
HIGH CONFIDENCE INFERENCE ON THE PROBABILITY AN INDIVIDUAL BENEFITS FROM TREATMENT USING EXPERIMENTAL OR OBSERVATIONAL DATA

Gabriel Ruiz*

Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
17ruiz17@gmail.com

Oscar Hernan Madrid Padilla†

Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
oscar.madrid@stat.ucla.edu

ABSTRACT

We seek to understand the probability an individual benefits from treatment (PIBT), an inestimable quantity that must be bounded in practice. Given the innate uncertainty in the population-level bounds on PIBT, we seek to better understand the margin of error for their estimation in order to discern whether the estimated bounds on PIBT are tight or wide due to random chance or not. Toward this goal, we present guarantees to the estimation of bounds on marginal PIBT, with any threshold of interest, for a randomized experiment setting or an observational setting where propensity scores are known. We also derive results that permit us to understand heterogeneity in PIBT across learnable sub-groups delineated by pre-treatment features. These results can be used to help with formal statistical power analyses and frequentist confidence statements for settings where we are interested in assumption-lean inference on PIBT through the target bounds with minimal computational complexity compared to a bootstrap approach. Through a real data example from a large randomized experiment, we also demonstrate how our results for PIBT can allow us to understand the practical implication and goodness of fit of an estimate for the conditional average treatment effect (CATE), a function of an individual's baseline covariates.

1 Introduction

1.1 Motivation

The premise of this work, in a vein similar to predictive inference with quantile regression, is that observations may lie far away from their conditional expectation. In the context of causal inference, due to the missing-ness of one outcome, it is difficult to check whether an individual's treatment effect lies close to its prediction given by the estimated average treatment effect or conditional average treatment effect (CATE). With the aim of augmenting the inference with these estimands in practice, we further study a distribution-free framework for the plug-in estimation of bounds on the probability an individual benefits from treatment (PIBT), a generally inestimable quantity that would concisely summarize an intervention's efficacy if it could be known.

*gabriel-ruiz.github.io

†<https://hernanmp.github.io/>

In order to study PIBT, we are interested in the individual treatment effect:

$$Y_i(1) - Y_i(0) \quad (i = 1, \dots, n).$$

Here, $Y_i(w)$ ($w = 0, 1$) is known as a potential outcome (Splawa-Neyman et al., 1990; Rubin, 1974) and equivalently as a counterfactual (Pearl, 2009; Hernán and Robins, 2020). The random variable $Y_i(w)$ corresponds to the outcome Y_i of interest when a hypothetical experimenter intervenes with nature to force the binary exposure indicator W_i for individual i to be little $w \in \{0, 1\}$, typically through random assignment in an experiment. This intervention can also be denoted by Pearl (2009)'s "do" operator: $Y_i(w)$ is the outcome we observe under $\text{do}(W_i = w)$.

Suppose a large value of the observed outcome $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$ is "good" for an individual. Assuming we wish to show $W_i = 1$ is effective at accomplishing this, we will consider an individual such that

$$Y_i(1) - Y_i(0) > \delta \quad (1)$$

to have benefited from treatment. We can take δ to be any relevant threshold we would like, such as $\delta = 0$. We may also reverse the inequality if more appropriate. Just as well, the vocabulary and notational semantics of the inequality in (1) can instead be with respect to whether an individual is harmed by an intervention as we discuss in Appendix D.4. Should the potential outcomes be binary, we may also use the inequality in (1) with $\delta = 0$ as we discuss in Appendix D.2. Moreover, if the outcome of interest is strictly positive, we may alternatively define benefiting from treatment in terms of the ratio of an individual's potential outcomes being above a threshold as we discuss in Appendix D.3. For simplicity of presentation and without any loss of generality for these alternative definitions which our results can be applied to, we will say an individual benefits from treatment when the inequality in (1) holds.

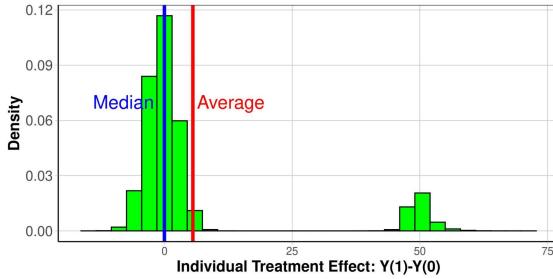


Figure 1: A hypothetical distribution for the individual treatment effects. Here, the mean is positive yet the probability an individual's treatment outcome is better (larger in value) than their control outcome is 49.75%.

As can be appreciated from the hypothetical distribution of the individual treatment effects in Figure 1, it is very well possible that the average of the individual treatment effect distribution is pulled by outliers and therefore may mislead us. This example motivates our interest in the probability an individual benefits from treatment (PIBT) as provided in Definition 1.1.

Definition 1.1 (Probability an individual benefits from treatment (PIBT)). To understand heterogeneity in treatment effect for individuals across differing strata of a pre-treatment covariate, X_i , denote

$$\text{pr}\{Y_i(1) - Y_i(0) > \delta \mid X_i = x\}$$

as the conditional probability an individual benefits from treatment in pre-treatment covariate stratum x . To understand the effect of treatment for all individuals regardless of strata, denote

$$\theta(\delta) := \text{pr}\{Y_i(1) - Y_i(0) > \delta\}$$

as the marginal probability an individual benefits from treatment.

Here, X_i is a pre-treatment covariate that is thought to deconfound variability in the outcome Y_i that is not due to W_i and exogenous noise alone (Rubin, 1974; Imbens and Rubin, 2015), as we state formally in Assumption 1.5. One can see that $\theta(\delta) = \mathbb{E}[\text{pr}\{Y_i(1) - Y_i(0) > \delta \mid X_i\}]$, where the expectation is taken with respect to the confounder X_i . Consider now two very similar quantities in Definition 1.2.

Definition 1.2. To understand heterogeneity in treatment's effect across differing individuals and across differing strata of a pre-treatment covariate, denote

$$\eta(\delta, x) := \text{pr}\{Y_i(1) - Y_j(0) > \delta \mid X_i = X_j = x; i \neq j\}$$

as the probability that a randomly selected individual in pre-treatment covariate stratum x has a treatment potential outcome that is better than the control potential outcome of a differing randomly selected individual also in stratum x . To understand treatment's effect across differing individuals overall, denote

$$\eta(\delta) := \text{pr}\{Y_i(1) - Y_j(0) > \delta; i \neq j\}$$

as the overall probability that a randomly selected individual has a treatment outcome that is better than a differing randomly selected individual's control outcome.

Of crucial importance, $\text{pr}\{Y_i(1) - Y_i(0) > \delta | X_i\}$ and $\theta(\delta)$ are not in general the same as $\eta(\delta, x)$ and $\eta(\delta)$ in Definition 1.2 (Hand, 1992; Fay et al., 2018; Greenland et al., 2020). This is because $Y_i(1)$ and $Y_i(0)$ are in general dependent random variables, while $Y_i(1)$ is independent of $Y_j(0)$ when $i \neq j$ under the standard Stable Unit Treatment Value Assumption (Assumption 1.3). Under appropriate identifiability assumptions, such as Assumption 1.5 below, $\eta(\delta, x)$ and $\eta(\delta)$ can be identified because we can generally sample from the conditional distribution of Y_i given $(W_i, X_i) = (w, x)$ and the conditional distribution of Y_i given $W_i = w$ ($w = 0, 1$). It is impossible to sample from the joint distributions of $(Y_i(0), Y_i(1))$ marginally or given $X_i = x$, because an individual cannot be in both the treatment group and the control group simultaneously. So we cannot generally identify PIBT. Not even if we are able to perfectly match individuals in opposite treatment groups based on pre-treatment covariates. This is what is known as the fundamental problem of causal inference.

The goal nonetheless is to reason about PIBT through estimators for bounds on this quantity. Our focus is on deriving closed-form, non-asymptotic margins of error (ε) on the bound estimators for an overall confidence band on PIBT of the form: with a desired frequentist confidence level, PIBT is contained between

$$(\text{estimator for lower bound on PIBT}) - \varepsilon \text{ and } (\text{estimator for upper bound on PIBT}) + \varepsilon. \quad (2)$$

1.2 Overview of our contributions

For the bounds on the marginal probability an individual benefits from treatment which are estimated with data from a randomized experiment or observational study with known treatment probabilities, we derive in Section 2 a closed-form concentration inequality depending on only the sample size and the desired frequentist confidence level. As discussed in Section 2.4, this allows for a formal statistical power analysis, albeit conservative, but notably without the requirement of an asymptotic limiting distribution nor the specification of any unknown parameters (e.g. plausible effect sizes). Our main results are presented in a general manner in terms of sub-groups, delineated by pre-treatment features, and estimators for bounds on PIBT based on inverse probability of treatment weighting (IPTW) (Imbens and Rubin, 2015; Hernán and Robins, 2020). The inference on PIBT with a randomized experiment is handled as a specific case.

Different from the non-asymptotic margin of error that can be obtained with bootstrap re-sampling (Efron and Tibshirani, 1994; Bickel et al., 1997), our non-asymptotic margin of error will be closed-form and simultaneous for all thresholds δ that can be used to define PIBT, thus allowing for a form of sensitivity analysis on its definition, and even cherry-picking. Also different from a margin of error that can be obtained via Bootstrap, our approach can be up to an order M (the number of bootstrap samples) faster; see Remark 2.1 for details on the relatively low computational complexity of our approach.

We include in Section 3 an example application to a real-life randomized experiment dataset, Criteo AI Lab's benchmark data for uplift prediction (Diemert Eustache, Betlei Artem et al., 2018). In particular, this section points toward a useful combination of conditional average treatment effect (CATE) estimation and the inference we provide for PIBT in this work: through a partitioning of individuals in a sample based on their similar CATE prediction, we can estimate bounds on PIBT in each of these strata to better understand the implication of the CATE estimate. This is related to work on causal analysis involving the stratification on an interpretable score, such as a prognostic score, propensity score, or other diagnostic score (Abadie et al., 2018; Padilla et al., 2021; Ye et al., 2021; Yadlowsky et al., 2021). It is also related to recent efforts to provide valid and benchmarked causal inferences (Parikh et al., 2022; Xu and Yadlowsky, 2022; Crabbé et al., 2022; Gradu et al., 2022).

For the case that we would like to consider heterogeneity in PIBT beyond the sub-group level as we consider in Section 2, Proposition C.7 provides an application of general results in Appendix C (see extended discussion there) to the canonical linear regression model. Moreover, the Appendix D is an extended discussion on the scope of our results. For binary potential outcomes, we show in Proposition D.1 that the general approach we take to bound PIBT at the population-level using Makarov (1982) is equivalent to using the sharp Boole-Fréchet bounds (Fréchet, 1935; Rüschenhoff, 1981). Our results also extend to the case when we define benefiting from treatment in terms of the ratio of an individual's two positive potential outcomes as discussed in Section D.3. This section also discusses how our results can easily extend to reasoning about the proportion of individuals that are harmed by an intervention (Kallus, 2022).

Appendix E contains all the proofs of the main results presented here.

1.2.1 Existing work: bounding the distribution of individual treatment effects

The contribution of the present work, relative to the contribution of Fan and Park (2010) who discuss inference for only the randomized experiment setting, is the concentration inequality for the PIBT bound estimators. Under regularity conditions, Fan and Park (2010) show asymptotically that the plug-in bound estimators follow either a normal distribution (centered at the target bound), or a truncated normal distribution, or a point mass. Exactly which distribution this depends on the supremum difference between the two potential outcomes' cumulative distribution functions (CDFs), which is unknown. Even if we know that the asymptotic distribution of the estimator is Gaussian, a prospective power analysis further requires an estimator for the standard error to guarantee a target confidence level and margin of error (e.g. a maximum deviation of 0.05), but such an estimator is not discussed in Fan and Park (2010). This points to a strength of our main concentration result in the randomized experiment setting: despite the possibility that the plug-in estimator can have a non-trivial, possibly biased, sampling distribution in a finite sample, the confidence level we can have for a target margin of error depends only on sample size. The discussion around Proposition 2.2 and Fig. 2 gives more details.

Fay et al. (2018) also discuss the statistical inference technique of Fan and Park (2010) in conjunction with the quantity $1 - \eta(\delta)$ in Definition 1.2. Interestingly, it has been established that the Makarov bounds for the marginal CDF of $Y_i(1) - Y_i(0)$ studied in Fan and Park (2010) are point-wise but not uniformly sharp (Firpo and Ridder, 2010, 2019). While promising, we consider the estimation of these tightened bounds beyond the scope of this paper as it is not immediately clear that it is amenable to our analysis. For continuous outcomes in a randomized experiment setting, Frandsen and Lefgren (2021) works under a condition known as mutual stochastic increasing-ness of the potential outcomes ($Y_i(0), Y_i(1)$) (Lehmann, 1966). The plug-in estimation approach we use in the randomized experiment case does not make the assumption of positive correlation: it works for any joint distribution on $(Y_i(0), Y_i(1))$ (Fan and Park, 2010), including those with any type of negative association. Also in the context of a randomized experiment, Caughey et al. (2021) study PIBT under a randomization inference setup that is traditionally used to test the sharp null hypothesis that all individual treatment effects are constant (Fisher, 1935). The approach we take to bound PIBT assumes the existence of an infinite super-population that subjects in our sample at hand are drawn independent and identically distributed from and for which our plug-in estimators provide inference for. Caughey et al. (2021) appears to be a nice alternative under the differing assumption that randomness is solely due to random assignment of subjects to a treatment.

Of special note, the quantity $\theta(\delta)$ in Definition 1.1 when $0 \leq \delta < 1$ is equivalent to what is known as the “probability of necessity and sufficiency (PNS)” if $Y_i(0)$ and $Y_i(1)$ only take on binary values (Pearl, 1999; Tian and Pearl, 2000; Pearl, 2009; Mueller et al., 2022). In this case, PNS and what we call “marginal PIBT” are given by the joint probability $\text{pr}\{Y_i(1) = 1, Y_i(0) = 0\}$. As suggested by the intriguing use of prepositional logic terminology in its name, PNS informs us of an intervention’s effectiveness at achieving a strictly better outcome. See also Cuellar and Kennedy (2018), Cinelli and Pearl (2021), Dawid and Musio (2022), and Cuellar (2022) for recent discussions on related quantities, meticulously defined as conditional probabilities of one potential outcome given the other, that inform about treatment’s efficacy.

The use of IPTW to bound PNS when it is defined for inferring harm from an intervention is considered in Kallus (2022). Relative to their handling of PNS (binary PIBT), our results are for both real-valued and binary-valued outcomes. We discuss the sharpness of our target bounds for the binary case in Appendix D. Moreover, our non-asymptotic inference is complementary to the invocation of the Central Limit Theorem in Kallus (2022).

Our work is also complementary to inference on the inverse of PIBT (Fan and Park, 2012; Lei and Candès, 2021; Yin et al., 2021; Jin et al., 2021), such as with quantile regression (Koenker and Bassett, 1978; Romano et al., 2019).

1.3 Assumptions

We now state the conditions we will work with. Throughout this work, we will assume the stable unit treatment value assumption, commonly abbreviated as SUTVA, in Assumption 1.3. We will also assume consistency of the observed outcome throughout as given in Assumption 1.4. Moreover, our inference is valid under conditional ignorability conditions in Assumption 1.5.

Assumption 1.3 (Stable Unit Treatment Value Assumption). We quote Imbens and Rubin (2015): “The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.”

Assumption 1.4 (Consistency). The observed outcome is dictated by treatment receipt indicator W_i :

$$Y_i := W_i Y_i(1) + (1 - W_i) Y_i(0).$$

Assumption 1.5 (Strong Conditional Ignorability (Rubin, 1974; Imbens and Rubin, 2015)). For $i = 1, \dots, n$, we have that:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp W_i \mid X_i.$$

Let $e(x) := \text{pr}\{W_i = 1 \mid X_i = x\}$ —the propensity score. We have further that

$$0 < \underline{e} := \inf_x e(x) \text{ and } \bar{e} := \sup_x e(x) < 1.$$

2 PIBT bounds in sub-groups

2.1 Overview

Here, we aim to estimate bounds on

$$\theta(\delta, x) := \text{pr}\{Y_i(1) - Y_i(0) > \delta \mid \ell(X_i) = \ell(x)\}.$$

Here, $\ell(x) = l$ if $x \in \mathcal{X}_l$, where

$$\mathcal{X}_l \subseteq \mathcal{X} := \text{support}(X_i) (l = 1, \dots, L; L \geq 1).$$

We consider the mapping $\ell(\cdot)$ to be fixed. However, among other possibilities, $\ell(\cdot)$ could be the partition given by a decision tree (Breiman et al., 2017; Athey et al., 2019; Athey and Wager, 2021; Sverdrup et al., 2020) or uniform mass binning (Zadrozny and Elkan, 2002; Abadie et al., 2018; Kumar et al., 2019; Gupta et al., 2020; Gupta and Ramdas, 2021) during a pre-processing step with independent data. See § 3 below for an example. We may also trivially take $L = 1$ as a special case, like we do in Figure 2.

In order for us to identify these bounds, we will work under Assumption 1.5. Denote Makarov (1982)’s pointwise sharp lower bound and upper bound on PIBT as $\theta^L(\delta, x)$ and $\theta^U(\delta, x)$, which are such that

$$\theta^L(\delta, x) \leq \theta(\delta, x) \leq \theta^U(\delta, x).$$

Denote their corresponding estimators based on independent and identically distributed data as $\hat{\theta}^L(\delta, x)$ and $\hat{\theta}^U(\delta, x)$, respectively. Proposition 2.2 is the main result in this section, providing a guarantee for the accuracy of these estimators, which we now formally define. Remark 2.1 discusses the satisfactory computational complexity of our procedure, while Section 2.4 discusses how to use Proposition 2.2 for sample size determination (power analysis). Moreover, Section B.2 of the appendix contains a simulation study.

2.2 The target bounds on PIBT and their estimators

We refer the reader to Lemma A.1 in Appendix A for the formal statement of the point-wise sharp Makarov bounds (Makarov, 1982; Williamson and Downs, 1990) we make use of to bound PIBT. The bounds $\theta^L(\delta)$ and $\theta^U(\delta)$, our target parameters, are in terms of the potential outcomes’ marginal CDFs. The CDF of $Y_i(w)$ in sub-group $\ell(x)$ is:

$$F_w(y|x) := \text{pr}\{Y_i(w) \leq y \mid \ell(X_i) = \ell(x)\},$$

which is identified under Assumption 1.5. Denote the empirical cumulative distribution function (eCDF) for $Y_i(w)$, an unbiased plug-in estimator for $F_w(y|x)$, as ($w = 0, 1$):

$$\hat{F}_{wn}(y|x) := \frac{1}{n_{\ell(x)}} \sum_{i: \ell(X_i)=\ell(x)} \frac{1\{Y_i \leq y\} 1\{W_i = w\}}{we(X_i) + (1-w)[1 - e(X_i)]}. \quad (3)$$

Here, $1\{\cdot\}$ is the indicator function, while

$$n_{\ell(x)} := \#\{i : \ell(X_i) = \ell(x)\},$$

the number of units in sub-group $\ell(x)$. Unbiasedness of $\hat{F}_w(y|x)$ follows by noting that $\hat{F}_w(y|x)$ is an inverse probability of treatment weighted (IPTW) estimator (Imbens and Rubin, 2015; Hernán and Robins, 2020) for the conditional mean of an indicator:

$$F_w(y|x) = \mathbb{E}[1\{Y_i(w) \leq y\} \mid \ell(X_i) = \ell(x)].$$

Using Lemma A.1 in the Appendix A, the target parameters to bound $\theta(\delta, x)$ across any joint distribution of $\{Y_i(0), Y_i(1)\} \mid \ell(X_i) = \ell(x)$ are:

$$\theta^L(\delta, x) = -\min \left[\inf_y \{F_1(y + \delta/2 \mid x) - F_0(y - \delta/2 \mid x)\}, 0 \right]$$

for the lower bound, while for the upper bound we have:

$$\theta^U(\delta, x) = 1 - \max \left[\sup_y \{ F_1(y + \delta/2 | x) - F_0(y - \delta/2 | x) \}, 0 \right].$$

Correspondingly, we can obtain the bound estimators by plugging in the CDF estimators in analogy to Fan and Park (2010) who consider only the randomized experiment case:

$$\hat{\theta}^L(\delta, | x) := -\min \left[\inf_y \{ \hat{F}_{1n}(y + \delta/2 | x) - \hat{F}_{0n}(y - \delta/2 | x) \}, 0 \right] \quad (4)$$

and

$$\hat{\theta}^U(\delta, x) := 1 - \max \left[\sup_y \{ \hat{F}_{1n}(y + \delta/2 | x) - \hat{F}_{0n}(y - \delta/2 | x) \}, 0 \right] \quad (5)$$

are the lower bound and upper bound estimators, respectively.

Remark 2.1 (Computational complexity). For the computation of the lower and upper bound estimates at a point x , the number of total iterations across elements in two lists of size $n_{0\ell(x)}$ and $n_{1\ell(x)}$, respectively, is

$$O[(n_{0\ell(x)} + n_{1\ell(x)}) \{ \log(n_{0\ell(x)}) + \log(n_{1\ell(x)}) \}].$$

See Appendix B.1 for details. Here, $n_{w\ell(x)}$ ($w = 0, 1$) is the number of units in treatment group w and sub-group $\ell(x)$.

2.3 The main result

Given our estimators $\{\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x)\}$, we consider now their accuracy. Proposition 2.2 provides us with this understanding. Here, the symbol \vee denotes the maximum between the left and right arguments.

Proposition 2.2 (Inference on PIBT through bounds).

Let Assumption 1.5 hold.

Case 1: Consider evaluating the bound estimators at a random new test point, X_{n+1} , independent of our training data.

With probability at least $1 - \alpha$, we have:

$$\begin{aligned} & \sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta, X_{n+1}) - \theta^L(\delta, X_{n+1}) \right| \vee \left| \hat{\theta}^U(\delta, X_{n+1}) - \theta^U(\delta, X_{n+1}) \right| \right\} \\ & \leq n_{\ell(X_{n+1})}^{-1/2} \sum_{w=0,1} C_{w,\alpha/2}(X_{n+1}). \end{aligned}$$

Case 2: Now consider a uniform guarantee on the bound estimators across $x \in \mathcal{X}$.

Let $x_l \in \mathcal{X}_l$ be arbitrary ($l = 1, \dots, L$). With probability at least $1 - \alpha$, we have:

$$\begin{aligned} & \sup_{\delta,x} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \\ & \leq \max_l \left\{ n_l^{-1/2} \sum_{w=0,1} C_{w,\alpha/(2L)}(x_l) \right\}. \end{aligned}$$

In Proposition 2.2, we define $C_{w,\beta}$, where $\beta = \alpha/2$ (Case 1) or $\beta = \alpha/(2L)$ (Case 2), as follows.

- **Completely Randomized Experiment in sub-group $\ell(x)$:** Suppose $\ell(x) = l$. If $e(x') = n_{1l}/n_l$ for each $x' \in \mathcal{X}_l$, we have:

$$C_{w,\beta}(x) \doteq \left(2^{-1} n_{\ell(x)} n_{w\ell(x)}^{-1} \log(2/\beta) \right)^{1/2}.$$

- **General case with known propensity scores:** We have:

$$C_{w,\beta}(x) \doteq \frac{2 \left[2 \log^{\frac{1}{2}} (n_{\ell(x)} + 1) + \log^{\frac{1}{2}} (1/\beta) \right]}{n_{\ell(x)}^{1/2} (w\underline{e} + (1-w)(1-\bar{e}))}.$$

The exact choice of these values is explained in Lemma E.3, a key result that makes use of the Dvoretzky-Kiefer-Wolfowitz-Massart (DKWM) inequality (Dvoretzky et al., 1956; Massart, 1990; Naaman, 2021; D’Amour et al., 2021), a functional Hoeffding’s theorem (Wainwright, 2019), and techniques from empirical process theory (Wainwright, 2019; van der Vaart and Wellner, 1996).

In Remark 2.3, we can see one practical implication of Proposition 2.2. In Section 2.4, we further discuss its practical implication with respect to a statistical power analysis. In § 3, we apply Case 2 of Proposition 2.2.

Remark 2.3. Using Proposition 2.2, we can say that with confidence at least $(1 - \alpha) \times 100\%$, the probability an individual represented by our study will benefit from treatment,

$$\text{pr} \{Y_i(1) - Y_i(0) > \delta \mid \ell(X_{n+1})\}$$

for any threshold δ of interest, is between

$$\hat{\theta}^L(\delta, X_{n+1}) - n_{\ell(X_{n+1})}^{-1/2} \sum_{w=0,1} C_{w,\alpha/2}(X_{n+1})$$

and

$$\hat{\theta}^U(\delta, X_{n+1}) + n_{\ell(X_{n+1})}^{-1/2} \sum_{w=0,1} C_{w,\alpha/2}(X_{n+1}).$$

2.4 A power analysis with Proposition 2.2

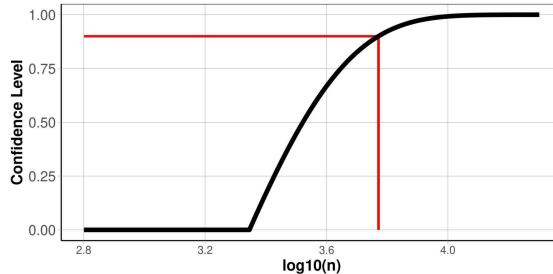


Figure 2: An application of Proposition 2.2. Here, $L = 1$, $n_0 = n_1$, and $n = n_0 + n_1$. The target margin of error is 0.05. The black curve, given by (6), corresponds to the possible confidence level we may have at a given n . From the two red line segments, we see that a sample size of $n \geq 5902$ guarantees the target margin of error with at least a 90% confidence level.

Figure 2 presents an example power analysis making use of Proposition 2.2 for the completely randomized experiment case. More generally, suppose our target margin of error for

$$\sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\}$$

in the completely randomized experiment case is $0 < \varepsilon < 1$. Solving for the significance level α_ε when we set ε equal to the margin of error in Proposition 2.2:

$$\varepsilon \doteq \left\{ \frac{\log(4/\alpha_\varepsilon)}{2} \right\}^{1/2} \left(n_{0\ell(x)}^{-1/2} + n_{1\ell(x)}^{-1/2} \right)$$

gives:

$$\alpha_\varepsilon = 4 \exp \left\{ -2 \left(n_{0\ell(x)}^{-1/2} + n_{1\ell(x)}^{-1/2} \right)^{-2} \varepsilon^2 \right\}. \quad (6)$$

The confidence level we can thus have for the target margin of error ε at any given sample size $(n_{0\ell(x)}, n_{1\ell(x)})$ is at least:

$$\begin{cases} 1 - \alpha_\varepsilon & \text{if } 0 \leq \alpha_\varepsilon \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

3 Application: Understanding the implication of conditional average treatment effects

3.1 One practical definition of sub-group level PIBT and its bounds

Using pre-treatment covariates, X_i , we will study heterogeneity and the implication of an average treatment effect estimate as follows. We first learn the conditional average treatment effect (CATE) function,

$$\tau(x) := \mathbb{E} \{Y_i(1) - Y_i(0) \mid X_i = x\}.$$

We then partition the feature space based on similar CATE predictions through uniform mass binning (Zadrozny and Elkan, 2002; Abadie et al., 2018; Kumar et al., 2019; Gupta et al., 2020; Gupta and Ramdas, 2021) as follows. Denote the quantiles of $\hat{\tau}(X_i)$ as

$$q_\alpha := \inf \{q : \text{pr} \{\hat{\tau}(X_i) \leq q\} \geq \alpha\} \quad (0 \leq \alpha \leq 1).$$

We will obtain estimators, \hat{q}_α using a second fold of randomly sampled rows. Now, define the discrete mapping $s_m : \mathcal{X} \rightarrow \{1, \dots, m\}$ as:

$$s_m(x) = k \text{ if } \hat{q}_{(k-1)/m} < \hat{\tau}(x) \leq \hat{q}_{k/m}; \quad k \in \{1, \dots, m\}.$$

The value $m \geq 2$ corresponds to the number of sub-groups that are created in the partitioning.

Finally, we estimate bounds on PIBT conditional on the partition, defined as

$$\theta(\delta, k) := \text{pr} \{Y_i(1) - Y_i(0) > \delta \mid s_m(X_i) = k\}. \quad (7)$$

The parameter $\theta(\delta, k)$ is the probability that treatment is beneficial in CATE stratum $k = 1, \dots, m$. When treatment is randomized independently of the baseline covariates X_i , as we assume in this section, we have the ignorability statement

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp W_i \mid s_m(X_i) = k \quad (k = 1, \dots, m).$$

This allows us to identify the Makarov lower and upper bounds on $\theta(\delta, k)$, which we will denote as $\theta^L(\delta, k)$ and $\theta^U(\delta, k)$, respectively. The mappings $s_m(\cdot)$ allow us to stratify on an interpretable univariate score. For our demonstrative choice of the mapping $s_m(\cdot)$, if the learned function CATE function $\hat{\tau}(\cdot)$ is indicative of benefiting from treatment, we would like to see a monotone increasing relation between $k = 1, \dots, m$ and the bound estimators $\{\hat{\theta}^L(\delta, k), \hat{\theta}^U(\delta, k)\}$ as supporting evidence.

3.1.1 Criteo's uplift prediction benchmark dataset

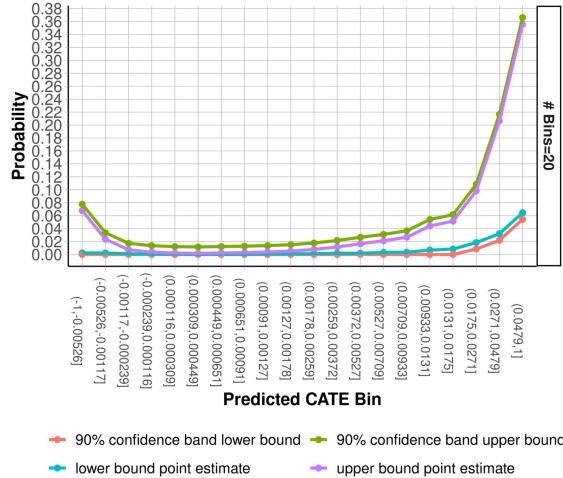


Figure 3: The 90% Bonferroni corrected lower and upper confidence bands on PIBT across bins of CATE predictions on the Criteo uplift dataset.

We now present an application to Criteo AI Lab's uplift prediction benchmark dataset (Diemert Eustache, Betlei Artem et al., 2018). According to the webpage that hosts the data,

This dataset is constructed by assembling data resulting from several incrementality tests, a particular randomized trial procedure where a random part of the population is prevented from being targeted by advertising. It consists of 25M rows, each one representing a user with [12] features, a treatment indicator and 2 labels (visits and conversions).

For this application, the proportion we will estimate bounds for should be understood in plain language as the proportion of the time that the advertiser benefits from presenting an advertisement on the website, rather than the probability an individual benefits from treatment. Equation (7) below is the formal statement of this proportion.

The available down sampled data consists of 13,979,592 observations. We focus on the effect treatment assignment, rather than treatment receipt, has on visits, making this an intent-to-treat analysis. The outcome of interest in our analysis is the indicator for whether a user visited the advertiser website during the test period of 2 weeks.

With 50,000 randomly sampled rows, we learn CATE through `grf :: causal_forest` with default options in R (Athey et al., 2019). Next, we obtain estimators for the $m - 1$ quantiles of the random variable $\hat{\tau}(X_i)$ using a second fold of 8,987,000 rows. The bounds are then estimated within each bin in the second fold.

Restricting $\delta \in (0, 1)$ and noting that the outcome is binary, Figure 3 provides a 90% Bonferroni corrected confidence band on

$$\theta(\delta, k) = \text{pr} \{Y_i(1) = 1, Y_i(0) = 0 \mid s_m(X_i) = k\}$$

that allows for simultaneous inference across CATE prediction bins $k = 1, \dots, 20$. The practical insight is that for an individual such that $0.0479 < \hat{\tau}(X_i) \leq 1$, the probability this individual will visit the advertiser's web-page when assigned treatment and otherwise not visit the website if untreated, is between 5.42% and 36.62% with 90% confidence. Given the Bonferroni correction and that all bins have equal amounts of subjects, we take an average of the upper confidence bound for every bin corresponding to a CATE prediction of 0.0271 or less. This gives that an individual with a CATE prediction of 0.0271 or less has a joint probability of interest of 4.18% or less with 90% confidence.

4 Discussion

Promising directions related to the work presented here include applying the results in Appendix C to a setting more general than that demonstrated in Proposition C.7. Alternatively, we may like to tailor these results to model specifications that are sufficient for interesting applications, such as involving generalized linear models (McCullagh and Nelder, 2019; Sur and Candès, 2019). For settings where comparing an individual's two potential outcomes in a ratio rather than a difference can provide interesting insight, such as studies where time to an event is of interest (Cox, 1972; Stitelman and van der Laan, 2010; Austin, 2014; Schober and Vetter, 2018; Cai and van der Laan, 2020), it seems worthwhile to extend the discussion in Appendix D.3.

Our probability bound formulation to reason about treatment effects, rather than the more common average formulation, is similar to recent model-free work that moves beyond an average in favor of controlling the probability of a type I error, or a false discovery, using a Neyman-Pearson paradigm (Tong et al., 2018; Li et al., 2021). Future work relating our present work to these advances could be of interest, for example as it relates to causal classification (Fernández-Loría and Provost, 2022) and mitigating harm from an intervention (Kallus, 2022). This direction may involve more rigorously choosing the appropriate threshold δ to define PIBT in Definition 1.1, in accordance with an appropriately formulated Neyman-Pearson objective. Recalling our uniform control across threshold values, we currently propose that δ be anything, or everything, a practitioner finds reasonable.

Acknowledgement

Gabriel would like to thank NSF DGE-1650604 for financial support, Qing Zhou for encouragement on this project, and members of the UCLA causal inference reading group for their feedback.

References

- Abadie, A., Chingos, M. M., and West, M. R. (2018). Endogenous Stratification in Randomized Experiments. *The Review of Economics and Statistics*, 100(4):567–580.
- Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.
- Athey, S. and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, 89(1):133–161.
- Austin, P. C. (2014). The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat. Med.*, 33(7):1242–1258.

- Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7(1):1–31.
- Boole, G. (1854). *An investigation of the laws of thought : on which are founded the mathematical theories of logic and probabilities / By George Boole*. Walton and Maberly, London.
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (2017). *Classification And Regression Trees*. Routledge.
- Burkhart, M. C. and Ruiz, G. (2022). Neuroevolutionary feature representations for causal inference. In Groen, D., de Matalier, C., Paszynski, M., Krzhizhanovskaya, V. V., Dongarra, J. J., and Sloot, P. M. A., editors, *Computational Science – ICCS 2022*, pages 3–10, Cham. Springer International Publishing.
- Cai, W. and van der Laan, M. J. (2020). One-step targeted maximum likelihood estimation for time-to-event outcomes. *Biometrics*, 76(3):722–733.
- Caughey, D., Dafoe, A., Li, X., and Miratrix, L. (2021). Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects.
- Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.
- Cinelli, C. and Pearl, J. (2021). Generalizing experimental results by leveraging knowledge of mechanisms. *Eur. J. Epidemiol.*, 36(2):149–164.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.
- Crabbé, J., Curth, A., Bica, I., and van der Schaar, M. (2022). Benchmarking heterogeneous treatment effect models through the lens of interpretability.
- Cuellar, M. (2022). *Causes of Effects and Effects of Causes*, pages 211–233. Springer International Publishing, Cham.
- Cuellar, M. and Kennedy, E. (2018). A nonparametric projection-based estimator for the probability of causation, with application to water sanitation in kenya. *SSRN Electron. J.*
- Dasgupta, S., Papadimitriou, C. H., and Vazirani, U. V. (2006). *Algorithms*. McGraw-Hill Professional, New York, NY.
- Dawid, A. P. and Musio, M. (2022). Effects of causes and causes of effects. *Annual Review of Statistics and Its Application*, 9(1):261–287.
- Diemert Eustache, Betlei Artem, Renaudin, C., and Massih-Reza, A. (2018). A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London, United Kingdom, August, 20, 2018*. ACM.
- Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669.
- D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. Chapman and Hall/CRC.
- Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951.
- Fan, Y. and Park, S. S. (2012). Confidence intervals for the quantile of treatment effects in randomized experiments. *Journal of Econometrics*, 167(2):330–344. Fourth Symposium on Econometric Theory and Applications (SETA).
- Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., and Gabriel, E. E. (2018). Causal estimands and confidence intervals associated with wilcoxon-mann-whitney tests in randomized experiments. *Statistics in Medicine*, 37(20):2923–2937.
- Fernández-Loría, C. and Provost, F. (2022). Causal classification: Treatment effect estimation vs. outcome prediction. *Journal of Machine Learning Research*, 23(59):1–35.
- Firpo, S. and Ridder, G. (2019). Partial identification of the treatment effect distribution and its functionals. *Journal of Econometrics*, 213(1):210–234. Annals: In Honor of Roger Koenker.
- Firpo, S. P. and Ridder, G. (2010). Bounds on functionals of the distribution treatment effects. Textos para discussão 201, FGV EESP - Escola de Economia de São Paulo, Fundação Getulio Vargas (Brazil).
- Fisher, R. A. (1935). *The design of experiments*. Oliver and Boyd, Edinburgh.
- Frandsen, B. R. and Lefgren, L. J. (2021). Partial identification of the distribution of treatment effects with an application to the knowledge is power program (kipp). *Quantitative Economics*, 12(1):143–171.

- Frank, M. J., Nelsen, R. B., and Schweizer, B. (1987). Best-possible bounds for the distribution of a sum — a problem of kolmogorov. *Probability Theory and Related Fields*, 74:199–211.
- Fréchet, M. (1935). Généralisation du théorème des probabilités totales. *Fundamenta mathematicae*, 25:379–387.
- Fréchet, M. (1960). Sur les tableaux dont les marges et des bornes sont données. *Revue de l'Institut international de statistique*, 28(1/2):10–32.
- Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., and Davidian, M. (2011). Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761–767.
- Gradu, P., Zrnic, T., Wang, Y., and Jordan, M. I. (2022). Valid inference after causal discovery.
- Greenland, S., Fay, M. P., Brittain, E. H., Shih, J. H., Follmann, D. A., Gabriel, E. E., and Robins, J. M. (2020). On causal inferences for personalized medicine: How hidden causal assumptions led to erroneous causal claims about the d-value. *Am. Stat.*, 74(3):243–248.
- Gupta, C., Podkopaev, A., and Ramdas, A. (2020). Distribution-free binary classification: prediction sets, confidence intervals and calibration. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3711–3723. Curran Associates, Inc.
- Gupta, C. and Ramdas, A. (2021). Distribution-Free Calibration Guarantees for Histogram Binning without Sample Splitting. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3942–3952. PMLR.
- Hailperin, T. (1986). *Boole's logic and probability : a critical exposition from the standpoint of contemporary algebra, logic, and probability theory / Theodore Hailperin*. Studies in logic and the foundations of mathematics ; v. 85. North-Holland Pub. Co., Amsterdam, Netherlands ;, 2nd ed., rev. and enl. edition.
- Hand, D. J. (1992). On comparing two treatments. *The American Statistician*, 46(3):190–192.
- Hernán, M. A. and Robins, J. (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
- Jin, Y., Ren, Z., and Candès, E. J. (2021). Sensitivity analysis of individual treatment effects: A robust conformal inference approach.
- Kallus, N. (2022). What's the harm? sharp bounds on the fraction negatively affected by treatment.
- Kennedy, E. H. (2020). Towards optimal doubly robust estimation of heterogeneous causal effects.
- Kneib, T., Silbersdorff, A., and Säfken, B. (2021). Rage against the mean – a review of distributional regression approaches. *Econometrics and Statistics*.
- Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.
- Koenker, R., Leorato, S., and Peracchi, F. (2013). Distributional vs. quantile regression. *SSRN Electronic Journal*.
- Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.
- Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5):1137–1153.
- Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B*, 83(5):911–938.
- Li, J. J., Chen, Y. E., and Tong, X. (2021). A flexible model-free prediction-based framework for feature ranking. *Journal of Machine Learning Research*, 22(124):1–54.
- Lu, J., Ding, P., and Dasgupta, T. (2015). Treatment effects on ordinal outcomes: Causal estimands and sharp bounds. *Journal of Educational and Behavioral Statistics*, 43:540 – 567.
- Makarov, G. D. (1982). Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & Its Applications*, 26(4):803–806.
- Massart, P. (1990). The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283.
- McCullagh, P. and Nelder, J. A. (2019). *Generalized Linear Models*. Routledge.

- Mueller, S., Li, A., and Pearl, J. (2022). Causes of effects: Learning individual responses from population data. In Raedt, L. D., editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2712–2718. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Naaman, M. (2021). On the tight constant in the multivariate dvoretzky–kiefer–wolfowitz inequality. *Statistics & Probability Letters*, 173:109088.
- Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.
- Padilla, O. H. M., Ding, P., Chen, Y., and Ruiz, G. (2021). A causal fused lasso for interpretable heterogeneous treatment effects estimation.
- Parikh, H., Varjao, C., Xu, L., and Tchetgen, E. T. (2022). Validating causal inference methods. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17346–17358. PMLR.
- Pearl, J. (1999). Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1):93–149.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Romano, Y., Patterson, E., and Candès, E. J. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems* 32, pages 3543–3553.
- Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.
- Rüsendorf, L. (1981). Sharpness of fréchet-bounds. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57:293–302.
- Schober, P. and Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: The tortoise and the hare. *Anesth. Analg.*, 127(3):792–798.
- Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12):371–421.
- Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statist. Sci.*, 5(4):465–472.
- Stitelman, O. M. and van der Laan, M. J. (2010). Collaborative targeted maximum likelihood for time to event data. *Int. J. Biostat.*, 6(1):Article 21.
- Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.
- Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., and Wager, S. (2020). policytree: Policy learning via doubly robust empirical welfare maximization over trees. *Journal of Open Source Software*, 5(50):2232.
- Tian, J. and Pearl, J. (2000). Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.*, 28(1/4):287–313.
- Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. (2019). *Conformal Prediction under Covariate Shift*. Curran Associates Inc., Red Hook, NY, USA.
- Tong, X., Feng, Y., and Li, J. J. (2018). Neyman-pearson classification algorithms and np receiver operating characteristics. *Science Advances*, 4(2):eaao1659.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, New York, NY.
- Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*.
- Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
- Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
- Williamson, R. C. and Downs, T. (1990). Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds. *International Journal of Approximate Reasoning*, 4(2):89–158.

- Xu, Y. and Yadlowsky, S. (2022). Calibration error for heterogeneous treatment effects. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 9280–9303. PMLR.
- Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., and Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects.
- Ye, S. S., Chen, Y., and Padilla, O. H. M. (2021). 2d score based estimation of heterogeneous treatment effects.
- Yin, M., Shi, C., Wang, Y., and Blei, D. M. (2021). Conformal sensitivity analysis for individual treatment effects.
- Zadrozny, B. and Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, page 694–699, New York, NY, USA. Association for Computing Machinery.

A The Makarov bounds

A.1 Formal Statement

We present the bounds as stated in Williamson and Downs (1990)'s Theorem 2.

Lemma A.1 (The Makarov Bounds). *Uniformly across all possible, unknown joint distributions*

$$(V_1, V_2) \sim \text{pr} \{V_1 \leq v_1, V_2 \leq v_2\}$$

having fixed marginal CDFs, $F_1(v_1) = \text{pr} \{V_1 \leq v_1\}$ and $F_2(v_2) = \text{pr} \{V_2 \leq v_2\}$, the CDF of $V_1 - V_2$ evaluated at $\delta \in \mathbb{R}$ satisfies:

$$F^L(\delta) \leq \text{pr} \{V_1 - V_2 \leq \delta\} \leq F^U(\delta), \quad (8)$$

where

$$F^L(\delta) = -\max \left[\sup_{a,b: a+b=\delta} \{F_1(a) - F_2(-b)\}, 0 \right]$$

and

$$F^U(\delta) = 1 + \min \left[\inf_{a,b: a+b=\delta} \{F_1(a) - F_2(-b)\}, 0 \right].$$

Lemma A.1 was first proved in Makarov (1982) to bound the distribution of a sum of two random variables. We present this result for subtraction, which is a simple extension, as $V_1 - V_2$ is technically the sum of two random variables V_1 and $(-V_2)$. Lemma A.1's proof was later extended in Frank et al. (1987) and Williamson and Downs (1990), who also seek bounds on the distribution of other binary operations on V_1 and V_2 , like their difference, product, and their ratio, under minimal distributional assumptions.

A.2 Equivalent forms of the bounds

The terms $F^L(\delta)$ and $F^U(\delta)$ in Lemma A.1 can be rewritten.

Consider a one-to-one change of variables $(a, b) \mapsto (u + \delta/2, -u + \delta/2)$. With it, we have equivalently:

$$F^L(\delta) = -\max \left[\sup_u \{F_1(u + \delta/2) - F_2(u - \delta/2)\}, 0 \right]$$

along with

$$F^U(\delta) = 1 + \min \left[\inf_u \{F_1(u + \delta/2) - F_2(u - \delta/2)\}, 0 \right].$$

This is in line with what we have written in Section 2 and C of the main text.

Consider instead the one-to-one change of variables $(a, b) \mapsto (u, -u + \delta)$. With it, we have equivalently:

$$F^L(\delta) = -\max \left[\sup_u \{F_1(u) - F_2(u - \delta)\}, 0 \right]$$

along with

$$F^U(\delta) = 1 + \min \left[\inf_u \{F_1(u) - F_2(u - \delta)\}, 0 \right].$$

This is in line with Fan and Park (2010)'s Lemma 2.1 and Equation (2) and (3), and it also agrees with the alternative form given in Equations (21) and (22) of Williamson and Downs (1990).

Moreover, it is straightforward to see that:

$$1 - F^U(\delta) \leq \text{pr} \{U_1 - U_0 > \delta\} \leq 1 - F^L(\delta).$$

We make use of this in the main text when bounding PIBT.

B Appendix to Section 2

B.1 Computational Complexity

The reasoning for the computational complexity claim in Remark 2.1 is as follows. Consider the mapping

$$\hat{G}_n(y, \delta, x) := \hat{F}_{1n}(y + \delta/2) | x - \hat{F}_{0n}(y - \delta/2 | x).$$

The bound estimators in (4) and (5) are each a functional of $\hat{G}_n(\cdot, \delta)$ for fixed (δ, x) . The knots of $\hat{G}(\cdot, \delta, x)$ are given by the set

$$\mathcal{Y}_{\delta, x, n} := \{Y_i(w) - \delta/2 : W_i = 0, \ell(X_i) = \ell(x)\} \cup \{Y_i(w) + \delta/2 : W_i = 1, \ell(X_i) = \ell(x)\}.$$

This means that the infimum and supremum used to specify the bound estimators in (4) and (5) are really a minimum and maximum, respectively, taken by indexing the finite set $\mathcal{Y}_{\delta, x, n}$.

Next, consider that sorting the list $\{Y_i : W_i = w, \ell(X_i) = \ell(x)\}$ has worst time computational complexity $O(n_w(x) \log(n_w(x)))$ with a procedure such as merge sort (Dasgupta et al., 2006). Given the lists $\{Y_i : W_i = 0, \ell(X_i) = \ell(x)\}$ and $\{Y_i : W_i = 1, \ell(X_i) = \ell(x)\}$ in sorted form, the evaluation of $\hat{G}(y, \delta, x)$ for one $y \in \mathcal{Y}_{\delta, x, n}$ is no worse than $O(\log(n_0(x)) + \log(n_1(x)))$ with a procedure making use of binary search (Dasgupta et al., 2006). Evaluating $\hat{G}(y, x, \delta)$ at each point $y \in \mathcal{Y}_{\delta, x, n}$ then brings us to the stated overall computational complexity of $O[(n_0(x) + n_1(x)) \{\log(n_0(x)) + \log(n_1(x))\}]$. Finally, the minimum and maximum calculation on the list of evaluations $\{\hat{G}(y, x, \delta) : y \in \mathcal{Y}_{\delta, x, n}\}$ add $O(n_0(x) + n_1(x))$ complexity, so we retain the overall complexity stated in Remark 2.1.

B.2 A simulation study

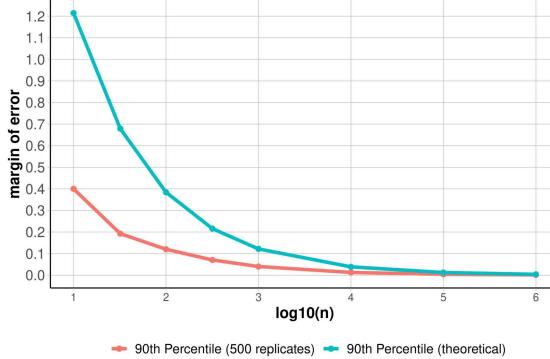


Figure 4: Across 500 replicates at each n , we compare the 90th percentile of the error defined in Equation (9) against the margin of error given by Proposition 2.2 at the 90% confidence level. We hold the signal to noise ratio and propensity score fixed.

We now conduct some simulations to demonstrate that Proposition 2.2 provides the appropriate coverage. First, we estimate bounds on PIBT at multiple thresholds δ . Then we fix δ and estimate PIBT based on synthetic data generated according to differing multiple signal to noise ratios and propensity scores.

In the first setting, our data is generated according to

$$Y_i(w) \sim \mathcal{N}(0, 1) \quad (w = 0, 1)$$

and $\text{pr}\{W_i = 1\} = 0.5$. Moreover, our bound estimators are based on sample size and threshold given by their respective choices: $n = 10, [10^{1.5}], 10^2, [10^{2.5}], 10^3, 10^4, 10^5, 10^6$ and $\delta = 0, \dots, 5$. For each n , we keep track of the estimation error

$$\max_{\delta=0, \dots, 5} |\hat{\theta}^L(\delta) - \theta^L(\delta)| \vee |\hat{\theta}^U(\delta) - \theta^U(\delta)| \quad (9)$$

across 500 replicates. Figure 4 summarizes these results by comparing the 90th percentile of this error across the 500 replicates to the margin of error given by Proposition 2.2 at the 90% confidence level. As suggested by this figure, our joint coverage of $\theta^L(\delta)$ and $\theta^U(\delta)$ for each $\delta = 0, 1, 2, 3, 4, 5$ using Proposition 2.2 appears to be conservative. In fact, across the 500 replicates for each n we tried, the actual joint coverage is no less than 99.8%.

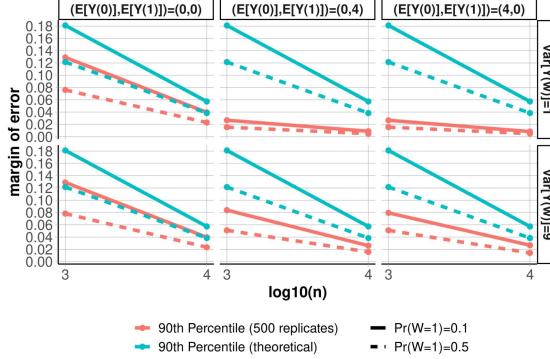


Figure 5: Across 500 replicates at each n and δ , we compare the 90th percentile of the error defined in Equation (10) against the margin of error given by Proposition 2.2 at the 90% confidence level. We fix $\delta = 0$, vary sample size, the signal to noise ratio, and the propensity score.

For the nominal coverage guarantee for the bound estimators using Proposition 2.2, we now seek to understand the influence of varying values of the average treatment effect, $\mu_1 - \mu_0$, differing amounts of potential outcomes' variation, σ^2 , and the propensity score, $\text{pr}\{W_i = 1\}$. We fix $\delta = 0$ in our estimation of the bounds on PIBT. For this second synthetic data scenario, we generate the potential outcomes marginally according to:

$$Y_i(w) \sim \mathcal{N}(\mu_w, \sigma^2).$$

We consider 500 replicates for every combination of simulation parameters denoted in Fig. 5 and keep track of the error

$$\left| \hat{\theta}^L(0) - \theta^L(0) \right| \vee \left| \hat{\theta}^U(0) - \theta^U(0) \right|. \quad (10)$$

As with the previous synthetic example, it seems that the theoretical margin of error for a 90% confidence level given by Proposition 2.2 is conservative compared to the 90th percentile of the bound estimates' margin of error. Again, the actual joint coverage of the bound estimators is at or very close to 100%.

Recall that the margin of error in Proposition 2.2 only depends on sample size and target confidence level. Correspondingly, the only change in the theoretical margin of error is when the propensity score changes between 0.1 and 0.5, corresponding to the bold-blue and dashed-blue lines in Fig. 5, respectively. Observe now the red lines in Fig. 5, corresponding to the 90th percentile of the estimation error across the 500 replicates. As expected, the propensity score also affects estimation error, with the balanced case of $\text{pr}\{W_i = 1\} = 0.5$ (dashed-red lines) generally giving better results compared to the imbalanced case of $\text{pr}\{W_i = 1\} = 0.1$ (bold-red lines). Denote signal to noise as

$$|\mathbb{E}\{Y_i(1) - Y_i(0)\}| / [\text{var}\{Y_i(0)\} + \text{var}\{Y_i(1)\}]^{1/2}.$$

We see from the red lines in Fig. 5 a smaller estimation error when the signal to noise is $2^{3/2}$ compared to when it is 0 or $2^{3/2}/3$.

C PIBT bounds with pre-treatment covariates

C.1 Overview

We now seek to estimate bounds on the unidentifiable probability an individual benefits from treatment (PIBT) in pre-treatment stratum $X_i = x$:

$$\theta(\delta, x) = \text{pr}\{Y_i(1) - Y_i(0) > \delta \mid X_i = x\}.$$

Could it be known, this quantity is helpful to understand whether the benefit of receiving treatment varies across pre-treatment covariate strata. Denote the target bounds as $\theta^L(\delta, x)$ and $\theta^U(\delta, x)$, the lower and upper bound, respectively. They satisfy:

$$\theta^L(\delta, x) \leq \theta(\delta, x) \leq \theta^U(\delta, x).$$

And denote the corresponding estimators as $\hat{\theta}^L(\delta, x)$ and $\hat{\theta}^U(\delta, x)$, respectively. We would like a guarantee about how close $\{\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x)\}$ is to $\{\theta^L(\delta, x), \theta^U(\delta, x)\}$.

Remark C.1 (Large enough sample at a covariate stratum?). Importantly, should a large enough sample be collected at stratum x of the pre-treatment covariates, Proposition 2.2 can be applied for a frequentist confidence statement about $\theta(\delta, x)$ using the analogous interpretation in Remark 2.3. The rest of this section is useful for the case that a large enough sample is not collected for some or all of the pre-treatment covariate strata of interest.

Proposition C.2 is the main non-asymptotic, model-free result for this setting. Proposition C.4 is the adaptation of Proposition C.2 to a case where we strategically use regression residuals to estimate the bounds on PIBT. For this model-free approach with regression residuals, we show how a confidence statement about the conditional PIBT bound estimators can be written in terms of a target confidence level that is adjusted according to how accurate the arbitrary regression function estimator is. In Corollary C.5, we demonstrate how the statement written in this manner implies that the conditional bound estimators are as statistically efficient as the regression function estimator of choice. Moreover, Proposition C.7 adopts the more general Proposition C.4 to the canonical linear regression case.

C.2 Existing work: PIBT conditional on pre-treatment covariates

Fan and Park (2010) also discuss the conditional bounds for the conditional individual treatment effect distribution at the population-level along with a brief discussion of possible estimation approaches. The appendix of Frandsen and Lefgren (2021) also mentions a generalisation of the mutual stochastic increasing-ness assumption in order to arrive at bounds for $1 - \theta(\delta, x)$. In the context of ordinal outcomes, using Makarov's bounds to study the individual treatment effect, Lu et al. (2015) also consider the case we would like to condition on covariates. All three works suggest some form of distributional regression, the semi-parametric estimation of a conditional CDF (Koenker et al., 2013; Chernozhukov et al., 2013; Kneib et al., 2021). We extend their discussion on covariate conditioning with a discussion on theoretical guarantees and how to conduct statistical inference with the bound estimators.

Related to our use of pre-treatment covariates, Lei and Candès (2021) develops prediction intervals for the individual treatment effect based on quantile regression (Koenker and Bassett, 1978) with strategic calibration using conformal inference (Vovk et al., 2005; Shafer and Vovk, 2008; Tibshirani et al., 2019). Moreover, this work is extended to scenarios where unobserved confounding is possible (Yin et al., 2021; Jin et al., 2021). Our work here is complementary to these advances, in analogy to the inverse relation between quantiles and the CDF of a distribution. With respect to theoretical guarantees, Proposition C.4 below is with respect to the supremum deviation across inputted pre-treatment covariate levels, whereas the conformal guarantees of Lei and Candès (2021) are with respect to any single randomly generated covariate level. This is a subtle but important difference: one may like inference about heterogeneity in individual treatments effects to extend simultaneously to multiple individuals with fixed (non-random) covariate levels, not necessarily a single random individual with a random pre-treatment covariate value. On the other hand, our work does not necessarily extend to a target population beyond that represented by our training sample, and one of the main results here (Proposition C.4) makes use of a regularity condition on regression residuals that quantile regression generally avoids.

Related to the use of quantiles to infer individual treatment effects, Fan and Park (2010) and Fan and Park (2012) discuss estimators for the sharp bounds on the quantiles of the individual treatment effect distribution.

C.3 The target bounds on conditional PIBT, their estimators, and the main result

The bounds $\theta^L(\delta, x)$ and $\theta^U(\delta, x)$ make use of the conditional CDFs ($w = 0, 1$):

$$F_w(y | x) := \text{pr} \{Y_i(w) \leq y | X_i = x\} = \text{pr} \{Y_i \leq y | X_i = x, W_i = w\},$$

with the second equality due to Assumption 1.5 and consistency. Explicitly, due to Lemma A.1 in the Appendix A, we have:

$$\theta^L(\delta, x) = - \min \left[\inf_y \{F_1(y + \delta/2 | x) - F_0(y - \delta/2 | x)\}, 0 \right]$$

along with

$$\theta^U(\delta, x) = 1 - \max \left[\sup_y \{F_1(y + \delta/2 | x) - F_0(y - \delta/2 | x)\}, 0 \right]$$

at the population-level. Denote

$$G(y, \delta, x) := F_1(y + \delta/2 | x) - F_0(y - \delta/2 | x),$$

and its corresponding estimator as $\hat{G}_n(y, \delta, x)$ based on the training sample. In practice, one may specify $\hat{G}_n(y, \delta, x)$ as the difference of two conditional CDF estimators as in Corollary C.3 below. The plug-in estimators for the lower

bound and upper bounds, respectively, will be:

$$\hat{\theta}^L(\delta, x) = -\min \left[\inf_y \left\{ \hat{G}(y, \delta, x) \right\}, 0 \right] \quad (11)$$

along with

$$\hat{\theta}^U(\delta, x) = 1 - \max \left[\sup_y \left\{ \hat{G}_n(y, \delta, x) \right\}, 0 \right]. \quad (12)$$

With respect to this choice of $\{\hat{\theta}^L(\cdot, \cdot), \hat{\theta}^U(\cdot, \cdot)\}$, Proposition C.2 is a result under the most general conditions. Corollary C.3 and Proposition C.4 give further concreteness for how exactly to guarantee the premise of Proposition C.2 with respect to $\hat{G}(y, \delta, x)$. The idea behind the generic statement in Proposition C.2 is to encourage extensions, especially those with the possibility of being more statistically efficient, with less restricted conditions than those in Proposition C.4, or with modeling assumptions that are tailored to the application at hand.

Proposition C.2 (A model-free inequality). *If Assumption 1.5 holds, we have for all $\delta \in \mathbb{R}$ and all $x \in \text{support}(X_i)$:*

$$\left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \leq \sup_y \left| \hat{G}_n(y, \delta, x) - G(y, \delta, x) \right|. \quad (13)$$

The implications of the deterministic inequality in (13) are interesting and perhaps a bit surprising. This inequality is stating that in a finite sample, the conditional bound estimators $\{\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x)\}$ in (11) and (12) are jointly no less accurate at estimating $\{\theta^L(\delta, x), \theta^U(\delta, x)\}$ as the choice of $\hat{G}_n(y, \delta, x)$ is for $G(y, \delta, x)$, whatever the choice may be.

We can also turn (13) into a statement of frequentist confidence: if, additionally, $\hat{G}_n(y, \delta, x)$ is such that there exists a value $t_\alpha(\delta, x)$ such that:

$$\text{pr} \left\{ \sup_y \left| \hat{G}_n(y, \delta, x) - G(y, \delta, x) \right| \leq t_\alpha(\delta, x) \right\} \geq 1 - \alpha,$$

then we have that:

$$\text{pr} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \leq t_\alpha(\delta, x) \right\} \geq 1 - \alpha.$$

The term $t_\alpha(\delta, x)$ need not vary with x or δ ; it can also be with respect to the concentration of $\hat{G}_n(y, \delta, x)$ uniformly across x or across δ if more appropriate. Proposition C.4 below is an example with such a uniform guarantee.

With regard to specifying $\hat{G}_n(y, \delta, x)$, one choice is to plug in estimators of $F_w(y | x)$ ($w = 0, 1$) to arrive at a concentration inequality for the conditional bound estimators as Corollary C.3 suggests. In doing so, provided the appropriate guarantee exists for the conditional CDF estimators, we actually get a strong guarantee for the bound estimators $\hat{\theta}^L(\delta, x)$ and $\hat{\theta}^U(\delta, x)$ that is simultaneous across all threshold values δ used to define PIBT in Definition 1.1.

Corollary C.3 (Concentration when plugging in conditional CDF estimators). *Let $\hat{F}_{wn}(y|x)$ denote an estimator for the conditional CDF, $F_{wn}(y|x)$ ($w = 0, 1$). If Assumption 1.5 holds, and the estimators $\hat{F}_{wn}(y|x)$ ($w = 0, 1$) are such that there exists a value $t_{w,\alpha}(x)$ satisfying:*

$$\text{pr} \left\{ \sup_y \left| \hat{F}_{wn}(y | x) - F_w(y | x) \right| \leq t_{w,\alpha}(x) \right\} \geq 1 - \alpha/2,$$

then we have that:

$$\text{pr} \left[\sup_\delta \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \leq \sum_{w=0,1} t_{w,\alpha}(x) \right] \geq 1 - \alpha.$$

C.4 More explicit conditional bounds with strategic use of regression residuals

The question now becomes how exactly to specify $\hat{F}_{wn}(y | x)$ in Corollary C.3, while guaranteeing the closeness between $\{\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x)\}$ and $\{\theta^L(\delta, x), \theta^U(\delta, x)\}$ at some target confidence level. We explore one such choice using regression residuals for which such a high confidence guarantee is possible as summarized in Proposition C.4. The motivation is that we would like something very similar to the plug-in estimator of $\text{pr}\{Y_i(w) \leq y\}$ given in (3) for the randomized experiment case. For the case that a locally linear approximation is desired, we give an actionable application of this result with regression residuals in § C.4.1.

Let $(X_1, W_1, Y_1(0), Y_1(1)), \dots, (X_n, W_n, Y_n(0), Y_n(1))$ be independent and identically distributed copies from a joint distribution. Denote the observable training data as:

$$\mathcal{T} := \{(X_i, W_i, Y_i)\}_{i=1}^n,$$

where $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. Consider partitioning \mathcal{T} into two independent splits \mathcal{T}_1 and \mathcal{T}_2 . Denote the corresponding training indices as $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, \dots, n\}$ for \mathcal{T}_1 and \mathcal{T}_2 , respectively. Denote $S_w := \{i : W_i = w\}$, the index set of individuals in the sample in treatment group $w = 0, 1$. Let $n_w := |S_w \cap \mathcal{I}_2|$, the sample size in treatment group $w = 0, 1$ coming from data split \mathcal{T}_2 .

Further, denote

$$\mu_w(x) := E \{Y_i(w) \mid X_i = x\},$$

the conditional expectation of the potential outcome as a function of the pre-treatment covariates. Denote the regression estimate using \mathcal{T}_1 as $\hat{\mu}_w(x)$. Importantly, we will be able to reason about the counterfactual quantity $\mu_w(x)$ under Assumption 1.5, because:

$$\mu_w(x) = E \{Y_i \mid W_i = w, X_i = x\},$$

the conditional expectation of the observed outcome in treatment group $w = 0, 1$. Now consider the residuals:

$$R_i(w) := Y_i(w) - \mu_w(X_i).$$

Denote the approximation of $R_i(w)$ using $\hat{\mu}_w(\cdot)$ as

$$\hat{R}_i(w) := Y_i(w) - \hat{\mu}_w(X_i).$$

Motivated by the use of independent and identically distributed draws used to define $\hat{F}_{wn}(y)$ in (3) for the marginal PIBT in a randomized experiment, we would like to approximate draws from the distribution

$$Y_i \mid X_i = x, W_i = w.$$

With this in mind, we will specify:

$$\hat{F}_{wn}(y \mid x) := \frac{1}{n_w} \sum_{i \in S_w \cap \mathcal{I}_2} 1 \left\{ \hat{\mu}_w(x) + \hat{R}_i(w) \leq y \right\}. \quad (14)$$

Considering that the definition of $R_i(w)$ means that $Y_i(w) = \mu_w(x) + R_i(w)$ conditional on $X_i = x$, it seems that using $\hat{\mu}_w(x) + \hat{R}_i(w)$ will make this choice of $\hat{F}_{wn}(y \mid x)$ a reasonable approximation to $F_w(y \mid x)$.

Noting the liberal use of the plug-in principle en route to the choice of $\{\hat{\theta}^L(\cdot, \cdot), \hat{\theta}^U(\cdot, \cdot)\}$ using the conditional CDF estimators in (14), a concern is now what the regularity conditions must be so that overall the conditional bound estimators are close to their true values. For any given value of x , $\hat{F}_{wn}(y \mid x)$ is reusing residuals for indices in split \mathcal{I}_2 corresponding to subjects that are not necessarily in stratum x . Implicit in this use is that the distribution of $\{R_i(0), R_i(1)\}$ is the same across values of x . That is, we are using the independence assumption:

$$\{R_i(0), R_i(1)\} \perp\!\!\!\perp X_i \quad (i = 1, \dots, n).$$

Beyond this regularity condition, we also require that the distribution of $\hat{R}_i(w)$ approximates well the distribution of $R_i(w)$, which in turn requires that $\hat{\mu}_w(\cdot)$ be close to $\mu_w(\cdot)$. This explains the correction to the confidence level in Proposition C.4 with respect to how likely a deviation, in a uniform sense, is to occur between the true regression curve and the estimated regression curve based on random training data.

In Proposition C.4, $X \in R^{|\mathcal{I}_1| \times p}$ is such that its rows are comprised of $X_i^T \in R^{1 \times p}$ for $i \in \mathcal{I}_1$.

Proposition C.4 (Justifying approach with regression residuals). *For the bound estimators in (11) and (12), let us specify $\hat{G}(y \mid x) = \hat{F}_{1n}(y \mid x) - \hat{F}_{0n}(y \mid x)$ using the conditional CDF estimators of (14). Let Assumption 1.5 hold, and assume further that the arbitrary joint distribution of $(R_i(0), R_i(1), X_i)$ is such that*

$$(R_i(0), R_i(1)) \perp\!\!\!\perp X_i \quad (i = 1, \dots, n).$$

Conditional on X , we have for any appropriate $0 \leq \alpha \leq 1$:

$$\begin{aligned} & \sup_{\delta, x} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \\ & \leq \sum_{w=0,1} \sup_r [\Pr \{r < R_i(w) \leq r + 2t_w\} \vee \Pr \{r - 2t_w < R_i(w) \leq r\}] \\ & \quad + \left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} \left(n_0^{-1/2} + n_1^{-1/2} \right) \end{aligned}$$

with probability at least

$$1 - \alpha - \sum_{w=0,1} \text{pr} \left\{ \sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_w \mid X \right\}.$$

Here, $t_0, t_1 \geq 0$ may depend on X . If they do not, we may remove the conditional statements. Appropriate values of α are those such that $1 - \alpha - \sum_{w=0,1} \text{pr} \{ \sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t \}$ is between 0 and 1.

The proof of Proposition C.4 involves a strategic application of the Dvoretzky-Kiefer-Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990; Naaman, 2021) that is tailored to the imputed draws from the conditional potential outcome distribution. Consider the following corollary to Proposition C.4.

Corollary C.5 (Efficiency of Conditional Bound Estimators in Proposition C.4). *Let \mathcal{F} be the function class containing our regression estimator, $\hat{\mu}_w(\cdot)$. Assume there exists a sequence $g_{n,\mathcal{F}}$, depending on n and the complexity of \mathcal{F} , like feature dimension or regularisation parameters, such that*

$$\max_{w=0,1} \sup_x |\hat{\mu}_w(x) - \mu_w(x)| \lesssim g_{n,\mathcal{F}}$$

with probability at least $1 - \alpha$. Then:

$$\sup_{\delta,x} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \lesssim g_{n,\mathcal{F}} + \left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right)$$

holds with probability at least $1 - 2\alpha$.

The rate of convergence of the regression estimator, $g_{n,\mathcal{F}}$, decreases down to 0 as n increases under proper specification. Should this rate satisfy

$$\left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right) \lesssim g_{n,\mathcal{F}},$$

Corollary C.5 means that we lose nothing with respect to asymptotic efficiency with the plug-in estimators in (11) and (12) that build on an estimator of $\{\mu_0(\cdot), \mu_1(\cdot)\}$. Provided the framework gives $\hat{\mu}_w(x)$ ($w = 0, 1$), one can apply Proposition C.4 with any meta-learning algorithms that are used to estimate the conditional average treatment effect function (CATE) (Künzel et al., 2019; Nie and Wager, 2020; Athey et al., 2019; Wager and Athey, 2018; Kennedy, 2020; Burkhardt and Ruiz, 2022).

C.4.1 An application to the linear Gaussian regression setup

We now explicitly apply Proposition C.4 to a case where we can derive non-asymptotic uniform guarantees on the regression function. It can be of interest when a locally linear approximation is desired, as made explicit by the following additional assumption.

Assumption C.6 (Restricted data generating mechanism). We assume that

$$Y_i(w) = \mu_w(X_i) + R_i(w) \quad (i = 1, \dots, n).$$

Here, $X_i \in \mathbb{R}^p$, $\mu_w(X_i) = \beta_w^T \Psi_w(X_i)$, and $\Psi_w : \mathbb{R}^p \rightarrow \mathbb{R}^d$ is a fixed mapping such that $\|\Psi_w(x)\|_2 \leq 1$ and $d < n$. Moreover, $(R_i(0), R_i(1))$ ($i = 1, \dots, n$) follow a joint distribution satisfying $(R_i(0), R_i(1)) \perp\!\!\!\perp X_i$ and such that marginally,

$$R_i(w) \sim \mathcal{N}(0, \sigma_w^2); \quad \sigma_w^2 > 0 \quad (w = 0, 1).$$

Denote the training data as:

$$\mathcal{T} := \{(X_i, W_i, Y_i)\}_{i=1}^n.$$

Consider partitioning \mathcal{T} into two splits indexed by the disjoint sets $\mathcal{I}_1, \mathcal{I}_2 \subseteq \{1, \dots, n\}$. Denote $S_w := \{i : W_i = w\}$, the index set of individuals in the sample in treatment group $w = 0, 1$. Let $n_w := |S_w \cap \mathcal{I}_2|$, the sample size in treatment group $w = 0, 1$ coming from data split \mathcal{T}_2 . Now, let $\Lambda_w \in \mathbb{R}^{|\mathcal{I}_1 \cap S_w| \times d}$ be such that its rows are made up by stacking $\Psi_w(X_i)^T$ for each $i \in \mathcal{I}_1 \cap S_w$. Further, let $Y_w \in \mathbb{R}^{|\mathcal{I}_1 \cap S_w| \times 1}$ contain entries for the corresponding observed outcome Y_i for each $i \in \mathcal{I}_1 \cap S_w$.

In Proposition C.7, Φ denotes the CDF of the standard normal distribution. For a matrix $A \in \mathbb{R}^{m \times q}$, denote its operator norm as:

$$\|A\|_{op} := \sup_{v \in \mathbb{R}^q: \|v\|_2=1} \|Av\|_2.$$

Proposition C.7. Assume $\text{rank}(\Lambda_w) = d$ ($w = 0, 1$) almost surely. Let $v_{d,\alpha}$ denote the $(1 - \alpha/2)$ th quantile for the χ_d^2 distribution. Under Assumption C.6, we have with confidence at least $(1 - 2\alpha) \times 100\%$ that uniformly across all pre-treatment covariate strata x ,

$$\theta(\delta, x) = \text{pr} \{Y_i(1) - Y_i(0) > \delta \mid X_i = x\}$$

is contained in the interval with starting point

$$\begin{aligned} \hat{\theta}^L(\delta, x) &= \sum_{w=0,1} \Phi \left\{ (v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \right\} \\ &\quad + \sum_{w=0,1} \Phi \left\{ -(v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \right\} \\ &\quad - \left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right) \end{aligned}$$

and end point

$$\begin{aligned} \hat{\theta}^U(\delta, x) &+ \sum_{w=0,1} \Phi \left((v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \right) \\ &- \sum_{w=0,1} \Phi \left(-(v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \right) \\ &+ \left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} \left(n_0^{-\frac{1}{2}} + n_1^{-\frac{1}{2}} \right). \end{aligned}$$

Although we do not prove that the term added and subtracted to the bound estimators $\{\hat{\theta}^L(\delta, x), \hat{\theta}^U(\delta, x)\}$ in Proposition C.7 goes to zero for a general feature matrix (Λ_0, Λ_1) as sample size increase, the illustration of Proposition C.7 in Figure 6 of the appendix gives empirical evidence of this desirable decay. The proof of the inequality in Proposition C.7 is contained in Appendix E.8.1. It is an application of Proposition C.4 in the appendix, which specifies the bound estimators in terms of arbitrary mean-regression estimators $\{\hat{\mu}_0(x), \hat{\mu}_1(x)\}$.

We believe Proposition C.7 may be regarded as partly representative of what can occur in practice while attempting to infer bounds on $\text{pr} \{Y_i(1) - Y_i(0) > \delta \mid X_i = x\}$ uniformly across x . The primary concerns for satisfactory margins of error are the typical concerns of regression: parsimony, multicollinearity, and feature dimension, all of which affect the convergence to zero of the terms involving Λ_w ($w = 0, 1$) in Proposition 6.

C.4.2 An example power analysis using Proposition C.4 and Proposition C.7

Under Assumption C.6, Fig. 6 illustrates the behavior of the margin of error for

$$\sup_{\delta, x} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\}$$

at the 90% confidence level using the result in Proposition C.7. As one can imagine, the distribution of X_i , the transformation Ψ_w in Assumption C.6, and the propensity score matter for an application of Proposition C.7. That is, there may very well exist cases where the operator norm of $(\Lambda_w^T \Lambda_w)^{-1/2}$ does not decrease with n , along with cases of the propensity score where this norm decreases slowly due to insufficient treated or control units in the sample. To study this, the example summarized in Fig. 6 generates data as follows.

We sample from a population in which each $X_{ij} \sim \text{Uniform}(0, 1)$ across $i = 1, \dots, n$ and $j = 1, \dots, p$. The degree $q = 1, 2$ polynomial transformation of X_i into $\Psi_w(X_i)$ includes all possible interaction terms of degree $1 \leq k \leq q$, and each entry is re-scaled by $1/d^{1/2}$ so that $\|\Psi_w(X_i)\|_2 \leq 1$ as in Assumption C.6. For example, $\Psi_w(X_i) = (X_{i1}, X_{i2}, X_{i1}X_{i2}, X_{i1}^2, X_{i2}^2)/5^{1/2}$ when $p = q = 2$. The propensity score is $\text{pr} \{W_i = 1 \mid X_i\} = \{(X_{i1} + X_{i2} + 0.5)/3\}^m$ ($m = 1, 6$). The case with $m = 1$ makes it so that the number of treated and control units are very close to each other in a random sample, while the case with $m = 6$ will make it so that control units are typically much more represented. Moreover, the sample splitting is such that half of the observations are used to estimate $\{\mu_0(\cdot), \mu_1(\cdot)\}$, while the other half of the observations' residuals are used to estimate $\{\theta^L(\cdot, \cdot), \theta^U(\cdot, \cdot)\}$.

At sample size n , what is the uniform Margin of Error on efficacy bounds at the 90% confidence level?

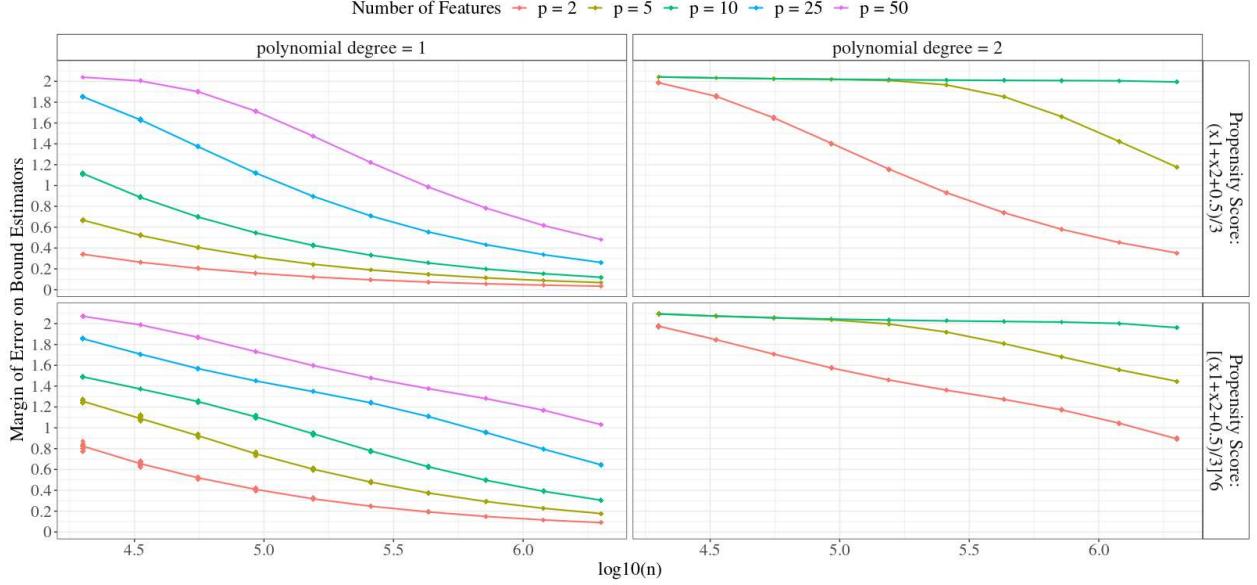


Figure 6: Power analysis based on Proposition C.7. Each curve is the median calculated margin of error across 30 Monte Carlo simulations at the points that are also plotted. For these Monte Carlo draws, we generate $X_{ij} \sim \text{Uniform}(0, 1)$ ($i = 1, \dots, n$; $j = 1, \dots, p$).

D More on the scope of our results

D.1 Important concerns

The contribution of this paper is in the quantification of the margin of error for the bounds on PIBT given by the Makarov bounds (Makarov, 1982; Frank et al., 1987; Williamson and Downs, 1990; Fan and Park, 2010). Three important questions may come up regarding the scope of our results. For binary potential outcomes, is the use of the Makarov bounds on par with the use of the sharp Fréchet-Boole bounds (Boole, 1854; Fréchet, 1935, 1960; Hailperin, 1986; Mueller et al., 2022)? Moreover, can we have an alternative definition of benefiting from treatment in terms of the ratio of potential outcomes when the potential outcomes are strictly positive? Additionally, can we bound the proportion who are harmed by an intervention? The answer is yes to all three questions as we discuss in the next three subsections, respectively.

D.2 When the potential outcomes are binary, the Makarov bounds on PIBT are the same as the Boole-Fréchet bounds

Recall that the probability of necessity and sufficiency (PNS) for binary potential outcomes (Pearl, 1999; Tian and Pearl, 2000; Pearl, 2009) is the joint probability,

$$\text{pr}\{Y_i(1) = 1, Y_i(0) = 0\}.$$

For two measurable events A and B , the Boole-Fréchet bounds on their joint probability are:

$$\max\{\text{pr}\{A\} + \text{pr}\{B\} - 1, 0\} \leq \text{pr}\{A \cap B\} \leq \min\{\text{pr}\{A\}, \text{pr}\{B\}\}.$$

These bounds are sharp (Rüschenhoff, 1981), meaning the bounds are attainable. For example, if $A \cap B = \emptyset$, then $\text{pr}\{A \cap B\} = \max\{\text{pr}\{A\} + \text{pr}\{B\} - 1, 0\}$ holds. Moreover, if $A \subseteq B$, then $\text{pr}\{A \cap B\} = \min\{\text{pr}\{A\}, \text{pr}\{B\}\}$ is the case.

We claim in Proposition D.1 that the Makarov bounds on PNS are the same as the Boole-Fréchet bounds. From this proposition, it follows that all the estimation results we have shown are as applicable to the binary potential outcome case as any estimation approach involving the Boole-Fréchet bounds.

Proposition D.1 (Boole-Fréchet bounds vs. Makarov bounds). *When the potential outcomes are binary, we have that the tightest Makarov bounds on PNS are:*

$$\sup_{0 \leq \delta < 1} \theta^L(\delta) \leq \text{pr}\{Y_i(1) = 1, Y_i(0) = 0\} \leq \inf_{0 \leq \delta < 1} \theta^U(\delta)$$

in the marginal case, along with

$$\sup_{0 \leq \delta < 1} \theta^L(\delta, x) \leq \text{pr}\{Y_i(1) = 1, Y_i(0) = 0 \mid X_i = x\} \leq \inf_{0 \leq \delta < 1} \theta^U(\delta, x)$$

in the conditional case. Moreover, these tightest Makarov bounds are the same as the Boole-Fréchet bounds:

$$\begin{aligned} \sup_{0 \leq \delta < 1} \theta^L(\delta) &= \max[\text{pr}\{Y_i(1) = 1\} + \text{pr}\{Y_i(0) = 0\} - 1, 0] \\ \inf_{0 \leq \delta < 1} \theta^U(\delta) &= \min[\text{pr}\{Y_i(1) = 1\}, \text{pr}\{Y_i(0) = 0\}] \end{aligned}$$

and

$$\begin{aligned} \sup_{0 \leq \delta < 1} \theta^L(\delta, x) &= \max[\text{pr}\{Y_i(1) = 1 \mid X_i = x\} + \text{pr}\{Y_i(0) = 0 \mid X_i = x\} - 1, 0] \\ \inf_{0 \leq \delta < 1} \theta^U(\delta, x) &= \min[\text{pr}\{Y_i(1) = 1 \mid X_i = x\}, \text{pr}\{Y_i(0) = 0 \mid X_i = x\}]. \end{aligned}$$

The proof of Proposition D.1 can be found in Appendix E.9. The general idea is to use that the CDFs of binary potential outcomes have only three values, 0, $\text{pr}\{Y_i(w) = 0\}$, and 1, across evaluation points $y \in \mathbb{R}$. We also use the fact that PIBT is the same as PNS when the threshold δ is in $[0, 1)$.

D.3 PIBT in terms of the ratio of potential outcomes

We also have the following simple extension to ratios of potential outcomes. It is motivated by Theorem 2 of Williamson and Downs (1990), which derives bounds for the CDF of a sum, difference, product, or ratio of two random variables with an unknown joint distribution.

Suppose that we are interested in strictly positive potential outcomes $\tilde{Y}_i(0)$ and $\tilde{Y}_i(1)$. For example, this can be in a setting where the time to an event is the outcome of interest (Cox, 1972; Stitelman and van der Laan, 2010; Austin, 2014; Schober and Vetter, 2018; Cai and van der Laan, 2020). Using a threshold $\tilde{\delta} > 0$ (e.g. $\tilde{\delta} = 1$), one may alternatively consider an individual to have benefited from treatment should the inequality

$$\tilde{Y}_i(1)/\tilde{Y}_i(0) > \tilde{\delta}$$

occur. That is, individual i is deemed to have benefited from treatment should their treatment outcome be larger than their control outcome by a factor larger than $\tilde{\delta}$. Correspondingly, one may be interested in bounding the unidentifiable probability

$$\tilde{\theta}(\tilde{\delta}) := \text{pr}\left\{\tilde{Y}_i(1)/\tilde{Y}_i(0) > \tilde{\delta}\right\}.$$

To do so, consider the variable transformation

$$Y_i(w) \doteq \log\left\{\tilde{Y}_i(w)\right\} (w = 0, 1).$$

Given our definition of $\theta(\delta)$ in Definition 1.1 and the one-to-one nature of the log transformation, we get that

$$\tilde{\theta}(\tilde{\delta}) = \theta(\delta)$$

when we set $\delta \doteq \log(\tilde{\delta})$. It follows that:

$$\theta^L(\delta) \leq \tilde{\theta}(\tilde{\delta}) \leq \theta^U(\delta).$$

We will simply have to work with the eCDF of $\log\{\tilde{Y}_i(w)\}$ when obtaining the estimators $\hat{\theta}^L(\delta)$ and $\hat{\theta}^U(\delta)$. Moreover, Proposition 2.2 here is still useful to conduct inference on $\tilde{\theta}(\tilde{\delta})$ under the independent and identically distributed assumption. Likewise, the results in Section C are also applicable when we work with the log-transformation.

D.4 Reasoning about the proportion harmed by an intervention

Two days following the posting of our first manuscript on arXiv, a similar manuscript by Kallus (2022) was also posted on arXiv. This work studies the probability an individual is harmed by an intervention (PIHI) in the case of binary potential outcomes. Supposing instead that $Y_i = 1$ is bad for an individual, while $Y_i = 0$ is good for an individual, we have that the PIHI is PNS:

$$\text{pr} \{Y_i(1) = 1, Y_i(0) = 0\}.$$

This is simply, but importantly, due to notational and vocabulary semantics. Moreover, if the potential outcomes are real-valued, then one can refer to the quantity

$$\text{pr} \{Y_i(1) - Y_i(0) > \delta\}$$

as PIHI instead of calling this quantity PIBT as we do in Definition 1.1. This holds analogously when we condition on $X_i = x$.

Given the discussion surrounding Proposition D.1, our model-free results extend to binary potential outcome cases for PIHI as well. Given that the Boole-Fréchet inequalities underlay the theoretical estimation results for binary potential outcomes in Kallus (2022), we believe that the contribution of their work compared to our present work is as follows. Most importantly, their results are for binary potential outcomes, while our results hold for binary, real-valued, or strictly positive potential outcomes. Next, their results include a doubly robust (Robins et al., 1994; Funk et al., 2011) estimation method which uses the estimated propensity score to adjust for covariates, possibly in an observational setting, when bounding marginal PIHI. This warrants further investigation for the case of non-binary potential outcomes. Moreover, their presentation of estimation theorems is in terms of big-O probability notation, i.e. statistical rates of convergence. We present results in terms of non-asymptotic concentration statements. The subtle difference is that our presentation helps provide nominal coverage guarantees for the confidence bands on PIBT (or PIHI). Further, confidence bands for PIHI presented in their work make use of a standard error: $n^{-\frac{1}{2}}$ times the standard deviation of a random variable being averaged to obtain the bound estimators. These standard-error-based confidence intervals, which are due to a central limit theorem, do not achieve nominal coverage until a fairly large sample size as demonstrated in their numerical results. See their Algorithm 1 and Fig. 3 for details. Finally, the non-asymptotic presentation of our results, having kept track of all constants, can also help with a statistical power analysis, as discussed in Sections 2.4 and C.4.2. However, keeping track of constants can prove difficult for some applications of Propositions C.2 and C.4. To extend the utility of Propositions C.2 and C.4, it may be interesting to study whether plausible bounds on these constants can be specified in such regression settings.

E Proofs of the main theoretical results

E.1 Overview

The existing mathematical results we exploit include the following. Key to establishing the population-level target bounds on PIBT, we use the Makarov bounds first introduced in Makarov (1982) and generalized in Frank et al. (1987) and Williamson and Downs (1990). These works establish a distribution-free bound on the cumulative distribution function (CDF) on the sum (or difference or product) of two or more random variables having any unknown joint distribution and fixed marginal distributions. For the non-asymptotic concentration results (the margin of error derivations), the novel contribution of this paper, we use the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990; Naaman, 2021). We will refer to it as the DKW inequality for readability. The DKW inequality gives a non-parametric, non-asymptotic deviation inequality for the supremum difference at any evaluation point between a target CDF and its empirical analogue estimated with a sample of independent and identically distributed random variables. Importantly, Massart (1990) and Naaman (2021) show that this inequality is tight under no additional assumptions.

E.2 Some key lemmas

Lemma E.1 (Plug-in estimation of Makarov (1982)'s conditional bounds). *Consider jointly distributed random variables (U_0, U_1, V) . Denote:*

$$\gamma^L(\delta, v) := -\min \left[\inf_u \{ \text{pr} \{U_1 \leq u + \delta/2 \mid V = v\} - \text{pr} \{U_0 \leq u - \delta/2 \mid V = v\} \}, 0 \right]$$

and

$$\gamma^U(\delta, v) := 1 - \max \left[\sup_u \{ \text{pr} \{U_1 \leq u + \delta/2 \mid V = v\} - \text{pr} \{U_0 \leq u - \delta/2 \mid V = v\} \}, 0 \right],$$

the Makarov (1982) and Williamson and Downs (1990) lower and upper bounds for $\text{pr}\{U_1 - U_0 > \delta \mid V = v\}$. Denote

$$H(u, \delta, v) := \text{pr}\{U_1 \leq u + \delta/2 \mid V = v\} - \text{pr}\{U_0 \leq u - \delta/2 \mid V = v\}.$$

Consider any estimator $\hat{H}(u, \delta, v)$ of $H(u, \delta, v)$ based on a sample

$$\{(U_{i0}, V_i)\}_{i=1}^{n_0} \cup \{(U_{i1}, V_i)\}_{i=1}^{n_1}$$

such that (U_{i0}, V_i) ($i = 1, \dots, n_0$) are independent and identically distributed copies of (U_0, V) , while (U_{i1}, V_i) ($i = 1, \dots, n_1$) are independent and identically distributed copies of (U_1, V) . Now let

$$\hat{\gamma}^L(\delta, v) := -\min \left[\inf_u \left\{ \hat{H}(u, \delta, v) \right\}, 0 \right]$$

and

$$\hat{\gamma}^U(\delta, v) := 1 - \max \left[\sup_u \left\{ \hat{H}(u, \delta, v) \right\}, 0 \right].$$

We claim for every $\delta \in \mathbb{R}$ and every $v \in \text{support}(V)$ that:

$$\begin{aligned} \hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v) &\leq \sup_u \left\{ H(u, \delta, v) - \hat{H}(u, \delta, v) + |\hat{H}(u, \delta, v) - H(u, \delta, v)| \right\} / 2; \\ \gamma^U(\delta, v) - \hat{\gamma}^U(\delta, v) &\leq \sup_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) + |\hat{H}(u, \delta, v) - H(u, \delta, v)| \right\} / 2; \end{aligned}$$

and

$$|\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v)| \vee |\hat{\gamma}^U(\delta, v) - \gamma^U(\delta, v)| \leq \sup_u |\hat{H}(u, \delta, v) - H(u, \delta, v)|.$$

Proof. For any real-valued function $g(t)$, denote its positive and negative parts, respectively, as:

$$g^+(t) = -\max[g(t), 0] \text{ and } g^-(t) = -\min[g(t), 0].$$

We have the following properties we will make use of:

$$\begin{aligned} g(t) &= g^+(t) - g^-(t); \\ |g(t)| &= g^+(t) + g^-(t); \\ g^+(t) &= [|g(t)| + g(t)] / 2; \\ g^-(t) &= [|g(t)| - g(t)] / 2. \end{aligned}$$

Below, we will use the positive and negative parts of

$$g_{\text{inf}}(\delta, v) := \inf_u H(u, \delta, v) \text{ and } g_{\text{sup}}(\delta, v) := \sup_u H(u, \delta, v)$$

along with

$$\hat{g}_{\text{inf}}(\delta, v) := \inf_u \hat{H}(u, \delta, v) \text{ and } \hat{g}_{\text{sup}}(\delta, v) := \sup_u \hat{H}(u, \delta, v).$$

Consider:

For the lower bound on $\text{pr}\{U_1 - U_0 > \delta \mid V = v\}$, we are bounding the difference

$$\begin{aligned} &\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v) \\ &= \hat{g}_{\text{inf}}^-(\delta, v) - g_{\text{inf}}^-(\delta, v) \\ &\stackrel{(i)}{=} \frac{1}{2} [g_{\text{inf}}(\delta, v) - \hat{g}_{\text{inf}}(\delta, v) + |\hat{g}_{\text{inf}}(\delta, v)| - |g_{\text{inf}}(\delta, v)|] \\ &\stackrel{(ii)}{\leq} \frac{1}{2} [g_{\text{inf}}(\delta, v) - \hat{g}_{\text{inf}}(\delta, v) + |\hat{g}_{\text{inf}}(\delta, v) - g_{\text{inf}}(\delta, v)|] \\ &= \frac{1}{2} \left[\inf_u H(u, \delta, v) - \inf_u \hat{H}(u, \delta, v) + \left| \inf_u \hat{H}(u, \delta, v) - \inf_u H(u, \delta, v) \right| \right]. \end{aligned}$$

In (i), we used the properties of the negative part of a function introduced above, while in (ii) we used reverse triangle inequality. Similarly, we have that:

$$|\gamma^L(\delta, v) - \hat{\gamma}^L(\delta, v)| \leq \left| \inf_u \hat{H}(u, \delta, v) - \inf_u H(u, \delta, v) \right|.$$

Consider that

$$\begin{aligned}
& \inf_u \hat{H}(u, \delta, v) - \inf_u H(u, \delta, v) \\
&= \inf_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) + H(u, \delta, v) \right\} - \inf_u H(u, \delta, v) \\
&\stackrel{(i)}{=} \inf_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) \right\} \\
&\leq \sup_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) \right\}
\end{aligned}$$

Equality (i) follows from the fact that $\inf\{a + b : a \in A, b \in B\} = \inf(A) + \inf(B)$. Similarly,

$$\begin{aligned}
& \inf_u H(u, \delta, v) - \inf_u \hat{H}(u, \delta, v) \\
&= \inf_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) + H(u, \delta, v) \right\} - \inf_u H(u, \delta, v) \\
&= \inf_u \left\{ H(u, \delta, v) - \hat{H}(u, \delta, v) \right\} \\
&\leq \sup_u \left\{ H(u, \delta, v) - \hat{H}(u, \delta, v) \right\}
\end{aligned}$$

The previous two equations imply that

$$\left| \inf_u H(u, \delta, v) - \inf_u \hat{H}(u, \delta, v) \right| \leq \sup_u \left| \hat{H}(u, \delta, v) - H(u, \delta, v) \right|.$$

So overall we have that

$$\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v) \leq \sup_u \left[H(u, \delta, v) - \hat{H}(u, \delta, v) + \left| \hat{H}(u, \delta, v) - H(u, \delta, v) \right| \right] / 2 \quad (15)$$

along with

$$\left| \hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v) \right| \leq \sup_u \left| \hat{H}(u, \delta, v) - H(u, \delta, v) \right|. \quad (16)$$

Here, (15) is the first desired conclusion.

For the upper bound on $\text{pr}\{U_1 - U_0 > \delta \mid V = v\}$, we are bounding the difference

$$\begin{aligned}
& \gamma^U(\delta, v) - \hat{\gamma}^U(\delta, v) \\
&= \hat{g}_{\sup}^+(\delta, v) - g_{\sup}^+(\delta, v) \\
&\stackrel{(i)}{=} \frac{1}{2} [\hat{g}_{\sup}(\delta, v) - g_{\sup}(\delta, v) + |\hat{g}_{\sup}(\delta, v)| - |g_{\sup}(\delta, v)|] \\
&\stackrel{(ii)}{\leq} \frac{1}{2} [\hat{g}_{\sup}(\delta, v) - g_{\sup}(\delta, v) + |g_{\sup}(\delta, v) - \hat{g}_{\sup}(\delta, v)|] \\
&= \frac{1}{2} \left\{ \sup_u \hat{H}(u, \delta, v) - \sup_u H(u, \delta, v) + \left| \sup_u H(u, \delta, v) - \sup_u \hat{H}(u, \delta, v) \right| \right\}.
\end{aligned}$$

In (i), we used the properties of the positive part of a function introduced above, while in (ii) we used triangle inequality followed by reverse triangle inequality. Noting that $\sup\{a + b : a \in A, b \in B\} = \sup(A) + \sup(B)$, we can arrive at the below equalities based on similar steps to the case with the lower bound:

$$\sup_u H(u, \delta, v) - \sup_u \hat{H}(u, \delta, v) = \sup_u \left\{ H(u, \delta, v) - \hat{H}(u, \delta, v) \right\}$$

and

$$\sup_u \hat{H}(u, \delta, v) - \sup_u H(u, \delta, v) = \sup_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) \right\}.$$

So overall we have that:

$$\gamma^U(\delta, v) - \hat{\gamma}^U(\delta, v) \leq \sup_u \left\{ \hat{H}(u, \delta, v) - H(u, \delta, v) + \left| \hat{H}(u, \delta, v) - H(u, \delta, v) \right| \right\} / 2, \quad (17)$$

along with

$$|\hat{\gamma}^U(\delta, v) - \gamma^U(\delta, v)| \leq \sup_u |\hat{H}(u, \delta, v) - H(u, \delta, v)|, \quad (18)$$

as with the lower bound estimate. Here, (17) is the second desired conclusion.

The inequalities in (16) and (18) give us the third desired conclusion:

$$\{|\hat{\gamma}^L(\delta, v) - \gamma^L(\delta, v)| \vee |\hat{\gamma}^U(\delta, v) - \gamma^U(\delta, v)|\} \leq \sup_u |\hat{H}(u, \delta, v) - H(u, \delta, v)|.$$

□

Lemma E.2 (Inequality for a sum of random variables). *Let U and V be arbitrary real-valued random variables, and let $u, v \in \mathbb{R}$ be non-random scalars. We have that:*

$$\text{pr}\{U + V > u + v\} \leq \text{pr}\{U > u\} + \text{pr}\{V > v\}.$$

Proof. Consider that we have the following:

$$\{U + V \leq u + v\} \supseteq \{U \leq u\} \cap \{V \leq v\} \iff \{U + V > u + v\} \subseteq \{U > u\} \cup \{V > v\}.$$

This containment of events holds because $U \leq u$ and $V \leq v$ implies that $U + V \leq u + v$. It follows that

$$\text{pr}\{U + V > u + v\} \leq \text{pr}\{\{U > u\} \cup \{V > v\}\} \leq \text{pr}\{U > u\} + \text{pr}\{V > v\},$$

where the second inequality is due to union bound. □

Lemma E.3 (Uniform estimation guarantee for the IPTW eCDF in Equation (3)).

We have that

$$\text{pr}\left\{\sup_y \left|\hat{F}_w(y|x) - F_w(y|x)\right| \leq n_{\ell(x)}^{-1/2} C_{w,\beta}(x)\right\} \geq 1 - \beta.$$

Here,

$$C_{w,\beta}(x) := \begin{cases} [2^{-1} n_{\ell(x)} n_{w\ell(x)}^{-1} \log(2/\beta)]^{1/2} & \text{if } e(x') = n_{1\ell(x)}/n_{\ell(x)} \text{ for each } x' \in \mathcal{X}_{\ell(x)} \\ \frac{2[2 \log^{\frac{1}{2}}(n_{\ell(x)}+1) + \log^{\frac{1}{2}}(1/\beta)]}{n_{\ell(x)}^{1/2} (w\bar{e} + (1-w)(1-\bar{e}))} & \text{otherwise} \end{cases}.$$

Proof.

Case 1: randomized experiment

For the case the $e(x') = n_{1\ell(x)}/n_{\ell(x)}$ for each $x' \in \mathcal{X}_{\ell(x)}$, the estimator in (3) reduces to:

$$\frac{1}{n_{w\ell(x)}} \sum_{i: W_i=w, \ell(X_i)=\ell(x)} 1\{Y_i \leq y\}.$$

When $(Y_1, W_1, X_1), \dots, (Y_n, W_n, X_n)$ are iid copies to begin with, it follows that Y_i is iid among indices such that $W_i = w$ and $\ell(X_i) = \ell(x)$. So the iid assumption of the Dvoretzky-Kiefer-Wolfowitz-Massart (DKWM) inequality holds (Dvoretzky et al., 1956; Massart, 1990; Naaman, 2021; D'Amour et al., 2021). This gives us the desired conclusion.

Case 2: general observational setting

Let $t > 0$. Let $S \in \{-1, 1\}$. Consider that:

$$\begin{aligned}
& \sup_y \left| \hat{F}_w(y | x) - F_w(y | x) \right| \\
&= \sup_{y,S} S \left\{ \frac{1}{n_{\ell(x)}} \sum_{i: \ell(X_i)=\ell(x)} \left(\frac{1 \{Y_i \leq y\} 1 \{W_i = w\}}{we(X_i) + (1-w)[1-e(X_i)]} - \mathbb{E} \left[\frac{1 \{Y_i \leq y\} 1 \{W_i = w\}}{we(X_i) + (1-w)[1-e(X_i)]} \mid \ell(X_i) = \ell(x) \right] \right) \right\} \\
&\stackrel{(i)}{\leq} t + \mathbb{E} \left[\sup_{y,S} S \left\{ \frac{1}{n_{\ell(x)}} \sum_{i: \ell(X_i)=\ell(x)} \left(\frac{1 \{Y_i \leq y\} 1 \{W_i = w\}}{we(X_i) + (1-w)[1-e(X_i)]} - \mathbb{E} \left[\frac{1 \{Y_i \leq y\} 1 \{W_i = w\}}{we(X_i) + (1-w)[1-e(X_i)]} \mid \ell(X_i) = \ell(x) \right] \right) \right\} \right] \\
&\stackrel{(ii)}{\leq} t + \frac{4 \log^{\frac{1}{2}}(n_{\ell(x)} + 1)}{n_{\ell(x)}^{\frac{1}{2}} (w\underline{e} + (1-w)(1-\bar{e}))}. \tag{19}
\end{aligned}$$

Here, inequality (i) holds with probability at least

$$1 - \exp \left(- \frac{n_{\ell(x)} t^2 [w\underline{e} + (1-w)(1-\bar{e})]^2}{4} \right).$$

It is by Functional Hoeffding's inequality (Wainwright, 2019); its premise holds: with probability 1,

$$0 \leq \sup_{y,S} S \left\{ \frac{1 \{Y_i \leq y\} 1 \{W_i = w\}}{we(X_i) + (1-w)[1-e(X_i)]} - \mathbb{E} \left[\frac{1 \{Y_i \leq y\} 1 \{W_i = w\}}{we(X_i) + (1-w)[1-e(X_i)]} \mid \ell(X_i) = \ell(x) \right] \right\} \leq \frac{1}{w\underline{e} + (1-w)(1-\bar{e})}.$$

Meanwhile, inequality (ii) holds with probability 1 due to Lemma E.5.

With some algebra, we see that:

$$\text{pr} \left\{ \sup_y \left| \hat{F}_w(y) - F_w(y) \right| \leq \frac{2 \left[2 \log^{\frac{1}{2}}(n_{\ell(x)} + 1) + \log^{\frac{1}{2}}(1/\beta) \right]}{n_{\ell(x)}^{1/2} (w\underline{e} + (1-w)(1-\bar{e}))} \right\} \geq 1 - \beta.$$

□

E.2.1 Inferring an Off Policy Outcome Distribution

Suppose the following generation process for our observed tuple of interest:

$$(Y_i, W_i, X_i) \stackrel{\text{i.i.d.}}{\sim} p(y|w, x)p(w|x)p(x); i = 1, \dots, n.$$

We are interested in inferring the outcome had we generated treatment with policy $\pi(w|x)$ rather than $p(w|x)$. Here, π is understood to be a conditional probability density (pdf) or probability mass function (pmf) for W_i —a vector, discrete, and/or continuous treatment variable. We define the generative process for such a random Variable in Definition E.4.

Definition E.4 (Off-Policy Outcome). Consider the generative process to obtain our outcome of interest,

$$Y_i(\pi),$$

under treatment assignment policy π :

1. $X_i \sim p(x)$.
2. Given $X_i = x, W_i \sim \pi(w|x)$.
3. Given $X_i = x$ and $W_i = w, Y_i \sim p(y|w, x)$.

The potential random variable $Y_i(\pi)$ of Definition E.4 is exactly Y_i when $\pi(w|x) = p(w|x)$. Additionally, when W_i is discrete and $\pi(w|x) = 1$ for a fixed value w , while $\pi(w'|x) = 0$ for $w' \neq w$, then

$$Y_i(\pi) \doteq Y_i(w)$$

is exactly what is known as a potential outcome (Imbens and Rubin, 2015) or counterfactual (Pearl, 2009; Hernán and Robins, 2020). This instance is the random variable we consider in the main text, but Definition E.4 is more general.

Denote the conditional CDF of $Y_i(\pi)$ given $\ell(X_i) = \ell(x)$ as

$$F_\pi(y|x) := \text{pr} \{Y_i(\pi) \leq y \mid \ell(X_i) = \ell(x)\}.$$

Here, $\ell(x) = l$ if $x \in \mathcal{X}_l$, where

$$\mathcal{X}_l \subseteq \mathcal{X} := \text{support}(X_i) \ (l = 1, \dots, L; L \geq 1).$$

Our estimator is:

$$\hat{F}_\pi(y|x) := \frac{1}{n_{\ell(x)}} \sum_{i: \ell(X_i)=\ell(x)} \frac{\text{1}\{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)}. \quad (20)$$

$\hat{F}_\pi(y|x)$ in (20) is unbiased for $F_\pi(y|x)$:

$$\begin{aligned} & \mathbb{E} \left[\hat{F}_\pi(t|x) \mid \ell(X_i) = l; i : \ell(X_i) = \ell(x) \right] \\ &= \frac{1}{n_l} \sum_{i: \ell(X_i)=l} \int_{(y,w,x)} \frac{\text{1}\{Y_i \leq t\} \pi(W_i|X_i)}{p(W_i|X_i)} p(y, w, x | \ell(x) = l) \\ &= \frac{1}{n_l} \sum_{i: \ell(X_i)=l} \int_x \int_w \int_y \text{1}\{y \leq t\} p(y|w, x) \frac{\pi(w|x)}{p(w|x)} p(w|x) p(x | \ell(x) = l) \\ &= \frac{1}{n_l} \sum_{i: \ell(X_i)=l} \int_x \int_w \int_y \text{1}\{y \leq t\} p(y|w, x) \pi(w|x) p(x | \ell(x) = l) \\ &= \frac{1}{n_l} \sum_{i: \ell(X_i)=l} \int_x \int_w \text{pr} \{Y_i \leq t \mid X_i = x, W_i = w\} \pi(w|x) p(x | \ell(x) = l) \\ &= \frac{1}{n_l} \sum_{i: \ell(X_i)=l} \int_x \text{pr} \{Y_i(\pi) \leq t \mid X_i = x\} p(x | \ell(x) = l) \\ &= \frac{1}{n_l} \sum_{i: \ell(X_i)=l} \text{pr} \{Y_i(\pi) \leq t \mid \ell(X_i) = l\} \\ &= F_\pi(t|x). \end{aligned} \quad (21)$$

In Lemma E.5, we now study the finite sample performance of the Off-policy estimator in (20) across possible evaluation points y .

Lemma E.5 (Bounding the expected uniform deviation of the eCDF estimator).

We have that

$$\begin{aligned} & \mathbb{E} \left[\sup_y \left| \hat{F}_\pi(y|x) - F_\pi(y|x) \right| \mid \ell(X_i) = \ell(x) \right] \\ & \leq 4 \left\| \frac{\pi(W_i|X_i)}{p(W_i|X_i)} \right\|_{x,\infty} \log^{\frac{1}{2}}(n_{\ell(x)} + 1) n_{\ell(x)}^{-\frac{1}{2}}. \end{aligned}$$

Here, $\|V_i\|_{x,\infty} := \inf \{t : \text{pr} \{V_i \leq t \mid \ell(X_i) = \ell(x)\} = 1\}$.

Proof.

Let $(Y_i, X_i, W_i) \stackrel{\text{i.i.d.}}{\sim} (Y_i^*, X_i^*, W_i^*)$ for $i = 1, \dots, n$. Independent of these, let $\sigma_i \stackrel{\text{i.i.d.}}{\sim} \text{Rademacher}$. Below, also let $S \in \{-1, 1\}$. Omitting the dependence on $\ell(X_i) = \ell(x)$, as it is a straightforward extension, we have:

$$\begin{aligned}
& \mathbb{E} \left[\sup_y \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} - \mathbb{E} \left[\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} \right] \right) \right| \right] \\
&= \mathbb{E} \left[\sup_{y,S} S \left\{ \frac{1}{n} \sum_{i=1}^n \left(\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} - \mathbb{E} \left[\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} \right] \right) \right\} \right] \\
&= \mathbb{E} \left[\sup_{y,S} S \left\{ \frac{1}{n} \sum_{i=1}^n \left(\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} - \mathbb{E} \left[\frac{1 \{Y_i^* \leq y\} \pi(W_i^*|X_i^*)}{p(W_i^*|X_i^*)} \right] \right) \right\} \right] \\
&\leq \mathbb{E} \left[\sup_y \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} - \mathbb{E} \left[\frac{1 \{Y_i^* \leq y\} \pi(W_i^*|X_i^*)}{p(W_i^*|X_i^*)} \right] \right) \right| \right] \\
&\leq \mathbb{E} \left[\sup_y \left| \frac{1}{n} \sum_{i=1}^n \left(\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} - \frac{1 \{Y_i^* \leq y\} \pi(W_i^*|X_i^*)}{p(W_i^*|X_i^*)} \right) \right| \right] \\
&= \mathbb{E} \left[\sup_y \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \left(\frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} - \frac{1 \{Y_i^* \leq y\} \pi(W_i^*|X_i^*)}{p(W_i^*|X_i^*)} \right) \right| \right] \\
&\leq 2 \mathbb{E} \left[\sup_y \left| \frac{1}{n} \sum_{i=1}^n \sigma_i \frac{1 \{Y_i \leq y\} \pi(W_i|X_i)}{p(W_i|X_i)} \right| \right] \\
&\stackrel{(i)}{\leq} 4 \left\| \frac{\pi(W_i|X_i)}{p(W_i|X_i)} \right\|_\infty \log^{\frac{1}{2}}(n+1)n^{-\frac{1}{2}}.
\end{aligned}$$

Up to right before inequality (i), we follow the standard symmetrization argument for bounding the expectation of an empirical process in terms of what is called the Rademacher complexity of a function class that indexes the empirical process. See for example, Wainwright (2019) and van der Vaart and Wellner (1996).

Inequality (i) follows due to *Lemma 4.14* in Wainwright (2019). This lemma covers what is called the Rademacher complexity, an expectation of the across independent random variables (T_1, \dots, T_n) and $(\sigma_1, \dots, \sigma_n)$:

$$\mathbb{E} \left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \sigma_i g(T_i) \right| \right].$$

The Lemma particularly covers the class of functions \mathcal{G} satisfying (22). Denote the evaluation of these functions on a finite set of points as:

$$\mathcal{G}(t_1, \dots, t_n) = \{g(t) : t \in \{t_1, \dots, t_n\}, g \in \mathcal{G}\}.$$

Lemma 4.14 in Wainwright (2019) requires the existence of some $\nu > 0$ such that the cardinality of $\mathcal{G}(t_1, \dots, t_n)$ is bounded by a polynomial as:

$$|\mathcal{G}(t_1, \dots, t_n)| \leq (n+1)^\nu, \quad (22)$$

a condition known as "Polynomial Discrimination."

In our case, the random variable $T_i = (Y_i, W_i, X_i)$ and our infinitely sized function class, indexed by y and evaluated at $t = (a, b, c)$, is:

$$\mathcal{G} = \left\{ \frac{1 \{a \leq y\} \pi(b|c)}{p(b|c)} : y \in \mathbb{R} \right\}.$$

Consider that for any $t = (a, b, c)$ and any y , it must be the case that

$$\frac{1 \{a \leq y\} \pi(b|c)}{p(b|c)} \in \left\{ 0, \frac{\pi(b|c)}{p(b|c)} \right\}.$$

Thus, (22) is satisfied with $\nu = 1$.

□

E.3 The proof of Proposition 2.2

Proof of Proposition 2.2. Let $x \in \mathcal{X}$ be fixed. We apply Lemma E.1 with $U_0 := Y_i(0)$, $U_1 := Y_i(1)$, and $V = \ell(X_i)$, which gives:

$$\begin{aligned}
& \sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \\
& \leq \sup_{\delta, y} \left| \left\{ \hat{F}_{1n}(y + \delta/2 \mid x) - \hat{F}_{0n}(y - \delta/2 \mid x) \right\} - \{F_1(y + \delta/2 \mid x) - F_0(y - \delta/2 \mid x)\} \right| \\
& \leq \sup_{\delta, y} \left| \hat{F}_{1n}(y + \delta/2 \mid x) - F_1(y + \delta/2 \mid x) \right| + \sup_{\delta, y} \left| \hat{F}_{0n}(y - \delta/2 \mid x) - F_0(y - \delta/2 \mid x) \right| \\
& = \sup_y \left| \hat{F}_{1n}(y \mid x) - F_1(y \mid x) \right| + \sup_y \left| \hat{F}_{0n}(y \mid x) - F_0(y \mid x) \right| \\
& \stackrel{(i)}{\leq} n_{\ell(x)}^{-1/2} \sum_{w=0,1} C_{w,\beta/2}(x).
\end{aligned} \tag{23}$$

Here, inequality (i) holds with probability at least $1 - \beta$ due to an application of Lemma E.3 along with Lemma E.2.

So far, x is fixed. We now consider the two cases as stated in the proposition.

Case 1: Consider evaluating the bound estimators at a random new test point, X_{n+1} , independent of our training data. We have:

$$\begin{aligned}
& \text{pr} \left\{ \sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta, X_{n+1}) - \theta^L(\delta, X_{n+1}) \right| \vee \left| \hat{\theta}^U(\delta, X_{n+1}) - \theta^U(\delta, X_{n+1}) \right| \right\} \leq n_{\ell(X_{n+1})}^{-1/2} \sum_{w=0,1} C_{w,\alpha/2}(X_{n+1}) \right\} \\
& \stackrel{(i)}{=} \sum_{l=1}^L \text{pr} \{ \ell(X_{n+1}) = l \} \\
& \quad \times \text{pr} \left\{ \sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta, x_l) - \theta^L(\delta, x_l) \right| \vee \left| \hat{\theta}^U(\delta, x_l) - \theta^U(\delta, x_l) \right| \right\} \leq n_{\ell(x_l)}^{-1/2} \sum_{w=0,1} C_{w,\alpha/2}(x_l) \right\} \\
& \stackrel{(ii)}{\geq} 1 - \alpha.
\end{aligned}$$

Equality (i) holds by law of total probability. Here, we insert arbitrary fixed points $x_l \in \mathcal{X}_l$ ($l = 1, \dots, L$). The conditioning statement in the second factor inside the sum is omitted since X_{n+1} is independent of the bound estimators after we insert the point x_l . Next, inequality (ii) holds by applying (23) at each point x_l with $\beta = \alpha/2$.

Case 2: Now consider a uniform guarantee on the bound estimators across $x \in \mathcal{X}$, in addition to across $\delta \in \mathbb{R}$.

We have:

$$\begin{aligned}
& \text{pr} \left\{ \sup_{\delta,x} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \leq t \right\} \\
& \geq 1 - \alpha
\end{aligned}$$

by union bound provided we set

$$t \doteq \max_l \left\{ n_l^{-1/2} \sum_{w=0,1} C_{w,\alpha/(2L)}(x_l) \right\}.$$

□

E.4 The proof of Proposition C.2

Proof of Proposition C.2. This result is a direct consequence of Lemma E.1 with $U_{i0} := Y_i(0)$, $U_{i1} := Y_i(1)$, and $V_i := X_i$. □

E.5 The proof of Corollary C.3

Proof of Corollary C.3. Specify

$$\hat{G}(y, \delta, x) = \hat{F}_{1n}(y + \delta/2 | x) - \hat{F}_{0n}(y - \delta/2 | x),$$

which simply plugs in the estimate of $F_1(y + \delta/2 | x)$ and $F_0(y - \delta/2 | x)$. Consider that:

$$\sup_y \left| \hat{F}_{wn}(y | x) - F_w(y | x) \right| = \sup_{a,y} \left| \hat{F}_{wn}(y + a | x) - F_w(y + a | x) \right|,$$

where the supremum is with respect to y is across the real line, while the supremum across (y, a) is across the euclidean plane. Therefore, when the inequality

$$\sup_y \left| \hat{F}_{wn}(y | x) - F_w(y | x) \right| \leq t_{w,\alpha}(x) \quad (24)$$

holds, triangle inequality and the definition of $G(y, \delta, x)$ tell us that

$$\begin{aligned} \sup_{\delta,y} \left| \hat{G}_n(y, \delta, x) - G(y, \delta, x) \right| &\leq \sup_{\delta,y} \left| \hat{F}_{1n}(y + \delta/2 | x) - F_w(y + \delta/2 | x) \right| \\ &\quad + \sup_{\delta,y} \left| \hat{F}_{0n}(y - \delta/2 | x) - F_0(y - \delta/2 | x) \right| \\ &= \sup_y \left| \hat{F}_{1n}(y | x) - F_w(y | x) \right| + \sup_y \left| \hat{F}_{0n}(y | x) - F_0(y | x) \right| \\ &\leq \sum_{w=0,1} t_{w,\alpha}(x). \end{aligned} \quad (25)$$

Combining inequality (25) and inequality (13) in Proposition C.2 (after taking the supremum on both sides with respect δ), we get:

$$\sup_{\delta} \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \leq \sum_{w=0,1} t_{w,\alpha}(x), \quad (26)$$

also. Now, if the inequality in (24) holds with high probability, which is the premise of this corollary, then so must (26). \square

E.6 The proof of Proposition C.4

Proof of Proposition C.4. Given inequality (13) in Proposition C.2, our specification for

$$\hat{G}(y, \delta, x) := \hat{F}_{1n}(y | x) - \hat{F}_{0n}(y | x),$$

and a similar argument to the proof of Corollary C.3 in Appendix E.5, we have that:

$$\sup_x \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} \leq \sum_{w=0,1} \sup_{y,x} \left| \hat{F}_{wn}(y | x) - F_w(y | x) \right|. \quad (27)$$

Consider $s_0, s_1 \geq 0$. Due to Lemma E.2 and (27), we have that:

$$\begin{aligned} &\text{pr} \left[\sup_x \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} > s_0 + s_1 \mid X \right] \\ &\leq \text{pr} \left\{ \sum_{w=0,1} \sup_{y,x} \left| \hat{F}_{wn}(y | x) - F_w(y | x) \right| > s_0 + s_1 \mid X \right\} \\ &\leq \sum_{w=0,1} \text{pr} \left\{ \sup_{y,x} \left| \hat{F}_{wn}(y | x) - F_w(y | x) \right| > s_w \mid X \right\}. \end{aligned} \quad (28)$$

Let $t_w \geq 0$. Lemma E.6 tells us that for

$$\begin{aligned} s_w &\doteq \sup_r [\text{pr} \{r < R_i(w) \leq r + 2t_w \mid X\} \vee \text{pr} \{r - 2t_w < R_i(w) \leq r \mid X\}] \\ &\quad + \left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} n_w^{-\frac{1}{2}}, \end{aligned}$$

we get:

$$\text{pr} \left\{ \sup_{y,x} \left| \hat{F}_{wn}(y|x) - F_w(y|x) \right| > s_w \mid X \right\} \leq \alpha/2 + \text{pr} \left\{ \sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_w \mid X \right\}. \quad (29)$$

Now, combining (28) and (29), we have conditional on X :

$$\begin{aligned} & \text{pr} \left[\sup_x \left\{ \left| \hat{\theta}^L(\delta, x) - \theta^L(\delta, x) \right| \vee \left| \hat{\theta}^U(\delta, x) - \theta^U(\delta, x) \right| \right\} > s_0 + s_1 \mid X \right] \\ & \leq \alpha + \sum_{w=0,1} \text{pr} \left\{ \sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_w \mid X \right\}, \end{aligned}$$

which is the desired conclusion. \square

E.7 A lemma for Proposition C.4: conditional CDF estimation with regression residuals

Lemma E.6 (Conditional CDF estimator with regression residuals). *Let $(U_1, V_1), \dots, (U_n, V_n)$ be independent and identically distributed random variables with $U_1, \dots, U_n \in \mathbb{R}$. Partition the training indices $\{1, \dots, n\}$ into two non-intersecting splits, \mathcal{I}_1 and \mathcal{I}_2 , respectively. Let $\mathcal{T}_j := \{(U_i, V_i)\}_{i \in \mathcal{I}_j}$ ($j = 1, 2$). Denote $\mu(v) := \mathbb{E}[U_i \mid V_i = v]$, and let $\hat{\mu}(v)$ denote its estimator based on \mathcal{T}_1 . Denote $R_i := U_i - \mu(V_i)$ along with its approximation given by $\hat{R}_i := U_i - \hat{\mu}(V_i)$ across $i = 1, \dots, n$. For $u \in \text{support}(U_i)$ and $v \in \text{support}(V_i)$, denote $F(u \mid v) := \text{pr}\{U \leq u \mid V = v\}$. Consider its estimator given by:*

$$\hat{F}_n(u \mid v) := \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{1} \left\{ \hat{\mu}(v) + \hat{R}_i \leq u \right\},$$

where $|\mathcal{I}_2|$ is the number of indices in \mathcal{I}_2 .

If $R_i \perp\!\!\!\perp V_i$, then conditional on $V = (V_i; i \in \mathcal{I}_1)$, we have that for any $t \geq 0$, where t is possibly dependent on V , that:

$$\begin{aligned} & \sup_{u,v} \left| \hat{F}(u \mid v) - F(u \mid v) \right| \\ & \leq \sup_r \left\{ \text{pr} \{r < R_i \leq r + 2t \mid V\} \vee \text{pr} \{r - 2t < R_i \leq r \mid V\} \right\} + \left\{ \frac{\log(4/\alpha)}{2} \right\}^{1/2} |\mathcal{I}_2|^{-\frac{1}{2}} \end{aligned}$$

with probability at least

$$1 - \alpha/2 - \text{pr} \left\{ \sup_v |\hat{\mu}(v) - \mu(v)| \geq t \mid V \right\}.$$

Proof. Below, let the probability statements be with respect to $i \notin \mathcal{I}_1$, where \mathcal{I}_1 is the set of indices used to train $\hat{\mu}$. This is important to note as we will use that $R_i \perp\!\!\!\perp (\mathcal{T}_1, V)$ later. Consider that:

$$\begin{aligned} & \sup_{u,v} \left| \hat{F}_n(u \mid v) - F(u \mid v) \right| \\ & = \sup_{u,v} \left| \hat{F}_n(u \mid v) - \text{pr} \{ \mu(v) + R_i \leq u \} \right| \\ & \leq \sup_{u,v} \left| \hat{F}_n(u \mid v) - \text{pr} \left\{ \hat{\mu}(v) + \hat{R}_i \leq u \mid \mathcal{T}_1 \right\} \right| \\ & \quad + \sup_{u,v} \left| \text{pr} \left\{ \hat{\mu}(v) + \hat{R}_i \leq u \mid \mathcal{T}_1 \right\} - \text{pr} \{ \mu(v) + R_i \leq u \} \right| \\ & \stackrel{(i)}{=} \sup_{u,v} \left| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} \mathbf{1} \{ U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v) \} - \text{pr} \{ U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v) \mid \mathcal{T}_1 \} \right| \\ & \quad + \sup_{u,v} \left| \text{pr} \left[R_i \leq u - \mu(v) + \left\{ \hat{b}(V_i) - \hat{b}(v) \right\} \mid \mathcal{T}_1 \right] - \text{pr} \{ R_i \leq u - \mu(v) \} \right|. \end{aligned} \quad (30)$$

For any v , define:

$$\hat{b}(v) := \hat{\mu}(v) - \mu(v), \quad (31)$$

a term that characterizes the bias in $\hat{\mu}(v)$ for any $v \in \text{support}(V_i)$. In equality (i), we used the identities $\hat{\mu}(v) = \mu(v) + \hat{b}(v)$ and $\hat{R}_i = R_i - \hat{b}(V_i)$, which are a consequence of the definition of $\hat{b}(\cdot)$.

Now, denote:

$$A := \sup_{u,v} \left| \frac{1}{|\mathcal{I}_w|} \sum_{i \in \mathcal{I}_2} 1 \{U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v)\} - \text{pr} \{U_i - \hat{\mu}(V_i) \leq u - \hat{\mu}(v) \mid \mathcal{T}_1\} \right|$$

and

$$B := \sup_{u,v} \left| \text{pr} \left[R_i \leq u - \mu(v) + \{\hat{b}(V_i) - \hat{b}(v)\} \mid \mathcal{T}_1 \right] - \text{pr} \{R_i \leq u - \mu(v)\} \right|.$$

Due to (30) and Lemma E.2:

$$\begin{aligned} & \text{pr} \left\{ \sup_{u,v} \left| \hat{F}_n(u \mid v) - F(u \mid v) \right| > a + b \mid V \right\} \\ & \leq \text{pr} \{A + B > a + b \mid V\} \\ & \leq \text{pr} \{A > a \mid V\} + \text{pr} \{B > b \mid V\}. \end{aligned} \tag{32}$$

We control the two terms in the latter inequality separately in (33) and (36) below. We also explain the choices

$$a \doteq \{2^{-1}(|\mathcal{I}_2|)^{-1} \log(4/\alpha)\}^{1/2}$$

and

$$b \doteq \sup_r \{\text{pr} \{r < R_i \leq r + 2t \mid V\} \vee \text{pr} \{r - 2t < R_i \leq r \mid V\}\}.$$

From these choices, we get the desired conclusion:

$$\sup_{u,v} \left| \hat{F}_n(u \mid v) - F(u \mid v) \right| \leq a + b.$$

with probability at least $1 - \alpha/2 - \text{pr} \{\sup_v |\hat{\mu}(v) - \mu(v)| \mid V\}$.

For the term $\text{pr} \{A > a \mid V\}$ in (32), notice that:

$$A = \sup_r \left| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} 1 \{U_i - \hat{\mu}(V_i) \leq r\} - \text{pr} \{U_i - \hat{\mu}(V_i) \leq r \mid \mathcal{T}_1\} \right|,$$

where the supremum with respect to $r = u - \hat{\mu}(v)$ is taken in the support of $u - \hat{\mu}(v)$.

The quantity $\hat{\mu}(v)$ is random, even if v is not, because it is an estimator based on \mathcal{T}_1 . However, conditional on (\mathcal{T}_1, V) , $r = u - \hat{\mu}(v)$ is a constant. Importantly, conditional on (\mathcal{T}_1, V) , we have that $\hat{R}_i = U_i - \hat{\mu}(V_i)$ is an independent and identically distributed random variable across $i \notin \mathcal{I}_1$. This means that the estimator

$$\frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} 1 \{U_i - \hat{\mu}(V_i) \leq r\}$$

for

$$\text{pr} \{U_i - \hat{\mu}(V_i) \leq r \mid \mathcal{T}_1, V\}$$

satisfies the independent and identically distributed sample condition for the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990; Naaman, 2021). The Dvoretzky–Kiefer–Wolfowitz inequality gives:

$$\begin{aligned} & \text{pr} \{A > a \mid \mathcal{T}_1, V\} \\ &= \text{pr} \left[\sup_r \left| \frac{1}{|\mathcal{I}_2|} \sum_{i \in \mathcal{I}_2} 1 \{U_i - \hat{\mu}(V_i) \leq r\} - \text{pr} \{U_i - \hat{\mu}(V_i) \leq r \mid \mathcal{T}_1\} \right| > a \mid \mathcal{T}_1, V \right] \\ &\leq 2e^{-2|\mathcal{I}_2|a^2}. \end{aligned}$$

Noting that the upper bound does not depend on (\mathcal{T}_1, V) , we get due to law of total expectation that:

$$\text{pr} \{A > a \mid V\} \leq 2e^{-2|\mathcal{I}_2|a^2} \tag{33}$$

and

$$\text{pr} \{A > a\} \leq 2e^{-2|\mathcal{I}_2|a^2}.$$

We will take

$$a \doteq \{2^{-1}(|\mathcal{I}_2|)^{-1} \log(4/\alpha)\}^{1/2},$$

so that we are guaranteed $\text{pr}\{A > a\} \leq \alpha/2$ and $\text{pr}\{A > a | V\} \leq \alpha/2$.

That covers the first important term $\text{pr}\{A > a | V\}$ in Equation 32. Now consider the second term, $\text{pr}\{B > b | V\}$. Here,

$$B := \sup_{u,v} \left| \text{pr} \left[R_i \leq u - \mu(v) + \{\hat{b}(V_i) - \hat{b}(v)\} | \mathcal{T}_1 \right] - \text{pr}\{R_i \leq u - \mu(v)\} \right|$$

is a random variable with respect to \mathcal{T}_1 . For t , possibly a function of V , consider the following event with respect to the random data in \mathcal{T}_1 :

$$E(t) := \left\{ \sup_v |\hat{\mu}(v) - \mu(v)| \leq t \right\}.$$

The event $E(t)$ concerns the deviation between $\hat{\mu}$ and μ uniformly across the support of V_i . Conditional on the event $E(t)$ and V , we have that the regression bias term defined in (31) satisfies $-t \leq \hat{b}(V_i) \leq t$, in spite of V_i being random, along with $-t \leq \hat{b}(v) \leq t$ for any v . This means that conditionally on $\{E(t), V\}$:

$$-2t \leq \hat{b}(V_i) - \hat{b}(v) \leq 2t \quad (34)$$

almost surely.

Now let $E(t)^C$ denote the complement of $E(t)$. Consider that:

$$\begin{aligned} \text{pr}\{B > b | V\} &= \text{pr}\{B > b | E(t), V\} \text{pr}\{E(t) | V\} + \text{pr}\{B > b | E(t)^C, V\} \text{pr}\{E(t)^C | V\} \\ &\leq \text{pr}\{B > b | E(t), V\} + \text{pr}\left\{ \sup_v |\hat{\mu}(v) - \mu(v)| > t | V \right\}, \end{aligned} \quad (35)$$

by the law of total probability, the fact that $0 \leq \text{pr}\{\cdot | V\}, \text{pr}\{\cdot | E(t), V\} \leq 1$, along with the definition of $E(t)^C$.

For a random event $D \perp\!\!\!\perp R_i$ (recall, $i \notin \mathcal{I}_1$), consider the four-argument function:

$$\nu(r, s, D, V) := |\text{pr}\{R_i \leq r + s | D, V\} - \text{pr}\{R_i \leq r\}|.$$

For $D = \emptyset$, we write:

$$\nu(r, s, \emptyset, V) := |\text{pr}\{R_i \leq r + s | V\} - \text{pr}\{R_i \leq r\}|.$$

Fixing (r, D, V) , we can see that $\nu(r, s, D, V)$ is an increasing function with respect to s in the domain $[0, 2t]$, and it is a decreasing function with respect to s in the domain $[-2t, 0]$. From this, it follows that for all $-2t \leq s \leq 2t$:

$$\nu(r, s, D, V) \leq \nu(r, -2t, D, V) \vee \nu(r, 2t, D, V).$$

Based on (34) and this property of $\nu(r, s, D, V)$, it follows that conditional on $E(t)$ and V we have:

$$0 \leq B \leq \sup_{u,v} \{\nu(u - \mu(v), -2t, E(t), V) \vee \nu(u - \mu(v), 2t, E(t), V)\}$$

almost surely. We have further that

$$\begin{aligned} &\sup_r \{\nu(r, -2t, \emptyset, V) \vee \nu(r, 2t, \emptyset, V)\} \\ &= \sup_{u,v} \{\nu(u - \mu(v), -2t, E(t), V) \vee \nu(u - \mu(v), 2t, E(t), V)\}. \end{aligned}$$

This is based on the properties of the supremum, the fact that t is fixed conditional on V , and because $R_i \perp\!\!\!\perp E(t)$. So we can re-write that conditional on $E(t)$ and V :

$$0 \leq B \leq \sup_r \{\nu(r, -2t, \emptyset, V) \vee \nu(r, 2t, \emptyset, V)\}$$

almost surely. We will strategically set:

$$b \doteq \sup_r \{\nu(r, -2t, \emptyset, V) \vee \nu(r, 2t, \emptyset, V)\}.$$

Moreover, $R_i \perp\!\!\!\perp V$ means that $\text{pr}\{R_i \leq r\} = \text{pr}\{R_i \leq r | V\}$, so we have:

$$b = \sup_r \{\text{pr}\{r < R_i \leq r + 2t | V\} \vee \text{pr}\{r - 2t < R_i \leq r | V\}\}.$$

With this choice of b , we have that $\text{pr}\{B > b | E(t), V\} = 0$. Using this in (35), we get that:

$$\text{pr}\{B > b | V\} \leq \text{pr}\left\{ \sup_v |\hat{\mu}(v) - \mu(v)| > t | V \right\}, \quad (36)$$

as we wanted. \square

E.8 The proof of Corollary C.5

Proof. Denote $f_w(r)$ as the marginal density of $R_i(w)$ ($w = 0, 1$). The key here is that

$$\begin{aligned} & \sup_r [\Pr\{r < R_i(w) \leq r + 2t_w \mid X\} \vee \Pr\{r - 2t_w < R_i(w) \leq r \mid X\}] \\ & \leq 2t_w \max_{w=0,1} \sup_r f_w(r), \end{aligned}$$

using that $\int_a^b h(u)du \leq \sup_u |h(u)||b-a|$ for any integrable function h . We also used that $R_i(w) \perp\!\!\!\perp X$. Under the regularity condition that the density $f_w(r)$ is non-negative and bounded away from infinity, we set $t_w \doteq g_{n,\mathcal{F}}(w = 0, 1)$. This inequality and Proposition C.4 allow us to arrive at the desired result. \square

E.8.1 The proof of Proposition C.7

Proof of Proposition C.7. Let $\Lambda_w \in R^{|\mathcal{I}_1 \cap S_w| \times d}$ be such that its rows are formed by stacking $\Psi_w(X_i)^T$ for each $i \in \mathcal{I}_1 \cap S_w$. Further, let $Y_w \in R^{|\mathcal{I}_1 \cap S_w| \times 1}$ contain the corresponding observed outcome Y_i for each $i \in \mathcal{I}_1 \cap S_w$. The ordinary least squares estimator for the coefficient vector β_w is

$$\hat{\beta}_w = (\Lambda_w^T \Lambda_w)^{-1} \Lambda_w^T Y_w.$$

We have that

$$\begin{aligned} \sup_x |\hat{\mu}_w(x) - \mu_w(x)| &= \sup_x \left| \Psi_w^T(x)(\hat{\beta}_w - \beta_w) \right| \\ &\leq \left\| \hat{\beta}_w - \beta_w \right\|_2 \\ &\leq \left\| \sigma_w (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \left\| \frac{1}{\sigma_w} (\Lambda_w^T \Lambda_w)^{1/2} (\hat{\beta}_w - \beta_w) \right\|_2. \end{aligned}$$

The first inequality is based on Hölder's inequality, while the second inequality is again due to Hölder's inequality and the definition of the operator norm. Conditional on Λ_w , the d -dimensional sampling distribution for $\hat{\beta}$ is:

$$\hat{\beta}_w | \Lambda_w \sim \mathcal{N}(\beta_w, \sigma_w^2 (\Lambda_w^T \Lambda_w)^{-1}).$$

This is based on a standard argument in low dimensional linear regression given the residual distribution assumption. This further implies that conditionally on Λ_w ,

$$\left\| \sigma_w (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op}^2 \left\| \frac{1}{\sigma_w} (\Lambda_w^T \Lambda_w)^{1/2} (\hat{\beta}_w - \beta_w) \right\|_2^2 | \Lambda_w \sim \sigma_w^2 \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op}^2 \times \chi_d^2,$$

a distribution generated by taking a chi-squared distributed random variable (d degrees of freedom) and multiplying it by a factor of $\sigma_w^2 \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op}^2$. This is because

$$\frac{1}{\sigma_w} (\Lambda_w^T \Lambda_w)^{1/2} (\hat{\beta}_w - \beta_w) | \Lambda_w \sim \mathcal{N}_d(0, \mathbb{I}) \implies \left\| \frac{1}{\sigma_w} (\Lambda_w^T \Lambda_w)^{1/2} (\hat{\beta}_w - \beta_w) \right\|_2^2 | \Lambda_w \sim \chi_d^2.$$

Let V have a χ_d^2 distribution conditional on Λ_w , and denote $v_{d,\alpha}$ as the $(1 - \alpha/2)$ th quantile of V 's distribution. It follows that setting

$$t_{w,\alpha} \doteq (v_{d,\alpha})^{1/2} \sigma_w \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op}$$

implies that

$$\begin{aligned} 2 \max_{w=0,1} \Pr \left\{ \sup_x |\hat{\mu}_w(x) - \mu_w(x)| > t_{w,\alpha} \mid X \right\} &\leq 2 \max_{w=0,1} \Pr \left\{ V > \frac{t_{w,\alpha}^2}{\sigma_w^2 \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op}^2} \mid \Lambda_w \right\} \\ &= \alpha. \end{aligned} \tag{37}$$

Now, consider that the distributional assumption on $R_i(w)$ implies the following. We have:

$$\begin{aligned}
& 2 \max_{w=0,1} \sup_r \Pr \{r < R_i(w) \leq r + 2t_{w,\alpha}\} \vee \Pr \{r - 2t_{w,\alpha} < R_i(w) \leq r\} \\
& \stackrel{(i)}{=} 2 \max_{w=0,1} \sup_r \Pr \{r < R_i(w) \leq r + 2t_{w,\alpha} \mid \Lambda_w\} \vee \Pr \{r - 2t_{w,\alpha} < R_i(w) \leq r \mid \Lambda_w\} \\
& \stackrel{(ii)}{=} 2 \max_{w=0,1} \sup_r \Pr \{r < R_i(w) \leq r + 2t_{w,\alpha} \mid \Lambda_w\} \\
& \stackrel{(iii)}{=} 2 \max_{w=0,1} \Pr \{-t_{w,\alpha} < R_i(w) \leq t_{w,\alpha} \mid \Lambda_w\} \\
& \stackrel{(iv)}{=} 2 \max_{w=0,1} \Pr \left\{ -(v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} < R_i(w)/\sigma_w \leq (v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \mid \Lambda_w \right\} \\
& \stackrel{(v)}{=} 2 \max_{w=0,1} \left[\Phi \left\{ (v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \right\} - \Phi \left\{ -(v_{d,\alpha})^{1/2} \left\| (\Lambda_w^T \Lambda_w)^{-1/2} \right\|_{op} \right\} \right]. \tag{38}
\end{aligned}$$

Here, Φ denotes the CDF for the standard normal distribution. Equality (i) holds due to the assumption that $R_i(w) \perp\!\!\!\perp \Lambda_w$. Equality (ii) holds due to the symmetry of the Gaussian density around its mean, which is zero in the case of $R_i(w)$ conditional on Λ_w . Moreover, (iii) holds because the biggest slice of area under the normal density of width $2t_{w,\alpha}$ is the one centered at its mean. Next, (iv) holds due to our choice of $t_{w,\alpha}$, while (v) holds since $R_i(w)/\sigma_w$ follows a standard normal distribution.

The final conclusion in Proposition C.7 follows by applying Proposition C.4 with (37) and (38). \square

E.9 The proof of Proposition D.1

Proof of Proposition D.1. The first claim is immediate from Lemma A.1 and our definition of $\theta^L(\delta)$ and $\theta^U(\delta)$ in Section 2. We also use the fact that PNS is the same as PIBT when $0 \leq \delta < 1$.

We now show why the second part of our claim holds. We will use that CDFs for binary potential outcomes satisfy:

$$F_w(y) = \begin{cases} 0 & \text{if } y < 0 \\ \Pr \{Y_i(w) = 0\} & \text{if } 0 \leq y < 1 ; w = 0, 1 \\ 1 & \text{if } y \geq 1 \end{cases}$$

For $0 \leq \delta < 1$, the Makarov lower bound on PIBT is:

$$\begin{aligned}
\theta^L(\delta) &= - \min \left[\inf_{y=0,1} \{F_1(y + \delta/2) - F_0(y - \delta/2)\}, 0 \right] \\
&= - \min \{F_1(\delta/2) - F_0(-\delta/2), F_1(1 + \delta/2) - F_0(1 - \delta/2), 0\} \\
&= \begin{cases} - \min [\Pr \{Y_i(1) = 0\} - \Pr \{Y_i(0) = 0\}, 0] & \text{if } \delta = 0 \\ - \min [\Pr \{Y_i(1) = 0\}, 1 - \Pr \{Y_i(0) = 0\}, 0] & \text{if } 0 < \delta < 1 \end{cases} \\
&= \begin{cases} - \min [1 - \Pr \{Y_i(1) = 1\} - \Pr \{Y_i(0) = 0\}, 0] & \text{if } \delta = 0 \\ 0 & \text{if } 0 < \delta < 1 \end{cases} \\
&= \begin{cases} \max [\Pr \{Y_i(1) = 1\} + \Pr \{Y_i(0) = 0\} - 1, 0] & \text{if } \delta = 0 \\ 0 & \text{if } 0 < \delta < 1 \end{cases}.
\end{aligned}$$

It follows that

$$\sup_{0 \leq \delta < 1} \theta^L(\delta) = \max [\Pr \{Y_i(1) = 1\} + \Pr \{Y_i(0) = 0\} - 1, 0],$$

as we wanted.

Similarly, for $0 \leq \delta < 1$, the Makarov upper bound on PIBT is:

$$\begin{aligned}
\theta^U(\delta) &= 1 - \max \left[\sup_{y=0,1} \{F_1(y + \delta/2) - F_0(y - \delta/2)\}, 0 \right] \\
&= 1 - \max [F_1(\delta/2) - F_0(-\delta/2), F_1(1 + \delta/2) - F_0(1 - \delta/2), 0] \\
&= \begin{cases} 1 - \max [\text{pr}\{Y_i(1) = 0\} - \text{pr}\{Y_i(0) = 0\}, 0] & \text{if } \delta = 0 \\ 1 - \max [\text{pr}\{Y_i(1) = 0\}, 1 - \text{pr}\{Y_i(0) = 0\}] & \text{if } 0 < \delta < 1 \end{cases} \\
&= \begin{cases} 1 - \max [1 - \text{pr}\{Y_i(1) = 1\} - \text{pr}\{Y_i(0) = 0\}, 0] & \text{if } \delta = 0 \\ 1 - \max [\text{pr}\{Y_i(1) = 0\}, \text{pr}\{Y_i(0) = 1\}] & \text{if } 0 < \delta < 1 \end{cases} \\
&= \begin{cases} \min [\text{pr}\{Y_i(1) = 1\} + \text{pr}\{Y_i(0) = 0\}, 1] & \text{if } \delta = 0 \\ \min [\text{pr}\{Y_i(1) = 1\}, \text{pr}\{Y_i(0) = 0\}] & \text{if } 0 < \delta < 1 \end{cases}
\end{aligned}$$

It follows that

$$\inf_{0 \leq \delta < 1} \theta^U(\delta) = \min [\text{pr}\{Y_i(1) = 1\}, \text{pr}\{Y_i(0) = 0\}],$$

as we wanted.

With essentially the same reasoning, we have the analogous claims for

$$\text{pr}\{Y_i(1) = 1, Y_i(0) = 0 \mid X_i = x\}.$$

□