sum_{i=1}^{2n} \tilde{\xi}_{n,i}\right), \end{aligned}$$

which converges to α since $\tilde{s}_n(P)/\text{se}_\tau(\hat{L}_k)$ converges to one uniformly over $P \in \mathcal{P}$, □

B.2 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 $\mathcal{N}_{d,n} = \{i: D_i = d, i \in \{1, \dots, n\}\}$ and let

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

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

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

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

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

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

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

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 $\{x: \|x\|_{\mathcal{X}} \leq 1\}$, with $k(x) \geq \underline{k}$ on $\{x: \max_{1 \leq k \leq p} |x_k| \leq \eta\}$ for some $\underline{k} > 0$. By the above display, the function $f_n(x, d) = \sum_{\tilde{x} \in \tilde{\mathcal{X}}_n(h, \eta)} (1 - d)k((x - \tilde{x})/h)$ is equal to zero for $(x, d) = (X_i, D_i)$ for all $i = 1, \dots, n$. Thus, it is observationally equivalent to the zero function conditional on $\{X_i, D_i\}_{i=1}^n$: $P_{f_n}(\cdot \mid \{X_i, D_i\}_{i=1}^n) = P_0(\cdot \mid \{X_i, D_i\}_{i=1}^n)$. Furthermore, we have

$$Lf_n = -\frac{1}{n} \sum_{i=1}^n \sum_{\tilde{x} \in \tilde{\mathcal{X}}_n(h, \eta)} k((X_i - \tilde{x})/h) \leq -\underline{k} \frac{\#\mathcal{I}_n(h)}{n}, \quad (44)$$

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 . Let ℓ be the greatest integer strictly less than γ and let D^r denote the derivative with respect to the multi-index $r = r_1, \dots, r_p$ 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,

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

which implies that $\tilde{f}_n \in \Sigma(C, \gamma)$ where $\tilde{f}_n(x, d) = \frac{h^\gamma C}{K/\eta^p} f_n(x, d)$. By (44), the CATE under \tilde{f}_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 is bounded from above by $-\varepsilon n^{-1/p}$ for large enough n with probability one. On this probability one event,

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

which gives the result.

B.3 Proofs of Theorems 4.2 and 4.3

We use the following result, which shows that it suffices to verify eq. (22) for the weights \tilde{k}_δ^* and $k_{\text{match}, M}$. The proof follows by Theorem F.1 in Armstrong and Kolesár (2018b) and standard arguments, and is deferred to the supplemental materials.

Lemma B.2. *Suppose condition (a) of Theorem 4.2 holds, and suppose that for J fixed, $\max_i C_n \|x_{\ell_J(i)} - x_i\|_{\mathcal{X}} \rightarrow 0$, where $\ell_J(i)$ is the J th nearest neighbor of i . Let \mathcal{C} be the CI in Remark 3.3 based on a linear estimator such that eq. (22) holds, with \hat{u}_i^2 given by the nearest neighbor estimator with J neighbors. Then $\liminf_{n \rightarrow \infty} \inf_{f \in \mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha$.*

Proof of Theorem 4.2. By boundedness of $\tilde{\sigma}(x_i, d_i)$ away from zero and infinity, eq. (22) for $k = \tilde{k}_\delta^*$ is equivalent to

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

where \tilde{f}_δ^* is the solution to the optimization problem defined by (12) and (14) 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 (12) 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 $\max_{1 \leq i \leq n} \tilde{f}_\delta^*(x_i, d_i)^2 \rightarrow 0$.

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, $|\tilde{f}_\delta^*(x, d_{i_n^*})| \geq \eta - C_n \|x - x_{i_n^*}\|_{\mathcal{X}}$ so that, for $\|x - x_{i_n^*}\|_{\mathcal{X}} \leq \eta/(2C_n)$, we have $|\tilde{f}_\delta^*(x, d_{i_n^*})| \geq \eta/2$. Thus, we have

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

infinitely often. This gives a contradiction so long as (21) holds. \square

Proof of Theorem 4.3. To show that (22) holds for $k_{\text{match},M}(x_i, d_i) = (1 + K_M(i))/n$, it is sufficient to show that $\max_{1 \leq i \leq n} K_M(i)^2/n \rightarrow 0$. To this end, let $U_M(x, d) = \|x_j - x\|_{\mathcal{X}}$ where x_j is the M th 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 (23) holds and n is large enough so that $n\bar{G}(a_n) \geq M$, 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 (23) then implies $K_M(i) \leq n\bar{G}(a_n)$. Thus, it suffices to show that $[n\bar{G}(a_n)]^2/n = n\bar{G}(a_n)^2 \rightarrow 0$.

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

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

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 $1/\sqrt{b((\log n)^2/n)}$). \square

B.4 Proof of 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} , $\inf_{w \in \mathcal{W}} \#\{i \in \{1, \dots, n\} : \|w - w_i\| \leq \varepsilon\} \rightarrow \infty$ 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 $B_\delta(\tilde{w}) \subseteq B_\varepsilon(w)$, and hence the quantity of interest is bounded from below by

$$\min_{w \in \widetilde{\mathcal{W}}_\delta} \#\{i \in \{1, \dots, n\} : \|w - w_i\| < \delta\}, \quad (45)$$

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 (45) converges to infinity almost surely.

References

- Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*, 88(1):265–296.
- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.

- Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.
- Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843.
- Abadie, A., Imbens, G. W., and Zheng, F. (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508):1601–1614.
- Armstrong, T. B. and Kolesár, M. (2018a). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. ArXiv: 1712.04594v2.
- Armstrong, T. B. and Kolesár, M. (2018b). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.
- Armstrong, T. B. and Kolesár, M. (2020). Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, forthcoming.
- Bailey, M. J. and Goodman-Bacon, A. (2015). The war on poverty’s experiment in public medicine: Community health centers and the mortality of older Americans. *American Economic Review*, 105(3):1067–1104.
- Beliakov, G. (2006). Interpolation of Lipschitz functions. *Journal of Computational and Applied Mathematics*, 196(1):20–44.
- Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York, NY, 3 edition.
- Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *The Review of Economics and Statistics*, 96(5):885–897.
- Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840.
- Chen, X., Hong, H., and Tarozi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2):808–843.
- Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.

- Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.
- Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921.
- Donoho, D. L., Liu, R. C., and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, 18(3):1416–1437.
- Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. J. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.
- Galiani, S., Gertler, P., and Schargrotsky, E. (2005). Water for life: The impact of the privatization of water services on child mortality. *Journal of Political Economy*, 113(1):83–120.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.
- Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294.
- Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4):605–654.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.
- Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.
- Khan, S. and Nekipelov, D. (2013). On uniform inference in nonlinear models with endogeneity. Working paper, Duke University.
- Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620.
- Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

- Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23(3):824–835.
- Noack, C. and Rothe, C. (2020). Bias-aware inference in fuzzy regression discontinuity designs. Unpublished manuscript, University of Mannheim.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer, New York, NY.
- Robins, J., Tchetgen, E. T., Li, L., and van der Vaart, A. W. (2009). Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321.
- Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(1-3):285.
- Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030.
- Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660.
- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal Of The American Statistical Association*, 74(366):318–328.
- Smith, J. A. and Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review*, 91(2):112–118.
- Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125(1–2):305–353.
- Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *The Review of Economics and Statistics*, 86(1):91–107.