The Annals of Statistics
2022, Vol. 50, No. 5, 2587–2615
https://doi.org/10.1214/22-AOS2195

BOUNDS ON THE CONDITIONAL AND AVERAGE TREATMENT EFFECT WITH UNOBSERVED CONFOUNDING FACTORS

BY STEVE YADLOWSKY[1,a], HONGSEOK NAMKOONG[2,b], SANJAY BASU[3,c], JOHN DUCHI[4,d] AND LU TIAN[5,e]

[1]*Google Research, Brain Team,* [a]*yadlowsky@google.com*
[2]*Decision, Risk, and Operations Division, Columbia Business School,* [b]*namkoong@gsb.columbia.edu*
[3]*Research and Development, Waymark,* [c]*sanjay.basu@waymarkcare.com*
[4]*Statistics and Electrical Engineering, Stanford University,* [d]*jduchi@stanford.edu*
[5]*Biomedical Data Science, Stanford University,* [e]*lutian@stanford.edu*

For observational studies, we study the sensitivity of causal inference when treatment assignments may depend on unobserved confounders. We develop a loss minimization approach for estimating bounds on the conditional average treatment effect (CATE) when unobserved confounders have a bounded effect on the odds ratio of treatment selection. Our approach is scalable and allows flexible use of model classes in estimation, including nonparametric and black-box machine learning methods. Based on these bounds for the CATE, we propose a sensitivity analysis for the average treatment effect (ATE). Our semiparametric estimator extends/bounds the augmented inverse propensity weighted (AIPW) estimator for the ATE under bounded unobserved confounding. By constructing a Neyman orthogonal score, our estimator of the bound for the ATE is a regular root-n estimator so long as the nuisance parameters are estimated at the $o_p(n^{-1/4})$ rate. We complement our methodology with optimality results showing that our proposed bounds are tight in certain cases. We demonstrate our method on simulated and real data examples, and show accurate coverage of our confidence intervals in practical finite sample regimes with rich covariate information.

1. Introduction. Consider a causal inference problem with treatment indicator $Z \in \{0, 1\}$ representing control or intervention, potential outcomes $Y(1) \in \mathbb{R}$ under intervention and $Y(0) \in \mathbb{R}$ under control, and a set of observed covariates $X \in \mathcal{X} \subseteq \mathbb{R}^d$. Our interest is in studying confounding bias in estimators of the *conditional average treatment effect* (CATE)

$$\tau(x) := \mathbb{E}\big[Y(1) - Y(0) \mid X = x\big],$$

and estimation and inference of the *average treatment effect* (ATE)

$$\tau := \mathbb{E}\big[Y(1) - Y(0)\big]$$

based on n i.i.d. observations $\{Y = Y(Z), Z, X\}$.[1] Many methods provide consistent estimators for the ATE [26] under the independence assumption

$$\big\{Y(1), Y(0)\big\} \perp\!\!\!\perp Z \mid X, \tag{1}$$

that all confounding factors are observed or equivalently, that observed covariates X account for all dependence between the potential outcomes and treatment assignments. Estimation of the CATE, $\tau(x)$, under the independence assumption (1) has recently generated substantial interest [3, 22, 30, 38, 55].

Received June 2020; revised February 2022.
MSC2020 subject classifications. Primary 62H12; secondary 62H15, 62F03, 62F30.
Key words and phrases. Causal estimation, semiparametric inference, sensitivity analysis, omitted variable.
[1]Together, these imply the stable unit treatment value assumption, which will be assumed throughout.

Confounding bias is ubiquitous in observational studies, and the assumption (1) is frequently too restrictive: in practice, there is almost always an unobserved confounding factor $U \in \mathcal{U}$ affecting both treatment selection and outcome. Consequently, we consider an unobserved confounding factor U such that

$$\{Y(1), Y(0)\} \perp\!\!\!\perp Z \mid X, U. \tag{2}$$

This allows there to be a common cause U of the treatment Z and potential outcomes $\{Y(0), Y(1)\}$ that contains the relevant information about the potential outcomes that influence the treatment assignment. More abstractly, it allows the treatment assignment Z to depend directly on the unobserved potential outcome; a multivariate unobserved confounder U satisfying condition (2) always exists by letting $U = (Y(1), Y(0))$. Under this assumption, neither the ATE τ nor the CATE $\tau(x)$ is identifiable, and traditional estimators can be arbitrarily biased [25, 41, 43]. Yet it may be plausible that there is not "too much" confounding, so it is interesting to provide bounds on the possible range of treatment effects under such scenarios. We take this approach to propose a sensitivity analysis linking the posited strength of unobserved confounding to the range of possible values of the ATE τ and CATE $\tau(x)$.

We consider unobserved confounders that have bounded influence on the odds of treatment assignment, following Rosenbaum's ideas [43].

DEFINITION 1. A distribution P over $\{Y(1), Y(0), X, U, Z\}$ satisfies the Γ-*selection bias* condition with $1 \le \Gamma < \infty$ if for P-almost all $u, \tilde{u} \in \mathcal{U}$ and $X \in \mathcal{X}$,

$$\frac{1}{\Gamma} \le \frac{P(Z=1 \mid X, U=u)}{P(Z=0 \mid X, U=u)} \frac{P(Z=0 \mid X, U=\tilde{u})}{P(Z=1 \mid X, U=\tilde{u})} \le \Gamma. \tag{3}$$

Condition (3) limits departures from the independence assumption (1), and is equivalent to a regression model for the treatment selection probability ([43], Proposition 12), where the log odds ratio for treatment is

$$\log \frac{P(Z=1 \mid X, U)}{P(Z=0 \mid X, U)} = \kappa(X) + \log(\Gamma) b(U, X), \tag{4}$$

for some function $\kappa : \mathcal{X} \to \mathbb{R}$ of observed covariates X and a bounded function $b : \mathcal{U} \times \mathcal{X} \to [0, 1]$ of the unobserved, and observed confounders, U and X, respectively. Such odds ratios are common, for example, in medicine, where they reflect associations between risk factors and outcomes [39]. Practice requires choosing a realistic value of Γ to interpret the sensitivity analysis; we discuss this in more detail in Section 5. One common approach by practitioners is to look at the level of Γ when bounds on the ATE τ crosses a certain level of interest (e.g., 0), which measures the robustness of the findings to unobserved confounding [43], and then consider how plausible that choice of Γ would be for the data generating process.

The ATE τ, and CATE $\tau(x)$, are partially identified under the Γ-selection bias condition (3), so we focus instead on estimating bounds for them. This perspective on sensitivity to unobserved confounding traces to Cornfield et al.'s analysis demonstrating that if an unmeasured hormone can explain the observed association between smoking and lung cancer, it would need to increase the probability of smoking by nine-fold (an unrealistic amount) [15]. Contemporary medical informatics and epidemiological studies focusing on small effect sizes require a more nuanced approach for estimating the causal effect in the presence of unobserved confounding than the simple one used by Cornfield et al. [15]. For example, observational data is often used for post-market drug surveillance, but Bosco et al. [6] shows that unobserved confounding presents a particularly high risk in these data, motivating the need for sensitivity analysis to contextualize findings. Coloma et al. [14] show that effect sizes are often small, as adverse events for approved drugs are relatively rare. Therefore,

to draw confident and precise conclusions when there is mild confounding, it is important to avoid applying an overly conservative sensitivity analysis. Motivated to provide the most precise conclusions possible in the presence of confounding, we seek methods that provide optimal (tight) bounds on the CATE and ATE under the Γ-selection bias condition (3).

1.1. *Bounding treatment effects*. In what follows, we bound the confounding bias using analogues of the plug-in treatment contrast estimator for the the CATE, and the augmented inverse probability weighted (AIPW) estimator for the ATE [5]. We treat each potential outcome separately, focusing on lower bounds on $\mu_1 = \mathbb{E}[Y(1)]$ (other cases are symmetric). Based on observed data, all parameters necessary to estimate μ_1 can be nonparametrically identified, except the conditional mean of the unobserved potential outcome, $\mathbb{E}[Y(1) \mid X, Z = 0]$. Since this quantity is not identifiable in the presence of unobserved confounding, we develop a worst-case bound under the Γ-selection bias condition (3), and develop estimators based on the observed data. Specifically, let

$$\theta_1(x) := \inf\big\{\mathbb{E}_Q\big[Y(1) \mid X = x, Z = 0\big] : Q \in \mathcal{Q}_x\big\}, \tag{5}$$

where $\mathcal{Q}_x$ is the set of all distributions for $(Y(0), Y(1), Z)$ conditional on $X = x$ satisfying the independence assumption (2) and the bound (3) for $X = x$, and matching the conditional distributions that are identified in the observed data P: $Q(Z = 1 \mid X) = P(Z = 1 \mid X)$ and $Q(Y(1) \in \cdot \mid Z = 1, X) = P(Y(1) \in \cdot \mid Z = 1, X)$. By definition, $\theta_1(x) \leq \mathbb{E}_P[Y(1) \mid X = x, Z = 0]$ under the bounded unobserved confounding (Γ-selection bias condition (3)). Lower bounds on $\mathbb{E}[Y(1) \mid X = x]$ and $\mathbb{E}[Y(1)]$ follow from plugging in $\theta_1(x)$ in place of the unknown $\mathbb{E}_P[Y(1) \mid X = x, Z = 0]$.

Our first main result (Section 2) shows that $\theta_1(x)$ can be expressed as the solution to the loss minimization problem with a reweighted squared loss

$$\underset{\theta(\cdot)}{\text{minimize}}\, \frac{1}{2}\mathbb{E}\big[\big(Y(1) - \theta(X)\big)_+^2 + \Gamma\big(Y(1) - \theta(X)\big)_-^2 \mid Z = 1\big],$$

where $a_+ = a\mathbf{1}\{a > 0\}$, $a_- = -a\mathbf{1}\{a < 0\}$, $a \in \mathbb{R}$ and $\mathbf{1}\{\cdot\}$ is the indicator function. The scalable loss minimization approach allows us to use flexible model classes to estimate the lower bound, including many nonparametric and machine learning methods. Intuitively, the preceding display upweights the penalty for negative residuals, therefore increasing the impact of smaller observed outcomes on the minimizer $\theta_1(x)$, correcting for the fact that selection bias from confounding may have decreased the frequency of smaller observed outcomes.

Our second main result defines a semiparametric estimator (29) for the lower bound on the expected outcome $Y(1)$ under the Γ-selection bias condition (3):

$$\mu_1^- := \mathbb{E}\big[ZY(1) + (1 - Z)\theta_1(X)\big] \leq \mathbb{E}\big[Y(1)\big]. \tag{6}$$

Our estimation approach (Section 3) builds out of a line of work [5, 13] for statistical inference on τ when all confounders are observed (1); we adapt Chernozhukov et al.'s [13] cross-fitting procedure to allow large model classes to estimate nuisance parameters. Our semiparametric estimator satisfies *Neyman orthogonality* [37], and is insensitive to estimation errors in nuisance parameters. By virtue of this orthogonality, our estimator is root-n consistent and asymptotically normal so long as the nuisance parameters are estimated at a slower-than-parametric $o_p(n^{-1/4})$ rate of convergence. Our result gives asymptotically exact confidence intervals (CIs) for the lower bound μ_1^- (6).

Coupling the asymptotic distribution for $\hat{\mu}_1^-$ with the symmetrically defined upper and lower bounds $\hat{\mu}_z^{\pm}$ for $\mathbb{E}[Y(z)]$, we can construct a CI for the ATE τ under the Γ-selection bias condition (3). In general, the boundary of our interval never shrinks to τ even in the

large sample limit due to unobserved confounding. However, when there is no unmeasured confounding ($\Gamma = 1$), our method is equivalent to the AIPW estimator for the ATE τ.

Our population-level bound is unimprovable for bounding each expected potential outcome and their conditional counterparts, $\mathbb{E}[Y(z)]$ and $\mathbb{E}[Y(z) \mid X = x]$, $z \in \{0, 1\}$, but may not be always optimal in bounding their difference, the ATE $\tau = \mathbb{E}[Y(1) - Y(0)]$. On the other hand, when the potential outcomes are symmetric in the sense that $Y(0) \overset{d}{=} C(1 - Y(1))$ for some constant C, then our bounds on the treatment effect are also unimprovable (Section 3.4), thereby guaranteeing that our CI converges (in the large sample limit) to the smallest possible interval containing τ under the Γ-selection bias condition (3).

Finally, we supplement our theoretical analysis with an experimental investigation of the proposed approaches in Section 4. On both simulated and real-world data, we show that the CIs have good coverage and reasonable length.

1.2. *Related work.* The semiparametric literature [5, 13] have shown that the augmented inverse probability weighted (AIPW) estimator allows the use of flexible nonparametric and machine learning models to estimate the nuisance parameters: conditional means $\mathbb{E}[Y(z) \mid X]$, $z \in \{0, 1\}$, and the propensity score $\mathbb{P}(Z = 1 \mid X)$. By exploiting certain orthogonality properties, Chernozhukov et al. [13] showed how to obtain root-n consistency and asymptotic normality for estimated τ even when involved estimates of the nuisance parameters converge at slower nonparametric rates. We generalize this approach under the Γ-selection bias condition.

A number of authors have studied nonparametric and semiparametric models for sensitivity analysis. These works consider alternatives to our choice of model (3) in characterizing the relationship between unobserved confounders, treatment and outcomes [8, 18, 40, 41, 50, 58]. We focus on the model of Rosenbaum [43] because of its appealing interpretation as a regression model (4).

Imbens [24] derived a sensitivity analysis for the treatment effect in the presence of unobserved confounding. His approach requires specifying parametric models for the effect of an unobserved confounder on both the treatment selection and outcome. Specifically, the relationship between the unmeasured confounder and treatment assignment is modeled via a logistic regression, which is a special case of condition (3).

Aronow and Lee [2] and Miratrix, Wager and Zubizarreta [33] study the bias due to unknown selection probabilities in survey analysis, with an approach similar to ours. In the survey setting, only surveyed individuals provide covariates X, so the papers [2, 33] consider a simplified model for selection bias,

$$\frac{1}{\Gamma} \leq \frac{P(Z = 1 \mid U = u)}{P(Z = 0 \mid U = u)} \frac{P(Z = 0 \mid U = \tilde{u})}{P(Z = 1 \mid U = \tilde{u})} \leq \Gamma. \tag{7}$$

Zhao, Small and Bhattacharya [58] and Shen et al. [50] consider the sensitivity of inverse probability weighted estimates of the ATE τ to unobserved confounding by varying the propensity score estimates around their estimated values. Zhao, Small and Bhattacharya [58] discuss the relationship between their model of bounded unobserved confounding—which they call the marginal sensitivity model—and that based on the Γ-selection bias (3). Compared to our semiparametric estimator, the complexity of the asymptotic distribution of their estimator necessitates using a bootstrap method for inference. A interesting future direction is to extend the methods in this paper to improve statistical inference under their model.

The most common approach to sensitivity analysis for the ATE under condition (3) is to use matched observations [17, 43, 45–47]. Unfortunately, exactly matched pairs rarely exist in practice, even for covariate vectors of moderate dimension; when considering continuous covariates, the probability of finding exactly matched pairs is zero. Abadie and Imbens [1]

show that under appropriate regularity conditions on the functions $\mu_z(x)$ and $e_1(x)$ (defined in equations (8) and (11)), estimators of τ using approximately matched pairs can have a bias of order $\Omega(n^{-1/d})$ for d-dimensional continuous covariates. For these data, the AIPW method is a more appropriate statistical analysis tool. The AIPW estimator and other semiparametric methods can provide $\sqrt{n}$-consistent estimates of the ATE without unmeasured confounding [13, 21, 25, 48]. The semiparametric approach for the lower bound on the ATE that we present in Section 3 is $\sqrt{n}$-consistent under analogous regularity conditions. Therefore, when analyzing an observational study using the AIPW estimator, one should perform a sensitivity analysis using the semiparametric method we provide here. When finding good matched pairs is feasible, many analysts prefer matching due to the transparency of the results and the simplicity of confounding adjustment. If analyzing an observational study using matching methods, it would be natural to also use a matching-based method for sensitivity analysis, such as the ones described above. In summary, our proposed method and matching based sensitivity analysis approaches can be coupled with different main analyses in practice, and are complementary to each other.

Most work [3, 22, 30, 38, 55] directly studies estimation of the CATE $\tau(x) = \mu_1(x) - \mu_0(x)$ assuming that all confounders are observed. More recently, Kallus and Zhou [28] present an approach to learning personalized decision policy in the presence of unobserved confounding, and a contemporaneous work with this paper [27] derive bounds on the CATE; their methods are based on the marginal sensitivity model of Zhao, Small and Bhattacharya [58].

Notation. We use $\mathbb{P}_n$ and $\mathbb{P}_n(\cdot \mid Z = z)$ to represent the empirical probabilities of $\{(Y_i(Z_i), X_i, Z_i)\}_{i=1}^n$ and $\{(Y_i(Z_i), X_i) \mid Z_i = z\}$, respectively, and $\mathbb{E}_n[\cdot \mid Z = z]$ is the expectation with respect to $\mathbb{P}_n(\cdot \mid Z = z)$ for $z = 0, 1$. We let $n_z = \sum_{i=1}^n \mathbf{1}\{Z_i = z\}$ be the count of observations with $Z_i = z$, where $\mathbf{1}\{\cdot\}$ is the indicator function. For a distribution P and function $f : \mathcal{X} \to \mathbb{R}$, we use $\|f\|_{2,P} = (\int_{\mathcal{X}} f^2(x)\,\mathrm{d}P(x))^{1/2}$. For functions $f : \Omega \to \mathbb{R}$ and $g : \Omega \to \mathbb{R}$ with arbitrary domain Ω, we write $f \lesssim g$ if there exists constant $C < \infty$ such that $f(t) \le Cg(t)$ for all $t \in \Omega$, and $f \asymp g$ if $g \lesssim f \lesssim g$. We use P_z and $\mathbb{E}_z$ to denote the conditional distribution $P(\cdot \mid Z = z)$ and associated expectation, respectively. We write $\mathbb{E}_Q$ for the expectation under the probability Q, and omit the subscript under the data-generating distribution P.

2. Bounds on conditional average treatment effect. To bound the CATE $\tau(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$, we begin by separately bounding

$$\mu_1(X) = \mathbb{E}\big[Y(1) \mid X\big] \quad \text{and} \quad \mu_0(X) = \mathbb{E}\big[Y(0) \mid X\big]. \tag{8}$$

We focus on $\mu_1(\cdot)$ as these two cases are symmetric. Henceforth, our statements hold for P-almost every X and P_z-almost every Y (where z should be inferred from context).

2.1. *Bounding the unobserved potential outcome.* Decompose $\mu_1(\cdot)$ into observed and unobserved components

$$\mu_1(X) = \mathbb{E}\big[Y(1) \mid Z = 1, X\big]P(Z = 1 \mid X) + \mathbb{E}\big[Y(1) \mid Z = 0, X\big]P(Z = 0 \mid X). \tag{9}$$

The mean functions and the nominal propensity score,

$$\mu_{z,z}(X) = \mathbb{E}\big[Y(z) \mid Z = z, X\big], \tag{10}$$

$$e_z(X) = P(Z = z \mid X), \tag{11}$$

are standard regression functions estimable based on observed data [25, 38, 48, 56]. The key difficulty in estimating the CATE is that one potential outcome is always unobserved; we never observe data to directly estimate $\mathbb{E}[Y(1) \mid Z = 0, X]$.

We begin by reformulating the worst-case lower bound (5), $\theta_1(\cdot)$, based on the likelihood ratio between the observed and unobserved potential outcomes. We take a worst-case optimization approach over likelihood ratios to bound the unobserved conditional mean. Using Lemma 2.1 to come, the conditional distribution $P(Y(1) \in \cdot \mid X, Z = 1)$ is absolutely continuous with respect to $P(Y(1) \in \cdot \mid X, Z = 0)$ under condition (3), so

$$\mathbb{E}\big[Y(1) \mid Z = 0, X\big] = \mathbb{E}\big[Y(1)L\big(Y(1), X\big) \mid Z = 1, X\big], \tag{12}$$

where L is the likelihood ratio

$$L(y, x) = \frac{\mathrm{d}P(Y(1) \in \cdot \mid Z = 0, X = x)}{\mathrm{d}P(Y(1) \in \cdot \mid Z = 1, X = x)}(y). \tag{13}$$

While L is unknown, the Γ-selection bias condition (3) constrains it, inducing a lower bound on the unobserved quantity (12).

LEMMA 2.1. *Let P satisfy the Γ-selection bias condition* (3), *and the conditional independence* (2). *Then $P_{Y(1)|Z=0,X=x}$ is absolutely continuous with respect to $P_{Y(1)|Z=1,X=x}$, and the likelihood ratio* (13) *satisfies $0 \le L(y, x) \le \Gamma L(\tilde{y}, x)$ for almost every y, $\tilde{y}$ and x.*

Furthermore, for any likelihood ratio L satisfies $0 \le L(y, x) \le \Gamma L(\tilde{y}, x)$ for almost every y, $\tilde{y}$ and x, there is a distribution P satisfying the Γ-selection bias condition (3), *and the independence assumption* (2), *such that equation* (13) *holds.*

See Appendix A.1 for a proof of the absolute continuity. The rest of the results are illuminating, so we provide them here, assuming absolute continuity.

PROOF. For simplicity in notation and without loss of generality, we assume there are no covariates x. Define the likelihood ratio for the unobserved U by $r(u) := \frac{q_0(u)}{q_1(u)}$, where $q_z(u)$ is the probability density function for $U \mid Z = z$. Note that by applying Bayes rule in the inequality (3), for any u, $\widetilde{u}$,

$$r(u) \le \Gamma r(\widetilde{u}). \tag{14}$$

Then, for any set $B \in \sigma(Y(1))$, the sigma algebra of $Y(1)$, we have

$$\mathbb{E}\big[\mathbf{1}\{B\} \mid Z = 0\big] = \mathbb{E}\big[\mathbb{E}\big[r(U) \mid Y(1), Z = 1\big]\mathbf{1}\{B\} \mid Z = 1\big],$$

so that almost everywhere, the likelihood ratio $L(y) = \frac{dP_{Y(1)|Z=0}}{dP_{Y(1)|Z=1}}(y)$ satisfies

$$L(y) = \mathbb{E}\big[r(U) \mid Y(1) = y, Z = 1\big] \tag{15}$$

by the Radon–Nikodym theorem. Now, for an arbitrary $\epsilon > 0$, and y, $\tilde{y}$ satisfying the equality (15), let u_0 be such that $r(u_0) \le \inf_u r(u) + \epsilon$. Then

$$L(y) \overset{(i)}{=} \mathbb{E}\big[r(U) \mid Y(1) = y, Z = 1\big] = r(u_0)\mathbb{E}\left[\frac{r(U)}{r(u_0)} \,\middle|\, Y(1) = y, Z = 1\right] \overset{(ii)}{\le} \Gamma r(u_0),$$

where equality (i) is simply equation (15) and inequality (ii) follows from the bound (14). We also have $L(\tilde{y}) \ge \inf_u r(u) \ge r(u_0) - \epsilon$ by equality (15). Consequently, $L(y) \le \Gamma r(u_0) \le \Gamma(L(\tilde{y}) + \epsilon)$, and as ϵ was arbitrary, this completes the proof.

The converse follows easily as well: given a likelihood ratio satisfying the above constraint, the Γ-selection bias (3) condition and the independence $\{Y(1), Y(0)\} \perp\!\!\!\perp Z \mid X, U$ is satisfied for $U := (Y(1), Y(0))$, and $P(Z = 1 \mid Z = z, U = u)$ only depending on the $Y(1)$ component of U, and defined by applying Bayes rule to the likelihood ratio. □

Lemma 2.1 implies that the lower bound $\theta_1(x)$ from equation (5) on the unobserved conditional expectation $\mathbb{E}[Y(1) \mid X, Z = 0]$ is

$$\theta_1(X) = \inf\{\mathbb{E}[Y(1)L(Y(1)) \mid Z = 1, X] : L \in \mathscr{L}\}, \tag{16}$$

where

$$\mathscr{L} = \left\{ L : \mathbb{R} \to \mathbb{R} \text{ measurable} : \begin{array}{l} 0 \le L(y) \le \Gamma L(\tilde{y}) \text{ for all } y, \tilde{y}, \\ \mathbb{E}[L(Y(1)) \mid Z = 1, X] = 1 \end{array} \right\}.$$

The first constraint in $\mathscr{L}$ comes from the Γ-selection bias condition (Lemma 2.1), and the second normalization constraint guarantees that L is a likelihood ratio; the objectives and constraints are linear in L. Applying Lagrangian duality to these constraints and simplifying the resulting dual problem shows that the solution to this optimization problem is the solution to an estimating equation in terms of the function

$$\psi_\theta(y) := (y - \theta)_+ - \Gamma(y - \theta)_-. \tag{17}$$

LEMMA 2.2. *Let $\theta_1(X)$ be defined as in (16). If $|\theta_1(X)| < \infty$, then $\theta_1(X)$ solves*

$$\mathbb{E}[\psi_{\theta_1(X)}(Y(1)) \mid Z = 1, X] = 0$$

whenever this solution is unique. If the solution is not unique,

$$\theta_1(X) = \sup\{\mu \in \mathbb{R} : \mathbb{E}[\psi_\mu(Y(1)) \mid Z = 1, X] \ge 0\}. \tag{18}$$

While $\theta_1(\cdot)$ could be estimated using a local estimating equation approach (e.g., as in [34] and [4]) for the equations $\mathbb{E}[\psi_{\theta_1(X)}(Y(1)) \mid Z = 1, X] = 0$ for each X, we go further to provide an alternative loss minimization method to estimate $\theta_1(\cdot)$. This enables the application of a broad class of computationally and statistically efficient estimators.

The lower bound $\theta_1(\cdot)$ is the solution to the convex loss minimization problem

$$\underset{\theta(\cdot)}{\text{minimize}}\, \mathbb{E}[\ell_\Gamma(\theta(X), Y(1)) \mid Z = 1], \tag{19}$$

where ℓ_Γ is the weighted squared loss

$$\ell_\Gamma(\theta, y) := \frac{1}{2}[(y - \theta)_+^2 + \Gamma(y - \theta)_-^2], \tag{20}$$

illustrated in Figure 1. Noting that $\frac{\mathrm{d}}{\mathrm{d}\theta}\ell_\Gamma(\theta, y) = -\psi_\theta(y)$, we have the following lemma on the uniqueness properties and structure of θ_1 solving the optimization problem (19).

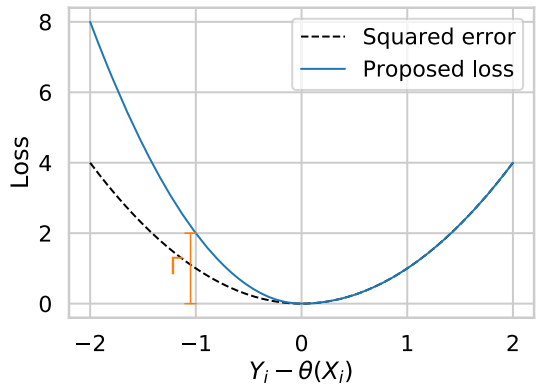

FIG. 1. *Loss function (20) to minimize to lower bound conditional mean of unobserved potential outcome under the Γ-selection bias condition. Illustrated here for $\Gamma = 2$. This loss penalizes negative residuals more than positive residuals, to account for the the fact that confounding could be already up-weighting positive residuals.*

LEMMA 2.3. *Assume* $(t, x) \mapsto \mathbb{E}[\ell_\Gamma(t, Y(1)) \mid X = x, Z = 1]$ *is continuous on* $\mathbb{R} \times \mathcal{X}$. *If* $\mathbb{E}_1[\ell_\Gamma(\theta_1(X), Y(1))] < \infty$, *then* $\theta_1(\cdot)$ *solves* $\mathbb{E}[\psi_\theta(Y(1)) \mid X = x, Z = 1] = 0$ *for almost every* x *if and only if it solves* (19). *Such a minimizer* $\theta_1(\cdot) : \mathcal{X} \to \mathbb{R}$ *exists and is unique up to measure-*0 *sets*.

See Appendix A.3 for proof. Our approach allows both classical techniques, such as sieves, and flexible use of modern machine learning methods to estimate $\theta_1(x)$; in our experiments, we demonstrate how to approximately solve the loss minimization problem (19) using gradient boosted decision trees.

2.2. *Nonparametric estimation with sieves.* To obtain concrete nonparametric guarantees, we consider the method of sieves [19], which considers an increasing sequence $\Theta_1 \subset \Theta_2 \subset \cdots \subset \Theta$ of spaces of (smooth) functions, where Θ denotes all measurable functions. Here, for a sample size n, we take the estimator $\widehat{\theta}_1(\cdot)$ solving

$$\underset{\theta \in \Theta_n}{\text{minimize}}\, \mathbb{E}_n\big[\ell_\Gamma\big(\theta(X), Y(1)\big) \mid Z = 1\big]. \tag{21}$$

With appropriate choices of the function spaces Θ_n, it is possible to provide strong approximation and estimation guarantees. As the loss $\theta \mapsto \ell_\Gamma(\theta(x), y)$ is convex, the empirical optimization problem (21) is convex when Θ_n is a finite-dimensional linear space (e.g., polynomials, splines), which facilitates efficient computation [7].

In Appendix B, we adapt results for sieve estimators [10] to show convergence rates for the solution $\widehat{\theta}_1(\cdot)$ to the empirical problem (21). When $\theta_1(X)$ belongs in a p-smooth Hölder space, in Theorem B.1 of the Supplementary Material [57], we prove that the empirical solution $\widehat{\theta}_1(\cdot)$ is consistent and achieves the following convergence rate (up to logarithmic factors):

$$\big\|\widehat{\theta}_1(\cdot) - \theta_1(\cdot)\big\|_{2,P_1} = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{p}{2p+d}}\right).$$

In the interest of space, we defer a comprehensive treatment to Appendix B.

2.3. *Bounding the CATE.* Since $\theta_1(\cdot)$ satisfies $\theta_1(X) \le \mathbb{E}[Y(1) \mid X, Z = 0]$ under the Γ-selection bias condition, altogether $\mu_1^-(\cdot)$ defined below provides the lower bound

$$\mu_1^-(X) = \mu_{1,1}(X)e_1(X) + \theta_1(X)e_0(X) \le \mu_1(X).$$

By symmetry, letting $\mu_0^+(X) = \mu_{0,0}(X)e_0(X) + \theta_0(X)e_1(X)$ where

$$\begin{aligned} \theta_0(X) = \sup_{L \text{ measurable}} \quad & \mathbb{E}\big[Y(0)L\big(Y(0)\big) \mid Z = 0, X\big] \\ \text{s.t.} \quad & 0 \le L(y) \le \Gamma L(\tilde{y}) \quad \text{all } y, \tilde{y}, \mathbb{E}\big[L\big(Y(1)\big) \mid Z = 0, X\big] = 1, \end{aligned} \tag{22}$$

we have the parallel conclusion that $\mu_0^+(X) \ge \mu_0(X)$ holds under Γ-selection bias condition. Similar to the above, $\theta_0(\cdot)$ is a unique minimizer of $\mathbb{E}[\ell_{\Gamma^{-1}}(\theta(X), Y(0)) \mid Z = 0]$.[2]

Thus, under the Γ-selection bias condition (3), a valid lower bound on the CATE is simply

$$\tau^-(X) = \mu_1^-(X) - \mu_0^+(X). \tag{23}$$

We summarize our developments in the theorem below.

[2]Convergence results for sieve estimators of $\theta_0(\cdot)$ again fall out of our results in Section B.

THEOREM 2.1. *Let* $\Gamma \geq 1$ *and* $\{Y(1), Y(0), Z, X, U\}$ *satisfy condition* (3) *and the conditional independence assumption* (2). *Let* $\tau^-(X)$ *in* (23) *use* $\theta_1(X)$ *and* $\theta_0(X)$ *solving the optimization problems* (16) *and* (22) *with the same* Γ. *When* $\mathbb{E}[|Y(z)| \mid X] < \infty$ *for* $z = 0, 1$ *and* $0 < e_1(X) < 1$,

$$\tau^-(X) \leq \mathbb{E}\big[Y(1) - Y(0) \mid X\big].$$

A natural estimator for $\tau^-(x)$ is the difference in conditional expected potential outcomes

$$\begin{aligned}
\widehat{\tau}^-(x) &= \widehat{\mu}_1^-(x) - \widehat{\mu}_0^+(x), \\
\widehat{\mu}_1^-(x) &= \widehat{\mu}_{1,1}(x)\widehat{e}_1(x) + \widehat{\theta}_1(x)\widehat{e}_0(x) \quad \text{and} \\
\widehat{\mu}_0^+(x) &= \widehat{\mu}_{0,0}(x)\widehat{e}_0(x) + \widehat{\theta}_0(x)\widehat{e}_1(x),
\end{aligned}$$

where $\widehat{e}_z(\cdot)$ and $\widehat{\mu}_{z,z}(\cdot)$ are suitable estimators for the nominal propensity score $e_z(\cdot)$ and the observed potential outcome's mean function $\mu_{z,z}(\cdot)$, respectively. A variety of classical nonparametric methods and machine learning methods can estimate these regression functions [12, 13, 56]. To understand convergence of $\widehat{\tau}^-(\cdot)$, consider the convergence of these regression estimates. Specifically, assume that the estimators $\widehat{e}_1(\cdot)$ and $\widehat{\mu}_{z,z}(\cdot)$ satisfy that

$$\begin{aligned}
\big\|\widehat{e}_1(\cdot) - e_1(\cdot)\big\|_{2,P} &= O_P(r_{n,1}), \\
\big\|\widehat{\mu}_{1,1}(\cdot) - \mu_{1,1}(\cdot)\big\|_{2,P_1} &= O_P(r_{n,2}), \\
\big\|\widehat{\mu}_{0,0}(\cdot) - \mu_{0,0}(\cdot)\big\|_{2,P_0} &= O_P(r_{n,3}), \\
\big\|\widehat{\theta}_1(\cdot) - \theta_1(\cdot)\big\|_{2,P_1} &= O_p(r_{n,4}), \\
\big\|\widehat{\theta}_0(\cdot) - \theta_0(\cdot)\big\|_{2,P_0} &= O_p(r_{n,5}),
\end{aligned}$$

where $r_{n,j}$ depend on the model assumptions and estimation method. We assume $0 < \epsilon \leq e_1(x) \leq 1 - \epsilon$, so $\|\cdot\|_{2,P_1} \asymp \|\cdot\|_{2,P_0} \asymp \|\cdot\|_{2,P}$. Then $\widehat{\tau}^-(\cdot)$ is a consistent estimator, and

$$\big\|\widehat{\tau}^-(\cdot) - \tau^-(\cdot)\big\|_{2,P} = O_p(r_{n,1} + r_{n,2} + r_{n,3} + r_{n,4} + r_{n,5}).$$

Under assumptions stated in Appendix B (A4–A6, including that θ_z belongs in a p-smooth Hölder space), our sieve estimators (21) for θ_z achieves the asymptotic convergence rate

$$\|\widehat{\theta}_z - \theta_z\|_{2,P_z} = \widetilde{O}_P\big(n^{-\frac{p}{2p+d}}\big) \quad z \in \{0, 1\},$$

where the notation $\widetilde{O}_P(\cdot)$ hides logarithmic factors. Under similar smoothness and regularity assumptions, Chen and White [12] establish that sieve estimators $\widehat{e}_z(\cdot)$ and $\widehat{\mu}_{z,z}(\cdot)$ for e_z and $\mu_{z,z}$ can also achieve a convergence rate of $r_{n,j} = \widetilde{O}(n^{-\frac{p}{2p+d}})$. Consequently,

$$\big\|\widehat{\tau}^-(\cdot) - \tau^-(\cdot)\big\|_{2,P} = \widetilde{O}_P\big(n^{-\frac{p}{2p+d}}\big),$$

where the convergence rates reflect typical behavior of (minimax optimal) nonparametric estimators of a regression function [36, 51]. These constitute the high order terms of the approximation error for estimating the CATE $\tau(x)$ without unobserved confounding [30], if the smoothness of the CATE $\tau(\cdot)$ is of a similar order to the individual parameters $\theta_z(\cdot)$, $\mu_z(\cdot)$ and $e_z(\cdot)$. Interesting future work would be to develop a method that adapts to the complexity of $\tau^-(\cdot)$, itself, as done by Nie and Wager [38] and Kennedy [29].

3. Bounds on the average treatment effect. Given the bounds developed in Section 2 for the conditional average treatment effect $\tau(\cdot)$, we now turn to bounding the average treatment effect (ATE) τ by marginalizing over X,

$$\tau^- := \mathbb{E}\big[\tau^-(X)\big] = \mathbb{E}\big[\mu_1^-(X) - \mu_0^+(X)\big]. \tag{24}$$

Because $\tau^-(x) \le \tau(x)$ for any x, $\tau^- \le \tau$ is a lower bound of the ATE. Rewriting τ^- as

$$\tau^- = \mu_1^- - \mu_0^+ \quad \text{where} \quad \begin{cases} \mu_1^- = \mathbb{E}\big[\mu_1^-(X)\big] = \mathbb{E}\big[ZY(1) + (1-Z)\theta_1(X)\big], \\ \mu_0^+ = \mathbb{E}\big[\mu_0^+(X)\big] = \mathbb{E}\big[(1-Z)Y(0) + Z\theta_0(X)\big], \end{cases} \tag{25}$$

we estimate μ_1^- and μ_0^+ separately and combine the resulting estimators.

In Section 3.1, we construct a semiparametric estimator for the bound τ^- that is conceptually similar to the AIPW estimator under unconfoundedness. We show in Section 3.2 that it achieves $\sqrt{n}$-consistency even when (nonparametric) estimates of the nuisance parameters (e.g., $e_z(\cdot)$, $\mu_{z,z}(\cdot)$, $\theta_1(\cdot)$) only converge at slower rates. We focus on lower bounds for the potential outcome $Y(1)$ as other cases are symmetric. We conclude our theoretical discussion by complementing our methodological developments with optimality guarantees (Section 3.4). We show that our approach is asymptotically unimprovable for testing a null of no treatment effect and unobserved confounding satisfying the Γ-selection bias condition against a positive alternative.

3.1. *Estimation procedure.* We construct a score $T(V;\eta)$ to estimate μ_1^- similar to the AIPW estimator of the ATE in the absence of unobserved confounding, where $V = (X, Y, Z)$ and η represents a set of nuisance parameters defined below. The score $T(V;\eta)$ comes from calculating the semiparametric influence function for μ_1^- from representation in (25) using the method described by Newey [35], and augmenting the representation with the influence function. To this end, by computing the pathwise derivative of the functional in (25) with respect to a parametric subfamily of the nonparametric model, and matching to the form derived by Newey [35], we see that the remaining term in the influence function is

$$\alpha_1(V;\theta_1,e_1,\nu_1) = Z\frac{\psi_{\theta_1(X)}(Y)(1-e_1(X))}{\nu_1(X)e_1(X)},$$

which depends on the nuisance parameters $\theta_1(x)$ and $e_1(x)$, and a new nuisance parameter,

$$\nu_1(x) = P\big(Y \ge \theta_1(x) \mid Z=1, X=x\big) + \Gamma\, P\big(Y < \theta_1(x) \mid Z=1, X=x\big), \tag{26}$$

which serves as a weight normalization factor. In this, the function $\psi_\theta(y)$ refers to the one defined in equation (17). Adding the term $\alpha_1(V;\eta)$ to the representation in (25) gives the augmented score

$$T(X,Y,Z;\theta_1,e_1,\nu_1) := ZY + (1-Z)\theta_1(X) + Z\frac{\psi_{\theta_1(X)}(Y)(1-e_1(X))}{\nu_1(X)e_1(X)}, \tag{27}$$

that we use for estimation. We have $E_P[T(X,Y,Z;\theta_1,e_1,\nu_1)] = \mu_1^-$ since $\mathbb{E}_P[\psi_{\theta_1(X)}(Y) \mid Z=1, X] = 0$. By virtue of its augmented form, the score $T(\cdot;\cdot)$ is insensitive to estimates in the nuisance parameters, formalized by the Neyman orthogonality condition [37].

DEFINITION 2. Let $Q, \eta \mapsto \mathbb{E}_Q[S(V;\eta)]$ be a statistical functional with Q a distribution over V, and nuisance parameter $\eta \in \Lambda$, where we take Λ to be a subset of a normed vector space containing the true nuisance parameter η_0. The score S is Neyman orthogonal at P if for all $\eta \in \Lambda$, the derivative $\frac{d}{dr}S(P;\eta_0 + r(\eta-\eta_0))$ exists for $r \in [0,1)$, and is zero at $r=0$.

As Chernozhukov et al. ([13], Section 2.2.5) shows, a score formed by adding the influence function adjustment $\alpha_1(v)$ from the pathwise derivative as in Newey [35] is Neyman orthogonal. Therefore, we expect Neyman orthogonality of the functional (27) constructed in this way; we verify this formally in the proof of Theorem 3.2 in the Supplementary Material.

We construct a semiparameteric plug-in estimator for the augmented functional, and show that estimation errors of the nuisance parameters multiply to reduce their influence on our final estimator. Concretely, we prove that our augmented esitmator preserves $\sqrt{n}$ consistency provided that our nuisance estimates converge at a rate of $o_P(n^{-1/4})$ in $\|\cdot\|_{2,P}$ norm. This draws important connections to the classical doubly-robust AIPW estimator under no unobserved confounding. Recalling the definitions (10) and (11) of $\mu_{z,z}(x)$ and $e_1(x)$ (respectively), the standard AIPW estimator for $\mathbb{E}[Y(1)]$ is

$$\widehat{\mu}_{1,\mathrm{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{\mu}_{1,1}(X_i) + \frac{Z_i}{\widehat{e}_1(X_i)}\big(Y_i - \widehat{\mu}_{1,1}(X_i)\big)\right]. \tag{28}$$

Assuming all confounding variables are observed (1), the AIPW (28) is an asymptotically efficient estimator of μ_1 [23]. The AIPW also satisfies the Neyman orthogonality condition, which Chernozhukov et al. [13] used to show that the AIPW estimator (28) with cross-fitting (described below) enjoys the root n rate so long as the nuisance parameters can be estimated at the rate $o_p(n^{-1/4})$. Our approach generalize the AIPW estimator (28) under the Γ-selection bias condition, and reduces to the AIPW when $\Gamma = 1$.

We use an efficient sample-splitting recipe for constructing an augmented estimator for μ_1^- by adapting Chernozhukov et al.'s [13] cross-fitting meta-procedure for Neyman-orthogonal functionals to our augmented score $T(\cdot)$. Letting $K \in \mathbb{N}$ be the number of folds for cross-fitting, randomly split the data into K folds of approximately equal size. With slight abuse of notation, let $\mathcal{I}_k$ be the indices corresponding to the observations in the kth part as well as the corresponding observation themselves.

For each k, using the sample $\mathcal{I}_{-k}$ of observations *not* in the kth fold, we compute:

1. an estimator of $\theta_1(x)$, denoted by $\widehat{\theta}_{1,k}(x)$, using the procedure described in Section 2;
2. an estimator of $e_1(x)$, denoted by $\widehat{e}_{1,k}(x)$, and let $\widehat{e}_{0,k}(x) = 1 - \widehat{e}_{1,k}(x)$;
3. an estimator of $\nu_1(\cdot)$, denoted by $\widehat{\nu}_{1,k}(\cdot)$, using the procedure described in Section 3.3.

Estimating $\nu_1(\cdot)$ in the last step is more involved, as it depends on $\theta_1(\cdot)$, so we defer the construction of $\widehat{\nu}_{1,k}(\cdot)$ to Section 3.3. Under appropriate regularity conditions—for example, sufficient smoothness of $\theta_1(x)$, $e_1(x)$ and $\nu_1(x)$—these estimators attain $o_P(n^{-1/4})$ convergence in $\|\cdot\|_{2,P}$. In the end, our proposed cross-fitting estimator of μ_1^- is

$$\widehat{\mu}_1^- = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left\{Z_iY_i + (1-Z_i)\widehat{\theta}_{1,k}(X_i) + Z_i\frac{\psi_{\widehat{\theta}_{1,k}(X_i)}(Y_i)\widehat{e}_{0,k}(X_i)}{\widehat{\nu}_{1,k}(X_i)\widehat{e}_{1,k}(X_i)}\right\}, \tag{29}$$

with an estimator $\widehat{\mu}_0^+$ for μ_0^+ constructed similarly. This estimator is natural; when $\Gamma = 1$, we recover the cross-fitting version of the standard doubly robust AIPW estimator (28). While the estimator satisfies the orthogonality conditions of Chernozhukov et al. [13] that imply a form of local robustness for $\widehat{\theta}_1(\cdot)$ near $\theta_1(\cdot)$, we explain below why it is not doubly robust.

3.2. *Asymptotic properties and inference.* To establish asymptotic normality of $\widehat{\mu}_1^-$, we require a few assumptions. Consistency of $\widehat{\mu}_1^-$ follows from weak regularity conditions and the consistency of $\widehat{\theta}_1(\cdot)$, which we address via Assumption A1. Asymptotic normality requires stronger conditions (Assumptions A2 and A3), in turn allowing us to establish Theorem 3.2 on the asymptotic normality of $\widehat{\mu}_1$.

ASSUMPTION A1. There exist $\epsilon > 0$ and $0 < c_{\text{low}} < c_{\text{hi}}$ such that (a) $\mathbb{E}[|Y(1)|] < \infty$, (b) $\|\widehat{\theta}_1(\cdot) - \theta_1(\cdot)\|_{1,P} \xrightarrow{p} 0$, (c) $e_1(X) \in [\epsilon, 1-\epsilon]$ almost surely, (d) $\mathbb{P}([\operatorname{ess\,inf} \widehat{e}_1(X), \operatorname{ess\,sup} \widehat{e}_1(X)] \subset [\epsilon, 1-\epsilon]) \to 1$ and (e) $\mathbb{P}(c_{\text{low}} \le \widehat{\nu}_1(x) \le c_{\text{hi}} \text{ for all } x) \to 1$.

Assumption A1(a)–(c) are slightly stronger than the usual assumptions for justifying consistency of the AIPW estimator for the ATE τ in the absence of unobserved confounding [5, 13]. When $\|\widehat{e}_1(\cdot) - e_1(\cdot)\|_{\infty,P} \xrightarrow{p} 0$, Assumption A1(c) implies Assumption A1(d), and similarly when $\|\widehat{\nu}_1(\cdot) - \nu_1(\cdot)\|_{\infty,P} \xrightarrow{p} 0$, $\nu_1(x) \in [1, \Gamma]$ implies Assumption A1(e).

Assumption A1(b) is necessary, and cannot be removed by alternatively assuming consistency of the other nuisance parameters. The $\alpha(V;\eta)$ with the true $\theta_1(\cdot)$ plugged in has mean zero regardless of the nominal propensity score used. In this case, the proposed estimator is consistent in estimating μ_1^-. However, if an incorrect $\theta_1(\cdot)$ is plugged in to $T(V;\eta)$, straightforward computation shows that $\mathbb{E}[T(V;\eta)]$ depends on the $\theta_1(\cdot)$ plugged in, even with the correct $e_1(\cdot)$ and $\nu_1(\cdot)$. Therefore, $\hat{\mu}_1^-$ is not globally doubly robust; the Neyman orthogonality condition only guarantees a local form of robustness.

THEOREM 3.1. *Under Assumption A1, the estimator (29) satisfies* $\widehat{\mu}_1^- \xrightarrow{p} \mu_1^-$.

See the Supplementary Material, Section 8.1, for the proof. We now turn to stronger regularity assumptions for the weak convergence of $\widehat{\mu}_1^-$.

ASSUMPTION A2. (a) There exist $q > 2$, and $C_q < \infty$ such that $\mathbb{E}[|Y(1)|^q] \le C_q$, and (b) $Y(1)$ has a conditional density $p_{Y(1)}(y \mid X = x, Z = 1)$ with respect to the Lebesgue measure and $\sup_{x,y} p_{Y(1)}(y \mid Z = 1, X = x) < \infty$.

ASSUMPTION A3. $\widehat{\eta}_1 = (\widehat{\theta}_1, \widehat{\nu}_1, \widehat{e}_1)$ is a consistent estimator of $\eta_1 := (\theta_1, \nu_1, e_1)$ and (a) $\|\widehat{\eta}_1(\cdot) - \eta_1(\cdot)\|_{2,P} = o_P(n^{-1/4})$, (b) $\|\widehat{\eta}_1(\cdot) - \eta_1(\cdot)\|_{\infty,P} = O_P(1)$.

Assumptions A2(a) is no stronger than the standard regularity conditions needed for existence of asymptotically normal estimators of the ATE without unobserved confounding [13]. Assumption A2(b) ensures that the term $\theta(\cdot) \mapsto \mathbb{E}[Z\psi_{\theta(X)}\{Y(1)\} \mid X]$ is sufficiently smooth to control fluctuations due to estimating $\theta_1(\cdot)$. Inspection of the proof of Theorem 3.2 to come shows that we may relax Assumption A2(b): if $\theta_1(x)$ and $\widehat{\theta}_1(x)$ have range $\mathcal{A}_1(x)$, we may replace A2(b) with

$$\operatorname*{ess\,sup}_X \sup_{y \in \mathcal{A}_1(X)} p_{Y(1)}(y \mid Z = 1, X) < \infty, \tag{30}$$

which is satisfied, for example, when the outcome $Y(z)$ is binary and $P(Y(z) = y \mid Z = z, X) < 1$ for $y \in \{0, 1\}$, because $\widehat{\theta}(X) \in (0, 1)$ eventually and $p(y \mid Z = 1, X) = 0$ for $y \notin \{0, 1\}$.

The convergence rate conditions for estimating nuisance parameters in Assumption A3 are relatively standard in semiparametric estimation [13, 35], but nonetheless this theoretical requirement can be restrictive and hard to achieve to certain applications. For example, while for $e_1(\cdot)$, the conditional mean of observed random variables, a variety of methods can provide $o_P(n^{-1/4})$ consistency, they still require the data generating distribution to meet appropriate conditions and the sample size to be large relative to the dimension of covariate [13, 55]. The estimators $\widehat{\theta}_1(\cdot)$ from Section 2 and $\widehat{\nu}_1(\cdot)$ from Section 3.1 achieve the convergence rates in Assumption A3 under appropriate smoothness conditions on $\theta_1(\cdot)$ and $\nu_1(\cdot)$. For instance, if Assumptions A4, A5 and A6 hold with $p > d/2$, then Theorem B.1 shows that estimating $\theta_1(x)$ as in Section 2 with linear sieves (see Examples 1 and 2) will satisfy

Assumption A3. Section 3.3 provides an efficient enough estimator of $\nu_1(\cdot)$ when $p > d/2$. Under these assumptions, the following theorem gives the asymptotic distribution of the estimator $\widehat{\mu}_1^-$ in (29), with asymptotic variance

$$\sigma_1^2 := \operatorname{Var}\left[ZY + (1-Z)\theta_1(X) + Z\frac{\psi_{\theta_1(X)}(Y)(1-e_1(X))}{\nu_1(X)e_1(X)} \right].$$

We use the following consistent estimator of the asymptotic variance:

$$\widehat{\sigma}_1^2 := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left[Z_iY_i + (1-Z_i)\widehat{\theta}_{1,k}(X_i) + Z_i\frac{\psi_{\widehat{\theta}_{1,k}(X_i)}(Y_i)\widehat{e}_{0,k}(X_i)}{\widehat{\nu}_{1,k}(X_i)\widehat{e}_{1,k}(X_i)} - \widehat{\mu}_1^- \right]^2.$$

THEOREM 3.2. *Let Assumptions A1, A2 and A3 hold. Then $\widehat{\mu}_1^-$ given in equation (29) is asymptotically normal with $\sqrt{n}(\widehat{\mu}_1^- - \mu_1^-) \xrightarrow{d} \mathsf{N}(0,\sigma_1^2)$. Furthermore, $\widehat{\sigma}_1^2 \xrightarrow{p} \sigma_1^2$, and $\frac{\sqrt{n}}{\widehat{\sigma}_1}(\widehat{\mu}_1^- - \mu_1^-) \xrightarrow{d} \mathsf{N}(0,1)$.*

See Section 8.2 in the Supplementary Material for a proof. To bound τ from below, let

$$\widehat{\tau}^- = \widehat{\mu}_1^- - \widehat{\mu}_0^+,$$

where $\tilde{\psi}_\theta(y) = \Gamma(y-\theta)_+ - (y-\theta)_-$,

$$\widehat{\mu}_0^+ = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\left[(1-Z_i)Y_i + Z_i\widehat{\theta}_{0,k}(X_i) + (1-Z_i)\frac{\tilde{\psi}_{\widehat{\theta}_{0,k}(X_i)}(Y_i)\widehat{e}_{1,k}(X_i)}{\widehat{\nu}_{0,k}(X_i)\widehat{e}_{0,k}(X_i)} \right],$$

and $\widehat{\nu}_{0,k}(\cdot)$ is the nonparametric estimator of

$$\nu_0(X) = P\big(Y \le \theta_0(X) \mid Z=0, X\big) + \Gamma\, P\big(Y > \theta_0(X) \mid Z=0, X\big)$$

based on data in $\mathcal{I}_{-k}$. A simple extension of Theorem 3.2 shows

$$\sqrt{n}\big(\widehat{\tau}^- - \tau^-\big) \to \mathsf{N}\big(0, \sigma_{\tau^-}^2\big),$$

as $n \to \infty$, where

$$\begin{aligned}\sigma_{\tau^-}^2 := \operatorname{Var}\Bigg[& ZY + (1-Z)\theta_1(X) + Z\frac{\psi_{\theta_1(X)}(Y)e_0(X)}{\nu_1(X)e_1(X)} \\ & - (1-Z)Y - Z\theta_0(X) - (1-Z)\frac{\tilde{\psi}_{\theta_0(X)}(Y)e_1(X)}{\nu_0(X)e_0(X)} \Bigg].\end{aligned} \tag{31}$$

Furthermore, a consistent estimator of the variance $\sigma_{\tau^-}^2$ is

$$\begin{aligned}\widehat{\sigma}_{\tau^-}^2 = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{I}_k}\Bigg[& Z_iY_i + (1-Z_i)\widehat{\theta}_{1,k}(X_i) + Z_i\frac{\psi_{\widehat{\theta}_{1,k}(X_i)}(Y_i)\widehat{e}_{0,k}(X_i)}{\widehat{\nu}_{1,k}(X_i)\widehat{e}_{1,k}(X_i)} \\ & - (1-Z_i)Y_i - Z_i\widehat{\theta}_{0,k}(X_i) \\ & - (1-Z_i)\frac{\tilde{\psi}_{\widehat{\theta}_{0,k}(X_i)}(Y_i)\widehat{e}_{1,k}(X_i)}{\widehat{\nu}_{0,k}(X_i)\widehat{e}_{0,k}(X_i)} - \widehat{\tau}^- \Bigg]^2\end{aligned} \tag{32}$$

and $[\widehat{\tau}^- - z_{1-\alpha/2}\widehat{\sigma}_{\tau^-}/\sqrt{n}, \widehat{\tau}^- + z_{1-\alpha/2}\widehat{\sigma}_{\tau^-}/\sqrt{n}]$ is a $100(1-\alpha)\%$ asymptotic confidence interval for τ^-. The proof is *mutatis mutandis* identical to that of Theorem 3.2.

Importantly, our bounds define a confidence set for $\tau = \mathbb{E}[Y(1) - Y(0)]$. The same approach as in Section 2, but upweighting large values of $Y(1)$ and small values of $Y(0)$, provides an estimate $\widehat{\tau}^+$ of τ^+ that upper bounds the ATE. The limiting distribution of $\widehat{\tau}^+$ is also normal. With these estimators, we may construct a confidence interval for the ATE,

$$\widehat{\mathrm{CI}}_\tau = \left[\widehat{\tau}^- - z_{1-\alpha/2}\frac{\widehat{\sigma}_{\tau^-}}{\sqrt{n}}, \widehat{\tau}^+ + z_{1-\alpha/2}\frac{\widehat{\sigma}_{\tau^+}}{\sqrt{n}}\right], \tag{33}$$

where $\widehat{\sigma}^2_{\tau^+}$ is a consistent estimator of the variance of $\sqrt{n}(\widehat{\tau}^+ - \tau^+)$. Because $\tau^- \le \tau \le \tau^+$, this confidence interval has appropriate asymptotic coverage.

COROLLARY 3.1. *Let P satisfy the Γ-selection bias condition* (3), *conditional independence* (2) *and Assumptions* A1–A3. *Let $\widehat{\mathrm{CI}}_\tau$ be defined as in* (33). *For $\tau = \mathbb{E}[Y(1) - Y(0)]$, we have*

$$\liminf_{n\to\infty} P(\tau \in \widehat{\mathrm{CI}}_\tau) \ge 1 - \alpha.$$

REMARK 1. It is possible to extend Theorem 3.2 to provide confidence intervals uniform over $\mathcal{P}$. In other words, the coverage probability of the relevant confidence intervals converge to the desired level uniformly over all the distributions in $\mathcal{P}$. To do so, Assumption A3 must be uniform over a class of distributions $\mathcal{P}$ satisfying Assumption A2, for instance by assuming there exists sequences $\Delta_n \to 0$ and $\delta_n \to 0$ such that

$$\sup_{P\in\mathcal{P}} P\big(\|\widehat{\eta}_1(\cdot) - \eta_1(\cdot)\|_{2,P} > n^{-1/4}\delta_n\big) < \Delta_n.$$

Previous work [11] shows that series estimators for the conditional regression function (example 1 in Section 2) converge uniformly; extending these results to the estimation of $\theta_1(\cdot)$ and $\nu_1(\cdot)$ is beyond the scope of the present work.

3.3. *Construction of $\widehat{\nu}_{1,k}(\cdot)$ and its asymptotic properties.* The above results assumed access to a well-behaved estimate of the the weighted probability $\nu_1(X) = 1 + (\Gamma - 1)\mathbb{P}(Y(1) \ge \theta_1(X) \mid Z = 1, X)$. Here, we describe a nonparametric estimator via a loss function: defining

$$\bar{\ell}_\Gamma(\nu, \theta, y) := \frac{1}{2}\big[1 + (\Gamma - 1)\mathbf{1}\{y \ge \theta\} - \nu\big]^2,$$

ν_1 uniquely solves the optimization problem

$$\underset{\nu(\cdot)\text{ measurable}}{\text{minimize}}\ \mathbb{E}\big[\bar{\ell}_\Gamma\big\{\nu(X), \theta_1(X), Y(1)\big\} \mid Z = 1\big]. \tag{34}$$

The natural sieve estimator for $\nu_1(\cdot)$ minimizes the empirical version of (34) under finite-dimensional sieves. However, this requires knowledge of $\theta_1(\cdot)$, which itself must be estimated. Therefore, consider the following (nested) cross-fitting approach:

1. Partition the sample $\mathcal{I}_{-k}$ into two *independent* sets, $\mathcal{I}_{-k,1}$ and $\mathcal{I}_{-k,2}$.
2. Let $\widehat{\theta}^{\nu_1}_{1k}(\cdot)$ be an estimator of $\theta_1(\cdot)$ based on the first subset $\mathcal{I}_{-k,1}$.
3. For a sequence of sieve parameter spaces $\Pi_1 \subseteq \cdots \subseteq \Pi_n \subseteq \cdots \subseteq \Pi$, estimate $\widehat{\nu}_{1,k}$ minimizing the plug-in version of the population problem (34),

$$\underset{\nu(\cdot)\in 1+(\Gamma-1)\Pi_n}{\text{minimize}}\ \mathbb{E}^{(k)}_{n,2}\big[\bar{\ell}_\Gamma\big(\nu(X), \widehat{\theta}^{\nu_1}_{1k}(X), Y\big) \mid Z = 1\big], \tag{35}$$

where $\mathbb{E}^{(k)}_{n,2}$ is the empirical expectation with respect to the second subset $\mathcal{I}_{-k,2}$.

When $\nu_1(X)$ belongs in a q-smooth Hölder space, in Proposition B.1 in the Supplementary Material, we prove that the empirical solution $\widehat{\nu}_{1,k}(\cdot)$ to the problem (35) achieves the minimax optimal nonparametric rate (up to logarithmic factors)

$$\|\widehat{\nu}_{1,k}(\cdot) - \nu_1(\cdot)\|_{2,P_1} = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{q}{2q+d}}\right).$$

If $q > d/2$, then $\|\widehat{\nu}_{1,k} - \nu_1\|_{2,P} = o_P(n^{-1/4})$, satisfying the assumptions in Theorem 3.2. We defer a rigorous treatment to Appendix B.2 as our results heavily build on the standard theory of sieve estimation [10]. In Proposition B.1, we demonstrate sufficient conditions for the convergence of $\widehat{\nu}_{1,k}$ needed for the lower bound estimator (29) and its asymptotic normality via Theorem 3.2: with sufficient smoothness of ν_1, it is possible to efficiently estimate lower and upper bounds on the average treatment effect.

3.4. *Design sensitivity and optimality of our bound on the ATE.* We complement our methodological development so far with optimality results for our worst-case bounds. By construction, our approach yields a tight bound on the mean of each unobserved potential outcome. We extend these results to the ATE by constructing an instance where our bound is tight. That is, we construct a family of data generating distributions such that whenever our bounds cannot infer the sign of the ATE, the ability to test whether or not the ATE is positive is intrinsically difficult. To this end, we study a pointwise asymptotic level α hypothesis test for the composite null

$$H_0(\Gamma) : \mathbb{E}\big[Y(1)\big] \leq \mathbb{E}\big[Y(0)\big] \quad \text{and the } \Gamma\text{-selection bias condition (3) holds} \tag{36}$$

under Assumptions A1–A3, and analyze its design sensitivity [45]. Let $H_1 : Q$ be an alternative with a positive average treatment effect $\tau = \mathbb{E}_Q[Y(1) - Y(0)] > 0$ and no confounding ($\Gamma = 1$ in equation (3)). Let $t_n^\Gamma = t_n^\Gamma\{(Y_i, Z_i, X_i)_{i=1}^n\} \in \{0, 1\}$ be a pointwise asymptotic level α test for the null hypothesis (36), where $t_n^\Gamma = 1$, if the null hypothesis $\tau \leq 0$ is rejected. The *design sensitivity* [45, 46] of the sequence $\{t_n^\Gamma\}$ is the threshold Γ_{design} such that the power $Q(t_n^\Gamma = 1) \to 0$ for $\Gamma > \Gamma_{\text{design}}$ and the power $Q(t_n^\Gamma = 1) \to 1$ for $\Gamma < \Gamma_{\text{design}}$. In other words, if the selection bias satisfies $\Gamma > \Gamma_{\text{design}}$, the test cannot differentiate the alternative $\tau > 0$ from the null $\tau \leq 0$ regardless of the sample size; if $\Gamma < \Gamma_{\text{design}}$, the test always rejects the null under the alternative Q for sufficiently large n (we define $\Gamma_{\text{design}} = \infty$ when no such threshold exists). Given the confidence interval for τ described in Section 3.2, a natural asymptotic level α test for $H_0(\Gamma)$, the hypothesis (36) is

$$\psi_n^\Gamma\{(Y_i, Z_i, X_i)_{i=1}^n\} := \mathbf{1}\left\{\widehat{\tau}^- > z_{1-\alpha}\frac{\widehat{\sigma}_{\tau^-}}{\sqrt{n}}\right\}. \tag{37}$$

We consider the design sensitivity of ψ_n^Γ in the simplified setting without covariates, which allows us to demonstrate its optimality. In this case, $\{Y(0), Y(1)\} \perp\!\!\!\perp Z \mid U$, the simplified Γ-selection bias condition (7) holds, $\{Y(0), Y(1)\} \perp\!\!\!\perp Z$ under the alternative Q (recall equation (1)) and $\theta_1, \theta_0 \in \mathbb{R}$ are constants defined in equation (16) and equation (22).

PROPOSITION 3.1. *Let ψ_n^Γ be defined as in equation (37), so that ψ_n^Γ is asymptotically level α for $H_0(\Gamma)$ in (36). For an alternative $H_1 = \{Q\}$, define*

$$\tau^-(\Gamma) := \mathbb{E}_Q\big[ZY(1) + (1 - Z)\theta_1 - (1 - Z)Y(0) - Z\theta_0\big],$$

where θ_1, θ_0 solve (16) and (22), respectively, at level Γ for the distribution Q. Then either the design sensitivity Γ_{design} of ψ_n^Γ is infinite or it uniquely solves the equation $\tau^-(\Gamma) = 0$.

See Section 8.3 in the Supplementary Material for proof. While there is no simplified expression for Γ_{design} in general, it can be derived explicitly for some special alternatives Q. For instance, in the Supplementary Material, Section 8.3.2, we prove the following result for Gaussian alternatives.

COROLLARY 3.2. *Let ψ_n^Γ be as in equation (37). For the alternative $H_1(Q)$,*

$$\left\{Y(1) \sim \mathsf{N}\left(\frac{\tau}{2}, \sigma^2\right), Y(0) \sim \mathsf{N}\left(-\frac{\tau}{2}, \sigma^2\right), Z \sim \text{Bernoulli}\left(\frac{1}{2}\right)\right\},$$

ψ_n^Γ has design sensitivity

$$\Gamma_{\text{design}}^{\text{gauss}} := -\frac{\int_0^\infty y \exp(-\frac{(y-\tau)^2}{2\sigma^2})\,\mathrm{d}y}{\int_{-\infty}^0 y \exp(-\frac{(y-\tau)^2}{2\sigma^2})\,\mathrm{d}y} = \frac{\phi(\frac{\tau}{\sigma}) + \frac{\tau}{\sigma}\Phi(\frac{\tau}{\sigma})}{\phi(\frac{\tau}{\sigma}) - \frac{\tau}{\sigma}\Phi(\frac{\tau}{\sigma})}, \tag{38}$$

where Φ and ϕ denote the standard Gaussian CDF and density, respectively.

The next proposition shows that the test ψ_n^Γ is optimal for alternative $H_1(Q)$ given in Corollary 3.2, as any asymptotic level α test of $H_0(\Gamma)$ has design sensitivity $\geq \Gamma_{\text{design}}^{\text{gauss}}$ (see the Supplementary Material, Section 8.4, for proof).

PROPOSITION 3.2. *Let $H_0(\Gamma)$ be as in (36). There exists $a \in [1/(1+\sqrt{\Gamma}), \sqrt{\Gamma}/(1+\sqrt{\Gamma})]$ such that for the alternative $H_1(Q)$:*

$$\left\{Y(1) \sim \mathsf{N}\left(\frac{\tau}{2}, \sigma^2\right), Y(0) \sim \mathsf{N}\left(-\frac{\tau}{2}, \sigma^2\right), Z \sim \text{Bernoulli}(a)\right\},$$

if $\Gamma \geq \Gamma_{\text{design}}^{\text{gauss}}$, there exists a probability measure $P \in H_0(\Gamma)$ for $\{Y(1), Y(0), Z, U\}$, such that for all $n \in \mathbb{N}$, all tests t_n, and (Y_i, Z_i) i.i.d.,

$$P\big(t_n\{(Y_i, Z_i)_{i=1}^n\} = 1\big) = Q\big(t_n\{(Y_i, Z_i)_{i=1}^n\} = 1\big).$$

REMARK 2. Our proof uses a specific choice of a to simplify the algebra; solving a system of nonlinear equations for the distribution of $P_{Z|U}$ allows for any marginal $P(Z=1)$.

REMARK 3. The above optimality results for ψ_n^Γ extend to alternatives beyond Gaussian distributions, so long as $Y(0) \overset{d}{=} C(1 - Y(1))$, for some constant $C > 0$. The proof relies on this symmetry in the potential outcomes to construct a distribution under $H_0(\Gamma)$ matching Q over the observed data, $\{(Y_i(Z_i), Z_i), i = 1, \ldots, n\}$. This symmetry is unnecessary if one is interested in the mean (or conditional mean) of a single potential outcome $\mathbb{E}[Y(1)]$ (or $\mathbb{E}[Y(1) \mid X = x]$, in which case the test ψ_n^Γ achieves the optimal design sensitivity for any alternative for which the proposed method is consistent.

4. Numerical experiments. To complement our theoretical analysis in Section 3, we examine the performance of the method using Monte-Carlo simulation and a real data set from an observational study examining the effect of fish consumption on blood mercury levels. We evaluate two implementations of the methodology developed in Sections 2 and 3—one based on the sieve estimators studied in Section 2.2 and the other based on gradient boosted trees fit to minimize the weighted squared loss (19).

The Monte-Carlo simulations support the validity of the inference procedure in realistic settings. We find that the semiparametric approach presented in Section 3 accurately bounds the average treatment effect under unobserved confounding, when our assumptions about the extent of confounding Γ hold. We show that by using machine learning to optimize the loss

function in (20), our method can scale to reasonably high-dimensional data. Additionally, we show that the bounds on the ATE are tight in practice, and empirically compare their conservativeness to that of the matching-based approach from Rosenbaum [47]. Finally, we confirm our findings on a real observational study, demonstrating that our semiparametric approach provides valid yet narrow bounds on the ATE τ.

4.1. *Method implementations*. When implementing an estimator to bound the ATE τ using the method developed in Section 3, one must choose estimators of the nuisance parameters $e_z(\cdot)$, $\theta_z(\cdot)$ and $\nu_z(\cdot)$, and select their hyperparameters. In the first implementation, we stay close to the estimators used in our theoretical analysis with formal convergence guarantees: we estimate the propensity score $\widehat{e}_1(\cdot)$ by a random forest [4], and $\widehat{\theta}_z(\cdot)$ (resp., $\widehat{\nu}_z(\cdot)$) by the nonparametric estimator from Section 2 (resp., Section 3.3) using the polynomial (power series) sieve. The sieve size and regularization were selected via 10-fold cross-validation, and then used with 10-fold cross-fitting for the semiparametric estimation. To estimate $\widehat{\nu}_z(\cdot)$, we use an iterative, instead of nested, form of cross-fitting that sacrifices some independence between folds to be more computationally efficient, described in Section 6 of the Supplementary Material. Nonparametric estimation of the propensity score $e_1(\cdot)$ leads to variability that requires weight clipping to stabilize the semiparametric estimates [31, 53]. We clipped weights worth more than $1/20$ of the total weight of the samples.

In one experiment below, we use a variant of this implementation where we fit $\widehat{e}_{1,k}(\cdot)$ via a simple logistic regression; the logistic regression model for the propensity score is misspecified, so the lower-order statistical bias from the Neyman orthogonality will not hold; the statistical bias of the estimator will depend on the convergence rate of the nonparametric estimator of $\widehat{\theta}_z(\cdot)$, which will not converge sufficiently quickly. As a result, we expect that the statistical bias will dominate the convergence of $\widehat{\tau}^-$ to τ^-.

In the second implementation, we use `xgboost` [9] to fit a machine learning estimator for all of the nuisance parameters, emphasizing the generality and scalability of our methods. `xgboost` is a gradient boosted tree method that performs well with tabular data, despite having little formal theory regarding its convergence guarantees. Therefore, we used the simulations discussed below as a way to assess its appropriateness as a nuisance parameter estimator for our semiparametric method from Section 3.1. In this implementation, we fit the estimator $\widehat{\theta}_z(\cdot)$ to minimize the weighted squared loss (19), and fit the remaining nuisance parameters to minimize the log loss for predicting a binary target (treatments or the targets $\mathbf{1}\{Y_i \geq \theta\}$ for estimating $\nu_z(\cdot)$). As with the previous implementation, $\widehat{\nu}_z(\cdot)$ are fit with the iterative cross-fitting described in Section 6 of the Supplementary Material. Similarly, all tuning parameters (boosting iterations, regularization, subsampling fraction, minimum node size) are selected via 10-fold cross-validation. We found that when estimating a generic nuisance parameter $\eta(\cdot)$, representing either $\theta_z(\cdot)$, $\nu_z(\cdot)$, or $e_1(\cdot)$, adding an additional intercept term as follows improved performance signficantly: After fitting $\widehat{\eta}_z(\cdot)$ using `xgboost`, we fit β_0 in the model $\widehat{\eta}_z(X) + \beta_0$ using the appropriate loss function for the nuisance parameter.

4.2. *Simulations*. The purpose of the simulation study is to demonstrate the good coverage of the proposed confidence intervals for reasonable choices of sample size n and covariate dimension d, and to understand some of the practical properties of the proposed methods relative to existing methods for sensitivity analysis, such as matching methods [47]. In all of the simulations, we generate the data as follows for a randomly chosen set of coefficients β and μ: draw $X \sim \text{Uniform}[0,1]^d$, and conditional on $X = x$, draw

$$U \sim \mathsf{N}\left\{0, \left(1 + \frac{1}{2}\sin(2.5x_1)\right)^2\right\}, \qquad Y(0) = \beta^\top x + U, \qquad Y(1) = \tau + \beta^\top x + U.$$

We draw the treatment assignment according to

$$Z \sim \text{Bernoulli}\left\{\frac{\exp(\alpha_0 + x^\top \mu + \log(\Gamma_{\text{data}})\mathbf{1}\{u > 0\})}{1 + \exp(\alpha_0 + x^\top \mu + \log(\Gamma_{\text{data}})\mathbf{1}\{u > 0\})}\right\},$$

where α_0 is a constant controlling the overall treatment assignment ratio. This model satisfies the Γ_{data}-selection bias condition, since

$$\frac{P(Z=1 \mid X=x, U=u)}{P(Z=0 \mid X=x, U=u)} \frac{P(Z=0 \mid X=x, U=\tilde{u})}{P(Z=1 \mid X=x, U=\tilde{u})} = \Gamma_{\text{data}}^{\mathbf{1}\{u>0\}-\mathbf{1}\{\tilde{u}>0\}} \in \left[\Gamma_{\text{data}}^{-1}, \Gamma_{\text{data}}\right].$$

Across all experiments, we set $\tau = 1$ and $\Gamma_{\text{data}} = \exp(1)$. Unless otherwise stated, we used the same Γ in our sensitivity analysis as the level of confounding Γ_{data} used to generate the data. Here, unobserved confounding inflates estimates that assume unconfoundedness: when $Z = 1$, U is more likely to be positive than when $Z = 0$, which inflates the mean of treated units, that is, $\mathbb{E}[Y(1) \mid Z = 1, X = x] > \mathbb{E}[Y(1) \mid X = x]$. We expect that the upper bound from the sensitivity analysis is above the true ATE, while the lower bound is only slightly below the truth, assuming that we choose $\Gamma \geq \Gamma_{\text{data}}$, but not by too much.

In the first set of simulations, we simulate data with a moderate number of observed covariates ($d = 20$), where we observe the proposed sensitivity analysis procedure quickly approaches its asymptotic behavior as sample size grows. For these simulations, we use the `xgboost` implementation, validating the performance of our semiparametric method when the nuisance parameters are estimated well, even if lacking in formal convergence guarantees.

Table 1 summarizes the empirical performance of the `xgboost` implementation based on 500 simulations. As expected, the average lower bound estimator $\widehat{\tau}^-$ is close to the true ATE, while the average upper bound estimator $\widehat{\tau}^+$ is higher than the true ATE to account for unmeasured confounding. The estimators of the standard errors of $\widehat{\tau}^-$ and $\widehat{\tau}^+$ are fairly accurate when $n \geq 1000$. When n is small, they slightly underestimate the true standard errors. The empirical coverage probability of the confidence interval of ATE is conservative because of unobserved confounding. As the unobserved confounding introduces upward bias, the lower bound $\tau^- \approx \tau$, and we expect that the coverage probability of the confidence interval of τ is close to 97.5% for large n, which is confirmed by the simulation results in Table 1.

In the second set of simulations, the dimension d of the covariates, sample size n and marginal treatment probability $P(Z = 1)$ match those from the real observational study on fish consumption and blood mercury levels in the next subsection ($d = 8$, $n = 1100$, $P(Z = 1) = 0.21$), so that we can validate our approach before interpreting the results on real data. We use the nonparametric sieve implementation for estimating the nuisance parameters in the real observational study, and so we use this implementation here. As estimation with sieves is challenging in this setting due to the eight covariates and a nonlinear model, in Table 2 we observe that the variance estimates $\widehat{\sigma}^{\pm}$ underestimate the standard deviation of $\widehat{\tau}^{\pm}$ by

TABLE 1

Simulation results of the proposed method with 20 *observed covariates.* $\widehat{\tau}^-$, *the empirical average of* $\widehat{\tau}^-$; $\widehat{\sigma}_{\tau^-}$, *the empirical average of* $\widehat{\sigma}_{\tau^-}$; *SD. of* $\widehat{\tau}^-$, *the empirical standard deviation of* $\widehat{\tau}^-$; $\widehat{\tau}^+$, *the empirical average of* $\widehat{\tau}^+$; $\widehat{\sigma}_{\tau^+}$, *the empirical average of* $\widehat{\sigma}_{\tau^+}$; *SD. of* $\widehat{\tau}^+$, *the empirical standard deviation of* $\widehat{\tau}^+$; *and coverage, the empirical coverage probability of the* 95% *confidence intervals* $\widehat{\text{CI}}_\tau$. ($ATE = \tau = 1$ *and* $\Gamma_{\text{data}} = \exp(1)$)

n	$\widehat{\tau}^-$	SD. of $\widehat{\tau}^-$	$\widehat{\sigma}_{\tau^-}$	$\widehat{\tau}^+$	SD. of $\widehat{\tau}^+$	$\widehat{\sigma}_{\tau^+}$	Coverage
500.0	1.008	0.085	0.081	1.424	0.082	0.077	0.952
1000.0	1.000	0.059	0.057	1.404	0.058	0.053	0.978
2000.0	0.998	0.042	0.040	1.395	0.040	0.038	0.966
4000.0	0.995	0.029	0.028	1.387	0.027	0.027	0.980

TABLE 2

Simulation results of the proposed method (parametric and nonparametric) and the existing matching method with eight observed covariates. $\widehat{\tau}^-$, the empirical average of $\widehat{\tau}^-$; $\widehat{\sigma}_{\tau^-}$, the empirical average of $\widehat{\sigma}_{\tau^-}$; SD. of $\widehat{\tau}^-$, the empirical standard deviation of $\widehat{\tau}^-$; $\widehat{\tau}^+$, the empirical average of $\widehat{\tau}^+$; $\widehat{\sigma}_{\tau^+}$, the empirical average of $\widehat{\sigma}_{\tau^+}$; SD. of $\widehat{\tau}^+$, the empirical standard deviation of $\widehat{\tau}^+$; and coverage, the empirical coverage probability of the 95% confidence intervals $\widehat{\text{CI}}_\tau$. ($ATE = \tau = 1$ and $\Gamma_{\text{data}} = \exp(1)$)

Approach	$\widehat{\tau}^-$	$\widehat{\sigma}_{\tau^-}$	SD. of $\widehat{\tau}^-$	$\widehat{\tau}^+$	$\widehat{\sigma}_{\tau^+}$	SD. of $\widehat{\tau}^+$	Coverage
Nonparametric	0.995	0.073	0.081	1.775	0.069	0.076	0.960
Misspecified	0.988	0.071	0.081	1.775	0.068	0.076	0.970
Matching	0.869	–	0.097	2.125	–	0.097	0.996

approximately 10%. We also evaluate the performance when the propensity score estimator is misspecified, as discussed in Section 4.1.

We compare our semiparametric methods to the M-estimator based matching method `sensitivitymw` [47]. Note that our simulation uses a constant treatment effect, as assumed by matching methods. The confidence intervals for the matching approach is conditional on the design (and assumes exact matched pairs), whereas our intervals are unconditional. The confidence intervals for the ATE from the matching method appear conservative, coming from having a lower design sensitivity and larger standard errors (Table 2). The larger standard errors could potentially be reduced using covariate adjustment in matching [44].

In the third set of simulations, we include only a single covariate ($d = 1$), and evaluate the performance of the semiparametric method with the `xgboost` implementation, and the matching method described above over a range of sample sizes. One of the challenges with interpreting the above simulations is that the results will include a mixture of errors—statistical error from having finite observations, and population-level uncertainty on the treatment effect. With one covariate, the semiparametric and approximate matching methods should have a small statistical bias relative to their standard errors, so the average of the point estimates from simulations with a large sample size should approximate the asymptotic sensitivity bounds well. This allows us to compare the asymptotic behavior of the semiparametric method and matching methods, over a variety of values of Γ used in analysis (while holding Γ_{data} used in the data-generation fixed). Like previous settings, Table 3 shows that the bounds from matching are more conservative than the semiparametric approach.

4.3. *Real observational data.* We apply our method to analyzing an observational study to infer the effect of fish consumption on blood mercury levels and compare our result to that of a prior analysis based on covariate matching [58]. The data consist of observations from 2512 adults in the United States who participated in a single cross-sectional wave of the National Health and Nutrition Examination Survey (2013–2014). All participants answered a questionnaire regarding their demographics and food consumption and had their blood mercury concentration measured (data available in the R package CrossScreening).

High fish consumption is defined as individuals who reported > 12 servings of fish or shellfish in the previous month per their questionnaire, low fish consumption as 0 or 1 servings of fish. The outcome of interest is $\log_2$ of total blood mercury concentration (ug/L). The primary objective is to study if fish consumption causes higher mercury concentration. To match prior analysis [58], we excluded one individual with missing education level and seven individuals with missing smoking status from the analysis, and imputed missing income data for 175 individuals using the median income. In addition, we created a supplementary binary covariate to indicate whether the income data were missing. There are a total of 234 treated individuals (those with high fish consumption), 873 control individuals (low

TABLE 3

Simulation results of the proposed method and matching with 1 *observed covariate. For each method,* 0.025*-quantile and average of the lower bound, followed by average and* 0.975*-quantile of the upper bound, and the coverage of the confidence interval are reported. Comparing the average bounds for each method shows that the semiparametric method has a less conservative lower bound as* Γ *varies, but is still below the true ATE when the appropriate* Γ *is used, which is* 1 *in this simulation; the coverage shows that it still covers the true ATE at the appropriate level. Varying the sample size shows that the statistical bias of both methods is already negligible with very small sample sizes.* ($ATE = \tau = 1$ *and* $\Gamma_{\text{data}} = \exp(1)$)

	Semiparametric Method					Matching Method				
	Lower 0.025-quantile	Lower	Upper	Upper 0.975-Quantile	Coverage	Lower 0.025-quantile	Lower	Upper	Upper 0.975-Quantile	Coverage
Γ					Fixing $n = 1000$					
1	1.08	1.18	1.18	1.29	0.06	1.23	1.42	1.42	1.65	0.00
exp(0.5)	1.00	1.09	1.27	1.37	0.56	0.93	1.11	1.73	1.96	0.61
exp(1)	0.90	1.00	1.35	1.46	0.97	0.58	0.80	2.05	2.30	1.00
exp(2)	0.71	0.81	1.52	1.64	1.00	−0.13	0.17	2.69	3.01	1.00
exp(3)	0.51	0.63	1.69	1.82	1.00	−0.90	−0.48	3.35	3.75	1.00
exp(4)	0.30	0.46	1.85	2.01	1.00	−1.67	−1.16	4.02	4.49	1.00
n					Fixing $\Gamma = \exp(1)$ as in simulation					
100.0	0.65	1.00	1.37	1.69	0.97	0.11	0.82	2.05	2.83	0.99
1000.0	0.90	1.00	1.35	1.46	0.97	0.58	0.80	2.05	2.30	1.00
4000.0	0.94	1.00	1.35	1.41	0.98	0.68	0.81	2.05	2.19	1.00

fish consumption). The data include eight covariates (gender, age, income, whether income is missing, race, education, ever smoked, and number of cigarettes smoked last month). Our approach uses the same Γ-selection bias model as the previous matched-pair analysis in [58], so results for our proposed method and the analysis based on these 234 matched pairs are nearly comparable. However, the confidence intervals constructed for matching are conditional on the covariates and choice of matched pairs. As Table 4 shows (see also Figure 2), when $\Gamma > \exp(1)$, our method achieves tighter confidence intervals around the effect of fish consumption on blood mercury level: our confidence intervals are nested within those based on the matching method. For example, when $\Gamma = \exp(3)$ (representing a relatively large selection bias), the 95% confidence interval for the increase in average $\log_2$-transformed blood mercury concentration caused by high fish consumption is $[0.47, 3.29]$ based on our new

TABLE 4

Comparison to sensitivity results of [58] *using the same data set. Because the same sensitivity model as the matched analysis was used, results can be compared directly. We demonstrate that the method can achieve tighter bounds on the average treatment effect both in point estimates and confidence intervals*

	Semiparametric Method					Matching Method				
Γ	Lower 95% CI	Lower	Upper	Upper 95% CI	Length of CI	Lower 95% CI	Lower	Upper	Upper 95% CI	Length of CI
1	1.51	1.74	1.74	1.97	0.46	1.9	2.08	2.08	2.25	0.35
exp(0.5)	1.31	1.53	2.03	2.26	0.95	1.57	1.75	2.41	2.59	1.02
exp(1)	1.07	1.27	2.27	2.47	1.4	1.25	1.45	2.74	2.94	1.89
exp(2)	0.74	0.91	2.77	2.89	2.15	0.58	0.87	3.36	3.65	3.07
exp(3)	0.47	0.6	3.19	3.29	2.82	−0.23	0.28	3.97	4.48	4.71
exp(4)	0.18	0.29	3.55	3.63	3.45	–	–	–	–	–

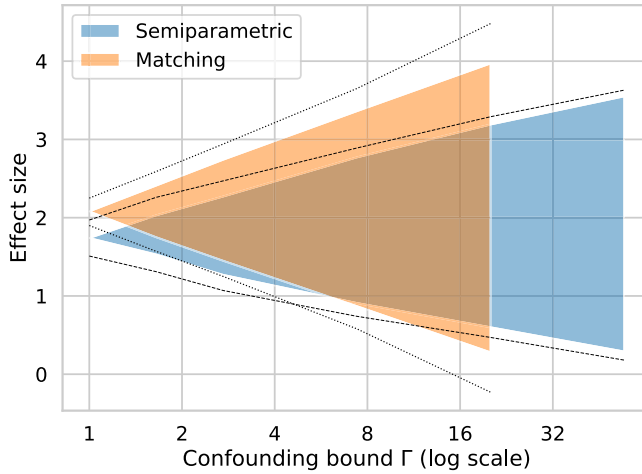

FIG. 2. *Visual comparison to sensitivity results of matching method in* [58] *using the same data set. See numerical details in Table 4. The filled areas represent the estimated bounds on the average treatment effect, whereas the dotted/dashed lines represent their confidence intervals. For values of* Γ *larger than* $\exp(0.5)$, *our approach produces intervals with shorter length.*

method and $[-0.24, 4.48]$ based on the matching method. While the former excludes zero, suggesting a significant association in the presence of unknown confounding, the latter includes the null association and is not statistically significant. The confidence intervals for our method are always shorter except when $\Gamma = 1$, that is, under unconfoundedness.

5. Discussion. The Γ-selection bias model (3) relaxes the unconfoundedness assumption (1) required for the identification of causal treatment effects. We propose estimators $\widehat{\tau}^{\pm}(\cdot)$ for upper and lower bounds on the CATE $\tau(x)$ and $\widehat{\tau}^{\pm}$ for the ATE τ under the Γ-selection bias condition (3) and derive their asymptotic properties. Our loss minimization approach is practical and scalable, allowing the use of flexible machine learning methods. Theoretically, we demonstrate the statistical advantages of our approach, replicating the advantageous $o_p(n^{-p/(2p+d)})$ convergence of series estimation procedures [36] and root n consistency of doubly robust semiparametric estimates [5, 13] in the absence of unobserved confounding (1). Our simulation studies and experimental evidence from real observational data confirm these advantages exist in practical finite sample regimes as well.

Our bounds demonstrate a few important phenomena for understanding the robustness of causal inference with observational data. First, as we note in Section 3, the estimator $\widehat{\tau}^{-}$ reduces to the AIPW estimator (28) when $\Gamma = 1$. Therefore, for any $\Gamma > 1$, the confidence interval for τ estimated in (33) includes the AIPW estimate (28), which serves as the center of the interval bounding the ATE. Second, the estimator $\widehat{\theta}_1(\cdot)$ minimizes a weighted squared error loss function (19), while the estimator $\widehat{\mu}_{1,1}(\cdot)$ minimizes a unweighted mean squared error loss. When the residual noise $Y - \mu_{1,1}(X)$ is small, the difference between weighted and unweighted loss functions also tends to be small. Therefore, the effect of selection bias on the bias of the ATE τ or CATE $\tau(x)$ estimated under the no unobserved confounding assumption (1) depends on the magnitude of these residuals; when these residuals are close to zero, the risk of unobserved confounding is mitigated.

Our bounds on the ATE τ and CATE $\tau(x)$ depend on bounding the conditional mean of the potential outcomes $\mu_1(x) = \mathbb{E}[Y(1) \mid X = x]$ and $\mu_0(x) = \mathbb{E}[Y(0) \mid X = x]$. The proposed $\widehat{\tau}^{-}$ and $\widehat{\tau}^{-}(x)$ employ a worst-case reweighting scheme (such as in (16) and (19)) to bound them separately. Section 3.4 establishes the optimality of this approach under a specific symmetry condition on the distributions of the potential outcomes. In general, our approach may

not be optimal; an optimal estimator may require worst-case treatment assignments that depend on both potential outcomes simultaneously, consistent with the independence assumption (2) and Γ-selection bias condition (3). Such joint consideration of $\mu_1(x)$ and $\mu_0(x)$ complicates the estimation procedure but is an important direction of future research.

In practice, choosing an appropriate level of Γ in the sensitivity analysis is important. Rosenbaum ([43], Chapter 6) discusses using known relationships between a treatment and an auxiliary measured outcome to detect the presence and magnitude of hidden bias. For example, suppose a drug is approved with an unbiased estimate of its effect on a primary outcome based on a randomized clinical trial, and drug surveillance investigates the potential adverse events associated with the drug use in real world. The difference between the estimated treatment effect on the primary outcome based on observational data and that based on a randomized clinical trial can serve as an indication of the magnitude of Γ, the hidden bias in the observational data. It may then be appropriate to perform a sensitivity analysis for adverse events with the same level of Γ. However, in many settings, there is no such surrogate for estimating Γ. In discussions with clinicians who often conduct biomedical studies, we find it helpful to provide results for a number of different values of Γ to help contextualize the strength of evidence, rather than present a single bound with undue certainty. While our result is valid for each fixed Γ, providing uniform inference results over a set of Γ would allow estimation of the smallest value of Γ consistent with zero treatment effect in the data (a *sensitivity value* analogous to the *E value* for risk ratios from VanderWeele and Ding [54]).

APPENDIX A: PROOFS FOR BOUNDS ON THE CATE

A.1. Proof of absolute continuity in Lemma 2.1. PROOF. Here, we only prove the absolute continuity result. The rest of the proof is in Section 2 after the statement of Lemma 2.1. Let $U \in \mathcal{U}$ be the unobserved confounder satisfying $Y(1) \perp\!\!\!\perp Z \mid X, U$ and (3). Then for any set $A \subset \mathcal{U}$,

$$\begin{aligned}&\frac{P(U \in A \mid Z=0, X=x)}{P(U \in A \mid Z=1, X=x)}\\ &= \frac{P(Z=0 \mid X=x, U \in A)}{P(Z=1 \mid X=x, U \in A)} \cdot \frac{P(Z=1 \mid X=x)}{P(Z=0 \mid X=x)} \in \left[\Gamma^{-1}, \Gamma\right]\end{aligned} \tag{39}$$

by condition (3) and the quasiconvexity of the ratio mapping $(a, b) \mapsto a/b$. Letting q_z denote the density of U (with respect to a base measure μ) conditional on $Z = z$, we then have $q_0(u \mid x)/q_1(u \mid x) \in [\Gamma^{-1}, \Gamma]$, and for any measurable set $A \subset \mathbb{R}$,

$$\begin{aligned}&\frac{P(Y(1) \in A \mid Z=0, X=x)}{P(Y(1) \in A \mid Z=1, X=x)}\\ &= \frac{\int P(Y(1) \in A \mid Z=0, U=u, X=x) q_0(u \mid x)\, d\mu(u)}{\int P(Y(1) \in A \mid Z=1, U=u, X=x) q_1(u \mid x)\, d\mu(u)}\\ &\overset{(i)}{=} \frac{\int P(Y(1) \in A \mid U=u, X=x) q_0(u \mid x)\, d\mu(u)}{\int P(Y(1) \in A \mid U=u, X=x) q_1(u \mid x)\, d\mu(u)} \overset{(ii)}{\in} \left[\Gamma^{-1}, \Gamma\right],\end{aligned}$$

where equality (i) is a consequence of $Y(1) \perp\!\!\!\perp Z \mid X, U$, and inequality (ii) follows again from the quasiconvexity of the ratio. This yields the absolute continuity claim. □

A.2. Proof of Lemma 2.2.

PROOF. As everything is conditional on x, we it without loss of generality, letting $\mathbb{E}_1[\cdot] = \mathbb{E}[\cdot \mid Z = 1]$ for shorthand. We first develop a simple duality argument. The set

$$\mathcal{L}_\Gamma := \left\{L : \mathcal{Y} \to \mathbb{R}_+, L \text{ measurable}, L(y) \le \Gamma L(\tilde{y}) \text{ for all } y, \tilde{y}\right\}$$

is convex, contains the constant function $L \equiv 1$ in its interior and for $L \equiv 1$ we have $\mathbb{E}_1[L(Y(1))] = 1$. Thus, strong duality ([32], Theorem 8.6.1 and Problem 8.7) implies

$$\inf_{L\in\mathcal{L}_\Gamma}\{\mathbb{E}_1[L(Y(1))] \mid \mathbb{E}_1[L(Y(1))] = 1\} = \sup_{\mu\in\mathbb{R}}\inf_{L\in\mathcal{L}_\Gamma}\{\mathbb{E}_1[(Y(1)-\mu)L(Y(1))] + \mu\}. \tag{40}$$

Now, we show that for each $\mu \in \mathbb{R}$,

$$L^*(y) \propto \Gamma\mathbf{1}\{y-\mu \le 0\} + \mathbf{1}\{y-\mu > 0\} \tag{41}$$

attains the minimum value of $\inf_{L\in\mathcal{L}_\Gamma}\mathbb{E}_1[(Y(1)-\mu)L(Y(1))]$. That is, the minimizer takes on only the values $L^*(y) \in \{c, c\Gamma\}$ for some $c \ge 0$. The constraint $L \in \mathcal{L}_\Gamma$ guarantees that $L^*(y) \in [c, c\Gamma]$ for some $c \ge 0$. Assume that $c \le L(y) \le c\Gamma$, but $L(y) \notin \{c, c\Gamma\}$. Then letting $L^\star(y) = c$ if $(y-\mu) > 0$ and $L^\star(y) = c\Gamma$ if $(y-\mu) \le 0$, we have $(y-\mu)L^\star(y) \le (y-\mu)L(y)$, with strict inequality if $y \ne \mu$. Thus, any function $L \in \mathcal{L}_\Gamma$ can be modified to be of the form (41) without increasing the objective $\mathbb{E}_1[(Y(1)-\mu)L(Y(1))]$.

Substituting the minimizer (41) into the right objective (40), we recall that $\psi_t(y) = (y-t)_+ - \Gamma(y-t)_-$ to obtain

$$\theta_1(x) = \sup_\mu \inf_{c\ge 0}\{\mathbb{E}_1[c\psi_\mu(Y(1)) \mid X = x] + \mu\}.$$

This gives the final result (18), as

$$\inf_{c\ge 0}\mathbb{E}_1[c\psi_\mu(Y(1)) \mid X = x] = \begin{cases} -\infty & \text{if } \mathbb{E}_1[\psi_\mu(Y(1)) \mid X = x] < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Since $\theta \mapsto \mathbb{E}[\psi_\theta(Y(1)) \mid Z = 1, X]$ is a decreasing function, $\theta_1(X)$ is the only zero crossing of the function for almost every X. □

A.3. Proof of Lemma 2.3.

LEMMA 2.3. *Assume $(t,x) \mapsto \mathbb{E}[\ell_\Gamma(t, Y(1)) \mid X = x, Z = 1]$ is continuous on $\mathbb{R}\times\mathcal{X}$. If $\mathbb{E}_1[\ell_\Gamma(\theta_1(X), Y(1))] < \infty$, then $\theta_1(\cdot)$ solves $\mathbb{E}[\psi_\theta(Y(1)) \mid X = x, Z = 1] = 0$ for almost every x if and only if it solves (19). Such a minimizer $\theta_1(\cdot) : \mathcal{X} \to \mathbb{R}$ exists and is unique up to measure-0 sets.*

PROOF. Let $\overline{\mathbb{R}} = \mathbb{R}\cup\{+\infty\}$. Normal integrand theory ([42], Section 14.D) allows swapping integrals and infimum over measurable mappings. A map $f : \mathbb{R}\times\mathcal{X} \to \overline{\mathbb{R}}$ is a normal integrand if its epigraphical mapping $x \mapsto S_f(x) := \operatorname{epi} f(\cdot; x) = \{(t,\alpha) \in \mathbb{R}\times\mathbb{R} : f(t;x) \le \alpha\}$ is closed-valued and measurable, that is, for $\mathcal{A}$ the Borel sigma-algebra on $\mathbb{R}$, $S_f^{-1}(O) \in \mathcal{A}$ for all open $O \subset \mathbb{R}^2$. We have the following.

LEMMA A.1 (Rockafellar and Wets [42], Theorem 14.60). *If $f : \mathbb{R}\times\mathcal{X} \to \overline{\mathbb{R}}$ is a normal integrand, and $\int_\mathcal{X} f(\theta_1(x); x)\,dP(x) < \infty$ for some measurable θ_1, then*

$$\inf_\theta\left\{\int_\mathcal{X} f(\theta(x); x)\,dP(x) \mid \theta : \mathcal{X} \to \mathbb{R} \text{ measurable}\right\} = \int_\mathcal{X}\inf_{t\in\mathbb{R}} f(t;x)\,dP(x).$$

If this common value is not $-\infty$, a measurable function $\theta^* : \mathcal{X} \to \mathbb{R}$ attains the minimum of the left-hand side iff $\theta^*(x) \in \operatorname{argmin}_{t\in\mathbb{R}} f(t;x)$ for P-almost every $x \in \mathcal{X}$.

Let $f(t,x) := \frac{1}{2}\mathbb{E}_1[(Y(1)-t)_+^2 + \Gamma(Y(1)-t)_-^2 \mid X = x]$. Since $(t,x) \mapsto f(t,x)$ is continuous by assumption, f is a normal integrand [42], Example 14.31. Rewrite the minimization problem (19) using the tower property

$$\inf_\theta\{\mathbb{E}_1[\mathbb{E}_1[\ell_\Gamma(\theta; (X, Y(1)))|X]] = \mathbb{E}_1[f\{\theta(X), X\}] \mid \theta : \mathcal{X} \to \mathbb{R} \text{ measurable}\}.$$

Apply Lemma A.1 to obtain $\theta_1(x) = \operatorname{argmin}_{t\in\mathbb{R}} f(t;x)$. Since $t \mapsto f(t,x)$ is convex, the first- order condition $\frac{d}{dt} f(t;x) = 0$ shows that $\theta_1(x)$ solves $\mathbb{E}_1[\psi_{\theta(x)}(Y(1)) \mid X = x] = 0$. The uniqueness (up to measure-zero transformations) of θ_1 is immediate by the strong convexity of $t \mapsto \ell_\Gamma(t, y)$. $\square$

APPENDIX B: SIEVE ESTIMATION

B.1. Convergence rates for $\widehat{\theta_1}$, the empirical minimizer (21). In this section, we establish asymptotic convergence rates for minimizers $\widehat{\theta}_1(\cdot)$ of (21). We consider two examples to make this concrete.

EXAMPLE 1 (Polynomials). Let $\mathrm{Pol}(J)$ be the space of Jth order polynomials on $[0, 1]$,

$$\mathrm{Pol}(J) := \left\{ [0,1] \ni x \mapsto \sum_{k=0}^{J} a_k x^k : a_k \in \mathbb{R} \right\}.$$

Define the sieve $\Theta_n := \{x \mapsto \Pi_{k=1}^d f_k(x_k) \mid f_k \in \mathrm{Pol}(J_n), k = 1, \ldots, d\}$, for $J_n \to \infty$.

EXAMPLE 2 (Splines). Let $0 = t_0 < \cdots < t_{J+1} = 1$ be knots that satisfy

$$\frac{\max_{0\le j\le J}(t_{j+1} - t_j)}{\min_{0\le j\le J}(t_{j+1} - t_j)} \le c$$

for some $c > 0$. Then the space of rth order splines with J knots is

$$\mathrm{Spl}(r, J) := \left\{ x \mapsto \sum_{k=0}^{r-1} a_k x^k + \sum_{j=1}^{J} b_j (x - t_j)_+^{r-1}, x \in [0,1] : a_k, b_k \in \mathbb{R} \right\}.$$

Define the sieves $\Theta_n := \{x \mapsto f_1(x_1) f_2(x_2) \ldots f_d(x) \mid f_k \in \mathrm{Spl}(r, J_n), k = 1, \ldots, d\}$ for some integer $r \ge \lfloor p \rfloor + 1$ and $J_n \to \infty$.

We require (standard) regularity conditions. Let $\Lambda_c^p(\mathcal{X})$ denote the Hölder class of p-smooth functions, defined for $p_1 = \lceil p \rceil - 1$ and $p_2 = p - p_1$ by

$$\Lambda_c^p(\mathcal{X}) := \left\{ h \in C^{p_1}(\mathcal{X}) : \sup_{\substack{x\in\mathcal{X} \\ \sum_{l=1}^d \alpha_l < p_1}} \left| D^\alpha h(x) \right| + \sup_{\substack{x\ne x'\in\mathcal{X} \\ \sum_{l=1}^d \beta_l = p_1}} \frac{|D^\beta h(x) - D^\beta h(x')|}{\|x - x'\|^{p_2}} \le c \right\},$$

where $C^{p_1}(\mathcal{X})$ denotes the space of p_1-times continuously differentiable functions on $\mathcal{X}$, and $D^\alpha = \frac{\partial^\alpha}{\partial^{\alpha_1}\ldots\partial^{\alpha_d}}$, for any d-tuple of nonnegative integers $\alpha = (\alpha_1, \ldots, \alpha_d)$. We make a few concrete assumptions on smoothness and other properties of parameters of interest.

ASSUMPTION A4. Let $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ be the Cartesian product of compact intervals $\mathcal{X}_1, \ldots, \mathcal{X}_d$, and assume $\theta_1 \in \Lambda_c^p(\mathcal{X}) =: \Theta$ for some $c > 0$.

ASSUMPTION A5. There exists $\sigma^2_{\text{shift}} < \infty$ such that for all $x \in \mathcal{X}$, $\mathbb{E}[\{Y(1) - \theta_1(X)\}^2 \mid Z = 1, X = x] \le \sigma^2_{\text{shift}}$.

ASSUMPTION A6. $P_{X|Z=1}$ has a density $p_1(x)$ with respect to the Lebesgue measure and $0 < \inf_{x\in\mathcal{X}} p_1(x) \le \sup_{x\in\mathcal{X}} p_1(x) < \infty$.

Assumption A4 assumes that $\theta_1(\cdot)$ is in a p-smooth Hölder space. Sufficient conditions for satisfying this assumption include when the conditional mean function $\mu_z(x) = \mathbb{E}[Y(z) \mid X = x]$ is in a p-smooth Hölder space, and the residuals $Y(1) - \mu_1(x)$ are homoskedastic or $Y(1)$ is binary: in these cases, $\theta_1(x)$ is a simple affine transformation of $\mu_1(x)$, preserving its smoothness. Assumption A4 allows for more general models where the residuals may be heteroskedastic but θ_1 is still smooth. Assumption A5 is a standard condition to ensure convergence of the empirical loss function by bounding the second moment. Finally, Assumption A6 asserts that $P_{X|Z=1}$ has upper- and lower-bounded density, so that it is equivalent to the Lebesgue measure on $\mathcal{X}$. This assumption, along with the symmetric assumption on $P_{X|Z=0}$ needed to estimate $\theta_0(\cdot)$, implies strong ignorability, as well as bounds on the marginal density of P_X under the Lebesgue measure. Assumption A6 allows us to relate the $L^2(P)$ norm, $\|\cdot\|_{2,P}$, to the supremum norm, $\|\cdot\|_{\infty,P}$, of $\widehat{\theta}_1 - \theta_1 \in \Lambda_c^p(\mathcal{X})$, which is important for proving the convergence of sieve estimators [10]. Although outside the scope of this paper, adapting other nonparametric estimators such as the partitioning estimates described in Györfi et al. [20] may admit good $\|\cdot\|_{2,P}$ convergence rates without this assumption.

The tradeoff between the random estimation error and approximation precision of the sieve space Θ_n (see Lemma 7.1 in the Supplementary Material) dictates the accuracy of $\widehat{\theta}_1(\cdot)$. The following theorem guarantees that finite-dimensional linear sieves considered yield standard nonparametric rates for estimating $\theta_1(\cdot)$ by balancing different sources of error.

THEOREM B.1. *For $\mathcal{X} = [0,1]^d$, let Θ_n be given by the finite- dimensional linear sieves in Examples 1 or 2 with $J_n \asymp n^{\frac{1}{2p+d}}$. Define $\epsilon_n = (\frac{\log n}{n})^{\frac{p}{2p+d}}$. Let Assumptions A4, A5 and A6 hold, and let $\widehat{\theta}_1$ satisfy*

$$\mathbb{E}_n\big[\ell_\Gamma\big(\widehat{\theta}_1(X), Y(1)\big) \mid Z = 1\big] \le \inf_{\theta \in \Theta_n} \mathbb{E}_n\big[\ell_\Gamma\big(\theta(X), Y(1)\big) \mid Z = 1\big] + O_P\big(\epsilon_n^2\big).$$

Then $\|\widehat{\theta}_1 - \theta_1\|_{2,P_1} = O_P(\epsilon_n)$ and $\|\widehat{\theta}_1 - \theta_1\|_{\infty,P_1} = O_P(\epsilon_n^{\frac{2p}{2p+d}})$.

See the Supplementary Material, Section 7.1, for proof. The key property of the function spaces Θ_n in Examples 1 and 2 is that $\inf_{\theta \in \Theta_n} \|\theta - \theta_1\|_\infty = O(J_n^{-p})$ (cf. [52], Section 5.3.1, or [49], Theorem 12.8), which allows appropriate balance between approximation and estimation error. Similar guarantees hold for wavelet bases and other finite-dimensional sieves [10, 16], allowing generalization of Theorem B.1 beyond the explicit examples provided.

B.2. Convergence rates for $\widehat{\nu}_{1,k}$, the empirical minimizer (35). To show convergence of the empirical minimizer (35), $\widehat{\nu}_{1,k}$, we need two assumptions.

ASSUMPTION A7. There exist $q, r > 0$ and a set $S \subset \Lambda_c^p(\mathcal{X})$ with $\theta_1 \in S$ such that (a) $P(\widehat{\theta}_{1k}^{\nu_1} \in S \mid Z = 1) \to 1$ as $n \to \infty$, (b) for all $\theta \in S$, $x \mapsto P(Y(1) \ge \theta(x) \mid Z = 1, X = x)$ belongs to $\Pi := \Lambda_r^q(\mathcal{X}) \cap \{\nu : \mathcal{X} \to [0,1]\}$.

ASSUMPTION A8. Let S be as in Assumption A7. There is a constant $L_\nu < \infty$ such that for $f, g \in S$,

$$\int \big[P\big(Y(1) \ge f(X) \mid Z = 1, X = x\big) - P\big(Y(1) \ge g(X) \mid Z = 1, X = x\big)\big]^2 \, \mathrm{d}P_1(x) \le L_\nu^2 \|f - g\|_{2,P_1}^2.$$

Assumption A7 ensures that the map $x \mapsto P(Y(1) \geq \theta(x) \mid Z = 1, X = x)$ is sufficiently smooth for functions $\theta(\cdot)$ close to $\theta_1(\cdot)$. In the current case, since $\|\widehat{\theta}_{1k}^{\nu_1} - \theta_1\|_{2,P_1} = o_p(1)$, S in Assumption A7 can be a $\|\cdot\|_{2,P_1}$-neighborhood of $\theta_1 \in \Lambda_c^p(\mathcal{X})$. This condition is necessary, as $\hat{\nu}_{1,k}(x)$ solves an empirical version of the optimization problem (34) using the estimator $\widehat{\theta}_{1k}^{\nu_1}(\cdot)$ instead of the true $\theta_1(\cdot)$. Assumption A8 guarantees that the map $\theta \mapsto P(Y(1) \geq \theta(x) \mid Z = 1, X = x)$ is also sufficiently smooth. A simple sufficient condition for Assumption A8 is that $Y(1) \mid Z = 1, X = x$ has a bounded density for almost every $x \in \mathcal{X}$. If the density in certain regions is not bounded, but we know a priori that $\theta(x) \neq y$ in these regions, then we choose S so that $\widehat{\theta}_{1k}^{\nu_1} \neq y$ in these regions, as well. For instance, if $Y(1) \in \{0, 1\}$, then unless it is deterministic, $\theta_1(x) \in (0, 1)$, so $\theta_1 \in S = \Lambda_r^q(\mathcal{X}) \cap \{f : \mathcal{X} \to (0, 1)\}$, and $P_{Y(1)|X=x,Z=1}$ has a density $p_{Y(1)|X=x,Z=1}(\theta_1(x)) = 0$, which implies Assumption A8.

Under these additional assumptions, the following proposition gives the convergence rate of the proposed sieve estimator. See the proof in the Supplementary Material, Section 7.2.

PROPOSITION B.1. *For $\mathcal{X} = [0, 1]^d$, let Π_n be the finite-dimensional linear sieves considered in Examples 1 or 2. Let $\epsilon_n = (\frac{\log n}{n})^{\frac{q}{2q+d}}$ and $J_n \asymp \epsilon_n^{-1/q}$. Let Assumptions A6, A7 and A8 hold. Assume that $\|\widehat{\theta}_{1k}^{\nu_1} - \theta_1\|_{2,P_1} = O_p(\epsilon_n)$, and let $\widehat{\nu}_{1,k}$ satisfy*

$$\mathbb{E}_{n,2}^{(k)}\big[\bar{\ell}_\Gamma\big(\widehat{\nu}_{1,k}(X), \widehat{\theta}_{1k}^{\nu_1}(X), Y(1)\big)\big] \leq \inf_{\nu \in 1+(\Gamma-1)\Pi_n} \mathbb{E}_{n,2}^{(k)}\big[\bar{\ell}_\Gamma\big(\nu(X), \widehat{\theta}_{1k}^{\nu_1}(X), Y(1)\big)\big] + O_p\big(\epsilon_n^2\big).$$

Then $\|\widehat{\nu}_{1,k} - \nu_1\|_{2,P} = O_p(\epsilon_n)$.

Acknowledgments. We would like to acknowledge useful comments, discussions and feedback from Stefan Wager and Nigam Shah on this work. SY would also like to thank Mike Baiocchi for his discussions and encouragement to study unobserved confounding.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding. SY was partially supported by a Stanford Graduate Fellowship, and HN was partially supported by Samsung Fellowship. JCD was partially supported by NSF CAREER Award CCF-1553086 and the Office of Naval Research Young Investigator Award N00014-19-1-2288.

Research reported in this publication was supported by the National Heart, Lung, and Blood Institute; the National Institute Of Diabetes And Digestive And Kidney Diseases; and the National Institute On Minority Health And Health Disparities of the National Institutes of Health under Award Numbers R01HL144555, R01DK116852, R21MD012867, DP2MD010478 and U54MD010724.

SUPPLEMENTARY MATERIAL

Supplement to "Bounds on the conditional and average treatment effect with unobserved confounding factors" (DOI: 10.1214/22-AOS2195SUPP; .pdf). Supplementary information.

REFERENCES

[1] ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325 https://doi.org/10.1111/j.1468-0262.2006.00655.x

[2] ARONOW, P. M. and LEE, D. K. K. (2013). Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika* **100** 235–240. MR3034337 https://doi.org/10.1093/biomet/ass064

[3] ATHEY, S. and IMBENS, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **113** 7353–7360. MR3531135 https://doi.org/10.1073/pnas.1510489113

[4] ATHEY, S., TIBSHIRANI, J. and WAGER, S. (2019). Generalized random forests. *Ann. Statist.* **47** 1148–1178. MR3909963 https://doi.org/10.1214/18-AOS1709

[5] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

[6] BOSCO, J. L., SILLIMAN, R. A., THWIN, S. S., GEIGER, A. M., BUIST, D. S., PROUT, M. N., YOOD, M. U., HAQUE, R., WEI, F. et al. (2010). A most stubborn bias: No adjustment method fully resolves confounding by indication in observational studies. *J. Clin. Epidemiol.* **63** 64–74.

[7] BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

[8] BRUMBACK, B. A., HERNÁN, M. A., HANEUSE, S. J. P. A. and ROBINS, J. M. (2004). Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Stat. Med.* **23** 749–767.

[9] CHEN, T. and GUESTRIN, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16* 785–794. ACM, New York, NY, USA.

[10] CHEN, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handb. Econom.* **6** 5549–5632.

[11] CHEN, X. and CHRISTENSEN, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics* **188** 447–465. MR3383220 https://doi.org/10.1016/j.jeconom.2015.03.010

[12] CHEN, X. and WHITE, H. (1999). Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Trans. Inf. Theory* **45** 682–691. MR1677026 https://doi.org/10.1109/18.749011

[13] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[14] COLOMA, P. M., TRIFIRÒ, G., SCHUEMIE, M. J., GINI, R., HERINGS, R., HIPPISLEY-COX, J., MAZZAGLIA, G., PICELLI, G., CORRAO, G. et al. (2012). Electronic healthcare databases for active drug safety surveillance: Is there enough leverage? *Pharmacoepidemiol. Drug Saf.* **21** 611–621.

[15] CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Natl. Cancer Inst.* **22** 173–203.

[16] DAUBECHIES, I. (1992). *Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics* **61**. SIAM, Philadelphia, PA. MR1162107 https://doi.org/10.1137/1.9781611970104

[17] FOGARTY, C. B. and SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.* **111** 1820–1830. MR3601738 https://doi.org/10.1080/01621459.2015.1120675

[18] FRANKS, A. M., D'AMOUR, A. and FELLER, A. (2020). Flexible sensitivity analysis for observational studies without observable implications. *J. Amer. Statist. Assoc.* **115** 1730–1746. MR4189753 https://doi.org/10.1080/01621459.2019.1604369

[19] GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414. MR0653512

[20] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer, Berlin.

[21] HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560

[22] HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

[23] HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826 https://doi.org/10.1111/1468-0262.00442

[24] IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

[25] IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **86** 4–29.

[26] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge Univ. Press, New York. MR3309951 https://doi.org/10.1017/CBO9781139025751

[27] KALLUS, N., MAO, X. and ZHOU, A. (2019). Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics* 2281–2290.

[28] KALLUS, N. and ZHOU, A. (2018). Confounding-robust policy improvement. Available at https://papers.nips.cc/paper/2018/hash/3a09a524440d44d7f19870070a5ad42f-Abstract.html.

[29] KENNEDY, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. arXiv preprint. Available at arXiv:2004.14497 [math.ST].

[30] KÜNZEL, S. R., SEKHON, J. S., BICKEL, P. J. and YU, B. (2017). Meta-learners for estimating heterogeneous treatment effects using machine learning. Available at https://www.pnas.org/doi/10.1073/pnas.1804597116.

[31] LEE, B. K., LESSLER, J. and STUART, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE* **6** e18174. https://doi.org/10.1371/journal.pone.0018174

[32] LUENBERGER, D. G. (1969). *Optimization by Vector Space Methods*. Wiley, New York. MR0238472

[33] MIRATRIX, L. W., WAGER, S. and ZUBIZARRETA, J. R. (2018). Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika* **105** 103–114. MR3768868 https://doi.org/10.1093/biomet/asx077

[34] NEWEY, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* **10** 233–253. MR1293201 https://doi.org/10.1017/S0266466600008409

[35] NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349–1382. MR1303237 https://doi.org/10.2307/2951752

[36] NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *J. Econometrics* **79** 147–168. MR1457700 https://doi.org/10.1016/S0304-4076(97)00011-0

[37] NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. *Probab. Stat.* **416**.

[38] NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319. MR4259133 https://doi.org/10.1093/biomet/asaa076

[39] NORTON, E. C., DOWD, B. E. and MACIEJEWSKI, M. L. (2018). Odds ratios-current best practice and use. *JAMA* **320** 84–85. https://doi.org/10.1001/jama.2018.6971

[40] RICHARDSON, A., HUDGENS, M. G., GILBERT, P. B. and FINE, J. P. (2014). Nonparametric bounds and sensitivity analysis of treatment effects. *Statist. Sci.* **29** 596–618. MR3300361 https://doi.org/10.1214/14-STS499

[41] ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (*Minneapolis, MN*, 1997). *IMA Vol. Math. Appl.* **116** 1–94. Springer, New York. MR1731681 https://doi.org/10.1007/978-1-4612-1284-3_1

[42] ROCKAFELLAR, R. T. and WETS, R. J.-B. (1998). *Variational Analysis. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **317**. Springer, Berlin. MR1491362 https://doi.org/10.1007/978-3-642-02431-3

[43] ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR1899138 https://doi.org/10.1007/978-1-4757-3692-2

[44] ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487 https://doi.org/10.1214/ss/1042727942

[45] ROSENBAUM, P. R. (2010). *Design of Observational Studies. Springer Series in Statistics*. Springer, New York. MR2561612 https://doi.org/10.1007/978-1-4419-1213-8

[46] ROSENBAUM, P. R. (2011). A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics* **67** 1017–1027. MR2829236 https://doi.org/10.1111/j.1541-0420.2010.01535.x

[47] ROSENBAUM, P. R. (2014). Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc.* **109** 1145–1158. MR3265687 https://doi.org/10.1080/01621459.2013.879261

[48] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

[49] SCHUMAKER, L. L. (2007). *Spline Functions: Basic Theory*, 3rd ed. *Cambridge Mathematical Library*. Cambridge Univ. Press, Cambridge. MR2348176 https://doi.org/10.1017/CBO9780511618994

[50] SHEN, C., LI, X., LI, L. and WERE, M. C. (2011). Sensitivity analysis for causal inference using inverse probability weighting. *Biom. J.* **53** 822–837. MR2861489 https://doi.org/10.1002/bimj.201100042

[51] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR0594650

[52] TIMAN, A. F. (1963). *Theory of Approximation of Functions of a Real Variable. A Pergamon Press Book*. The Macmillan Company, New York. MR0192238

[53] TSIATIS, A. A. and DAVIDIAN, M. (2007). Comment: Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data [MR2420458]. *Statist. Sci*. **22** 569–573. MR2420466 https://doi.org/10.1214/07-STS227B

[54] VANDERWEELE, T. J. and DING, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Ann. Intern. Med*. **167** 268–274. https://doi.org/10.7326/M16-2607

[55] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc*. **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

[56] WAGER, S. and WALTHER, G. (2015). Adaptive concentration of regression trees, with application to random forests. Available at arXiv:1503.06388 [math.ST].

[57] YADLOWSKY, S., NAMKOONG, H., BASU, S., DUCHI, J. and TIAN, L. (2022). Supplement to "Bounds on the Conditional and Average Treatment Effect with Unobserved Confounding Factors." https://doi.org/10.1214/22-AOS2195SUPP

[58] ZHAO, Q., SMALL, D. S. and BHATTACHARYA, B. B. (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **81** 735–761. MR3997099