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

Timothy B. Armstrong[†] Michal Kolesár[‡]

Yale University Princeton University

July 20, 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 typically assumed that the treatment is unconfounded given a set of pretreatment covariates. This assumption implies that systematic differences in outcomes between treated and control units with the same values of the covariates are attributable to the treatment. When the covariates are continuously distributed, it is not possible to perfectly match the treated and control units based on their covariate values, and estimation of the ATE requires nonparametric regularization methods such as kernel, series or sieve estimators, or matching estimators that allow for imperfect matches.

The standard approach to comparing estimators and constructing confidence intervals (CIs) in this setting is based on the theory of semiparametric efficiency bounds. If, in addition to unconfoundedness, one also assumes overlap of the covariate distributions of treated and untreated subpopulations, as well as enough smoothness of either the propensity score or the conditional mean of the outcome given the treatment and covariates, many regularization methods lead to estimators that are $\sqrt{n}$-consistent, asymptotically unbiased and normally distributed, with variance that achieves the semiparametric efficiency bound (see, among others, Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003; Chen et al., 2008). One can then construct CIs based on any such estimator by adding and subtracting its standard deviation times the conventional 1.96 critical value (for nominal 95% CIs).

However, in many applications, the overlap is limited, which can have drastic effects on finite-sample performance (Busso et al., 2014; Rothe, 2017) and leads to an infinite semiparametric efficiency bound (Khan and Tamer, 2010).[1] Furthermore, even under perfect overlap, the standard approach requires a large amount of smoothness: one typically assumes continuous differentiability of the order $p/2$ at minimum (e.g. Chen et al., 2008), and often of the order $p+1$ or higher (e.g. Hahn, 1998; Heckman et al., 1998; Hirano et al., 2003), where p is the dimension of the covariates. Unless p is very small, such assumptions are hard to evaluate, and may be much stronger than the researcher is willing to impose. Finally, as argued in, for instance, Robins and Ritov (1997), the standard approach may not provide a good description of finite-sample behavior of estimators and CIs: in finite samples, regularization leads to bias, and different estimators have different finite-sample biases even if they are asymptotically equivalent. The bias may in turn lead to undercoverage of the CIs.

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

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 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 appear to have 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).

We show that 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, we do not need to make any overlap assumptions, or even require that the average treatment effect be point identified: our results cover both the regular case (in which $\sqrt{n}$-inference is possible) and the irregular case (in which $\sqrt{n}$-inference may not be possible, due to lack of perfect overlap, or due to low regularity of the regression function relative to the dimension of covariates).[4] In the

[2] In settings with asymmetric restrictions, we allow for the estimators to be affine rather than linear, taking the form $a + \sum_{i=1}^{n} k_i Y_i$ where a is an intercept term which may also depend on the covariates and treatments. See Appendix A.

[3] The 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.

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

latter case, the coverage of conventional CIs, which assume $\sqrt{n}$-convergence and do not account for bias, converges to zero asymptotically. 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 reason for asymptotic validity of our CIs is simple: because they are based on a linear estimator and account for its finite-sample bias, the CIs will be asymptotically valid—even in irregular and possibly set-identified cases—so long as the estimator is asymptotically normal when normalized by its standard deviation, which in turn holds if the estimator doesn't put too much weight k_i on any single observation. However, since the weights solve a bias-variance tradeoff, no single observation can receive too much weight—otherwise, in large samples, a large decrease in variance could be achieved at a small cost to bias. On the other hand, asymptotic normality may fail under limited overlap for other estimators: we show that asymptotic normality for matching estimators generally requires strong overlap. In our empirical application, we 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.

To formally show that conventional $\sqrt{n}$-asymptotics cannot be used when the dimension of the covariates is large relative to the smoothness of the regression function, we show that for $\sqrt{n}$-inference to be possible, one needs to assume a bound on the derivative of the conditional mean of order at least $p/2$. If one only assumes a bound on derivatives of lower order, the bias will asymptotically dominate the variance—in contrast to some nonparametric settings such as estimation of a conditional mean at a point, it is not possible to "undersmooth", and valid CIs need to take the bias into account. The smoothness condition is essentially the same as when one does not condition on treatment and covariates (Robins et al., 2009), and when no smoothness is imposed on the propensity score. Intuitively, by conditioning on the treatment and covariates, we take away any role that the propensity score may play in increasing precision of inference.

We illustrate the results in an application to the National Supported Work (NSW) Demonstration. We find that finite-sample optimal CIs are substantially different from those based on conventional $\sqrt{n}$-asymptotic theory, with bias determining a substantial portion of the CI width. Furthermore, our finite-sample approach allows us to investigate several questions that are moot under $\sqrt{n}$-asymptotic theory, due to the asymptotic equivalence of different estimators that achieve the semiparametric efficiency bound. For example, we examine how the weights that optimize the RMSE criterion differ from those that optimize CI length, and we find that, in our application, for constructing CIs, the optimal weights oversmooth slightly relative to the RMSE-optimal weights.

Our results rely on the key insight that, once one conditions on treatment assignments and pretreatment variables, the ATE is a linear functional of a regression function. This puts the problem in the general framework of Donoho (1994) and Cai and Low (2004) and allows us to apply 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. Whereas simple strategies, such as gridding, are infeasible unless the dimension of covariates p is very small, we show that, for Lipschitz smoothness, the problem can be reduced to convex optimization in a finite number of variables and constraints, which depend only on the sample size and not on p.[5] Furthermore, our solution path algorithm uses insights from Rosset and Zhu (2007) on computation of penalized regression estimators to further speed up computation. In independent and contemporaneous work, Kallus (2020) computes optimal linear weights using a different characterization of the optimization problem.

In contrast, if one does not condition on treatment assignments and pretreatment variables, the problem becomes more difficult: while upper and lower bounds have been developed that bound the optimal rate (Robins et al., 2009), computing efficiency bounds that are sharp in finite samples (or even bounds on the asymptotic constant in non-regular cases) remains elusive. Whether one should condition on treatment assignments and pretreatment 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) provide some discussion in related settings. Since our CIs are valid unconditionally, they can be used in either setting.[6]

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

[5] While restricting attention to small p (say ≤ 2) would be severely limiting in our setting, it is not restrictive in some other problems, such as regression discontinuity. For computation of optimal weights in other settings with small p, see Heckman (1988) and Imbens and Wager (2019).

[6] While 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 3.3.

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

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

We are interested in the conditional average treatment effect (CATE)[7], which, under the assumption of unconfoundedness, $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$, is given by

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

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

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

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

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 centrosymmetric, 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 consider classes

[7] We 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.

that place Lipschitz constraints on $f(\cdot, 0)$ and $f(\cdot, 1)$:

$$\mathcal{F}_{\text{Lip}}(C) = \{f \colon |f(x,d) - f(\tilde{x},d)| \leq C\|x - \tilde{x}\|_{\mathcal{X}},\, d \in \{0,1\}\}, \tag{3}$$

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

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

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

2.2 Linear estimators

We start by showing how to construct CIs based on an estimator that is linear in the outcomes Y_i,

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

For now, we treat the weights k as given—in Section 2.3, we will show how to choose these weights 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.[8] 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

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

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 maximum bias

$$\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) = \sup_{f\in\mathcal{F}} E_f(\hat{L}_k - Lf) = \sup_{f\in\mathcal{F}} \left[\sum_{i=1}^{n} k(x_i, d_i) f(x_i, d_i) - Lf\right]. \tag{7}$$

[8] Nonlinear 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.

and variance $\operatorname{sd}(\hat{L}_k)^2 = \sum_{i=1}^n k(x_i, d_i)^2 \sigma^2(x_i, d_i)$. By centrosymmetry of $\mathcal{F}$, $\inf_{f\in\mathcal{F}} E_f(\hat{L}_k - Lf) = -\overline{\operatorname{bias}}_{\mathcal{F}}(\hat{L}_k)$, and if the minimum bias obtains at f^*, then the maximum bias obtains at $-f^*$.

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

$$\hat{c} = \hat{L}_k - \overline{\operatorname{bias}}_{\mathcal{F}}(\hat{L}_k) - \operatorname{sd}(\hat{L}_k) z_{1-\alpha}, \tag{8}$$

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

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

$$\left\{ \hat{L}_k \pm \operatorname{cv}_\alpha(\overline{\operatorname{bias}}_{\mathcal{F}}(\hat{L}_k)/\operatorname{sd}(\hat{L}_k)) \cdot \operatorname{sd}(\hat{L}_k) \right\}. \tag{9}$$

Note that $\operatorname{cv}_\alpha(0) = z_{1-\alpha/2}$, so that if $\hat{L}_k$ is unbiased, the critical value reduces to the usual critical value based on standard normal quantiles. For positive values of the worst-case bias-standard deviation ratio, it will be larger: for $b \geq 1.5$ and $\alpha \leq 0.2$, $\operatorname{cv}_\alpha(b) \approx b + z_{1-\alpha}$ up to three decimal places.[9] For large values of b, the CI is therefore approximately given by adding and subtracting $\overline{\operatorname{bias}}_{\mathcal{F}}(\hat{L}_k) + z_{1-\alpha} \operatorname{sd}(\hat{L}_k)$ from $\hat{L}_k$.

Following Donoho (1994), we refer to the CI (9) as a fixed-length confidence interval (FLCI), since its length does not depend on the realized outcomes—it only depends on the known variance function $\sigma^2(\cdot, \cdot)$ and the realized treatment and covariate values $\{x_i, d_i\}_{i=1}^n$.

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),

$$R_{\text{RMSE},\mathcal{F}}(\hat{L}_k) = \left(\sup_{f\in\mathcal{F}} E_f(\hat{L}_k - Lf)^2 \right)^{1/2} = \left(\overline{\operatorname{bias}}_{\mathcal{F}}(\hat{L}_k)^2 + \operatorname{sd}(\hat{L}_k)^2 \right)^{1/2}. \tag{10}$$

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

One-sided CIs can be compared using the maximum β-quantile of excess length, for a given β (see Appendix A). Finally, to evaluate a FLCI that satisfies (4), we simply consider its length, $2\,\mathrm{cv}_{\alpha}(\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)/\,\mathrm{sd}(\hat{L}_k))\cdot\mathrm{sd}(\hat{L}_k)$. Since the length of the CI is fixed—it doesn't depend on the data $\{Y_i\}_{i=1}^n$—choosing the weights k to minimize the length does not affect the coverage properties of the resulting CI.

While in general, the weights k that minimize FLCI length will be different from the one that minimizes RMSE, both performance criteria depend on k only through $\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)$, and $\mathrm{sd}(\hat{L}_k)$, and they are increasing in both quantities (this is also true for 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 $\overline{\mathrm{bias}}_{\mathcal{F}}(\hat{L}_k)$ subject to a bound on variance. We can then vary the bound to find the optimal bias-variance tradeoff for a given performance criterion (FLCI or RMSE). It follows from Donoho (1994) and Low (1995) that this bias-variance frontier can be traced out by solving a certain convex optimization indexed by δ. 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)[10]. Intuitively, since $L\iota_{\kappa}=\kappa$, the set of functions $\{\iota_{\kappa}\}_{\kappa\in\mathbb{R}}$ is the smoothest set of functions that span the potential values of the CATE parameter Lf, so that this assumption will typically hold unless $\mathcal{F}$ places constraints on the possible values Lf. For a given $\delta>0$, let f_{δ}^* solve

$$\max_{f\in\mathcal{F}} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^{n}\frac{f(x_i,d_i)^2}{\sigma^2(x_i,d_i)}\leq\frac{\delta^2}{4}, \tag{11}$$

and, with a slight abuse of notation, define

$$\hat{L}_{\delta}=\hat{L}_{k_{\delta}^*}, \qquad k_{\delta}^*(x_i,d_i)=\frac{f_{\delta}^*(x_i,d_i)/\sigma^2(x_i,d_i)}{\sum_{j=1}^{n}d_j f_{\delta}^*(x_j,d_j)/\sigma^2(x_j,d_j)}. \tag{12}$$

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

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

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

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

by

$$\mathrm{sd}(\hat{L}_\delta)^2 = \frac{\delta^2}{4(\sum_{j=1}^n d_j f_\delta^*(x_j, d_j)/\sigma^2(x_j, d_j))^2}.$$

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

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

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 (7). Computing the RMSE-optimal estimator, and the optimal FLCI requires solving the optimization problem (11), and then varying δ to find the optimal bias-variance tradeoff. Both of these optimization problems require optimizing over the set $\mathcal{F}$, which, in nonparametric settings, is infinite-dimensional. We now focus on the Lipschitz class $\mathcal{F} = \mathcal{F}_{\mathrm{Lip}}(C)$, and show that in this case, the solution to (7) can be found by solving a finite-dimensional linear program. The optimization problem (11) 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).

First, observe that in both optimization problems (7) and (11), the objective and constraints depend on f only through its value at the points $\{(x_i, 0), (x_i, 1)\}_{i=1}^n$; the value of f at other points does not matter. Furthermore, it follows from Beliakov (2006, Theorem 4) that if the Lipschitz constraints hold at these points, then it is always possible to find a function $f \in \mathcal{F}_{\mathrm{Lip}}(C)$ that interpolates these points (see Lemma A.1). Consequently, in solving the optimization problems (7) and (11), we 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 $f \in \mathcal{F} = \mathcal{F}_{\mathrm{Lip}}(C)$ with $2n(n-1)$ linear inequality constraints

$$f(x_i, d) - f(x_j, d) \leq C\|x_i - x_j\|_{\mathcal{X}} \quad d \in \{0, 1\},\ i, j \in \{1, \dots, n\}. \tag{13}$$

This leads to the following result:

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

$$\sum_{i=1}^n d_i k(x_i, d_i) = 1, \quad \text{and} \quad \sum_{i=1}^n (1-d_i)k(x_i, d_i) = -1. \tag{14}$$

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

$$\max_{f \in \mathbb{R}^{2n}} \left\{ \sum_{i=1}^n k(x_i, d_i) f(x_i, d_i) - \frac{1}{n} \sum_{i=1}^n [f(x_i, 1) - f(x_i, 0)] \right\}, \tag{15}$$

where the maximum is taken subject to (13). 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 the following subset of the constraints in (13):

$$f(x_i, 1) \leq f(x_j, 1) + C\|x_i - x_j\|_{\mathcal{X}}, \quad \text{all } i, j \text{ with } d_i = 1,\ d_j = 0 \text{ and } k(x_i, 1) > 1/n, \tag{16}$$

$$f(x_i, 0) \leq f(x_j, 0) + C\|x_i - x_j\|_{\mathcal{X}}, \quad \text{all } i, j \text{ with } d_i = 1,\ d_j = 0 \text{ and } k(x_j, 0) < -1/n. \tag{17}$$

The assumption that $\hat{L}_k$ satisfies (14) 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 by the value of (15).

The last part of the theorem follows by checking that the remaining constraints in (13) are automatically satisfied at the optimum. The conditions (16) and (17) give at most $2n_0n_1$ inequalities, where n_d is the number of observations with $d_i = d$. The condition on the weights k holds, for example, for the matching estimator given in (6). Since for the matching estimator $k(x_i, d_i) = (2d_i - 1)/n$ if observation i is not used as a match, the theorem says that one only needs to impose the constraint (13) for pairs of observations with opposite treatment status, and for which one of the observations is used as a match. Consequently, in settings with imperfect overlap, in which many observations are not used as a match, the number of constraints will be much lower than $2n_0n_1$.

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

Theorem 2.2. *Given $\delta > 0$, the value of the maximizer f_δ^* of (11) is given by the solution to the convex program*

$$\max_{f \in \mathbb{R}^{2n}} 2Lf \quad s.t. \quad \sum_{i=1}^n \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4} \quad \text{and s.t. (13).} \tag{18}$$

Furthermore, if $\sigma^2(x, d)$ doesn't depend on x, it suffices to impose the constraints (13) for $i, j \in \{1, \dots, n\}$ with $d_i = 0$ and $d_j = 1$, and the solution path $\{f_\delta^*\}_{\delta>0}$ can be computed by *the piecewise linear algorithm given in Appendix A.3.*

Theorem 2.2 shows that the infinite-dimensional program (11) can be replaced by a quadratic optimization problem in $\mathbb{R}^{2n}$ with $2n(n-1)$ linear constraints, one quadratic constraint and a linear objective function. 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 we discuss in more detail in Appendix A.3, it follows from the algorithm that, similarly to the matching estimator (see eq. (6)), 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 (13) 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 observations with opposite treatment status that are close to i according to the matrix of effective distances. Thus, observations for which there exist more good matches receive relatively more matches, since this decreases the variance of the estimator at little cost in terms of bias. Also, since the weight $k_\delta^*(x_j, d_j) = \frac{1}{n}(1-2d_j)(1+\sum_{i=1}^{n} W_{\delta,ij})$ on j is increasing in the number of times it has been used as a match, using it more often as a match increases the variance of the estimator. Using the "effective distance" matrix trades off this increase in the variance against an increase in the bias that results from using a lower-quality match instead.

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

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

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

Theorem 2.3 does not mean that one should choose a large value of 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 one finds it 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 the one used to define the Lipschitz constraint. Zhao (2004) has argued that conditions on the regression function should be considered when defining the metric used for matching. Theorem 2.3 establishes a formal connection between conditions on the regression function and the optimal metric for matching. We investigate this issue further in the context of our empirical application by calculating the efficiency loss from matching with the "wrong" metric (see Section 5.4).

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 linear minimax estimator—the estimator $\hat{L}_{\delta_{\text{RMSE}}}$ achieving the lowest RMSE in the class of linear estimators (5)—is also highly efficient among all estimators. Its efficiency is at least $\sqrt{80\%} = 89.4\%$ (i.e. one cannot reduce the RMSE by more than 10.6% by considering non-linear estimators), and, in particular applications, its efficiency can be shown to be even higher.

For FLCIs, an even stronger result obtains, addressing two concerns that one may have. 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 the whole parameter space), especially if this leads to substantial reduction in expected length when $f \in \mathcal{G}$. When such a CI also simultaneously 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 $\mathcal{F}$, including any smoothness constants such as the Lipschitz constant C if $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$. This rules out data-driven procedures that try to implicitly or explicitly estimate C from the data.

To address these concerns, 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 form of 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, we recommend reporting estimates and CIs for a range of choices of the Lipschitz constant C when implementing the FLCI in practice to see how assumptions about the parameter space affect the results. We adopt this approach in the empirical application in Section 5. This also mirrors the common practice of reporting results for different specifications of the regression function in parametric regression problems.

The key assumption needed for these efficiency bounds is that the parameter space $\mathcal{F}$ be convex and centrosymmetric. This holds for the function class $\mathcal{F}_{\text{Lip}}(C)$, and, more generally, for parameter spaces that place bounds on derivatives of f. If additional restrictions such as monotonicity are used that break either convexity or centrosymmetry, then some degree of adaptation may be possible. While we leave the full exploration of this question for future research, we note that the approach in Section 2.3 can still be used when the centrosymmetry assumption is dropped. As an example, we show how optimal FLCIs can be computed when $\mathcal{F}$ imposes Lipschitz and monotonicity constraints in Appendix A.2.

3 Practical implementation

We now discuss implementation of feasible versions of the estimators and CIs introduced 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, and practical implementation issues. In Section 3.3, we show how our approach can be used 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:[11]

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 $\mathcal{F} = \mathcal{F}_{\mathrm{Lip}}(C)$, 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, $\widetilde{\mathrm{sd}}_\delta^2 = \sum_{i=1}^n \tilde{k}^*_\delta(x_i,d_i)^2\tilde{\sigma}^2(x_i,d_i)$ denote its variance computed using $\tilde{\sigma}^2(x,d)$ as the variance function, and let $\overline{\mathrm{bias}}_\delta = \overline{\mathrm{bias}}_{\mathcal{F}}(\tilde{L}_\delta)$ denote its worst-case bias (which doesn't depend on the variance specification).

3. Compute the minimizer $\tilde{\delta}_{\mathrm{RMSE}}$ of $\overline{\mathrm{bias}}_\delta^2 + \widetilde{\mathrm{sd}}_\delta^2$. Compute the standard error $\mathrm{se}(\tilde{L}_{\tilde{\delta}_{\mathrm{RMSE}}})$ using the robust variance estimator

$$\mathrm{se}(\tilde{L}_\delta)^2 = \sum_{i=1}^n \tilde{k}^*_\delta(x_i,d_i)^2\hat{u}_i^2, \tag{19}$$

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

$$\left\{\tilde{L}_\delta \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_\delta / \mathrm{se}(\tilde{L}_\delta))\,\mathrm{se}(\tilde{L}_\delta)\right\}, \tag{20}$$

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

To estimate the conditional variance in (19), we can take $\hat{u}_i^2 = (Y_i - \hat{f}(x_i,d_i))^2$, where $\hat{f}(x,d)$ is a consistent estimator of $f(x,d)$, or the nearest-neighbor variance estimator of Abadie and Imbens (2006) $\hat{u}_i = J/(J+1)\cdot(Y_i - \hat{f}(x_i,d_i))^2$, where $\hat{f}(x_i,d_i)$ average outcome of J

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

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. (20) depends on the variance estimates, it is no longer fixed, in contrast to the infeasible FLCI.

In general, the point estimate $\tilde{L}_{\tilde{\delta}_{\mathrm{RMSE}}}$ will differ from the estimate $\tilde{L}_{\tilde{\delta}_{\mathrm{FLCI}}}$ used to form the CI. Since reporting multiple estimates can be cumbersome, one can simply compute the CI (20) at $\delta = \tilde{\delta}_{\mathrm{RMSE}}$. 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 1.1% longer than the one that reoptimizes δ for CI length.

Remark 3.1 (Specification of $\mathcal{F}$)**.** In forming our CI, we need to choose the function class $\mathcal{F}$. In case of the Lipschitz class, we still need to complete the definition of $\mathcal{F}$ by choosing the constant C and the norm on x used to define the Lipschitz condition. The results discussed in Section 2.5 imply that it is not possible to make these choices automatically in a data-driven way. Thus, we recommend that these choices be made using problem-specific knowledge wherever possible, and that CIs be reported for a range of plausible values of C as a form of sensitivity analysis. We consider these problems in more detail in the context of our application in Sections 5.1 and 5.4. 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 (18), 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 $\tilde{f}^*_{\delta,C} = C\tilde{f}^*_{\delta/C,1}$.

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 $\tilde{\delta}_{\mathrm{RMSE}}$ and $\tilde{\delta}_{\mathrm{FLCI}}$ depend on the initial guess $\tilde{\sigma}^2(x,d)$. Thus, the resulting CI in eq. (20) will not be optimal if this guess is incorrect. However, because the standard error estimator (19) 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, and in cases in which the CATE is not point identified (in which case our CI has asymptotically valid coverage for points in the identified set).

Second, the worst-case bias calculations do not depend on the error distribution. The feasible CI thus reflects the finite-sample impact of the empirical distribution of the covariates and treatment (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.[12] 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, while the CIs are asymptotically valid under heteroskedastic and non-normal errors.

3.2 CIs based on other estimators

To form a feasible CI based on a linear estimator $\hat{L}_k = \sum_{i=1}^n k(x_i, d_i)Y_i$, one can mirror the steps in the baseline implementation, using the weights $k(x_i, d_i)$ in eq. (19), and computing the worst-case bias by solving the optimization problem in Theorem 2.1. If $\hat{L}_k$ is an estimator that is asymptotically unbiased under conventional asymptotics, so that the conventional critical value $z_{1-\alpha/2}$ is asymptotically justified, one can still compute the critical value $\operatorname{cv}_\alpha(\overline{\operatorname{bias}}_{\mathcal{F}}(\hat{L}_k)/\operatorname{se}(\hat{L}_k))$ as a form of sensitivity analysis: if it's substantially larger than $z_{1-\alpha/2}$, this indicates that conventional asymptotics may not work well for the sample at hand unless one further restricts the parameter space for f.

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

[12] The 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 error distributions includes normal errors.

3.3 CI for the population average treatment effect

We now show how our approach can be adapted to construct CIs for the PATE based on estimators $\hat{L}_k$ of the form (5) that are linear in the outcomes. To do so, let us 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, as before, $Lf = \frac{1}{n}\sum_i E[Y_i(1) - Y_i(0) \mid X_i = x_i]$ denotes the CATE.

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. (9). 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 $\hat{L}_k$,

$$\{\hat{L}_k \pm (\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) + z_{1-\alpha/2}\,\text{se}_\tau(\hat{L}_k))\}. \tag{21}$$

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. (19), 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, one can use the estimator $\frac{1}{n^2}\sum_{i=1}^n[(\hat{f}(x_i,1) - \hat{f}(x_i,0))^2 - \hat{L}_k^2]$ where $\hat{f}(x,d)$ is the estimator of $f(x,d)$ used in eq. (19). Alternatively, if we use the nearest neighbor variance estimator $\hat{u}_i^2$ in (19), 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 can use the nearest neighbor estimator suggested by Abadie and Imbens (2006, Theorem 7), $\frac{1}{n^2}\sum_{i=1}^n(\hat{f}(x_i,1) - \hat{f}(x_i,0) - \hat{L}_k)^2 - \frac{1}{n^2}\sum_{i=1}^n(1 + \sum_{j=1}^n W_{ji}^2)\hat{u}_i^2$. In Appendix B.1, we provide formal asymptotic coverage results for the CI in (21) and its one-sided analog.

4 Asymptotic results

We now consider the asymptotic validity of feasible CIs with unknown error distribution, as well as bounds on the rate of convergence of estimators and CIs. In Section 4.1, we show formally that $\sqrt{n}$-inference is impossible when the dimension of covariates is high enough relative to the order of smoothness imposed by $\mathcal{F}$. In Section 4.2, we show that our feasible CIs are asymptotically valid and centered at estimators that are asymptotically normal when scaled by their standard deviation. Sections 4.3 and 4.4 give conditions for asymptotic validity and optimality, respectively, 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 is based on estimators that are $\sqrt{n}$-consistent and asymptotically normal, with asymptotically negligible bias. CIs based on such estimators are constructed by simply adding and subtracting their standard deviation times the usual critical value, $z_{1-\alpha/2}$. We now show that if the dimension of the (continuously distributed) covariates p is high enough relative to the smoothness of $\mathcal{F}$, this standard approach based on $\sqrt{n}$-consistency 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 $\sum_{j=1}^{p} k_j = \ell$, $\left| \frac{d^\ell}{dx_1^{k1} \cdots dx_p^{kp}} f(x) - \frac{d^\ell}{dx_1^{k1} \cdots dx_p^{kp}} f(x') \right| \leq C \|x - x'\|_{\mathcal{X}}^{\gamma-\ell}$, where ℓ is the greatest integer strictly less than γ and $\|\cdot\|_{\mathcal{X}}$ denotes the Euclidean norm on $\mathbb{R}^p$. Note that $f \in \mathcal{F}_{\text{Lip}}(C)$ is equivalent to $f(\cdot, 1), f(\cdot, 0) \in \Sigma(1, C)$.

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 (2) 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 with asymptotic coverage at least $1 - \alpha$ under $\Sigma(\gamma, C)$ for the CATE conditional on $\{X_i, D_i\}_{i=1}^n$:*

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

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

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

almost surely.

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

On the other hand, when $\gamma/p > 1/2$, Chen et al. (2008) show that, for example, series estimators do achieve the semiparametric efficiency bound. In other words, they are regular, $\sqrt{n}$-consistent and asymptotically normal and unbiased, with the lowest possible asymptotic variance (while Chen et al. 2008 do not condition on treatments and pretreatment variables, their arguments appear to extend to the conditional case).

4.2 Asymptotic validity of feasible CIs

The following theorem gives sufficient conditions for the asymptotic validity of the feasible CIs given in Section 3.1 based on the estimator $\tilde{L}_\delta$ when $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C)$. To allow us to better capture the finite-sample bias of the estimator in the asymptotic approximation, we allow $C = C_n \to \infty$ as $n \to \infty$. For concreteness, we restrict attention to particular forms of standard errors, based on nearest neighbor or uniform kernel estimates.

Theorem 4.2. *Consider the model (1) with $1/K \leq Eu_i^2 \leq K$ and $E|u_i|^{2+1/K} \leq K$ for some constant K. Suppose that*

$$\text{for all } \eta > 0 \ \min_{1 \leq i \leq n} \#\{j \in \{1, \dots, n\} \colon \|x_j - x_i\|_{\mathcal{X}} \leq \eta/C_n,\, d_i = d_j\} \to \infty, \tag{22}$$

and that the variance function $\sigma^2(x, d)$ is uniformly continuous in x for $d \in \{0, 1\}$. Let $\mathcal{C}$ be the CI in eq. (20) based on the feasible estimator $\tilde{L}_\delta$ with $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_n)$, with δ fixed and $\tilde{\sigma}^2(x, d)$ a nonrandom function bounded away from zero and infinity. Suppose the estimator $\hat{u}_i^2$ in (19) is the nearest-neighbor variance estimator based on a fixed number of nearest neighbors J, or that $\hat{u}_i^2 = (Y_i - \hat{f}(x_i, d_i))^2$, where $\hat{f}(x_i, d_i)$ the Nadaraya-Watson estimator with uniform kernel and a bandwidth sequence h_n with $h_n C_n$ converging to zero slowly enough. Then $\liminf_{n\to\infty} \inf_{f \in \mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1 - \alpha$.

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

Lemma 4.1. *Suppose that (X_i, D_i) is drawn i.i.d. from a distribution where X_i has bounded support and $0 < P(D_i = 1) < 1$, and that $C_n = C$ is fixed. Then (22) holds almost surely.*

In particular, Theorem 4.2 shows that the feasible CIs are asymptotically valid in irregular cases in which conventional CIs suffer from asymptotic undercoverage. This includes settings in which the covariate dimension p is high enough so that Theorem 4.1 applies and $\sqrt{n}$-inference is impossible, and settings with imperfect overlap (as in Khan and Tamer, 2010) including set identification due to complete lack of overlap. The estimator $\tilde{L}_\delta$ 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 $k_\delta^*(x_j, d_j)$ on any individual observation in the sense that for $k = k_\delta^*$, as $n \to \infty$,

$$\mathrm{Lind}(k) = \frac{\max_{1 \leq j \leq n} k(x_j, d_j)^2}{\sum_{i=1}^n k(x_i, d_i)^2} \to 0. \tag{23}$$

To give intuition for why (23) 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 (22), 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 $\mathrm{Lind}(k_\delta^*)$ and checking whether it is close to 0. We do this in our application in Section 5.[13]

4.3 Asymptotic validity 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 $\overline{G}\colon \mathbb{R}^+ \to \mathbb{R}^+$ and $\underline{G}\colon \mathbb{R}^+ \to \mathbb{R}^+$ be functions with $\lim_{t \to 0} \frac{\overline{G}(\underline{G}^{-1}(t))^2}{t/\log t^{-1}} = 0$. Suppose that, for any sequence a_n with $n\underline{G}(a_n)/\log n \to \infty$, we have*

$$\underline{G}(a_n) \leq \frac{\#\{i : \|x_i - x\|_{\mathcal{X}} \leq a_n,\, d_i = d\}}{n} \leq \overline{G}(a_n) \quad \text{all } x \in \mathcal{X},\, d \in \{0, 1\} \tag{24}$$

[13] One can also directly ensure that $\mathrm{Lind}(k_\delta^*)$ be 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 $\mathrm{Lind}(k_\delta^*)$ 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 Section 3.2 based on the matching estimator with a fixed number of matches M, and $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C_n)$. Then $\liminf_{n\to\infty} \inf_{f\in\mathcal{F}_{\text{Lip}}(C_n)} P_f(Lf \in \mathcal{C}) \geq 1-\alpha$.

Theorem 4.3 is related to results of Abadie and Imbens (2006) on asymptotic properties of matching estimators with a fixed number of matches. Abadie and Imbens (2006) note that, when p is large enough, the bias term will dominate, so that conventional CIs based on matching estimators will not be valid. In contrast, the CIs in Theorem 4.3 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 (24). 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 $\lim_{t\to 0} \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}] = 0$ 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 $P(D_i = 1 \mid X_i)$ is bounded away from zero and one on the support of the covariates.

On the other hand, (24) fails if there is not sufficient overlap, and this can lead to failure of asymptotic normality 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. (23) will be bounded away from zero for this observation, and the CLT will fail. In contrast, by Lemma 4.1, the estimator $\tilde{L}_\delta$ 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 $\mathrm{Lind}(k_{\mathrm{match},M})$. 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, using methods such as those proposed in Rothe (2017).[14]

4.4 Asymptotic efficiency of matching estimator with one match

By Theorem 2.2, the matching estimator with $M = 1$ is efficient in finite samples if the Lipschitz constant C is large enough. We now give conditions for its asymptotic optimality.

Theorem 4.4. *Suppose that the assumptions of Theorem 4.1 hold, and that $\sigma^2(x, d)$ is bounded away from zero and infinity. Suppose that, for some functions $\overline{G} \colon \mathbb{R}^+ \to \mathbb{R}^+$ and $\underline{G} \colon \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{t\to 0} \overline{G}(\underline{G}^{-1}(t))^2/[t/\log t^{-1}]^{2/p+1} = 0$,*

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

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

If X_i has sufficiently regular support and the conditional density of X_i given D_i is bounded away from zero on the support of X_i for both $D_i = 0$ and $D_i = 1$, then the conditions of Theorem 4.4 hold with $\underline{G}(a)$ and $\overline{G}(a)$ both given by constants times a^p, so that $\overline{G}(\underline{G}(a))$ decreases like a as $a \to 0$. Thus, the conditions of Theorem 4.4 hold so long as $p > 2$ and there is sufficient overlap. 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, as noted in Abadie and Imbens (2006), matching with $M = 1$ is suboptimal when $p = 1$. In addition, the conditions of Theorem 4.4 fail if there is insufficient overlap, and this may lead to asymptotic inefficiency.

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

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).[15] 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.[16] 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).

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 (3). We consider weighted ℓ_q norms of the form

$$\|x\|_{A,q} = \left(\sum_{j=1}^{p} |A_{jj} x_j|^q \right)^{1/q}, \tag{25}$$

where A is a diagonal matrix. Let us now describe our main specification; we discuss other specifications in Section 5.4. To make the restrictions on f implied by the choice of norm and C interpretable, in our main specification, we take $A = A_{\text{main}}$, with the diagonal elements of A_{main} given in Table 1, and set $C = 1$. The jth 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}.

[15] We use the data from Rajeev Dehejia's website, http://users.nber.org/~rdehejia/nswdata2.html.

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

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 the 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 at 10 thousand dollars are 1.5% and 6%, respectively, which is 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 (19) (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

To illustrate the bias-variance trade-offs in forming estimators and CIs, Figure 1 plots the estimator $\tilde{L}_\delta$ along with its standard deviation, worst-case bias, RMSE and CI length as a function of the smoothing parameter δ (recall that δ determines the relative weight on variance in the bias-variance tradeoff; it plays a role analogous to a bandwidth parameter). It shows that while the bias is increasing in δ and the variance is decreasing, the optimal resolution of the bias-variance trade-off depends on the optimality criterion. Interestingly, the optimal δ is *smaller* for the RMSE criterion than for optimizing CI length: it is cheaper, in terms of CI length, to use an estimator with *larger* bias and smaller variance than the RMSE-optimal estimator, and to take this bias into account by widening the CI. For comparison, Figure 2 plots the analog of Figure 1 for the matching estimator as a function of M, the number of matches.

Table 2 reports the point estimates and CIs at values of δ and M that optimize the RMSE and CI length criteria in Figures 1 and 2. There are three aspects of the results worth highlighting. First, in all cases, 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 (23) are low for the feasible estimators $\tilde{L}_\delta$ 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). This suggests that 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 (21), using the nearest neighbor variance estimator to estimate the marginal variance. The robust standard error based on the marginal variance of the estimator is only slightly bigger than the conditional (on treatment and covariates) standard error reported in Panel A. As a result, in all columns, the CIs for the PATT are only slightly longer than those for the CATT.[17]

To examine sensitivity of the results to the specification of the parameter space $\mathcal{F}$, Figure 3 plots the estimator $\tilde{L}_\delta$ 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.

[17] The bias-standard deviation ratio is above 1.5 in these specifications. Thus, as discussed in Section 2.2, the CI for the CATT is approximately given by adding and subtracting $\overline{\text{bias}}(\hat{L}_k) + z_{1-\alpha}\,\text{se}(\hat{L}_k)$ from the estimator. Most of the increase in length of the PATT CI comes from the fact that in eq. (21), we use a two-sided critical value $z_{1-\alpha/2}$, rather than the slightly larger marginal standard error $\text{se}_\tau(\hat{L}_k)$.

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.[18] 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 from the NSW non-experimental data is too ambitious, unless one imposes substantially stronger smoothness assumptions. Furthermore, differences between the estimates reported here and the experimental estimate may also arise from (1) failure of the selection on observables assumption; and (2) the sampling error in the experimental and non-experimental estimates.

5.4 Other choices of distance

A disadvantage of the distance based on $A = A_{\text{main}}$ is that it requires prior knowledge of the relative importance of different pretreatment variables in explaining the outcome variable.

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

An alternative is to specify the distance using moments of the pretreatment variables in a way that ensures invariance to scale transformations. For example, Abadie and Imbens (2011) form matching estimators using the weighted Euclidean norm (so $q = 2$) with $A = A_{\text{ne}} \equiv \text{diag}(1/\text{std}(x_1), \dots, 1/\text{std}(x_p))$, where std denotes sample standard deviation. Table 1 shows the diagonal elements of A_{ne}. It can be seen that this distance is most likely not the best way of encoding a researcher's prior beliefs about Lipschitz constraints. For example, the bound on the difference in average earnings between blacks and non-black non-Hispanics is substantially smaller than the bound on the difference in average earnings between Hispanics and non-black non-Hispanics.

If the constant C is to be chosen conservatively, the derivative of $f(x, d)$ with respect to each of these variables must be bounded by C times the corresponding element in this table. If one allows for somewhat persistent earnings, then C should be chosen in the range of 10 or above: to allow previous years' earnings to have a one-to-one effect, we would need to take $C = 1/\sqrt{.07^2 + .07^2} = 10.1$. For this C, when δ is chosen to optimize CI length, the resulting CI is given by 1.72 ± 7.63, which is much wider than the CIs reported in Table 2.

In Theorem 2.3, we showed that the matching estimator with a single match is optimal for C large enough. For this result, it is important that the norm used to construct the matches is the same as the norm defining the Lipschitz class. To illustrate this point, consider a matching estimator considered in Abadie and Imbens (2011), that uses $q = 2$ and $A = A_{\text{ne}}$. The RMSE efficiency of this estimator under our main specification (A_{main}, $q = 1$ and $C = 1$) is 77.5%; for CI length, its efficiency is 74.6%. This is considerably lower than the efficiencies of the matching estimator that matched on the norm defining the Lipschitz class reported in Section 5.2. Furthermore, the efficiency is never higher than 80.1%, even for large values of C.

Appendix A Finite-sample results: proofs and additional details

This appendix contains proofs and derivations in Section 2, as well as additional results. Appendix A.1 maps a generalization of the setup in Section 2.1 to the framework of Donoho (1994) and Armstrong and Kolesár (2018b), and specializes their general efficiency bounds and optimal estimator and CI construction to the current setting. This gives the formulas for optimal estimators and CIs given in Section 2.3, and the efficiency bounds discussed in Section 2.5. Appendix A.2 specializes the setup to the case with Lipschitz constraints, while allowing for possible additional monotonicity constraints, as well as allowing for the Lipschitz constraints to be 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 consider a generalization of the setup in Section 2.1 by letting the parameter of interest be a general weighted conditional average treatment effect of the form

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

where $\{w_i\}_{i=1}^n$ is a set of known weights that sum to one, $\sum_{i=1}^n w_i = 1$. Setting $w_i = 1/n$ gives the CATE, while setting $w_i = d_i/n_1$, gives the conditional average treatment effect on the treated (CATT). Here $n_d = \sum_{j=1}^n \mathbb{I}\{d_j = d\}$ gives the number of observations with treatment status equal to d. We retain the assumption that $\mathcal{F}$ is convex, but drop the centrosymmetry assumption. To handle non-centrosymmetric cases, we slightly generalize the class of estimators by considering 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

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

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. (8) and (9), with $\hat{L}_{k,a^*}$ in place of $\hat{L}_k$, with its RMSE given by eq. (10). 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 βth quantile of excess length over $\mathcal{G}$:

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

where $Lg - \hat{c}$ is the excess length of the CI $[\hat{c}, \infty)$, and $q_{g,\beta}(\cdot)$ denotes the βth quantile under the function g,

$$q_\beta(\hat{c}, \mathcal{G}) = 2\,\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_{k,a^*}) + \text{sd}(\hat{L}_{k,a^*})(z_{1-\alpha} + z_\beta),$$

This follows from the fact that the worst-case βth quantile of excess length over $\mathcal{G}$ is taken at the function $g \in \mathcal{G}$ that achieves $\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_{k,a})$ (i.e. when the estimate is biased downward as much as possible). 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.[19] 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)

$$\omega(\delta) = \sup_{f,g\in\mathcal{F}} \left\{ Lg - Lf \colon \sum_{i=1}^{n} \frac{(f(x_i, d_i) - g(x_i, d_i))^2}{\sigma^2(x_i, d_i)} \leq \delta^2 \right\}, \tag{26}$$

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 (11) 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

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

and

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

If the class $\mathcal{F}$ is translation invariant in the sense that $f \in \mathcal{F}$ implies $f + \iota_\kappa \in \mathcal{F}$[20], then by Lemma D.1 in Armstrong and Kolesár (2018b), the modulus is differentiable, with the

[19] In 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 $\mathcal{F}$.

[20] In 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 10, implies translation invariance.

derivative $\omega'(\delta) = \delta / \sum_{i=1}^n d_i(g_\delta^*(x_i, d_i) - f_\delta^*(x_i, d_i))/\sigma^2(x_i, d_i)$. The formula for $\hat{L}_\delta$ in the main text follows from this result combined with fact that, under centrosymmetry, $f_\delta^* = -g_\delta^*$. By Lemma A.1 in Armstrong and Kolesár (2018b), the maximum and minimum bias of $\hat{L}_\delta$ is attained at g_δ^* and f_δ^*, respectively, which yields $\overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) = -\underline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) = \frac{1}{2}(\omega(\delta) - \delta\omega'(\delta))$. Note that $\text{sd}(\hat{L}_\delta) = \omega'(\delta)$.

Corollary 3.1 in Armstrong and Kolesár (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 (26) with $\sum_{i=1}^n \frac{(f(x_i,d_i)-g(x_i,d_i))^2}{\sigma^2(x_i,d_i)} = \delta^2$, and let $\hat{c}_\delta^* = \hat{L}_\delta - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) - z_{1-\alpha}\,\text{sd}(\hat{L}_\delta)$. Then $[\hat{c}_\delta^*, \infty)$ is a $1-\alpha$ CI over $\mathcal{F}$, and it minimaxes the βth quantile of excess length among all $1-\alpha$ CIs for Lf, where $\beta = \Phi(\delta - z_{1-\alpha})$, and Φ denotes the standard normal cdf. (ii) Let δ_{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 (26) 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 (26) at $\delta = \delta_{RMSE}$. Then the estimator $\hat{L}_{\delta_{RMSE}}$ minimaxes RMSE among all affine estimators.

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

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

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

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

The one-dimensional subfamily argument used in Donoho (1994) to derive Theorem A.2 could also be used to obtain the minimax efficiency of the FLCI based on $\hat{L}_{\delta_{\mathrm{FLCI}}}$ 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 (26) with $\sum_{i=1}^n \frac{(f(x_i,d_i)-g(x_i,d_i))^2}{\sigma^2(x_i,d_i)} = \delta^2$, with $\delta = z_\beta + z_{1-\alpha}$, and define $\hat{c}^*_\delta$ as in Theorem A.1. Then the efficiency of $\hat{c}^*_\delta$ under the criterion $q_\beta(\cdot,\{g\})$ is given by*

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

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

$$\frac{\inf_{\{\mathcal{C}\colon \mathcal{C} \text{ satisfies } (4)\}} E_g\lambda(\mathcal{C})}{\inf_{\delta>0} 2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2}\right)\omega'(\delta)} = \frac{(1-\alpha)E\left[\omega(2(z_{1-\alpha}-Z)) \mid Z \leq z_{1-\alpha}\right]}{2\,\mathrm{cv}_\alpha\left(\frac{\omega(\delta_{FLCI})}{2\omega'(\delta_{FLCI})} - \frac{\delta_{FLCI}}{2}\right)\cdot\omega'(\delta_{FLCI})}$$
$$\geq \frac{z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha\Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha)}{z_{1-\alpha/2}}, \quad (27)$$

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 (27), 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). □

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 (27) 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 by[21] $\beta_d = \operatorname{argmin}_\beta \sum_{i=1}^n \sigma^{-2}(x_i, d_i)\mathbb{I}\{d_i = d\}(E[Y_i(d) \mid Z_i = z_i] - z_i'\beta)^2$. 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

$$\sum_{i=1}^{n} \mathbb{I}\{d_i = d\} \frac{z_i g(x_i, d)}{\sigma^2(x_i, d_i)} = 0, \quad d \in \{0, 1\}. \tag{28}$$

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

$$\mathcal{F}_{z,\text{Lip}}(C) = \{z'\beta_d + g(x, d) \colon g \in \mathcal{F}_{\text{Lip}}(C),\ g \text{ satisfies } (28),\ \beta_0, \beta_1 \in \mathbb{R}^p\}. \tag{29}$$

If z comprises just the intercept, then the constraint (28) 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 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

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

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 $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$ denote the set of functions $f \colon \{x_1, \dots, x_n\} \times \{0, 1\} \to \mathbb{R}$ such that $|f(x, d) - f(\tilde{x}, d)| \leq C\|x - \tilde{x}\|_{\mathcal{X}}$ for all $x, \tilde{x} \in \{x_1, \dots, x_n\}$ and each $d \in \{0, 1\}$. That is, $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$ 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 $f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(C)$. The following result shows that, given a function in $\widetilde{\mathcal{F}}_{\text{Lip},n}(C)$, one can always interpolate the points $x_1, \dots, x_n$ to obtain a function in $\mathcal{F}_{\text{Lip}}(C)$.

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

As a consequence, we obtain the following result:

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 (28) and subject to

$$g(x_i, d) - g(x_j, d) \leq C\|x_i - x_j\|_{\mathcal{X}}, \quad d \in \{0, 1\} \quad i, j \in \{1, \dots, n\}. \tag{30}$$

Lemma A.2 follows directly from Lemma A.1 and the fact that 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 $z = 1$ then yields the first part of Theorem 2.1 (where we use the observation above that in this case the constraint (28) 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 (30) 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 (30) holds with equality. Then (30) holds for all i, j. (ii) Suppose (30) 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 (30) holds with equality. Suppose also that for each j with $d_j = d$, there exists an i with $d_i = 1-d$ such that (30) holds with equality. Then (30) holds for all i, j.*

We'll now show that eq. (16) implies that (13) holds for $d = 1$ and all i, j. The argument that (13) also holds for $d = 0$ is analogous and omitted. Observe that if $k_i = w(1)$ for some i with $d_i = 1$, we can set $f(x_i, 1) = \min_{j\colon d_j=0}\{f(x_j, 1) + C\|x_i - x_j\|_{\mathcal{X}}\}$ without affecting the bias, so that we may suppose that eq. (16) holds for all i, j with $d_i = 1 - d_j = 1$. The result 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. (11) is given by the solution to the convex program*

$$\max_{g\in\mathbb{R}^{2n},\beta_0,\beta_1\in\mathbb{R}^{\dim(z_i)}} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^{n} \frac{g(x_i, d_i)^2 + (z_i'\beta_{d_i})^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4} \quad \text{and s.t. (30) and (28).} \tag{31}$$

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 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 $\operatorname{argmin}_{j\colon d_i \neq d_j} \|x_i - x_j\|_{\mathcal{X}}$ 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 $\hat{L}_\delta$ is given by the regression adjusted matching estimator with $M = 1$.*

The proof of Theorem A.5 is deferred to the end of this section. Theorem 2.3 follows directly from this theorem by setting $z = 1$.

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

We use a result on necessary and sufficient conditions for interpolation by monotone Lipschitz functions given by Beliakov (2005). For a vector $x \in \mathbb{R}^p$, let $x_{\mathcal{S}+}$ denote the vector where we replace elements in $\mathcal{S}$ that are negative with 0: $(x_{\mathcal{S}+})_j = \max\{x_j, 0\}$ if $j \in \mathcal{S}$ and $(x_{\mathcal{S}+})_j = x_j$ otherwise. Let $\widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S},n}(C)$ denote the set of functions $f : \{x_1, \ldots, x_n\} \times \{0, 1\} \to \mathbb{R}$ such that, for all $i, j \in \{1, \ldots, n\}$ and $d \in \{0, 1\}$

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

Lemma A.4. *(Beliakov, 2005, Proposition 4.1) Suppose that the norm $\|\cdot\|_{\mathcal{X}}$ satisfies $\|x\|_{\mathcal{X}} \geq \|x_{\mathcal{S}+}\|_{\mathcal{X}}$. Then for any function $f\colon \{x_1, \ldots, x_n\} \times \{0, 1\} \to \mathbb{R}$, we have $f \in \widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S},n}(C)$ if and only if there exists a function $h \in \mathcal{F}_{\text{Lip},\mathcal{S}}(C)$ such that $f(x, d) = h(x, d)$ for all $(x, d) \in \{x_1, \ldots, x_n\} \times \{0, 1\}$.*

The condition on the norm in Lemma A.4 holds for norms of the form in eq. (25) so long as A is diagonal. Using this result, the problem of computing the maximum bias of an affine estimator $\hat{L}_{k,a}$ that satisfies (14) can again be phrased as a finite-dimensional linear program of maximizing $a + \sum_{i=1}^n k(x_i, d_i) f(x_i, d_i) - Lf$ subject to $f \in \widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S},n}(C)$. The optimal estimator can be computed by solving (26) with $\mathcal{F} = \widetilde{\mathcal{F}}_{\text{Lip},\mathcal{S},n}(C)$, which is a finite-dimensional convex optimization problem.

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 (31) can be written as

$$\max_{\beta_0,\beta_1,m\in\mathbb{R}^n,r\in\mathbb{R}^n} 2Lf \quad \text{s.t.} \quad \sum_{i=1}^{n} \frac{m_i^2 + (z_i'\beta_{d_i})^2}{\sigma^2(x_i, d_i)} \leq \frac{\delta^2}{4} \tag{32}$$

subject to

$$Z_d' \Sigma_d^{-1} m_d = 0, \quad \text{for } d \in \{0, 1\} \tag{33}$$

and subject to (30). Now, note that if $w(d) = 0$, setting $r_d = 0$ and $m_d = 0$ is optimal and the constraints (30) hold trivially. If $w(d) > 0$, observe that

$$r_i - m_j \leq C\|x_i - x_j\|_{\mathcal{X}} \quad \text{for all } i, j \text{ with } d_i = 1 - d_j = d, \tag{34}$$

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 (31), we can replace the constraints in (30) with (34) and

$$m_i - m_j \leq C\|x_i - x_j\|_{\mathcal{X}} \quad \text{for all } i, j \text{ with } d_i = d_j = d. \tag{35}$$

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 (32) implies (35). For δ/C small enough, it thus suffices to maximize (32) subject to (33) and (34). This is a convex optimization problem and constraint qualification holds since $m = 0$ satisfies Slater's condition (see Boyd and Vandenberghe, 2004, p. 226). Thus, the solution (or set of solutions) is the same as the solution to the Lagrangian,

$$\begin{aligned} -Lf + \frac{\lambda}{2}\left(\sum_{d=0}^{1}(m_d'\Sigma_d^{-1}m_d + \beta_d' Z_d' \Sigma_d^{-1} Z_d \beta_d) - \frac{\delta^2}{4}\right) + \nu_0' Z_0' \Sigma_0^{-1} m_0 + \nu_1 Z_1' \Sigma_1^{-1} m_1 + \\ \sum_{i,j\colon d_i = 1 - d_j = 1} [\Lambda_{ij}^0 (r_i - m_j - C\|x_i - x_j\|_{\mathcal{X}}) + \Lambda_{ij}^1 (r_j - m_i - C\|x_i - x_j\|_{\mathcal{X}})], \end{aligned}$$

where ν_0, ν_1 are vectors of Lagrange multipliers associated with the constraints (33), and Λ^0, Λ^1 are matrices of Lagrange multipliers with dimension $n_1 \times n_0$ associated with the constraints (34). The first-order conditions are given by

$$\Sigma_d^{-1} m_d = \frac{1}{\lambda}(w(d)\iota_d + \Lambda^{d'}\iota_{1-d} - \Sigma_d^{-1} Z_d \nu_d) \tag{36}$$

$$w(d)\iota_d = \Lambda^{1-d}\iota_{1-d} \tag{37}$$

$$\beta_d = \frac{2d-1}{\lambda}(Z_d \Sigma_d^{-1} Z_d)^{-1}(w(1)Z_1'\iota_1 + w(0)Z_0'\iota_0), \tag{38}$$

for $d \in \{0, 1\}$. Furthermore, the constraint $Z_d'\Sigma_d^{-1} m_d = 0$, combined with (36) 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. (12), the optimal weights for the treated and untreated units, respectively, take the form

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

with the optimal estimator given by

$$\hat{L}_\delta = \sum_{i=1}^{n} (2d_i - 1)\left[w(d_i)(Y_i - z_i'\hat{\beta}_{1-d_i}) - \sum_j \Lambda_{ij}^{1-d_i}(Y_j - z_i'\hat{\beta}_{d_i}) \right].$$

The result then follows from the fact that if $w(d) = 0$, $\Lambda^{1-d} = 0$ at optimum by eq. (37) 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. (34). 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. (37), for this closest match j, $\Lambda_{ij}^{1-d} = w(d)$, which yields the result. □

A.3 Proof of Theorem 2.2

The dual problem to (18) 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

$$\min_{f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(C)} \frac{1}{2} \sum_{i=1}^{n} \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} - \mu Lf/C = \min_{f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)} \frac{C^2}{2} \sum_{i=1}^{n} \frac{f(x_i, d_i)^2}{\sigma^2(x_i, d_i)} - \mu Lf, \tag{39}$$

where we use the observation that if $f \in \widetilde{\mathcal{F}}_{\text{Lip},n}(C)$, then $f/C \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)$. Let g_μ^* denote the solution to the minimization problem on the right-hand side of (39). Because for each $\delta > 0$, the program (18) 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 $g \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)$ can be replaced with the constraint

$$r_j \leq m_i + \|x_i - x_j\|_{\mathcal{X}}, \quad d_i \neq d_j, \tag{40}$$

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. (40) 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 $g \in \widetilde{\mathcal{F}}_{\text{Lip},n}(1)$ with eq. (40) and

$$m_i \leq m_{i'} + \|x_i - x_{i'}\|_{\mathcal{X}}, \quad d_i = d_{i'}. \tag{41}$$

If we only impose the constraints in (40), the Lagrangian for the program (39) can be written as

$$\frac{1}{2} \sum_{i=1}^{n} \frac{m_i^2}{\sigma^2(d_i)} - \mu \left(\sum_{i=1}^{n} w(d_i)(m_i + r_i) \right) \\ + \sum_{i,j\colon d_i = 1 - d_j = 1} \left[\Lambda_{ij}^0 (r_i - m_j - \|x_i - x_j\|_{\mathcal{X}}) + \Lambda_{ij}^1 (r_j - m_i - \|x_i - x_j\|_{\mathcal{X}}) \right]. \tag{42}$$

This Lagrangian implies that (41) 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 $\Lambda^1_{ij}(r_j - m_i - \|x_i - x_j\|_{\mathcal{X}}) = 0$ then implies that $\sum_j \Lambda^1_{ij} = 0$, 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 (42). For $d \in \{0, 1\}$, let $N^d(\mu) \in \mathbb{R}^{n_1 \times n_0}$ denote a matrix with elements $N^d_{ij}(\mu) = 1$ if the constraint associated with $\Lambda^d_{ij}(\mu)$ is active, and $N^d_{ij}(\mu) = 0$ 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 $G^0_{jj'} = \mathbb{I}\{\sum_i N^0_{ij}(\mu) N^0_{ij'}(\mu) > 0\}$, and $G^1_{ii'} = \mathbb{I}\{\sum_j N^1_{ij}(\mu) N^1_{i'j}(\mu) > 0\}$. Then G^0 defines a graph (adjacency matrix) of a network in which j and j' are linked if the constraints associated with Λ^0_{ij} and $\Lambda^0_{ij'}$ are both active for some i. Similarly, G^1 defines a graph of a network in which i and i' are linked if the constraints associated with $\Lambda^1_{i'j}$ and Λ^1_{ij} are both active for some j. Let $\{\mathcal{M}^0_1, \dots, \mathcal{M}^0_{K_0}\}$ denote a partition of $\{1, \dots, n_0\}$ according to the connected components of G^0, so that if $j, j' \in \mathcal{M}^0_k$ then there exists a path from j to j'. Let $\{\mathcal{R}^0_1, \dots, \mathcal{R}^0_k\}$ be a corresponding partition of $\{1, \dots, n_1\}$, defined by $\mathcal{R}^0_k = \{i \in \{1, \dots, n_1\} \colon N^0_{ij}(\mu) = 1 \text{ for some } j \in \mathcal{M}^0_k\}$. Similarly, let $\{\mathcal{M}^1_1, \dots, \mathcal{M}^1_{K_1}\}$ denote a partition of $\{1, \dots, n_1\}$ according to the connected components of G^1, and let $\mathcal{R}^1_k = \{j \in \{1, \dots, n_0\} \colon N^1_{ij}(\mu) = 1 \text{ for some } i \in \mathcal{M}^1_k\}$.

In the supplemental materials, we show that the solution path for $m(\mu)$ is piecewise linear in μ, with points of non-differentiability when either a new constraint becomes active, or else the Lagrange multiplies $\Lambda^d_{ij}(\mu)$ 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 $D^d_{ij} = \|x_i - x_j\|_{\mathcal{X}}$, $d \in \{0, 1\}$, $d_i = 1 - d_j = 1$. Let r be a vector with elements $r_j = \min_{i=1,\dots,n_1}\{D^1_{ij}\}$, if $d_j = 0$, and $r_i = \min_{j=1,\dots,n_0}\{D^0_{ij}\}$, if $d_i = 1$. Let $N^0, N^1 \in \mathbb{R}^{n_1 \times n_0}$ be matrices with elements $N^0_{ij} = \mathbb{I}\{D^0_{ij} = r_i\}$ and $N^1_{ij} = \mathbb{I}\{D^1_{ij} = r_j\}$.

2. While $\mu < \infty$:

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

   (b) Calculate directions Δ^d for Λ^d by setting $\Delta^d_{ij} = 0$ if $N^d_{ij} = 0$, with the remaining elements given by a solution to the systems of n equations (i) $\sum_{i=1}^{n_0} \Delta^1_{ij} =$

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

(c) Calculate step size s as $s = \min\{s^0_1, s^0_2, s^1_1, s^1_2\}$, where

$$s^0_1 = \min\{s \geq 0 \colon r_i + \delta_{r,i} s = \delta_j s + D^0_{ij} \text{ some } (i,j) \text{ s.t. } N^0_{ij} = 0,\ \delta_j > \delta_{r,i}\}$$
$$s^1_1 = \min\{s \geq 0 \colon r_j + \delta_{r,j} s = \delta_i s + D^1_{ij} \text{ some } (i,j) \text{ s.t. } N^1_{ij} = 0,\ \delta_i > \delta_{rj}\}$$
$$s^d_2 = \min\{s \geq 0 \colon \Lambda^d_{ij} + s\Delta^d_{ij} = 0 \text{ among } (i,j) \text{ with } N^d_{ij} = 1 \text{ and } \Delta^d_{ij} < 0\}$$

(d) Update $\mu \mapsto \mu + s$, $m \mapsto m + s\delta$, $r \mapsto r + s\delta_r$ $\Lambda^d \mapsto \Lambda^d + s\Delta^d$, $D^0_{ij} \mapsto D^0_{ij} + s\delta_j$, $D^1_{ij} \mapsto D^1_{ij} + s\delta_i$ If $s = s^d_1$, then update $N^d_{ij} = 1$, where (i,j) is the index defining s^d_1. If $s = s^d_2$, update $N^d_{ij} = 0$, where (i,j) is the index defining s^d_2.

Given the solution path $\{m(\mu)\}_{\mu>0}$, the optimal estimator $\hat{L}_\delta$ and its worst-case bias can then be easily computed. For simplicity, we specialize to the 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 (42) (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 $\hat{f}_\mu(x_j, 1) = \sum_i n\Lambda^1_{ij}(\mu)/\mu Y_i$ 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 $\hat{f}_\mu(x_i, 0) = \sum_j n\Lambda^0_{i,j}(\mu)/\mu Y_j$ 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 $\hat{L}_\delta$ as a matching estimator with a variable number of matches, observe that $\sum_i n\Lambda^1_{ij}(\mu)/\mu = \sum_j n\Lambda^0_{ij}(\mu)/\mu = 1$. Also, $N^0_{ij}(\mu) = 0$ and hence $\Lambda^0_{ij}(\mu) = 0$ unless $D^0_{ij}(\mu) = \min_\ell D^0_{i\ell}(\mu)$. Similarly, $\Lambda^1_{ij}(\mu) = 0$ unless $D^1_{ij}(\mu) = \min_\ell D^1_{\ell j}(\mu)$. Thus, the counterfactual outcome for each observation i is given by a weighted average of outcomes for observations with opposite treatment status that are closest to it in terms of the "effective distance" matrices $D^0_{ik}(\mu)$ (if $d_i = 1$) or $D^1_{ki}(\mu)$ (if $d_i = 0$). Since $D^0_{ik}(\mu) = m_k(\mu) + \|x_i - x_k\|_{\mathcal{X}}$ $D^1_{ki}(\mu) = m_i(\mu) + \|x_i - x_k\|_{\mathcal{X}}$, 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.

Appendix B Proofs for asymptotic results

This appendix gives additional details and proofs for the results in Sections 3.3 and 4.

B.1 Proofs and details for Section 3.3

We now analyze the asymptotic coverage properties of the CI in (21) and its one-sided analog $[\hat{c}, \infty)$, where

$$\hat{c} = \hat{L}_k - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) - z_{1-\alpha}\, \text{se}_\tau(\hat{L}_k). \tag{43}$$

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 we 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

$$\liminf_n \inf_{P \in \mathcal{P}} P\left(\tau(P) \in \mathcal{C}\right) \geq 1 - \alpha. \tag{44}$$

To construct the class $\mathcal{P}$, we assume that $f_P \in \mathcal{F}$ for all $P \in \mathcal{P}$, where $\mathcal{F}$ is a convex and centrosymmetric regularity class for the conditional expectation function, such as the Lipschitz class that is the focus of most of this paper. 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 treat the class $\mathcal{F}$ as known, and our CI for the PATE will require knowledge of this class including any smoothness constants. However, we do not need to know the constants governing the regularity of u_i and X_i. The weighting function k in (5) may depend on the entire set $\{X_i, D_i\}_{i=1}^n$ of treatments and covariates, 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.

In the remainder of this appendix, we provide asymptotic coverage results for the CIs in eqs. (21) and (43). We first give a general result under high-level conditions in Theorem B.1 below. We then give primitive conditions for the matching estimator. Discussion of consistency of standard errors is deferred to the supplemental appendix.

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)*

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

uniformly over $P \in \mathcal{P}$.

Under this condition, the asymptotic coverage results follows from a CLT, which obtain via a martingale representation, following the martingale representation for matching estimators noted by Abadie and Imbens (2012).

Theorem B.1. *Suppose that Assumption B.1 holds, and that* $[V_{1,n}(P)+V_{2,n}(P)]/\operatorname{se}_\tau(\hat{L}_k)^2 \xrightarrow{p} 1$ *uniformly over* $P \in \mathcal{P}$*. In addition, suppose that for some* $\eta > 0$, $E_P|f_P(X_i, 0)|^{2+\eta} < 1/\eta$ *and* $E_P|f_P(X_i, 1)|^{2+\eta} < 1/\eta$ *for all* $P \in \mathcal{P}$*. Consider the confidence sets in eqs. (21) and (43). If the set* $\mathcal{F}$ *used to construct these confidence sets satisfies* $f_P \in \mathcal{F}$ *for all* $P \in \mathcal{P}$*, then these confidence sets satisfy the coverage criterion (44).*

Proof. Given a sequence P_n be a sequence 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 (21) is given by the intersection of two one-sided CIs, its asymptotic validity then follows immediately from Bonferroni's inequality.

For $i = 1, \ldots, n$, let $\xi_{n,i} = \frac{1}{n}(f(X_i, 1) - f(X_i, 0) - \tau(P_n))$ and let $\mathcal{H}_{n,i}$ be the sigma algebra generated by $D_1, \ldots, D_n$ and $X_1, \ldots, X_i$. For $i = n + 1, \ldots, 2n$, let $\xi_{n,i} = k(X_{i-n}, D_{i-n})u_{i-n}$, and let $\mathcal{H}_{n,i}$ be the sigma algebra generated by $\mathcal{H}_{n,n}$ and $u_1, \ldots, u_{i-n}$. Then $\sum_{i=1}^j \xi_{n,i}$ is a martingale with respect to the filtration $\mathcal{H}_{n,j}$. Let $s_n(P_n)^2 = s_n(P_n; \{X_i, D_i\}_{i=1}^n)^2 = \sum_{i=1}^{2n} E_{P_n}(\xi_{n,i}^2 \mid \mathcal{H}_{n,i-1}) = V_{2,n}(P_n) + \sum_{i=1}^n k(X_i, D_i)^2 \sigma_P^2(X_i, D_i)$ 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 $i = 1, \ldots, n$, $\tilde{\xi}_{n,i}^2 = (f_{P_n}(X_i, 1) - f(X_i, 0) - \tau(P_n))^2/(n^2\tilde{s}_n(P_n)^2) = W_i^2/[nE_{P_n}(W_i^2) + n^2V_{1,n}(P_n)] \leq W_i^2/[nE_{P_n}(W_i^2)] = \tilde{W}_i^2/n$ 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 $\varepsilon > 0$ $\sum_{i=1}^n \tilde{\xi}^2 \mathbb{I}\{\tilde{\xi}^2 > \varepsilon\} \leq E_{P_n}\tilde{W}_i^2 \mathbb{I}\{\tilde{W}_i^2 > n\}$ which converges to zero by the uniform $2+\eta$ moment bounds in part (i) of the assumptions. 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

$$\begin{aligned}
P_n(\tau(P_n) \notin [\hat{\underline{c}}_\alpha, \infty)) &= P_n\left(\tau(P_n) < \hat{L}_k - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_k) - z_{1-\alpha} \operatorname{se}_\tau(\hat{L}_k)\right) \\
&\leq P_n\left(\tau(P_n) < \sum_{i=1}^n k(X_i, D_i)u_i + \frac{1}{n}\sum_{i=1}^n (f(X_i, 1) - f(X_i, 0)) - z_{1-\alpha} \operatorname{se}_\tau(\hat{L}_k)\right) \\
&= P_n\left(z_{1-\alpha} \operatorname{se}_\tau(\hat{L}_k) < \sum_{i=1}^n k(X_i, D_i)u_i + \frac{1}{n}\sum_{i=1}^n (f(X_i, 1) - f(X_i, 0) - \tau(P_n))\right)
\end{aligned}$$

$$= P_n \left(z_{1-\alpha} < (\tilde{s}_n(P_n)/\operatorname{se}_\tau(\hat{L}_k)) \cdot \sum_{i=1}^{2n} \tilde{\xi}_{n,i} \right) \to \alpha,$$

where we use the assumption that $\tilde{s}_n(P)/\operatorname{se}_\tau(\hat{L}_k)$ converges to one uniformly over $P \in \mathcal{P}$. □

For matching estimators with a fixed number of matches we use results from Abadie and Imbens (2006) and Abadie and Imbens (2016) to verify Assumption B.1. Since such results appear to be available only for the case where X_i is scalar, we restrict ourselves to this case, and we leave the question of verifying Assumption B.1 when X_i is multivariate for future research. Since these results are stated for a single underlying distribution, we restrict attention to the case where the distribution of (X_i, D_i) is fixed over $P \in \mathcal{P}$ (but where the conditional expectation function f_P is allowed to vary over the given class $\mathcal{F}$).

Theorem B.2. *Suppose that the class $\mathcal{P}$ is such that the marginal distribution of (X_i, D_i) and the conditional variance function $\sigma_P^2(x,d)$ is the same for all $P \in \mathcal{P}$, and such that the following conditions hold: (i) X_i is scalar, and is supported on a compact interval $[a,b]$ with continuous density (ii) $\sigma_P^2(x,d)$ is continuous and uniformly bounded away from zero and infinity (iii) $0 < P(D_i = 1) < 1$ and letting $g(x|d)$ denote the density of X_i given D_i, $g(x|1)/g(x|0)$ is uniformly bounded from above and below away from zero on $[a,b]$. Suppose, in addition, that, for some η, $E_P(u_i^{2+\eta}|X_i = x, D_i = d) \leq 1/\eta$ for $d \in \{0,1\}$, all x and all $P \in \mathcal{P}$. Then Assumption B.1 holds for the weights $k(X_i, D_i) = \frac{1}{n}(2D_i - 1)\left(1 + \frac{K_M(i)}{M}\right)$ for the matching estimator with M matches.*

Proof. Part (i) of Assumption B.1 follows from Lemma S.11 in Abadie and Imbens (2016). The formula for $V_{1,n}(P)$ follows from this lemma as well, and is given by a constant times $1/n$ (where, under our assumptions, the constant is strictly positive and does not depend on P). Thus, to verify part (ii) of Assumption B.1, it suffices to show this condition with $V_{1,n}(P)$ replaced by $1/n$. To this end, note that replacing $V_{1,n}(P)$ with $1/n$ in this condition gives

$$n^2 E_P[k(X_i, D_i)^2 u_i^2 \mathbb{I}\{k(X_i, D_i)^2 u_i^2 > \varepsilon/n\}] = E_P[(1+K_M(i)/M)^2 u_i^2 \mathbb{I}\{(1+K_M(i))^2 u_i^2 > \varepsilon \cdot n\}].$$

This will converge to zero by the standard arguments showing that the Lyapunov condition implies the Lindeberg condition, so long as $E_P[(1+K_M(i)/M)^{2+\eta} u_i^{2+\eta}]$ is uniformly bounded. Indeed, the bound on the conditional $2+\eta$ moment of u_i implies that this is bounded by a constant times $E_P[(1 + K_M(i)/M)^{2+\eta}]$, which is bounded uniformly in i and n by Lemma S.8 in Abadie and Imbens (2016). □

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 $\mathcal{N}_{d,n} = \{i \colon D_i = d, i \in \{1,\dots,n\}\}$ and let

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

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

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

Proof. Let $A_n = \{x \in [a,b]^p | \text{there exists } j \text{ such that } D_j = 0 \text{ and } \|x - X_j\|_{\mathcal{X}} \le 2h\}$. Then $\#\mathcal{I}_n(h) = \sum_{i \in \mathcal{N}_{1,n}} [\mathbb{I}\{X_i \in [a,b]^p\} - \mathbb{I}\{X_i \in A_n\}]$. Note that, conditional on $\mathcal{E}$, the random variables $\mathbb{I}\{X_i \in A_n\}$ with $i \in \mathcal{N}_{1,n}$ are i.i.d. Bernoulli(ν_n) with $\nu_n = P(X_i \in A_n|\mathcal{E}) = \int \mathbb{I}\{x \in A_n\} f_{X|D}(x|1)\, dx \le K\lambda(A_n)$ where $f_{X|D}(x|1)$ is the conditional density of X_i given $D_i = 1$, λ is the Lebesgue measure and K is an upper bound on this density. Under the assumption that $\limsup_n h_n n^{1/p} \le \eta$, we have $\lambda(A_n) \le (4h_n)^p n \le 8^p \eta^p$ where the last inequality holds for large enough n. Thus, letting $\overline{\nu} = 8^p \eta^p K$, we can construct random variables Z_i for each $i \in \mathcal{N}_{1,n}$ that are i.i.d. Bernoulli($\overline{\nu}$) conditional on $\mathcal{E}$ such that $\mathbb{I}\{X_i \in A_n\} \le Z_i$. Applying the strong law of large numbers, it follows that

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n [f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,1) - f_{n,\{X_i,D_i\}_{i=1}^n}(X_i,0)]& \\ = -\frac{1}{n}\sum_{i=1}^n \sum_{\tilde{x}\in\tilde{\mathcal{X}}_n(h,\eta)} k((X_i-\tilde{x})/h) &\leq -\underline{k}\frac{\#\mathcal{I}_n(h)}{n}, \qquad (45)\end{aligned}$$

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

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

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

which implies that $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n} \in \Sigma(C,\gamma)$ where $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}(x,d) = \frac{h^\gamma C}{K/\eta^p} f_{n,\{X_i,D_i\}_{i=1}^n}(x,d)$. By (45), the CATE under $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}$ is bounded from above by $-\underline{k}\frac{h^\gamma C}{K/\eta^p}\frac{\#\mathcal{I}_n(h)}{n}$, which, by Lemma B.1, is bounded from above by a constant times h_n^γ for large enough n on a probability one event for h_n a small enough multiple of $n^{-1/p}$. Thus, there exists $\varepsilon > 0$ such that the CATE under $\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}$ is bounded from above by $-\varepsilon n^{-1/p}$ for large enough n with probability one. On this probability one event,

$$\liminf_n P_0\left(\hat{c}_n \leq -\varepsilon n^{-\gamma}|\{X_i,D_i\}_{i=1}^n\right) = \liminf_n P_{\tilde{f}_{n,\{X_i,D_i\}_{i=1}^n}}\left(\hat{c}_n \leq \varepsilon n^{-\gamma}|\{X_i,D_i\}_{i=1}^n\right)$$

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

which gives the result.

B.3 Proofs of Theorems 4.2 and 4.3

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

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

Theorems 4.2 and 4.3 follow from verifying the high level conditions of Theorem F.1 in Armstrong and Kolesár (2018b). In particular, we need to show that the weights k ($\tilde{k}_\delta^*$ for Theorem 4.2 and $k_{\text{match},M}$ for Theorem 4.3) are such that $\sum_{i=1}^n k(x_i,d_i)u_i/\operatorname{sd}_k$ converges in distribution to $N(0,1)$ (condition (S13) in Armstrong and Kolesár, 2018b) and $\sum_i \hat{u}_i^2 k(x_i,d_i)^2/\operatorname{sd}_k^2$ converges in probability to 1, uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$ (S14), where $\operatorname{sd}_k^2 = \sum_{i=1}^n \sigma^2(x_i,d_i)k(x_i,d_i)$. We claim that both (S13) and (S14) hold if the weights satisfy (23).

Under the moment bounds on u_i, eq. (23) directly implies the Lindeberg condition that is needed for condition (S13) to hold. To show that it also implies (S14), note that (S14) is equivalent to the requirement that $\sum_{i=1}^n \hat{u}_i^2 a_{ni} - \sum_{i=1}^n \sigma^2(x_i,n_i)a_{ni}$ converges to zero uniformly over $f \in \mathcal{F}_{\text{Lip}}(C_n)$, where

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

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

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

$$\leq \max_{1\leq i\leq n} a_{ni}^{1/(2K)} E|u_i^2 - \sigma^2(x_i, d_i)|^{1+1/(2K)} \cdot \sum_{i=1}^n a_{ni}.$$

Note that, by boundedness of $\sigma(x, d)$ away from zero and infinity, $\sum_{i=1}^n a_{ni}$ is uniformly bounded. Furthermore, it follows from (23), that $\max_{1\leq i\leq n} a_{ni} \to 0$. From this and the moment bounds on u_i, it follows that the above display converges to zero. It therefore suffices to show that $\sum_{i=1}^n(\hat{u}_i^2 - u_i^2)a_{ni}$ converges to zero. For the nearest-neighbor variance estimator, this follows from Lemma B.2. We therefore just need to show that this holds for the Nadaraya-Watson estimator with uniform kernel and bandwidth h_n. Denote this estimator by $\hat{u}_i^2 = (Y_i - \hat{f}(x_i, d_i))^2$ where $\hat{f}(x_i, d_i) = \sum_{j\in\mathcal{N}_i} Y_i/\#\mathcal{N}_i$ and $\mathcal{N}_i = \{j \in \{1, \dots, n\} : \|x_j - x_i\|_{\mathcal{X}} \leq h_n,\ d_i = d_j\}$. Write

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

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

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

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

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

it follows that

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

If condition (22) holds for all $\eta > 0$, then the same condition also holds with η replaced by a sequence η_n converging to zero. It follows that, under this condition, there exists a bandwidth sequence h_n with $h_n C_n \to 0$ such that $\min_{1\leq i\leq n} \#\mathcal{N}_i \to \infty$, so that under this bandwidth sequence, the above display converges to zero.

Proof of Theorem 4.2. We need to verify that (23) holds for the weights $\tilde{k}_\delta^*$. By boundedness of $\tilde{\sigma}(x_i, d_i)$ away from zero and infinity, (23) is equivalent to showing that

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

where $\tilde{f}^*_\delta$ is the solution to the optimization problem defined by (11) and (13) 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 (11) binds, the denominator is bounded from above and below by constants that depend only on δ and the upper and lower bounds on $\tilde{\sigma}^2(x_i, d_i)$. Thus, it suffices to show that

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

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

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

infinitely often. This gives a contradiction so long as (22) holds. This completes the proof of Theorem 4.2. □

Proof of Theorem 4.3. We need to show that (23) holds for the weights $k_{\text{match},M}(x_i, d_i) = (1 + K_M(i))/n$. For this, it is sufficient to show that $\max_{1\le i\le n} K_M(i)^2/n \to 0$. To this end, let $U_M(x, d) = \|x_j - x\|_{\mathcal{X}}$ where x_j is the Mth closest observation to x among observations i with $d_i = d$, so that $K_M(i) = \#\{j : d_j \ne d_i, \|x_j - x_i\|_{\mathcal{X}} \le U_M(x_j, d_i)\}$. When (24) holds and n is large enough so that $n\underline{G}(a_n) \ge M$, we will have $U_M(x, d) \le a_n$ for all $x \in \mathcal{X}$. By definition of $K_M(i)$, the upper bound in (24) then implies $K_M(i) \le n\overline{G}(a_n)$. Thus, it suffices to show that $[n\overline{G}(a_n)]^2/n = n\overline{G}(a_n)^2 \to 0$.

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

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

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

B.4 Proof of Lemma 4.1

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

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

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

B.5 Proof of Theorem 4.4

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

$$1\le\frac{\overline{\text{bias}}^2_{\text{match},1}+\text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta\text{RMSE},n}+\text{sd}^2_{\delta\text{RMSE},n}}\le\frac{\overline{\text{bias}}^2_{\delta\text{RMSE},n}+\text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta\text{RMSE},n}+\text{sd}^2_{\delta\text{RMSE},n}}\le1+\frac{\text{sd}^2_{\text{match},1}}{\overline{\text{bias}}^2_{\delta\text{RMSE},n}+\text{sd}^2_{\delta\text{RMSE},n}}.$$

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

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

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

References

Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometica*, 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. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Économie et de Statistique*, (91/92):175–187.

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. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

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. ArXiv: 1808.07387.

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. (2005). Monotonicity preserving approximation of multivariate scattered data. *BIT Numerical Mathematics*, 45(4):653–677.

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.

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

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 Tarozzi, 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 Schargrodsky, 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.

Heckman, N. E. (1988). Minimax estimates in a semiparametric model. *Journal of the American Statistical Association*, 83(404):1090–1096.

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.

Imbens, G. and Wager, S. (2019). Optimized regression discontinuity designs. *Review of Economics and Statistics*, 101(2):264–278.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.

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.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

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.

Table 1: Diagonal elements of the weight matrix A in definition of the norm (25) for the main specification, A_{main}, and alternative specification, A_{ne}.

						Earnings		Employed	
	Age	Educ.	Black	Hispanic	Married	1974	1975	1974	1975
A_{main}	0.15	0.60	2.50	2.50	2.50	0.50	0.50	0.10	0.10
A_{ne}	0.10	0.33	2.20	5.49	2.60	0.07	0.07	2.98	2.93

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

	Feasible estimator $\tilde{L}_\delta$		Matching estimator	
	(1)	(2)	(3)	(4)
Criterion	RMSE	CI length	RMSE	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	1.64	1.78	1.48	2.21
Rob. std. error	1.01	0.94	1.09	0.89
Critical value ($\text{cv}_{0.05}$)	3.26	3.55	3.01	4.13
95% conf. interval	$(-2.33, 4.26)$	$(-2.38, 4.27)$	$(-1.88, 4.66)$	$(2.41, 4.93)$
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 smoothing parameter chosen to optimize the criterion listed under "Criterion". In columns (1) and (3), δ and M are chosen to optimize the RMSE of the estimator, and in columns (2) and (4), they are chosen to optimize the length of the CI for the CATT. Lind(k) corresponds to the maximum Lindeberg weight given in eq. (23).

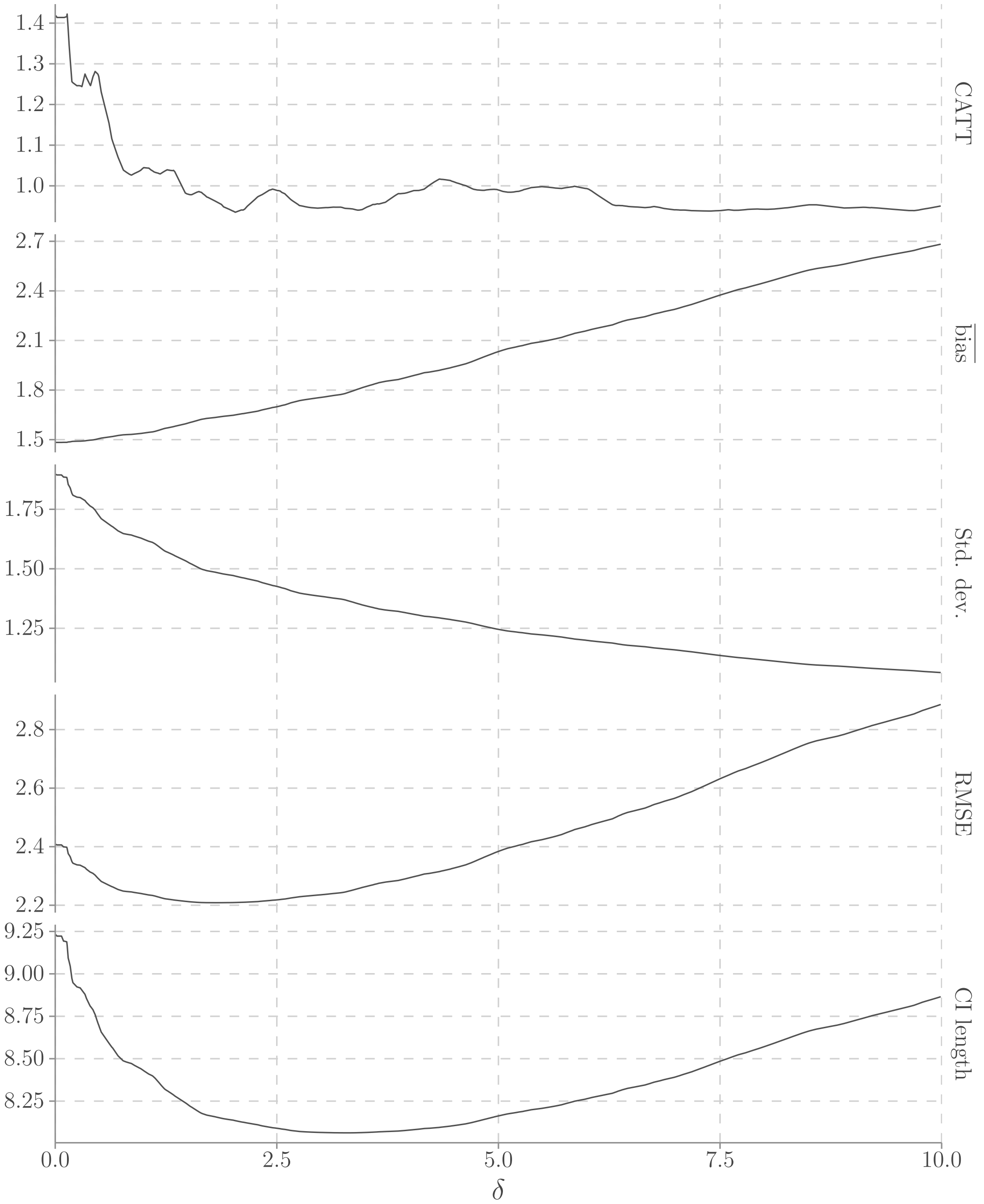

Figure 1: NSW application: performance of the feasible estimator $\tilde{L}_\delta$ of the CATT as a function of the smoothing parameter δ. CATT gives the value of the point estimate of the CATT, and $\overline{\text{bias}}$ gives the worst-case bias. The standard deviation, RMSE and CI length are computed under the assumption of homoskedasticity, so that the standard deviation is given by the square root of $\hat{\sigma}^2 \sum_{i=1}^{n} \tilde{k}^*_\delta(x_i, d_i)^2$.

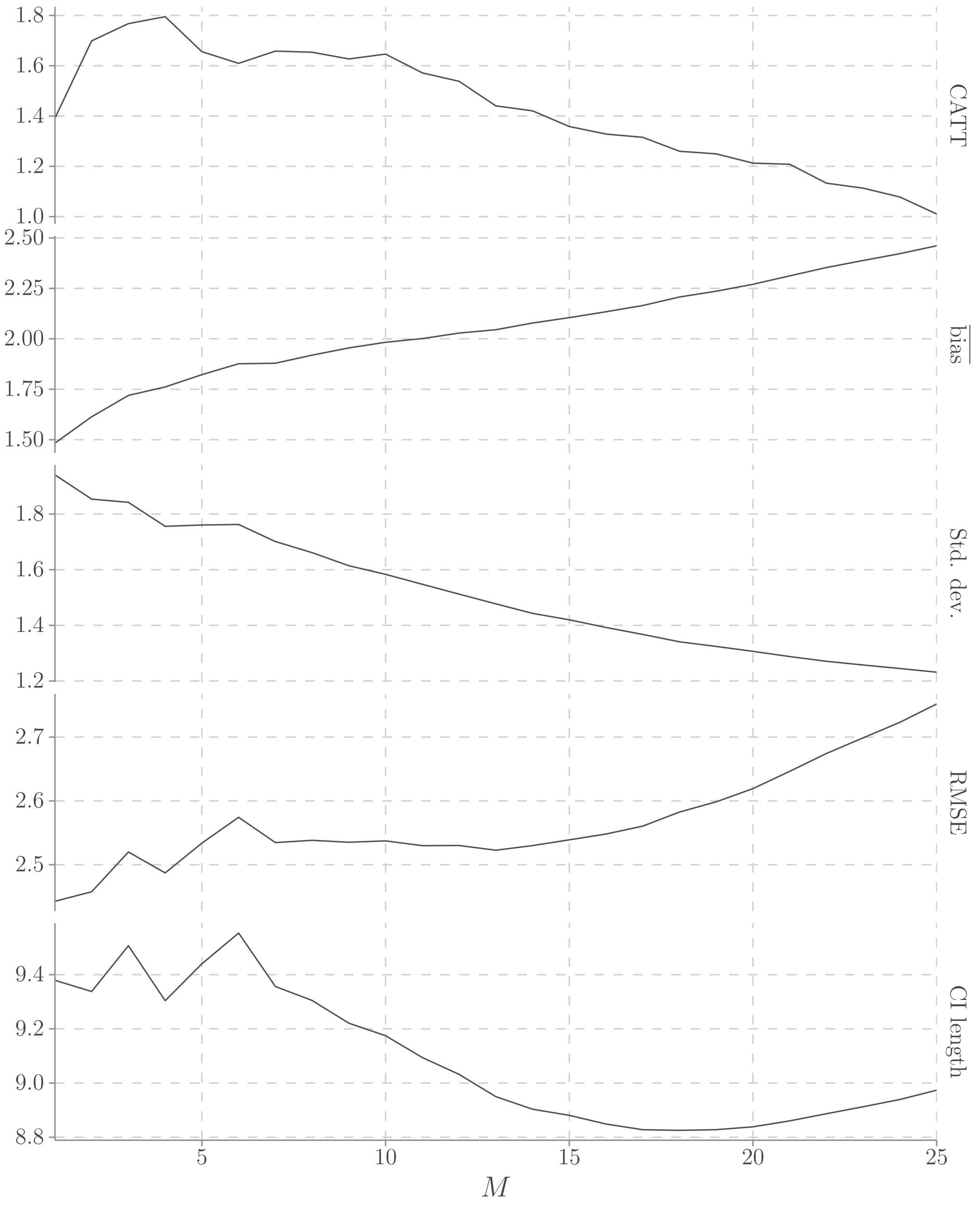

Figure 2: NSW application: performance of the matching estimator of the CATT as a function of the number of matches M. CATT gives the value of the point estimate of the CATT, and $\overline{\text{bias}}$ gives the worst-case bias. The standard deviation, RMSE and CI length are computed under the assumption of homoskedasticity, so that the standard deviation is given by the square root of $\hat{\sigma}^2 \sum_{i=1}^{n} k_{\text{match},M}(x_i, d_i)^2$.

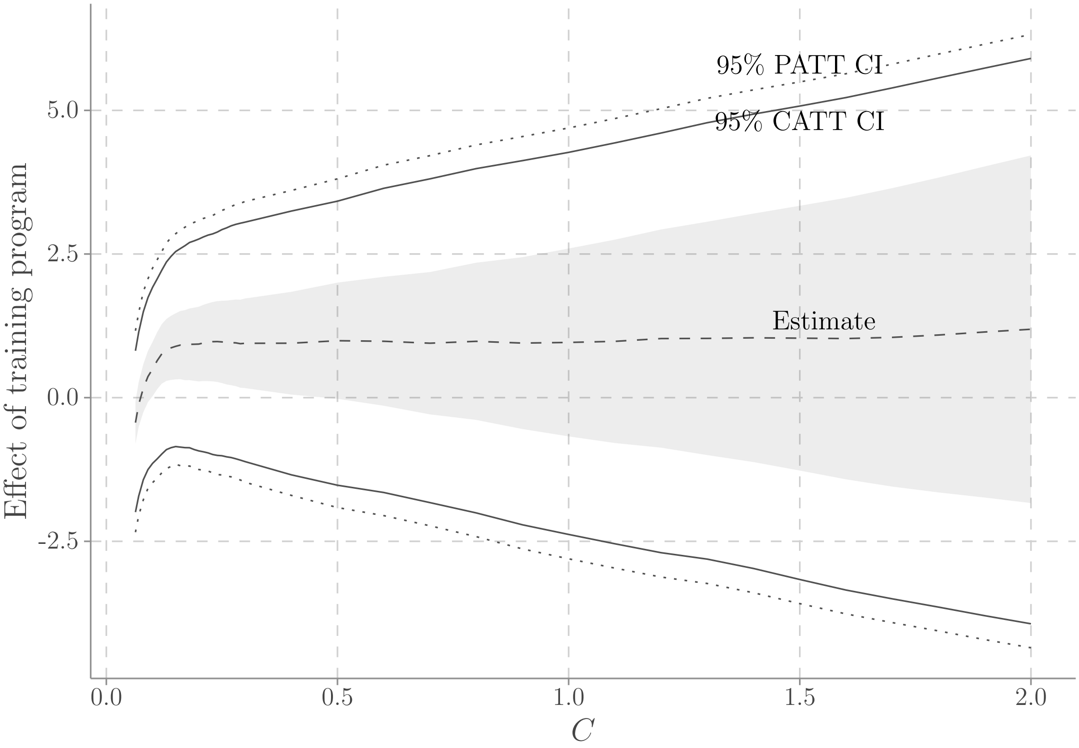

Figure 3: Feasible estimator and CIs for CATT and PATT in NSW application as a function of the Lipschitz constant C. Dashed line corresponds to point estimator $\tilde{L}_{\delta_{\mathrm{RMSE}}}$, shaded region denotes the estimator ± its worst-case bias. Solid lines give 95% robust CIs for the CATT based on the estimator $\tilde{L}_{\delta_{\mathrm{FLCI}}}$, and dotted lines give CIs for the PATT.