Proximal Causal Learning of Conditional Average Treatment Effects

Erik Sverdrup [1] **Yifan Cui** [2]

Abstract

Efficiently and flexibly estimating treatment effect heterogeneity is an important task in a wide variety of settings ranging from medicine to marketing, and there are a considerable number of promising conditional average treatment effect estimators currently available. These, however, typically rely on the assumption that the measured covariates are enough to justify conditional exchangeability. We propose the *P-learner*, motivated by the *R-* and *DR-learner*, a tailored two-stage loss function for learning heterogeneous treatment effects in settings where exchangeability given observed covariates is an implausible assumption, and we wish to rely on proxy variables for causal inference. Our proposed estimator can be implemented by off-the-shelf loss-minimizing machine learning methods, which in the case of kernel regression satisfies an oracle bound on the estimated error as long as the nuisance components are estimated reasonably well.

1. Introduction

The conditional average treatment effect (CATE) measures the net benefit, such as the decrease in blood pressure, a certain subset of a population experiences by being assigned a certain intervention, such as a drug. Obtaining accurate estimates of CATEs is important in order to understand for example which parts of a population should be assigned a treatment, if any. A large body of work has made tremendous advances in designing flexible estimators for the CATE. Examples include Hill (2011); Alaa & van der Schaar (2017); Hahn et al. (2020) for Bayesian approaches, Athey & Imbens (2016); Wager & Athey (2018) for tree-based methods, Johansson et al. (2016); Shalit et al. (2017); Yoon et al. (2018); Shi et al. (2019) for adopting neural networks, and Künzel et al. (2019); Nie & Wager (2021) for combinations thereof.

To identify causal effects, the aforementioned approaches operate under the exchangeability assumption, i.e., the assertion that conditional on observed covariates, the treatment assignment is as good as random. We propose a CATE estimator, which using the framework of Tchetgen Tchetgen et al. (2020), allows one to estimate causal effects in settings where conditional exchangeability *fails*, but one has measured a set of sufficient *proxy* variables. Our practical approach is motivated by the generic Neyman-orthogonal (Chernozhukov et al., 2018a) loss function from Nie & Wager (2021) and Kennedy (2020) that decouples nuisance estimation and CATE estimation into two stages that can be estimated (and tuned with cross-validation) by flexible loss-minimizing machine learning tools, where the latter stage is to first order less sensitive to estimation error arising from the first stage. Our contribution is to extend this flexible CATE estimation strategy to the proximal causal inference framework (Tchetgen Tchetgen et al., 2020). The proposed loss relies on doubly robust scores (Robins et al., 1994; Rotnitzky et al., 1998; Scharfstein et al., 1999; Chernozhukov et al., 2018a; Cui et al., 2023) which can also be re-purposed to enable semi-parametric efficient estimation and inference on lower-dimensional summaries of the CATEs, such as best linear projections (Semenova & Chernozhukov, 2021), or rank-weighted average treatment effects (Yadlowsky et al., 2021).

1.1. Proxy Variables and Unmeasured Confounding

Conditional exchangeability (often also referred to as unconfoundedness (Imbens & Rubin, 2015)), is a crucial identifying assumption that underlies many popular methodologies for estimating causal effects from observational data, including most CATE estimators. Loosely stated, it requires the investigator to have collected a sufficient set of covariates, such that controlling for these, the treatment assignment is as good as random. Given some additional regularity assumptions, this allows the investigator to estimate a difference in potential outcomes, without having access to a randomized control trial (Imbens & Rubin, 2015). Naturally, the quality of the causal estimates hinges on whether the collected covariates sufficiently account for confounding, and a large body of work, going back to for example Rosen-

[1]Graduate School of Business, Stanford University, Stanford, USA. [2]Center for Data Science, Zhejiang University, Hangzhou, China. Correspondence to: Erik Sverdrup <erikcs@stanford.edu>, Yifan Cui <cuiyf@zju.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

baum & Rubin (1983) has pioneered measures for assessing the sensitivity of a causal estimate to this assumption.

The proximal causal inference framework (Tchetgen Tchetgen et al., 2020) departs from this classical approach by instead asking: even with the presence of unmeasured confounding, are the alternative and realistic assumptions that can be made in order to estimate causal effects? The answer is yes: if the investigator has access to auxiliary variables that satisfy certain assumptions. These auxiliary variables are so-called *proxy* variables that augment the set of controls with additional variables that are either treatment-inducing or outcome-inducing. Consider a simple example where we are interested in the causal effect of the treatment A on the outcome Y and have collected a set of covariates L that is related to both the treatment and outcome. Unfortunately, due to the presence of an unmeasured confounder U, conditional exchangeability fails. Figure 1 (top) shows this scenario as a causal DAG (Pearl, 2009). The proximal causal learning framework relaxes the conditional exchangeability assumption and instead operates under the premise that the variables L can be partitioned into three specific groups: common causes of the treatment and outcomes (X), treatment-inducing confounding proxies (Z), and outcome-inducing confounding proxies (W). Figure 1 (bottom) show the covariates L partitioned into these three groups and suggest that with reasonable assumptions on the interdependencies between treatment, outcomes, confounders, and proxies, one may still learn causal effects. The intuition is that we may use this structure to back out the net effect of the unobserved confounder through its relation with Z and W, then remove this confounding bias to arrive at the effect of A on Y. As prudently pointed out in Tchetgen Tchetgen et al. (2020), many observational datasets exhibit a certain structure whereby the data collected was not measured with the precise intent of quantifying a certain source of confounding. Rather, depending on the question the investigator is trying to answer, these collected variables serve as noisy measures of confounding. The promise of proximal causal learning is that with the right assumptions and structure, one may leverage a subset of these noisy variables as "proxies" which serve the purpose of backing out the net effect of the confounder U.

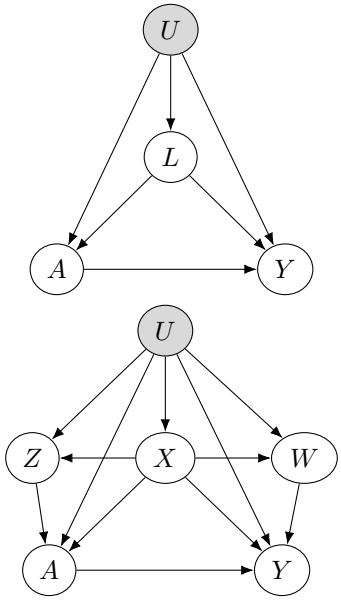

Figure 1. Top: an illustration of a violation of conditional exchangeability due to the presence of the unmeasured confounder U that affects both the treatment A and the outcome Y, even given observed covariates L. Bottom: if one is able to partition L into X (common causes of A and Y), Z (proxies for confounders that affect A), and W (proxies for confounders that affect Y), then these may be used to back out the implied bias caused by U. Note that exchangeability need not hold conditional on observed covariates.

1.2. Previous Work

There is a fast-growing literature on causal inference methods that leverage proxy variables to mitigate confounding (Lipsitch et al., 2010; Kuroki & Pearl, 2014; Deaner, 2018). On a high level, our work leverages the generalizations set forth by Tchetgen Tchetgen et al. (2020) who cast proximal causal learning in the potential outcomes framework and Miao et al. (2018); Cui et al. (2023) who provide nonparametric identification results for average treatment effects. Dukes et al. (2021); Shi et al. (2021b); Shpitser et al. (2021); Ying et al. (2023; 2022) consider identification and estimation for other causal quantities.

A nascent body of work is developing new methods utilizing this framework to answer important questions. Singh (2020) develop kernel methods for nonparametric estimation of for example dose-response curves under unmeasured confounding, Li et al. (2022) use proximal causal learning to estimate vaccine effectiveness, Qi et al. (2023); Shen & Cui (2022) develop individualized treatment allocation rules under unmeasured confounding, and Imbens et al. (2021) propose novel methods for panel data with proxies. Besides causal inference, the core idea of the proximal causal inference framework has also been adopted in survival analysis (Ying, 2022) to address dependent censoring and reinforcement learning (Shi et al., 2021a; Bennett & Kallus, 2021) to tackle a partially observable Markov decision process.

Our practical approach for CATE estimation draws inspiration from Nie & Wager (2021) who cast the problem as a generic two-step loss minimization (the *R-learner*, named so to commemorate the work of Robinson (1988) for the statistical foundation of the approach) that can be implemented by off-the-shelf machine learning methods. The benefit of this decoupling is that it clearly separates the statistical tasks of estimating nuisance components, and that of

estimating treatment effects, which can be implemented and optimized (by standard cross-validation) through different learners (Wu & Yang (2022) extend the *R-learner* to handle data combination from observational and experimental trial data).

The final step of our approach takes the form of a pseudo outcome regression, where transformed outcomes are regressed on covariates, and this approach dates back to van der Laan (2006); Luedtke & van der Laan (2016) who under traditional unconfoundedness suggests it as a method for estimating CATEs, but without explicit error guarantees. Kennedy (2020) and Foster & Syrgkanis (2019) give error guarantees under general assumptions on the nuisance components (when estimated using sample splitting) and derive desirable properties for this approach to CATE estimation (in Section 2.2 we highlight the connection between Kennedy (2020)'s proposed *DR-learner* under unconfoundedness with the proximal *P-learner*, and how our approach can be seen as a generalization of existing methods for CATE estimation under unconfoundedness to proxies). Finally, classical approaches to causal estimation, when unconfoundedness fails, is to rely on instruments, and Syrgkanis et al. (2019) develop a powerful loss-based method to estimate treatment effects conditional on the units complying with the instrument.

2. CATE Estimation under Unmeasured Confounding

2.1. Setup

We operate under the potential outcomes framework and posit the existence of potential outcomes $Y_i(1), Y_i(0)$ corresponding to binary treatment assignment $A_i = \{0, 1\}$. We have access to a collection of covariates X_i that are potential common causes of both the treatment and the outcome and are interested in measuring the conditional average treatment effect (CATE) defined as $\tau^*(X_i) = \mathbb{E}\left[Y_i(1) - Y_i(0)|X_i = x\right]$. We assume that conditioning on the covariates X_i is not sufficient to guarantee conditional exchangeability, due to the presence of latent unmeasured confounders U_i. To identify $\tau^*(x)$ we rely on the existence of proxy variables Z_i and W_i where Z_i are treatment inducing and W_i outcome inducing, which satisfies

Assumption 1. $Y_i \perp Z_i | U_i, X_i, A_i$.

Assumption 2. $W_i \perp (Z_i, A_i) | U_i, X_i$.

Assumption 3. $(Y_i(1), Y_i(0)) \perp A_i | U_i, X_i$.

Assumption 4. *For any square-integrable function g and for any a, x,* $\mathbb{E}\left[g(U_i)|Z_i, A_i = a, X_i = x\right] = 0$ *almost surely if and only if* $g(U_i) = 0$ *almost surely.*

Assumption 5. *For any square-integrable function g and for any a, x,* $\mathbb{E}\left[g(U_i)|W_i, A_i = a, X_i = x\right] = 0$ *almost surely if and only if* $g(U_i) = 0$ *almost surely.*

The last two assumptions are completeness conditions that essentially ensure that the proxy variables carry sufficient information about the confounders U_i. Given consistency assumptions on potential outcomes, positivity assumptions on the treatment assignment A_i, and some technical regularity conditions (Miao et al., 2018; Cui et al., 2023), this setup allows for identification of causal effects through assuming the existence of the following integral equations:

$$\mathbb{E}\left[Y_i|Z_i, A_i, X_i\right] = \int h^*(w, A_i, X_i) dF(w|Z_i, A_i, X_i), \tag{1}$$

and

$$\mathbb{E}\left[q^*(Z_i, a, X_i)|W_i, A_i = a, X_i\right] = \frac{1}{f(A_i = a|W_i, X_i)}, \tag{2}$$

almost surely. In the proximal causal inference literature, Equations (1) and (2) are referred to as bridge functions and characterize a type of inverse problem (known as the Fredholm integral equation) that allows for the identification of counterfactual means with the presence of unmeasured confounders U_i. We defer a discussion of obtaining estimates of h^* and q^* to Section 2.4.

2.2. A Loss Function for Proximal CATEs

To estimate an *average* treatment effect $\mathbb{E}\left[Y_i(1) - Y_i(0)\right]$ under proximal causal learning, motivated by the semiparametric efficient influence function given in Cui et al. (2023), we consider the following doubly robust score,

$$\begin{aligned}\Gamma_i &= (-1)^{1-A_i} q^*(Z_i, A_i, X_i)\left(Y_i - h^*(W_i, A_i, X_i)\right) \\ &+ h^*(W_i, 1, X_i) - h^*(W_i, 0, X_i).\end{aligned} \tag{3}$$

The corresponding semiparametric efficient ATE estimate is the sample average of Γ_i. The key insight is to recognize that to efficiently estimate a *conditional* ATE (CATE), it suffices to learn the mapping from covariates X to pseudo-outcome Γ. This motivates the following procedure:

Algorithm 1. *(P-learner)*

Step 1. Split the data, $i = 1 \ldots n$, into C evenly sized folds. Estimate $h(w, a, x)$ and $q(z, a, x)$ with cross-fitting over the C folds, using tuning as appropriate.

Step 2. Form the scores (3) using cross-fit plug-in estimates of nuisance components $\hat{h}^{(-c(i))}(z, a, x)$ and $\hat{q}^{(-c(i))}(z, a, x)$, where the notation $c(\cdot)$ maps from sample to fold and $(-c(i))$ indicates predictions made without using the i-th sample for training. We estimate treatment effects by minimizing the following empirical loss

$$\hat{\tau}(\cdot) = \arg\min_{\tau} \left[\hat{L}_n(\tau(\cdot))\right], \tag{4}$$

where

$$\hat{L}_n(\tau(\cdot)) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\Gamma}_i^{(-c(i))} - \tau(X_i)\right)^2 \quad (5)$$

and $\hat{\Gamma}_i^{(-c(i))}$ are cross-fit estimates of the scores (3), i.e.,

$$\begin{aligned}\hat{\Gamma}_i^{(-c(i))} &= (-1)^{1-A_i}\hat{q}^{(-c(i))}(Z_i, A_i, X_i) \\ &\times \left(Y_i - \hat{h}^{(-c(i))}(W_i, A_i, X_i)\right) \\ &+ \hat{h}^{(-c(i))}(W_i, 1, X_i) - \hat{h}^{(-c(i))}(W_i, 0, X_i).\end{aligned}$$

The final stage model $\mathbb{E}\left[\hat{\Gamma}_i^{(-c(i))}|X_i = x\right]$ is purely a predictive problem that can leverage flexible non-parametric learners ranging from random forests to neural networks, etc (or combinations in the form of stacking), to simpler parametric models. As shown in Section 3, this approach can deliver accurate estimates of $\tau^*(\cdot)$ even when the nuisance components are subject to estimation error.

An interesting conceptual connection is that under the traditional unconfoundedness assumption, Equation (3) reduces to the celebrated Augmented Inverse-Probability Weighted (AIPW) score of Robins et al. (1994); Robins & Rotnitzky (1995), which forms the basis for the *DR-learner* proposed by Kennedy (2020) where the unconfounded efficient influence function for the ATE is used to learn a CATE function via a similar procedure.

Remark 1. *The empirical loss (5) can be used for learning other estimands of interest. For example, for the conditional average treatment effect on the treated $\mu^*(x) = \mathbb{E}\left[Y_i(1) - Y_i(0)|A_i = 1, X_i = x\right]$, suppose there exists h^* and q^* satisfying*

$$\mathbb{E}\left[Y_i|Z_i, A_i = 0, X_i\right] = \int h^*(w, X_i)dF(w|Z_i, A_i = 0, X_i),$$

and

$$\mathbb{E}\left[q^*(Z_i, X_i)|W_i, A_i = 0, X_i\right] = \frac{f(A_i = 1|W_i, X_i)}{f(A_i = 0|W_i, X_i)}.$$

Motivated by Cui et al. (2023), we define the following loss function,

$$\begin{aligned}\hat{L}_n^{CATOT}(\mu(\cdot)) &= \frac{1}{n}\sum_{i=1}^{n}[A_iY_i \\ &- (1 - A_i)\hat{q}^{(-c(i))}(Z_i, X_i)[Y_i - \hat{h}^{(-c(i))}(W_i, X_i)] \\ &- A_i[\hat{h}^{(-c(i))}(W_i, X_i) + \mu(X_i)]]^2.\end{aligned}$$

The rest of the learning procedure follows Algorithm 1.

2.3. Doubly Robust Scores and Semiparametric Inference

While the aforementioned procedure for learning $\tau(\cdot)$ satisfies desirable oracle properties, it is still a challenging statistical task to conduct pointwise inference on the estimated treatment effects (see for example Armstrong & Kolesár (2018) for fundamental limits to uncertainty quantification for a single point estimated nonparametrically). For this reason, it is often advisable to delegate uncertainty quantification to lower-dimensional summaries of the $\tau(\cdot)$ function. This is the approach advocated by Chernozhukov et al. (2018a) and can be attained by the doubly robust score construction in Equation (3). As a concrete example, if we in *Step 2* fit a linear parametric model, we recover the best linear projection of Semenova & Chernozhukov (2021); Chernozhukov et al. (2018b) that delivers standard errors with nominal coverage (demonstrated as an example in Section 5). Moreover, for assessing if the estimated $\tau(\cdot)$ function actually manages to stratify the population into groups that respond differently to treatment, the same score construction (3) can be used to construct the recently proposed RATE metric of Yadlowsky et al. (2021). We present an example of this approach in Section 5.1.

2.4. Estimating Nuisance Components

As mentioned in Section 2.1, two crucial ingredients for proximal causal learning are the bridge functions h^* and q^*. Equations (1) and (2) define challenging inverse problems known as Fredholm integral equations of the first kind. Timely and pioneering work by Dikkala et al. (2020) gives an empirical strategy for estimating these quantities by regularized minimax estimation, which Ghassami et al. (2022) draws on to propose a flexible kernel machine learning estimator of h^* and q^* (this is also the approach used by Kallus et al. (2021); Mastouri et al. (2021); Qi et al. (2023)).

For the purpose of simulated and practical illustrations of the *P-learner*, we rely on the flexible kernel estimator of Kallus et al. (2021); Ghassami et al. (2022), which can be implemented by off-the-shelf software using Gaussian kernels and tuned with cross-validation. In particular, we consider the following min-max optimization problem,

$$\begin{aligned}\min_{h\in\mathcal{H}}\max_{r\in\mathcal{R}} &\mathbb{P}_n[(\mathbb{I}\{A = a\}Y - \mathbb{I}\{A = a\}h(W, a, X))r(Z, X) \\ &- r^2(Z, X)] - \lambda_r||r||^2_{\mathcal{R}} + \lambda_h||h||^2_{\mathcal{H}}, \\ \min_{q\in\mathcal{Q}}\max_{s\in\mathcal{S}} &\mathbb{P}_n[(1 - \mathbb{I}\{A = a\}q(Z, a, X))s(W, X) \\ &- s^2(W, X)] - \lambda_s||s||^2_{\mathcal{S}} + \lambda_q||q||^2_{\mathcal{Q}},\end{aligned}$$

where $\mathcal{R}, \mathcal{H}, \mathcal{S}$, and $\mathcal{Q}$ are critic classes with norms represented by $||\cdot||_{\mathcal{R}}, ||\cdot||_{\mathcal{H}}, ||\cdot||_{\mathcal{S}}$, and $||\cdot||_{\mathcal{Q}}$, $\mathbb{P}_n$ denotes the sample average with respect to a training sample, and $\lambda^h_Q, \lambda^h_H, \lambda^q_H, \lambda^q_Q$ are tuning parameters. This problem has

a closed-form solution given by Propositions 9 and 10 in Dikkala et al. (2020).

3. Oracle Bound for *P-learner*

Our theory focuses on a *P-learner* based on penalized kernel regression. Regularized kernel learning has been thoroughly studied in the learning literature (Bartlett & Mendelson, 2006; Steinwart & Christmann, 2008; Mendelson & Neeman, 2010) and is also used in the theoretical analysis of the R-learner (Nie & Wager, 2021). Following Nie & Wager (2021), we study $||\cdot||_{\mathcal{C}}$-penalized kernel regression, where $\mathcal{C}$ is a reproducing kernel Hilbert space (RKHS) with a continuous, positive semi-definite kernel function.

Our main goal is to establish error bounds for *P-learner* that only depends on the complexity of $\tau^*(\cdot)$, and that match the error bounds we could achieve if we knew h^* and q^* a priori. We study the following cross-fitted estimator and its oracle analog

$$\hat{\tau}(\cdot) = \arg\min_{\tau} \Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\Gamma}_i^{(-c(i))} - \tau(X_i)\Big)^2 + \lambda_n(||\tau||_{\mathcal{C}}) : ||\tau||_\infty \leq 2M\Big],$$

$$\tilde{\tau}(\cdot) = \arg\min_{\tau} \Big[\frac{1}{n}\sum_{i=1}^{n}(\Gamma_i - \tau(X_i))^2 + \lambda_n(||\tau||_{\mathcal{C}}) : ||\tau||_\infty \leq 2M\Big],$$

respectively, where $\lambda_n(||\tau||_{\mathcal{C}})$ is a properly chosen penalty. Similar to the estimated loss given in (5), we define population and oracle losses

$$L(\tau(\cdot)) = \mathbb{E}\left[(\Gamma_i - \tau(X_i))^2\right],$$

$$\tilde{L}_n(\tau(\cdot)) = \frac{1}{n}\sum_{i=1}^{n}(\Gamma_i - \tau(X_i))^2,$$

respectively. We are interested in the regret bound $R(\tau) = L(\tau(\cdot)) - L(\tau^*(\cdot))$ for our *P-learner* $\hat{\tau}(\cdot)$.

Let

$$\mathcal{C}_\alpha = \{\tau : ||\tau||_{\mathcal{C}} \leq \alpha, ||\tau||_\infty \leq 2M\}$$

denote a radius-α ball of $\mathcal{C}$ capped by $2M$. We denote $\tau_\alpha^* = \arg\min\{L(\tau) : \tau \in \mathcal{C}_\alpha\}$ as the best approximation to τ^* within the working class $\mathcal{C}_\alpha$.

In addition, we also define the population, estimated, and oracle α-regret functions

$$R(\tau,\alpha) = L(\tau) - L(\tau_\alpha^*),$$
$$\hat{R}_n(\tau,\alpha) = \hat{L}_n(\tau) - \hat{L}_n(\tau_\alpha^*),$$
$$\tilde{R}_n(\tau,\alpha) = \tilde{L}_n(\tau) - \tilde{L}_n(\tau_\alpha^*),$$

respectively. Then we are ready to state the following proposition.

Proposition 1. *Suppose Assumptions 1-5 and Equations (1)-(2) hold. Further assume Assumptions 6-9 given in the Appendix hold, then we have*

$$\hat{R}(\tau,\alpha) - \tilde{R}(\tau,\alpha) \leq O_P\Bigg(\alpha^p n^{-1/2} R(\tau,\alpha)^{\frac{1-p}{2}} + \alpha^p \frac{\log(n)}{n^{3/4}} R(\tau,\alpha)^{\frac{1-p}{2}} + \frac{\alpha^p}{n^{3/4}}\sqrt{\log(\frac{\alpha n^{1/(1-p)}}{R(\tau,\alpha)})} R(\tau,\alpha)^{\frac{1-p}{2}} + \frac{\alpha^p}{n}\log(\frac{\alpha n^{1/(1-p)}}{R(\tau,\alpha)}) R(\tau,\alpha)^{\frac{1-p}{2}} + \frac{\alpha^p}{n^{5/4}} R(\tau,\alpha)^{\frac{1-p}{2}}\Bigg),$$

where p is defined in Assumption 6 given in the Appendix.

The proof of Proposition 1 is given in the Appendix. This key result provides the excess error bound for regret $\hat{R}(\tau,\alpha)$ using the cross-fitted learner over the regret bound $\tilde{R}(\tau,\alpha)$ using the oracle learner. By leveraging the above proposition, we have the following theorem.

Theorem 1. *Suppose the conditions of Proposition 1 hold. With a properly chosen penalty $\lambda_n(||\tau||)$, $\hat{\tau}(\cdot)$ satisfies the same regret bound $\tilde{\tau}(\cdot)$, that is, $R(\hat{\tau}) = R(\tilde{\tau}) = O_P(n^{-(1-2\beta)/[p+(1-2\beta)]})$, where β is defined in Assumption 7 given in the Appendix.*

Note that the *P-learner* objective is the following regression: $\hat{\tau}(\cdot) = \arg\min_{\tau\in\mathcal{C}_\alpha} \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{\Gamma}_i^{(-c(i))} - \tau(X_i)\Big)^2$. The proof of Theorem 1 is essentially the same as Theorem 3 of Nie & Wager (2021) and we omit it here. Theorem 1 implies that with penalized kernel regression, the cross-fitted *P-learner* can achieve a similar performance as the oracle learner that knows both confounding bridge functions a priori.

4. A Motivating Example

To illustrate the promise of the *P-learner* we consider a simple motivating example that highlights some salient features (to the best of our knowledge, we are not aware of other current proposals for proximal CATE estimation, making traditional benchmark comparisons challenging). We design a proximal data generating mechanism using the setup from Cui et al. (2023) where we incorporate treatment heterogeneity using the moderately complex CATE function $\tau^*(X) = \exp(X_{(1)}) - 3X_{(2)}$ used in Shen & Cui (2022) to learn proximal treatment regimes and add three additional irrelevant normally distributed covariates X (the complete setup is described in Appendix B). In the left-most plot in Figure 2 we train a *Causal Forest* (Athey et al., 2019), a popular method for estimating CATEs under conditional exchangeability on data with $n_{train} = 4000$ samples and

predict the estimated CATEs on a test set with $n_{test} = 2000$. Since unconfoundedness fails, the point estimates are considerably biased, as seen by estimated CATEs falling above the 45-degree line shown in red.

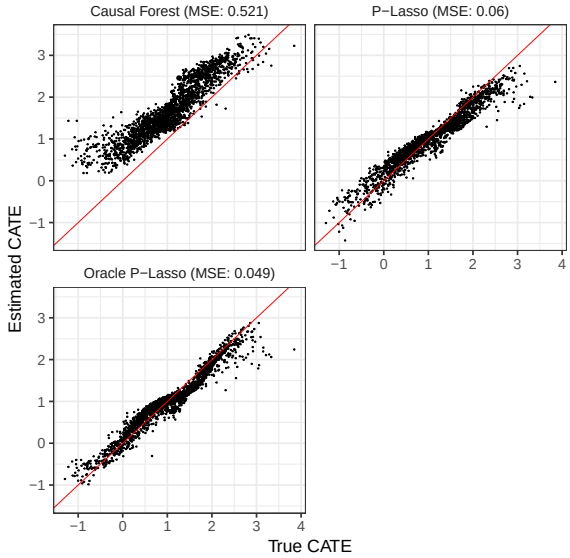

Figure 2. Top left: a *Causal Forest* (Athey et al., 2019) fit on simulated training data ($n_{train} = 4000$ and 5 covariates) where unconfoundedness fails, is used to predict estimated CATEs on a test set ($n_{test} = 2000$). Top right: estimated and true CATEs using the same training and test data with a *P-learner* using cross-validated Lasso (Friedman et al., 2010) and nuisance components h and q estimated with kernels (Ghassami et al., 2022). Bottom left: the same *P-learner* fit using oracle h and q. Mean square error (MSE) is defined as $\frac{1}{n}\sum_{i=1}^{n_{test}}(\hat{\tau}(X_i) - \tau^*(X_i))^2$.

In the right-hand panel in Figure 2 we use *glmnet* (Friedman et al., 2010; R Core Team, 2022) to fit a *P-learner* using cross-validated Lasso (Tibshirani, 1996) on squared and pairwise interactions of a 7-degree natural spline-based expansion on X where we use cross-fit estimates of nuisance components h and q using kernel estimation (Ghassami et al., 2022). This figure suggests the promise of the *P-learner*, as the point estimates are not far off from an oracle learner on h^* and q^* (Figure 2 bottom panel). In Figure 3 we consider two different final stage learners for $\tau^*(\cdot)$: Random Forest (Breiman, 2001), fit using an honest[1] regression forest as implemented in *grf* (Tibshirani et al., 2022) using default tuning parameters, and boosting using *XGBoost* (Chen & Guestrin, 2016) using tuning parameters selected by cross-validation. In this particular example *XGBoost* has slightly more trouble adapting to the $\tau(\cdot)$ signal (choosing a good grid of tuning parameters may sometimes be challenging), though paints a slightly more realistic picture in the sense that the deviation from the oracle MSE might be large for some realizations.

To conclude this section we caution that while the simulation example is intended to bear some resemblance to a real-world scenario where heterogeneity is present, but hidden beneath unmeasured confounding, it is but a toy example that has limited use in helping choose among different *P-learners* in practice. A machine learning algorithm fit to minimize the empirical loss (5) may do very well in minimizing test set error, regardless of whether treatment heterogeneity is actually present or not. Section 5.1 describes our suggested approach for evaluating the *practical* performance of a *P-learner* by using the recently developed RATE metric by Yadlowsky et al. (2021) which under considerable generality can be paired with the proximal doubly robust scores (3) to deliver a test set area-under-the-curve (AUC) measure of heterogeneity along with bootstrapped confidence intervals.[2]

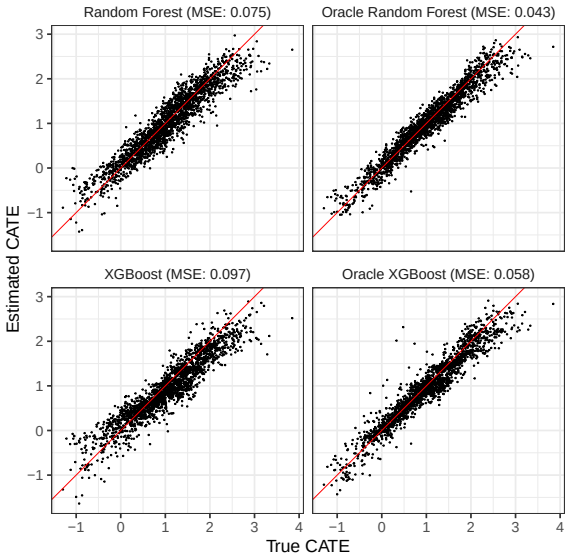

Figure 3. Top: *P-learner* implemented with honest regression forest (with default tuning parameters) (Athey et al., 2019) fit on simulated training data ($n_{train} = 4000$ and 5 covariates) where unconfoundedness fails. The left figure shows predictions on a test set ($n_{test} = 2000$) vs true simulated CATEs when using $\hat{h}$ and $\hat{q}$ (Ghassami et al., 2022) while the right-hand panel uses h^* and q^*. Bottom: *P-learner* implemented with *XGBoost* (Chen & Guestrin, 2016) (with tuning parameters selected by cross-validation). Mean square error (MSE) is defined as $\frac{1}{n}\sum_{i=1}^{n_{test}}(\hat{\tau}(X_i) - \tau^*(X_i))^2$.

[1] A honest regression forest use sample splitting to avoid estimation bias that arises from using the same observations to perform CART splitting as to form the leaf averages. For the example considered here, an honest regression forest performed better than a traditional random forest.

[2] For "real-world" simulation-based approaches to validating estimators see Athey et al. (2021) and for further discussion on the limitations of benchmarking in the context of CATE estimation see Curth et al. (2021).

5. Treatment Heterogeneity in the SUPPORT Study

To illustrate the *P-learner* in action we consider data from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments ("SUPPORT", Connors et al. (1996)), used in a series of papers (Tchetgen Tchetgen et al., 2020; Cui et al., 2023; Qi et al., 2023; Ying et al., 2022) as an example of the proximal causal inference framework. The treatment assignment in this study is a so-called right heart catheterization that was administered to certain patients when admitted to the intensive care unit. The outcome of interest is the number of days between admission, and death or censoring at 30 days. The SUPPORT study did not randomize treatment, however, an informative set of covariates were collected for each of the patients that were administered treatment (2184 patients) or control (3551 patients). Connors et al. (1996)'s original analysis concluded that heart catheterization was on average harmful to patient health. To account for latent confounding due to measurement error in the patient's physiological variables, Tchetgen Tchetgen et al. (2020) reanalyze the data by using proxy variables $Z = (\text{pafi1}, \text{paco21})$ and $W = (\text{ph1}, \text{hema1})$, a set of physiological status measures, and find evidence (using parametric modeling) that the harmful effect is even larger (ATE = -1.80 days) than previously reported. To investigate the possibility of heterogeneity in this harmful effect we fit *P-learners* using the covariates (age, sex, cat1_coma, cat2_coma, dnr1, surv2md1, aps1) also considered in Ghassami et al. (2022) for illustrating the non-parametric kernel estimation of h and q. Among those variables, cat1_coma, cat2_coma, and dnr1 are indicator variables indicating coma and "Do Not Resuscitate" status, while surv2md1 and aps1 are estimates of 2-month survival and severity-of-disease score respectively.

As a first step, we consider a linear CATE model, Table 1 show estimates of the best linear projection

$$\{\beta_0^*,\, \beta^*\} = \arg\min_{\beta_0,\beta} \mathbb{E}\left[(\tau^*(X_i) - \beta_0 - X_i\beta)^2\right], \quad (6)$$

using cross-fit estimates of h and q on all 5735 units to form the proximal scores (3). The second column of Table 1 is simply the projection onto a constant and recovers an ATE in line with the estimates from Ghassami et al. (2022) and Tchetgen Tchetgen et al. (2020).

Next, we consider the three *P-learners* described in Section 4. For the Lasso learner, we use the same spline-based featurization for the continuous covariates X_i and just interactions for the binary X_i. For all learners, we fit h and q using the non-parametric kernel estimator of Ghassami et al. (2022) (using their hyperparameters suggested for this dataset).

Table 1. Best linear projection (BLP) of the CATEs on covariates considered in Ghassami et al. (2022) for the SUPPORT data, as well as a doubly robust estimate of the average treatment effect (ATE), the BLP on a constant. All 5735 units are used to form cross-fit estimates of nuisance parameters h and q, using the kernel method of Ghassami et al. (2022) with associated suggested hyperparameters for this study. HC_3 (MacKinnon & White, 1985) standard errors are in parentheses.

	(BLP)	(ATE)
(Intercept)	4.55 (2.41)	-1.66 (0.27)***
age	-0.06 (0.02)***	
sex	-1.03 (0.54)	
cat1_coma	-0.28 (1.16)	
cat2_coma	3.08 (2.22)	
dnr1	1.00 (0.87)	
surv2md1	-3.93 (1.88)*	
aps1	0.01 (0.02)	

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

5.1. Assessing Treatment Heterogeneity with RATE

As pointed out in Section 4, a fundamental challenge with evaluating CATE estimators on real-world data is the lack of ground truth for calculating traditional error metrics, as treatment effects are fundamentally unobservable. Recent developments in the statistics literature offer an attractive solution to this problem in the form of a family of metrics called rank-weighted average treatment effects (RATE). A RATE metric's fundamental ingredient is a calibration curve inspired by the ROC, called the Targeting Operator Characteristic (TOC), defined as:

$$\text{TOC}(q) = \mathbb{E}\left[Y_i(1) - Y_i(0) | \hat{\tau}(X_i) \geq F^{-1}_{\hat{\tau}(X_i)}(1-q)\right] - \mathbb{E}\left[Y_i(1) - Y_i(0)\right],$$

and is a curve that ranks all observations on a test set according to $\hat{\tau}(X_i)$ estimated from a training set, and compares the ATE for the top q-th fraction of units prioritized by $\hat{\tau}(X_i)$ to the overall ATE. The RATE is the area under this curve and measures how well the CATE estimator stratifies the population in terms of units that benefit differently from treatment. If there is significant heterogeneity present, and the CATE estimator detects it, then the estimated RATE should be significantly different from zero. When selecting among CATE estimators, the one with the largest RATE metric is the best performing in the sense that it most successfully manages to stratify test set subjects according to different treatment benefits.

Inferential properties of the RATE extend to the proximal setting through the doubly robust proximal scores (3) (Yadlowsky et al., 2021, Theorem 4) which satisfy $\mathbb{E}\left[\hat{\Gamma}_i | X_i\right] \approx \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i\right]$ and motivates the following Algo-

rithm 2 to evaluate a *P-learner*.

Algorithm 2. *(Evaluate P-learner)*

Step 1. Randomly partition the data into a training and evaluation set.

Step 2. Using Algorithm 1, learn a $\tau(\cdot)$ function on the training data.

Step 3. Estimate cross-fit doubly robust scores (3) on the evaluation data and predict $\hat{\tau}(X_{evaluation})$ using $\tau(\cdot)$ learned in Step 2. Use these to compute the TOC and the area under the TOC.

Table 2 show estimated RATEs using three *P-learners* and indicate that all procedures learn a CATE function that manages to stratify units on a test set[3]. The largest estimated RATE is obtained by the Random Forest-based *P-learner* and Figure 4 shows the corresponding TOC curve. The RATE is the area under this curve and has a point estimate of -0.79 with a bootstrapped standard error of 0.31, and suggests there is heterogeneity in the response to right heart catheterization. For example, from Figure 4, the patients in the lowest 20-th quantile of estimated CATEs, die 3 days sooner than on average when administered right heart catheterization, indicating a considerably more harmful effect for certain parts of the population than previously reported for the population as a whole.

Table 2. Estimated RATEs and bootstrapped standard errors using CATE estimates obtained by training *P-learners* (selecting tuning-parameters with cross-validation) on a random half-sample of the SUPPORT study (with cross-fit kernel estimates of h^* and q^* with hyperparameters suggested in Ghassami et al. (2022)). Proximal doubly robust scores for evaluation are estimated on the held-out evaluation set.

	Lasso	Random Forest	XGBoost
AUTOC	-0.79	-1.07	-0.72
Std.err	(0.39)	(0.31)	(0.34)

6. Discussion and Extensions

Detecting and measuring treatment effect heterogeneity is an important task across a wide range of domains, and in many observational applications, the investigator does not have access to an unconfounded treatment assignment. We have shown how powerful approaches to CATE estimation under unconfoundedness that rely on a Neyman-orthogonal loss can be extended to the seminal proximal causal learning framework of Tchetgen Tchetgen et al. (2020).

[3]Since the ATE is negative we form the TOC by conditioning on the most negative CATEs first.

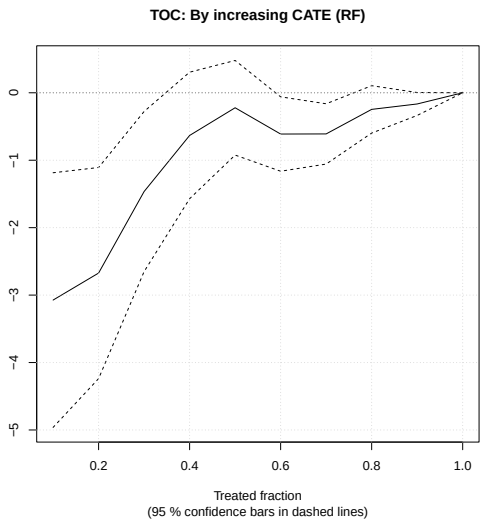

Figure 4. The Targeting Operator Characteristic (TOC) for CATEs estimated with a random forest-based *P-learner* fit on the SUPPORT data, using half of the dataset for CATE training, and the other half for evaluating the RATE. The area under the TOC curve (AUTOC) is -1.07 with a bootstrapped standard error of 0.31.

A natural extension is to ask for the CATE conditional on proxy variables Z_i and W_i: $\tau^*(x,z) = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x, Z_i = z\right]$ and $\tau^*(x,w) = \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x, W_i = w\right]$. Given doubly robust scores for $\tau^*(x,z)$ and $\tau^*(x,w)$, these could in principle be used in place of (3) in the *P-learner*. The derivation of these is an active research area (Tchetgen Tchetgen et al.)

Finally, a challenge with proximal causal learning is the estimation of the confounding bridge functions h^* and q^*. Kallus et al. (2021); Ghassami et al. (2022) make tremendously useful contributions in designing flexible regularized Gaussian kernel estimators for these components. Scaling this approach to large datasets, however, is challenging as kernel estimators typically require work on the order of $O(n^3)$. As pointed out by Dikkala et al. (2020), some promising approximation schemes such as Nyström's method can bring this down to the order of $O(n^2)$. Kompa et al. (2022) take a neural network approach to bridge function estimation and avoids the reliance on kernel methods. Alternate approaches like this may suggest other avenues for scaling. One example could be the connection to the general minimax optimization problem (Dikkala et al., 2020): this is a type of adversarial estimation problem that is similar to the ones encountered and successfully solved at large scales with Generative Adversarial Networks (Goodfellow et al., 2014; Arjovsky et al., 2017), and which have promising adaptions for certain statistical tasks (Kaji et al., 2020; Chernozhukov et al., 2020).

Acknowledgements

We are grateful to the ICML reviewers, as well as seminar participants at Stanford University, for their helpful and insightful comments. Yifan Cui gratefully acknowledges funding from the National Natural Science Foundation of China.

References

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems*, 30, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Armstrong, T. B. and Kolesár, M. Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683, 2018.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. Using Wasserstein Generative Adversarial networks for the design of Monte Carlo simulations. *Journal of Econometrics*, 2021.

Bartlett, P. L. and Mendelson, S. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006.

Bennett, A. and Kallus, N. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018a.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018b.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. Adversarial estimation of Riesz representers. *arXiv preprint arXiv:2101.00009*, 2020.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F. E., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., and Fulkerson, W. J. The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897, 1996.

Cucker, F. and Smale, S. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, pp. 1–12, 2023.

Curth, A., Svensson, D., Weatherall, J., and van der Schaar, M. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Deaner, B. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

Dukes, O., Shpitser, I., and Tchetgen Tchetgen, E. J. Proximal mediation analysis. *arXiv preprint arXiv:2109.11904*, 2021.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Ghassami, A., Ying, A., Shpitser, I., and Tchetgen Tchetgen, E. Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 7210–7239. PMLR, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Imbens, G., Kallus, N., and Mao, X. Controlling for unmeasured confounding in panel data using minimal bridge functions: From two-way fixed effects to factor models. *arXiv preprint arXiv:2108.03849*, 2021.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029. PMLR, 2016.

Kaji, T., Manresa, E., and Pouliot, G. An adversarial approach to structural estimation. *arXiv preprint arXiv:2007.06169*, 2020.

Kallus, N., Mao, X., and Uehara, M. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Kompa, B., Bellamy, D. R., Kolokotrones, T., Robins, J. M., and Beam, A. L. Deep learning methods for proximal inference via maximum moment restriction. *arXiv preprint arXiv:2205.09824*, 2022.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Li, K. Q., Shi, X., Miao, W., and Tchetgen Tchetgen, E. Double negative control inference in test-negative design studies of vaccine effectiveness. *ArXiv*, 2022.

Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383, 2010.

Luedtke, A. R. and van der Laan, M. J. Super-learning of an optimal dynamic treatment rule. *The International Journal of Biostatistics*, 12(1):305–332, 2016.

MacKinnon, J. G. and White, H. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3): 305–325, 1985.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pp. 7512–7523. PMLR, 2021.

Mendelson, S. and Neeman, J. Regularization in kernel learning. *The Annals of Statistics*, 38(1):526–565, 2010.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Pearl, J. *Causality*. Cambridge university press, 2009.

Qi, Z., Miao, R., and Zhang, X. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, pp. 1–14, 2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

Robins, J. M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Rosenbaum, P. R. and Rubin, D. B. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339, 1998.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

Semenova, V. and Chernozhukov, V. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 24(2):264–289, 2021.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Shen, T. and Cui, Y. Optimal individualized decision-making with proxies. *arXiv preprint arXiv:2212.09494*, 2022.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32, 2019.

Shi, C., Uehara, M., and Jiang, N. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. *arXiv preprint arXiv:2111.06784*, 2021a.

Shi, X., Miao, W., Hu, M., and Tchetgen Tchetgen, E. Theory for identification and inference with synthetic controls: a proximal causal inference framework. *arXiv preprint arXiv:2108.13935*, 2021b.

Shpitser, I., Wood-Doughty, Z., and Tchetgen Tchetgen, E. J. The proximal id algorithm. *arXiv preprint arXiv:2108.06818*, 2021.

Singh, R. Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*, 2020.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems*, 32, 2019.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., and Wright, M. *grf: Generalized Random Forests*, 2022. URL https://github.com/grf-labs/grf. R package version 2.2.1.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

van der Laan, M. J. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Wu, L. and Yang, S. Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pp. 904–926. PMLR, 2022.

Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*, 2021.

Ying, A. Proximal identification and estimation to handle dependent right censoring for survival analysis. *arXiv preprint arXiv:2208.07014*, 2022.

Ying, A., Cui, Y., and Tchetgen Tchetgen, E. J. Proximal causal inference for marginal counterfactual survival curves. *arXiv preprint arXiv:2204.13144*, 2022.

Ying, A., Miao, W., Shi, X., and Tchetgen Tchetgen, E. J. Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 03 2023.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

A. Proof of Proposition 1

Let P be a non-negative measure over the compact metric space $\mathcal{X} \subset \mathbb{R}^d$, and K be a kernel with respect to P. Let $T_K : L_2(P) \to L_2(P)$ be defined as $T_K(f)(\cdot) = \mathbb{E}\left[K(\cdot, X)f(X)\right]$. By Mercer's theorem (Cucker & Smale, 2002), there is an orthonormal basis of eigenfunctions $(\psi_j)_{j=1}^{\infty}$ of T_K with corresponding eigenvalues $\{\sigma_j\}_{j=1}^{\infty}$ such that $K(x, y) = \sum_{j=1}^{\infty} \sigma_j \psi_j(x)\psi_j(y)$.

We consider the function $\phi : \mathcal{X} \to l_2$ defined by $\phi(x) = \{\sqrt{\sigma_j}\psi_j(x)\}_{j=1}^{\infty}$. Following Mendelson & Neeman (2010), we define the RKHS $\mathcal{C}$ to be the image of l_2, that is, for every $t \in l_2$, we define the corresponding element in $\mathcal{C}$ by $f_t(x) = \langle \phi(x), t\rangle$, with the induced inner product $\langle f_s, f_t\rangle_{\mathcal{C}} = \langle t, s\rangle$.

Assumption 6. *Without loss of generality, we assume $K(x, x) \leq 1$ for any $x \in \mathcal{X}$. We further assume that for $0 < p < 1$, $\sup_{j\geq 1} j^{1/p}\sigma_j = G_1 < \infty$ and $\sup_j ||\psi_j||_\infty \leq G_2 < \infty$ for some constants G_1 and G_2.*

Assumption 7. *We assume that $||T_K^{\beta}(\tau^*(\cdot))||_{\mathcal{C}} < \infty$ for some $0 < \beta < 1/2$.*

Assumption 8. *For any a, we have that $\mathbb{E}\left[[\hat{h}(W, a, X) - h^*(W, a, X)]^2\right] = o(n^{-1/2})$, and $\mathbb{E}\left[[\hat{q}(Z, a, X) - q^*(Z, a, X)]^2\right] = o(n^{-1/2})$.*

Assumption 9. *We assume that $|Y_i| \leq M$, $\sup_{w,a,x} |h(w, a, x)| \leq M$, and $\sup_{z,a,x} |q(z, a, x)| \leq M$.*

To give some intuition on the role of p and β, $0 < \beta < 1/2$ essentially quantifies the amount of smoothing needed for τ^* to have finite $\mathcal{C}$-norm, where 0 corresponds to $||\tau^*||_{\mathcal{C}} < \inf$ and $1/2$ would mean τ^* is square-integrable. p captures how much the τ function can oscillate through the eigenvalues of the corresponding RKHS. If $\beta, p \to 0$, then Theorem 1 yields the familiar result that 4-th root nuisance rates are sufficient to yield $\sqrt{n}$ rates for a single target parameter.

Proof. We have the following decomposition,

$$
\begin{aligned}
&\hat{L}(\tau) - \hat{L}(\tau_\alpha^*) - \tilde{L}(\tau) + \tilde{L}(\tau_\alpha^*) \\
=&\frac{2}{n}\sum_{i=1}^{n} -A_i[\hat{h}(W_i, A_i, X_i) - h^*(W_i, A_i, X_i)][\hat{q}(Z_i, A_i, X_i) - q^*(Z_i, A_i, X_i)](\tau_h^*(X_i) - \tau(X_i)) \\
&+ A_i[\hat{q}(Z_i, A_i, X_i) - q^*(Z_i, A_i, X_i)][Y_i - h^*(W_i, A_i, X_i)](\tau_\alpha^*(X_i) - \tau(X_i)) \\
&+ [\hat{h}(W_i, 1, X_i) - h^*(W_i, 1, X_i)][1 - A_i q^*(Z_i, A_i, X_i)]](\tau_\alpha^*(X_i) - \tau(X_i)).
\end{aligned}
$$

By the Cauchy Schwarz inequality, the first term is bounded by

$$
\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n} - A_i[\hat{h}(W_i, A_i, X_i) - h^*(W_i, A_i, X_i)][\hat{q}(Z_i, A_i, X_i) - q^*(Z_i, A_i, X_i)](\tau_\alpha^*(X_i) - \tau(X_i)) \\
&\leq 2\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{h}(W_i, A_i, X_i) - h^*(W_i, A_i, X_i)]^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[\hat{q}(Z_i, A_i, X_i) - q^*(Z_i, A_i, X_i)]^2} \times ||\tau_\alpha^* - \tau||_\infty \\
&\leq \alpha^p R(\tau, \alpha)^{\frac{1-p}{2}} O_p(n^{-1/2}),
\end{aligned}
$$

where the last inequality holds by the following fact

$$
||\tau||_\infty \leq ||\tau||_{\mathcal{C}}^p ||\tau||_{L_2(P)}^{1-p},
$$

as provided in Lemma 5.1 of Mendelson & Neeman (2010).

Next, we consider the second term, and the third term can be bounded in a similar manner. For the c-th fold, we define

$$
\eta_c(\tau, \alpha) = \frac{2}{n_c}\sum_{i=1}^{n_c} A_i[\hat{q}^c(Z_i, A_i, X_i) - q^*(Z_i, A_i, X_i)][Y_i - h^*(W_i, A_i, X_i)](\tau_\alpha^*(X_i) - \tau(X_i)),
$$

where to ease notation, we use superscript c to denote $(-c(i))$ and n_c to denote the sample size of c-th fold. Note that to bound $\sup_{\tau\in\mathcal{C}_\alpha}\{\eta_c(\tau, \alpha)\}$, we essentially need to bound $\sup_{\tau\in\mathcal{C}_\alpha}\{\eta_c(\tau, \alpha) : ||\tau - \tau_\alpha^*||_{L_2(P)} \leq L\}$.

We denote the samples that are not included in the c-th fold by I^c. By cross-fitting, we have that

$$\begin{aligned}
&\mathbb{E}\left[\eta_c(\tau,\alpha)|I^c\right] \\
=&\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[A_i[\hat{q}^c(Z_i,A_i,X_i)-q^*(Z_i,A_i,X_i)][Y_i-h^*(W_i,A_i,X_i)](\tau_\alpha^*(X_i)-\tau(X_i))|I^c] \\
=&\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbb{E}[A_i[\hat{q}^c(Z_i,A_i,X_i)-q^*(Z_i,A_i,X_i)][Y_i-h^*(W_i,A_i,X_i)](\tau_\alpha^*(X_i)-\tau(X_i))|Z_i,A_i,X_i,I^c]I^c] \\
=&\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[A_i[\hat{q}^c(Z_i,A_i,X_i)-q^*(Z_i,A_i,X_i)]\mathbb{E}\left[Y_i-h^*(W_i,A_i,X_i)|Z_i,A_i,X_i\right](\tau_\alpha^*(X_i)-\tau(X_i))|I^c] \\
=&0.
\end{aligned}$$

Following Lemma 5 of Nie & Wager (2021), we further have that

$$\frac{\mathbb{E}\left[\sup_{\tau\in\mathcal{C}_\alpha}\{\eta_c(\tau,\alpha):||\tau-\tau_\alpha^*||_{L_2(P)}\leq L\}|I^c\right]}{\mathbb{E}\left[(A_i[\hat{q}^c(Z_i,A_i,X_i)-q^*(Z_i,A_i,X_i)])^2\right]^{1/2}}=O_p(\alpha^pL^{1-p}\frac{\log(n)}{n^{1/2}}),$$

and therefore, we have

$$\mathbb{E}\left[\sup_{\tau\in\mathcal{C}_\alpha}\{\eta_c(\tau,\alpha):||\tau-\tau_\alpha^*||_{L_2(P)}\leq L\}|I^c\right]=O_p(\alpha^pL^{1-p}\frac{\log(n)}{n^{3/4}}).$$

Note that for some constants $C_1,C_2>0$,

$$\sup_{\tau\in\mathcal{C}_\alpha}\{||A_i[\hat{q}^c(\cdot)-q^*(\cdot)][Y_i-h^*(\cdot)](\tau_\alpha^*(\cdot)-\tau(\cdot))||_\infty:||\tau-\tau_\alpha^*||_{L_2(P)}\leq L\}\leq C_1\alpha^pL^{1-p},$$

and

$$\begin{aligned}
&\sup_{\tau\in\mathcal{C}_\alpha}\{\mathbb{E}[\{A_i[\hat{q}^c(Z_i,A_i,X_i)-q^*(Z_i,A_i,X_i)][Y_i-h^*(W_i,A_i,X_i)](\tau_\alpha^*(X_i)-\tau(X_i))\}^2]:||\tau-\tau_\alpha^*||_{L_2(P)}\leq L\} \\
&\leq C_2\alpha^{2p}L^{2(1-p)}n^{-1/2}.
\end{aligned}$$

Then by Talagrand's concentration inequality, for any fixed $\alpha,L,\epsilon>0$,

$$\sup_{\tau\in\mathcal{C}_\alpha}\{\eta_c(\tau,\alpha)|I^c:||\tau-\tau_\alpha^*||_{L_2(P)}\leq L\}\leq O\left(\alpha^pL^{1-p}\frac{\log(n)}{n^{3/4}}+\frac{\alpha^pL^{1-p}}{n^{3/4}}\sqrt{\log(\frac{1}{\epsilon})}+\alpha^pL^{1-p}\frac{1}{n}\log(\frac{1}{\epsilon})\right)$$

holds with probability larger than $1-\epsilon$.

By a similar construction of Nie & Wager (2021), we have the following bound for any $\alpha>1$ and $L\leq 4M$.

$$\begin{aligned}
\sup_{\tau\in\mathcal{C}_\alpha}\{\eta_c(\tau,\alpha):||\tau-\tau_\alpha^*||_{L_2(P)}\leq L\}\leq O_P\Bigg(&\alpha^pL^{1-p}\frac{\log(n)}{n^{3/4}}+\frac{\alpha^pL^{1-p}}{n^{3/4}}\sqrt{\log(\frac{\alpha n^{1/(1-p)}}{L^2})} \\
&+\alpha^pL^{1-p}\frac{1}{n}\log(\frac{\alpha n^{1/(1-p)}}{L^2})+\frac{\alpha^pL^{1-p}}{n^{5/4}}\Bigg).
\end{aligned}$$

Finally, by combining three terms, we have that

$$\begin{aligned}
\hat{R}(\tau,\alpha)-\tilde{R}(\tau,\alpha)\leq O_P\Bigg(&\alpha^pn^{-1/2}R(\tau,\alpha)^{\frac{1-p}{2}}+ \\
&\alpha^p\frac{\log(n)}{n^{3/4}}R(\tau,\alpha)^{\frac{1-p}{2}}+\frac{\alpha^p}{n^{3/4}}\sqrt{\log(\frac{\alpha n^{1/(1-p)}}{R(\tau,\alpha)})}R(\tau,\alpha)^{\frac{1-p}{2}} \\
&+\frac{\alpha^p}{n}\log(\frac{\alpha n^{1/(1-p)}}{R(\tau,\alpha)})R(\tau,\alpha)^{\frac{1-p}{2}}+\frac{\alpha^p}{n^{5/4}}R(\tau,\alpha)^{\frac{1-p}{2}}\Bigg),
\end{aligned}$$

which completes the proof. □

B. Illustrative Data Generating Process

We adopt the setup from Cui et al. (2023). Covariates X are drawn from a normal distribution $N(0, 0.25I_{d\times d})$ where $I_{d\times d}$ is the identity matrix and $d = 5$; A is drawn from a Binomial distribution with success probability $(1 + \exp((0.125, 0.125, 0, 0, 0)^T X))^{-1}$; Z, W, U are drawn from a multivariate normal:

$$(Z, W, U)|A, X \sim MVN\left(\mu = \begin{bmatrix} 0.25 + 0.25A + (0.25, 0.25, 0, 0, 0)^T X \\ 0.25 + 0.125A + (0.25, 0.25, 0, 0, 0)^T X \\ 0.25 + 0.25A + (0.25, 0.25, 0, 0, 0)^T X \end{bmatrix}, \Sigma = \begin{bmatrix} 1 & 0.25 & 0.5 \\ 0.25 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}\right);$$

and Y is drawn from a normal with distribution with $\sigma = 0.25$ and conditional mean

$$\mathbb{E}\left[Y|W, U, A, Z, X\right] = 2 + \tau(X)A + (0.25, 0.25, 0, 0, 0)^T X + 2\mathbb{E}\left[W|U, X\right] + 2W,$$

where $\mathbb{E}\left[W|U, X\right] = 0.25 + (0.25, 0.25, 0, 0, 0)^T X + 0.5(U - 0.25 - (0.25, 0.25, 0, 0, 0)^T X)$ and the treatment effect $\tau^*(X) = \exp(X_{(1)}) - 3X_{(2)}$ is from Shen & Cui (2022).