Estimating Distributional Treatment Effects in Randomized Experiments: Machine Learning for Variance Reduction

Undral Byambadalai [1] **Tatsushi Oka** [2] **Shota Yasui** [1]

Abstract

We propose a novel regression adjustment method designed for estimating distributional treatment effect parameters in randomized experiments. Randomized experiments have been extensively used to estimate treatment effects in various scientific fields. However, to gain deeper insights, it is essential to estimate distributional treatment effects rather than relying solely on average effects. Our approach incorporates pre-treatment covariates into a distributional regression framework, utilizing machine learning techniques to improve the precision of distributional treatment effect estimators. The proposed approach can be readily implemented with off-the-shelf machine learning methods and remains valid as long as the nuisance components are reasonably well estimated. Also, we establish the asymptotic properties of the proposed estimator and present a uniformly valid inference method. Through simulation results and real data analysis, we demonstrate the effectiveness of integrating machine learning techniques in reducing the variance of distributional treatment effect estimators in finite samples.

1. Introduction

Randomized experiments have played a crucial role in understanding the effects of interventions and guiding policy decisions, ever since the seminal work by Fisher (1935). The estimation of causal effects through randomized experiments has found widespread application across various scientific disciplines (Rubin, 1974; Heckman et al., 1997; Imai, 2005; Imbens & Rubin, 2015) and has also become a standard practice within the technology sector (Tang et al., 2010; Bakshy et al., 2014; Xie & Aurisset, 2016; Kohavi et al., 2020).

When analyzing data from randomized experiments, one commonly used measure is the Average Treatment Effect (ATE). However, it is often the case that understanding the distributional treatment effects can provide a richer perspective than solely focusing on overall average effects. Furthermore, while randomized experiments simplify outcome-based analysis, pre-treatment auxiliary information is frequently available. This quest for a more comprehensive understanding of treatment effects, marked by supplementing auxiliary data, calls for new approaches to enhance precision through pre-treatment data incorporation.

In this work, we propose a novel regression-adjusted method to estimate a wide range of distributional parameters in the randomized experiment setup. Our approach draws inspiration from the generic Neyman-orthogonal moment condition (Chernozhukov et al., 2018), which facilitates the decoupling of nuisance parameter and treatment effect estimation into two stages. The nuisance parameters of our interest are the conditional outcome distributions given pre-treatment covariates, and we propose the use of machine learning models (e.g., LASSO, random forests, neural networks, etc.), allowing for complex data and distributional structures. By integrating these sophisticated machine learning techniques with cross-fitting, we reduce the sensitivity of our treatment effect estimator to errors arising from nuisance parameter estimation.

Our paper makes several noteworthy contributions. First, our approach expands the scope regression adjustment. While regression adjustment is commonly employed for variance reduction in the estimation of the ATE with mean regression (Deng et al., 2013; Poyarkov et al., 2016; Guo et al., 2021), our method leverages pre-treatment information under the distributional regression framework, incorporating machine learning methods. This enables us to conduct more powerful statistical inference for distributional parameters, including the Distributional Treatment Effect (DTE) and the Quantile Treatment Effect (QTE). Second, we provide theoretical validation for the regression-adjusted method by demonstrating the variance reduction of the esti-

[1]CyberAgent, Inc., Tokyo, Japan [2]Department of Economics, Keio University, Tokyo, Japan. Correspondence to: Undral Byambadalai <undral_byambadalai@cyberagent.co.jp>, Tatsushi Oka <tatsushi.oka@keio.jp>, Yasui Shota <yasui_shota@cyberagent.co.jp>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

mator of outcome distribution. Third, we establish asymptotic properties for the proposed treatment effect estimators and provide a uniformly valid inference method. Lastly, our simulation and real-data analysis highlight the significance and effectiveness of our method.

The rest of the paper is structured as follows. Section 2 describes related literature. We setup the problem and introduce notations in Section 3. Section 4 then introduces the regression-adjusted estimators for distributional parameters. We derive the asymptotic results in Section 5. Section 6 reports empirical results based on simulated experiments and real datasets. Section 7 concludes. The Appendix in the paper includes all proofs, as well as additional experimental details and results.

2. Related Work

Regression Adjustment There is an extensive literature investigating the use of pre-treatment covariates to reduce variance in estimating the ATE, dating back to Fisher (1932), followed by Cochran (1977); Yang & Tsiatis (2001); Rosenbaum (2002); Freedman (2008a;b); Tsiatis et al. (2008); Rosenblum & Van Der Laan (2010); Lin (2013); Berk et al. (2013); Ding et al. (2019); Negi & Wooldridge (2021), among others, in the case of low-dimensional asymptotics. In high-dimensional settings, this topic has been studied by Bloniarz et al. (2016); Wager et al. (2016); Lei & Ding (2021); Chiang et al. (2023), among others. Recent work by List et al. (2022) has linked regression adjustment to the semiparametric problem of estimating a low-dimensional parameter when a high-dimensional but orthogonal nuisance parameter is present, focusing on estimating the ATE. Our work extends those existing studies to estimate distributional parameters of treatment effects.

Conditional Average Treatment Effects To characterize the heterogeneity in treatment effects, an alternative approach is to condition on observed variables and estimate the Conditional Average Treatment Effect (CATE) (Imai & Ratkovic, 2013; Athey & Imbens, 2016; Johansson et al., 2016; Shalit et al., 2017; Alaa & Van Der Schaar, 2017; Wager & Athey, 2018; Chernozhukov et al., 2018; Künzel et al., 2019; Shi et al., 2019; Nie & Wager, 2021; Guo et al., 2023; Sverdrup & Cui, 2023; van der Laan et al., 2023). The CATE can be regarded as the ATE within subgroups defined by observed characteristics, such as gender, age, prior engagement with the platform, and more. Consequently, the CATE captures observed heterogeneity given the information available to the researchers, whereas our approach is valuable for quantifying unobserved heterogeneity and can be extended to estimate distributional parameters conditional on observed information.

Distributional Treatment Effects Distributional and quantile treatment effects have long been recognized as important parameters to estimate beyond the mean effects. The quantile treatment effect was first introduced by Doksum (1974) and Lehmann & D'Abrera (1975). Subsequently, estimation and inference methods for distributional and quantile treatment effects in various settings have been developed and applied in econometrics, statistics and machine learning community, including Heckman et al. (1997); Imbens & Rubin (1997); Abadie (2002); Abadie et al. (2002); Chernozhukov & Hansen (2005); Koenker (2005); Bitler et al. (2006); Athey & Imbens (2006); Firpo (2007); Chernozhukov et al. (2013); Koenker et al. (2017); Callaway et al. (2018); Callaway & Li (2019); Chernozhukov et al. (2019); Ge et al. (2020); Zhou et al. (2022); Kallus et al. (2024), among others. Some recent works, including Park et al. (2021) and Kallus & Oprescu (2023), explore the Conditional Distributional Treatment Effects as distributional analysis is useful even after conditioning on observed variables. However, there has been limited research on regression adjustment for unconditional distributional treatment effects. One exception is by Jiang et al. (2023), who consider quantile-regression adjustment for the QTE, but under covariate-adaptive randomization. Another exception is the study by Oka et al. (2024), which investigates the distribution regression approach using finite-dimensional covariates. We bridge this gap by proposing regression-adjusted estimators for various distributional parameters when data are obtained from randomized experiments with possibly high-dimensional covariates. Furthermore, our approach accommodates both discrete and mixed discrete-and-continuous outcome distributions, whereas quantile regression adjustment is specifically designed for continuous outcomes.

Semiparametric Estimation Our work is closely linked to the extensive literature on semiparametric estimation, which addresses the challenge of estimating low-dimensional parameters in the presence of high-dimensional nuisance parameters. This literature includes seminal contributions such as Klaassen (1987); Robinson (1988); Bickel et al. (1993); Andrews (1994a); Newey (1994); Robins & Rotnitzky (1995); Chernozhukov et al. (2018); Ichimura & Newey (2022), among others. Our setup can be framed as a semiparametric problem characterized by the Neyman-orthogonal moment condition, as outlined in Neyman (1959); Chernozhukov et al. (2018; 2022). Notably, cross-fitting is a commonly used technique in this literature. While our technical arguments share similarities with classical semiparametric methods, our research introduces a novel perspective by emphasizing the significance of flexible machine learning methods for estimating distributional treatment effects, within the framework of randomized experiments.

3. Setup and Parameters

3.1. Setup and Notation

We assume that our data are generated from a randomized experiment with K treatment arms. Let $Y \in \mathcal{Y} \subset \mathbb{R}$ denote the scalar-valued observed outcome, $W \in \mathcal{W} := \{1, \ldots, K\}$ denote the index of the treatment arm, and $X \in \mathcal{X} \subset \mathbb{R}^{d_x}$ denote pre-treatment covariates. We observe a size n random sample $\{Z_i\}_{i=1}^n := \{(X_i, W_i, Y_i)\}_{i=1}^n$ from the distribution of $Z := (X, W, Y)$. The probability of assignment to treatment arm w is denoted as $\pi_w := P(W_i = w)$ satisfying $\sum_{w \in \mathcal{W}} \pi_w = 1$, while n_w indicates the number of observations in treatment group w, satisfying $\sum_{w \in \mathcal{W}} n_w = n$.

We follow the potential outcome framework [e.g., Rubin (1974); Imbens & Rubin (2015)] and let $Y(1), \ldots, Y(K)$ denote the potential outcomes, which are hypothetical and represent what the outcome for an individual would be under each treatment scenario. These are unobserved variables and we only observe the outcome for the treatment that is actually administered to each individual. We assume no interference and impose Stable Unit Treatment Values Assumption (SUTVA), which gives us $Y = Y(W)$. In other words, treatment assigned to one unit does not affect the outcome for another unit, and so the potential outcome under any treatment is equal to its observed outcome. Throughout the paper, we also maintain the following two assumptions.

Assumption 3.1. $Y(1), \ldots, Y(K), X \perp\!\!\!\perp W$.

Assumption 3.2. $0 < \pi_w < 1$ for each $w \in \mathcal{W}$.

Assumption 3.1 states that the treatment indicator is independent of the potential outcomes and the pre-treatment covariates. Assumption 3.2 states that the treatment assignment probabilities are bounded away from 0 and 1. These assumptions are satisfied because we have a randomized experiment where the researcher assigns individuals to treatment groups randomly and have a control over the treatment assignment probabilities.

3.2. Parameters of interest

The parameters of our interest are based on (cumulative) distribution functions of potential outcomes, denoted by

$$F_{Y(w)}(y) := E[\mathbb{1}_{\{Y(w) \leqslant y\}}],$$

for $y \in \mathcal{Y} \subset \mathbb{R}$ and $w \in \mathcal{W}$, where $\mathbb{1}_{\{\cdot\}}$ represents the indicator function. In general, the potential outcomes $\{Y(w)\}_{w \in \mathcal{W}}$ are unobserved variables. However, under Assumptions 3.1 and 3.2, the potential outcome distribution $F_{Y(w)}(y)$ is the same as the outcome distribution $F_{Y|W}(y|w)$ under each treatment w. Therefore, they are identifiable given the data from $Z = (X, W, Y)$.

The results of this paper can be applied to estimate a range of distributional parameters, provided that they rely on (Hadamard) differentiable transformations of potential outcome distributions. We provide a few illustrative examples below.

Example 1: Distributional Treatment Effect Let $w, w' \in \mathcal{W}$ be two different treatment groups. We are interested in the Distributional Treatment Effect (DTE), which is defined as, for $y \in \mathcal{Y}$,

$$DTE_{w,w'}(y) := F_{Y(w)}(y) - F_{Y(w')}(y).$$

To contrast, the Average Treatment Effect (ATE) is defined as

$$ATE_{w,w'} := E[Y(w)] - E[Y(w')].$$

The DTE is a parameter that is indexed by a continuum of $y \in \mathcal{Y}$ and measures the effect of treatment on the whole distribution, whereas the ATE only quantifies the mean effect. As a special case, one can also be interested in the DTE at a certain threshold; i.e., $\mathcal{Y}$ can be defined to be a singleton set. One advantage of this measure is that it is well-defined for any type of outcome, including discrete, continuous, and mixed discrete-continuous variables.

Example 2: Probability Treatment Effect The DTE may not be straightforward to interpret since it measures the differences between two cumulative distributions. However, we can compute more intuitive measures based on these differences. Specifically, the DTE can be used to compute, what we will call, the Probability Treatment Effect (PTE), which is given by

$$\begin{aligned} PTE_{w,w'}(y,h) :=& (F_{Y(w)}(y+h) - F_{Y(w)}(y)) \\ & - (F_{Y(w')}(y+h) - F_{Y(w')}(y)), \end{aligned}$$

for $y \in \mathcal{Y}$ and some $h > 0$. The PTE measures the changes in the probability that the outcome falls in interval $(y, y+h]$. The PTE is also well-defined for any type of outcome, including discrete, continuous, and mixed discrete-continuous variables.

Example 3: Quantile Treatment Effect Another common measure used to characterize the entire distribution is the quantile function, defined as $F^{-1}_{Y(w)}(\tau) := \inf\{y : F_{Y(w)}(y) \geqslant \tau\}$ for $\tau \in (0,1)$. The Quantile Treatment Effect (QTE) for quantile $\tau \in (0,1)$ is then given by

$$QTE_{w,w'}(\tau) := F^{-1}_{Y(w)}(\tau) - F^{-1}_{Y(w')}(\tau).$$

The QTE quantifies the difference in quantiles between two potential outcome distributions across a continuum of $\tau \in (0,1)$. For example, one might be interested in the difference between the medians (when $\tau = 0.5$) of two groups. It is important to note that the QTE is only well-defined for continuous outcomes and may not be appropriate for discrete or mixed discrete-continuous outcomes.

4. Regression-Adjusted Estimator

As explained in the previous section, the potential outcome distribution serves as the fundamental building block for a broad range of distributional parameters. A simple estimator for the distribution function $F_{Y(w)}(y)$ is the empirical distribution function, given by:

$$\widehat{\mathbb{F}}^{simple}_{Y(w)}(y) := \frac{1}{n_w} \sum_{i:W_i=w} 1\!\!1_{\{Y_i \leqslant y\}},$$

for each treatment $w \in \mathcal{W}$. While this estimator is an unbiased and consistent estimator, we aim to enhance its precision by leveraging pre-treatment covariates.

To incorporate pre-treatment covariates X, we consider the distribution regression framework, in which the conditional distribution function $F_{Y(w)|X}(y|x)$ is regarded as the mean regression for a binary dependent variable $1\!\!1_{\{Y(w)\leqslant y\}}$. That is, for each $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, we can write

$$F_{Y(w)|X}(y|X) = E[1\!\!1_{\{Y(w)\leqslant y\}}|X].$$

For each location $y \in \mathcal{Y}$, the conditional mean function can be separetely estimated using various methods, such as linear regression, logistic regression, or other machine learning techniques (e.g., LASSO, random forests, boosted trees, deep neural networks, etc.). Additionally, the distribution regression is applicable for continuous, discrete, and mixed discrete-and-continuous outcome variable as explained in Chernozhukov et al. (2013).

For the regression-adjusted estimator of $F_{Y(w)}(y)$, the conditional distribution functions are nuisance parameters. We represent them as $\gamma_y^{(w)}(x) := F_{Y(w)|X}(y|x)$ for each $w \in \mathcal{W}$ and let $\widehat{\gamma}_y^{(w)}(\cdot)$ be an estimator for $\gamma_y^{(w)}(\cdot)$. We will explain the necessary conditions for the estimator $\widehat{\gamma}_y^{(w)}(\cdot)$ in the following section. The regression-adjusted estimator of $F_{Y(w)}(y)$ for each $w \in \mathcal{W}$ is then defined as

$$\widehat{\mathbb{F}}_{Y(w)}(y) := \underbrace{\frac{1}{n_w} \sum_{i:W_i=w} \left(1\!\!1_{\{Y_i \leqslant y\}} - \widehat{\gamma}_y^{(w)}(X_i)\right)}_{\text{averaged over observations in treatment group } w} + \underbrace{\frac{1}{n} \sum_{i=1}^{n} \widehat{\gamma}_y^{(w)}(X_i)}_{\text{averaged over all observations}}. \quad (1)$$

The regression-adjusted estimator is obtained by adjusting the empirical distribution function by subtracting $\widehat{\gamma}_y^{(w)}(\cdot)$ that is averaged over observations in that treatment group and adding $\widehat{\gamma}_y^{(w)}(\cdot)$ that is averaged over all observations.

This characterization of regression adjustment aligns closely with the concept of the augmented inverse propensity weighted estimator, as explored in Robins et al. (1994); Robins & Rotnitzky (1995). List et al. (2022) also consider a similar adjustment method for estimating the ATE. We extend this formulation to encompass distribution functions for any arbitrary outcome location $y \in \mathcal{Y}$. It is worth noting that this estimator also serves as an unbiased estimator for the distribution function, as the expected value of the adjustment terms cancels out to zero.

Moment condition problem We rewrite our problem as a moment condition problem. In what follows, we will simply write $\gamma_y^{(w)}$ to denote $\gamma_y^{(w)}(\cdot)$ and let $\gamma_y := (\gamma_y^{(1)}, \ldots, \gamma_y^{(K)})^\top$. Let $\theta_y^{(w)} := F_{Y(w)}(y)$ and $\theta_y := (\theta_y^{(1)}, \ldots, \theta_y^{(K)})^\top$. Define moment functions

$$\psi_y(Z; \theta_y, \gamma_y) := \left(\psi_y^{(1)}(Z; \theta_y, \gamma_y), \ldots, \psi_y^{(K)}(Z; \theta_y, \gamma_y)\right)^\top,$$

where, for each $w \in \mathcal{W}$,

$$\psi_y^{(w)}(Z; \theta_y, \gamma_y) := \frac{1\!\!1_{\{W=w\}} \cdot (1\!\!1_{\{Y\leqslant y\}} - \gamma_y^{(w)}(X))}{\pi_w} + \gamma_y^{(w)}(X) - \theta_y^{(w)}. \quad (2)$$

The following lemma shows what moment conditions are implied by our setup with a randomized experiment. Later, we will show how the regression-adjusted estimator, given in (1), can be seen as a method of moments estimator that solves the sample counterpart of these moment conditions.

Lemma 4.1 (Moment conditions)**.** *We have the following moment conditions for a continuum of $y \in \mathcal{Y}$:*

$$E[\psi_y(Z; \theta_y, \gamma_y)] = 0, \quad (3)$$

where $\psi_y^{(w)}(Z; \theta_y, \gamma_y)$ for each $w \in \mathcal{W}$ is given in (2).

Our parameter of interest θ_y for each $y \in \mathcal{Y}$ is identified as the solution to the moment condition in (3), where $Z = (X, W, Y)$ is the data and γ_y is the possibly infinite-dimensional nuisance parameter.

An important property of the moment conditions defined in (3) is that they are Neyman orthogonal with respect to the nuisance parameters (Neyman, 1959; Chernozhukov et al., 2018; 2022; Ichimura & Newey, 2022). More precisely, the derivative of its expectation with respect to the nuisance parameters vanishes when evaluated at the true parameter values. The following lemma states it formally.

Lemma 4.2 (Neyman Orthogonality)**.** *For the continuum of moment conditions defined in (2) for each $w \in \mathcal{W}$ and $y \in \mathcal{Y}$, we have*

$$\frac{\partial}{\partial t} E[\psi_y(Z; \theta_y, t)]\Big|_{t=\gamma_y} = 0, \qquad a.s.$$

Neyman orthogonality implies that the moment condition is first-order insensitive to the estimation errors of the nuisance parameters. This property, coupled with a form of

Algorithm 1 Regression-adjusted estimator

Input: Data $\{(X_i, W_i, Y_i)\}_{i=1}^n$ split randomly into L roughly equal-sized folds where $L > 1$; $\mathcal{S}$ a supervised learning algorithm
for $\ell = 1$ **to** L **do**
  Generate $\widehat{\gamma}_y^{(w)}(X_i)$ predicting $\mathbb{1}_{\{Y_i \leqslant y\}}$ given $W_i = w$ and X_i for each treatment group $w \in \mathcal{W}$ and each level $y \in \mathcal{Y}$, by training on data not in fold ℓ but in treatment group w, using $\mathcal{S}$.
end for
Compute $\widehat{F}_{Y(w)}(y)$, for each $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, according to (1).
Result: Regression-adjusted estimator $(\widehat{\theta}_y)_{y \in \mathcal{Y}}$.

sample-splitting called cross-fitting, allows us to derive the asymptotic distribution of the regression-adjusted estimator under mild conditions, even when the conditional distribution functions are estimated via machine learning (ML) methods.

The following lemma shows how the regression-adjusted estimator, given in (1), can be seen as an estimator that solves the sample counterpart of the moment conditions defined earlier.

Lemma 4.3 (Sample moment condition)**.** *For each $y \in \mathcal{Y}$, let the vector $\widehat{\gamma}_y$ denote the ML estimator of the vector γ_y. Then the regression-adjusted estimator $\widehat{\theta}_y := (\widehat{F}_{Y(1)}(y), \ldots, \widehat{F}_{Y(K)}(y))^\top$, where $\widehat{F}_{Y(w)}(y)$ for $w \in \mathcal{W}$ is defined in (1), is constructed as the solution to the following sample moment condition:*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_y(Z_i; \widehat{\theta}_y, \widehat{\gamma}_y) = 0.$$

Estimation procedure We now explain our algorithm, which is summarized in Algorithm 1. Our estimation procedure involves a sample-splitting method called cross-fitting [e.g., Chernozhukov et al. (2018)]. First, we split the data into L roughly equal-sized folds, where $L > 1$. Then, for every observation, we use a ML method and predict nuisance functions $\widehat{\gamma}_y^{(w)}(X_i)$ by training on data from treatment group w, excluding data points from the fold the observation belongs to. This ensures that the observation and the nuisance estimates are independent. Finally, we form a point estimate of $F_{Y(w)}(y)$ by plugging in estimates of the nuisances in (1). We do this for all $y \in \mathcal{Y}$ and $w \in \mathcal{W}$. Then we stack the estimators together to get a regression-adjusted estimator $(\widehat{\theta}_y)_{y \in \mathcal{Y}}$. We discuss the statistical inference in the next section.

Efficiency Gain To illustrate the potential efficiency gain from the proposed regression-adjusted method, consider the scenario where the true conditional distribution function, γ_y, is employed in (1), leading to the idealized form of the regression-adjusted estimator, denoted by $\widetilde{\theta}_y^{(w)} := \widetilde{F}_{Y(w)}(y)$. Let $\widetilde{\theta}_y := (\widetilde{\theta}_y^{(1)}, \ldots, \widetilde{\theta}_y^{(K)})^\top$. The following theorem highlights the efficiency improvements of this regression-adjusted estimator in comparison to the empirical distribution function. As demonstrated in the next section, our estimator asymptotically possess the same efficiency property in terms of variance.

Theorem 4.4. *Suppose that $n_w/n = \pi_w + o(1)$ as $n \to \infty$ for every $w \in \mathcal{W}$. Then, we have*
(a) for any $w \in \mathcal{W}$ and $y \in \mathcal{Y}$,

$$\mathrm{Var}\big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y)\big) \geq \mathrm{Var}\big(\widetilde{\mathbb{F}}_{Y(w)}(y)\big) + o(n^{-1}),$$

where the equality holds only if $F_{Y(w)|X}(y) = F_{Y(w)}(y)$,
(b) for any $y \in \mathcal{Y}$,

$$\mathrm{Var}\big(\widehat{\theta}^{simple}_y\big) \succeq \mathrm{Var}\big(\widetilde{\theta}_y\big) + o(n^{-1}),$$

where $\succeq$ denotes the positive semi-definiteness. When $\mathrm{Var}\big(F_{Y(w)}(y|X) - r \cdot F_{Y(w')}(y|X)\big) > 0$ for any distinct $w, w' \in \mathcal{W}$ and $r \in \mathbb{R}$, the positive definite result holds.

Theorem 4.4(a) shows the efficiency gains achieved by applying regression adjustment to distribution functions. Furthermore, Theorem 4.4(b) elaborates on these gains in terms of a vector of regression-adjusted estimators, indicating a marked improvement in the precision of the estimator for the DTE as a special case.

5. Asymptotic distribution

In this section, we derive the asymptotic distribution of the regression-adjusted estimator. These results are built upon the functional central limit theorem, functional delta method and other related results from Belloni et al. (2017).

Additional Notation We introduce some additional notations to state our results. For a vector $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$, $\|a\| = \sqrt{a^\top a}$ denotes the Euclidean norm of a. Let $\ell^\infty(\mathcal{Y})$ be the space of uniformly bounded functions mapping an arbitrary index set $\mathcal{Y}$ to the real line; $UC(\mathcal{Y})$ be the space of uniformly continuous functions mapping $\mathcal{Y}$ to the real line. $\mathbb{G}_{n,P} f$ denotes the empirical process $\sqrt{n}\sum_{i=1}^n (f(Z_i) - \int f(z)dP(z))$; but we will omit P and simply write $\mathbb{G}_n f$. Let $\mathcal{P}_n$ denote the set of probability measures, that is weakly increasing in n, i.e., $\mathcal{P}_n \subseteq \mathcal{P}_{n+1}$. We use $\rightsquigarrow$ to denote the convergence in distribution or law. Lastly, let $\mathbb{G}_P$ denote the P-Brownian bridge, as defined in Appendix Section D.1.

The following theorem shows that under regularity conditions, stated fully in Appendix D.3, our estimator $(\widehat{\theta}_y)_{y \in \mathcal{Y}}$ is asymptotically Gaussian. Since we employ cross-fitting,

the conditions required for the estimation of nuisance functions become much milder compared to not using any data-splitting. It is required that the estimators of nuisance functions attain sufficiently rapid rates of convergence τ_n, in particular $\tau_n = o(n^{-1/4})$ in smooth problems.

Theorem 5.1 (Uniform Functional Central Limit Theorem)**.** *Suppose Assumption D.11, D.12 and D.13 hold. Then, for an estimator $(\widehat{\theta}_y)_{y\in\mathcal{Y}}$ that is defined in Algorithm 1,*

$$\sqrt{n}(\widehat{\theta}_y - \theta_y)_{y\in\mathcal{Y}} = (\mathbb{G}_n\psi_y)_{y\in\mathcal{Y}} + o_p(1)$$

in $\ell^\infty(\mathcal{Y})^K$ uniformly in $P \in \mathcal{P}_n$, where

$$Z_{n,P} := (\mathbb{G}_n\psi_y)_{y\in\mathcal{Y}} \rightsquigarrow Z_P := (\mathbb{G}_P\psi_y)_{y\in\mathcal{Y}}$$

in $\ell^\infty(\mathcal{Y})^K$ uniformly in $P \in \mathcal{P}_n$, where the paths of $y \mapsto \mathbb{G}_p\psi_y$ are a.s. uniformly continuous on a semi-metric space $(\mathcal{Y}, d_\mathcal{Y})$ and

$$\sup_{P\in\mathcal{P}_n} E_P \sup_{y\in\mathcal{Y}} \|\mathbb{G}_P\psi_y\| < \infty,$$
$$\lim_{\delta\to 0}\sup_{P\in\mathcal{P}_n} E_P \sup_{d_\mathcal{Y}(y,y')\leqslant\delta} \|\mathbb{G}_P\psi_y - \mathbb{G}_P\psi_{y'}\| = 0.$$

Then, as a special case of the above theorem, for fixed $y \in \mathcal{Y}$, we have pointwise asymptotic normality, stated as $\sqrt{n}(\widehat{\theta}_y - \theta_y) \rightsquigarrow N(0, \mathrm{Var}(\psi_y))$. Note that, $\mathrm{Var}(\psi_y)$ can be consistently estimated via sample moment conditions using cross-fitting as well. The estimate of the asymptotic variance can then be used to construct the confidence intervals in a usual manner.

Functionals of θ The parameters we are interested in are functionals of potential outcome distributions. The examples include the DTE, the PTE and the QTE we discussed in Section 3. If, for instance, we are interested in the DTE between treatments 1 and 2, for each $y \in \mathcal{Y}$, we can calculate it as $(1, -1, 0, \ldots, 0)\theta_y$. Let $\phi(\theta^0) = \phi((\theta_y)_{y\in\mathcal{Y}})$. The following theorem shows the large sample law of the plug-in estimator $\phi(\widehat{\theta}) := \phi((\widehat{\theta}_y)_{y\in\mathcal{Y}})$. For complex objects, the inference can be facilitated by bootstrap. The validity of a multiplier bootstrap method (Giné & Zinn, 1984) is also shown in the theorem below.

Theorem 5.2 (Uniform Limit Theory and Validity of Multiplier Bootstrap for Smooth Functionals of θ)**.** *Suppose that for each $P \in \mathcal{P} := \cup_{n\geqslant n_0}\mathcal{P}_n$, $\theta^0 = \theta^0_P$ is an element of a compact set $\mathbb{D}_\theta$. Let $\phi : \mathbb{D}_\phi \subset \ell^\infty(\mathcal{Y})^K \longmapsto \ell^\infty(\mathcal{Q})$ be Hadamard-differentiable uniformly in $\theta \in \mathbb{D}_\theta \subset \mathbb{D}_\phi$ tangentially to $UC(\mathcal{Y})^K$ with derivative map ϕ'_θ. Then,*

$$\sqrt{n}(\phi(\widehat{\theta}) - \phi(\theta^0)) \rightsquigarrow T_P := \phi'_{\theta^0_P}(Z_P) \text{ in } \ell^\infty(\mathcal{Q}),$$

uniformly in $P \in \mathcal{P}_n$, where T_P is a zero mean tight Gaussian process, for each $P \in \mathcal{P}$. Moreover,

$$\sqrt{n}(\phi(\widehat{\theta}^*) - \phi(\widehat{\theta})) \rightsquigarrow_B T_P \text{ in } \ell^\infty(\mathcal{Q}),$$

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

Here $\rightsquigarrow_B$ denotes weak convergence of the bootstrap law in probability, as defined in Appendix D. $\phi(\widehat{\theta}^*) = \phi((\widehat{\theta}^*)_{y\in\mathcal{Y}})$ is the bootstrap version of $\phi(\widehat{\theta})$, and $\widehat{\theta}^*_y = \widehat{\theta}_y + n^{-1}\sum_{i=1}^n \xi_i\psi_y(Z_i; \widehat{\theta}_y, \widehat{\gamma}_y)$ is the multiplier bootstrap version of $\widehat{\theta}_y$. More details about the multiplier bootstrap procedure to obtain pointwise and uniform confidence bands can be found in the Appendix C. The assumption of Hadamard differentiability is imposed so that we can use the delta method. The formal definition can be found in the Appendix D.

6. Empirical results

In this section, we compare our regression-adjusted estimators to simple estimators in two types of experiments. In the first experiment, we use a synthetic dataset to assess the performance of our proposed method in finite samples. For the second experiment, we reanalyze data from a randomized experiment, conducted by Ferraro & Price (2013b), to compare the methods using real-world data.

6.1. Simulation Study

We conduct Monte Carlo simulation study to evaluate the performance of our adjusted estimators in finite samples. We compare the simple estimator with two regression-adjusted estimators - linear adjustment and ML adjustment. The simple estimator is based on empirical distribution functions. The regression-adjusted estimators are calculated according to the procedure in Algorithm 1 with 5 folds. For the linear adjustment, we use a linear regression for estimating $\widehat{\gamma}_y$. For the ML adjustment, we use logistic LASSO for estimating $\widehat{\gamma}_y$.

In this experiment, we generate i.i.d. sample of size $n \in \{500, 1000, 5000\}$ with covariates, a binary treatment and a continuous outcome. We design the data generating process such that the half of covariates are irrelevant to the outcome. Appendix E contains more details about the data generating process and describes the evaluation metrics.

The top figure of Figure 1 plots the bias of these estimators as a % of the true values of the DTE, across different quantiles under the sample sizes we consider. We confirm that the bias is small for all estimators across all quantiles. Even when sample size is small ($n = 500$), the bias is at most 2%. This is as expected since all estimators are unbiased estimators of the distribution functions and hence the DTE.

Next, we turn to the RMSE. The bottom figure of Figure 1 plots the RMSE reduction in % terms for the linear and ML adjustment, compared to the simple estimator. We see that the linear adjustment and ML adjustment estimators yield smaller RMSE compared to the simple estimator across all quantiles. Moreover, we see that ML adjustment out-

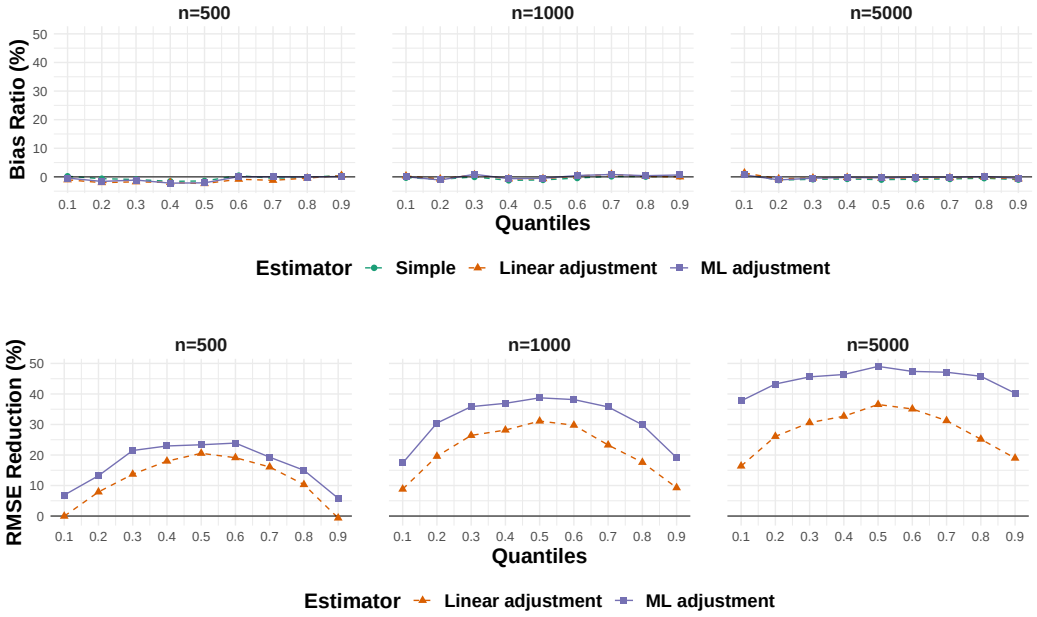

Figure 1. Bias (top figure), as a % of true value, of different DTE estimators and RMSE reduction in % (bottom figure) of adjusted estimators compared to simple DTE estimator, under sample sizes $\{500, 1000, 5000\}$, calculated over 1000 simulations. The simple estimator is calculated from empirical distribution functions. The regression-adjusted estimators (linear adjustment and ML adjustment based on LASSO) are implemented using 5-fold cross-fitting.

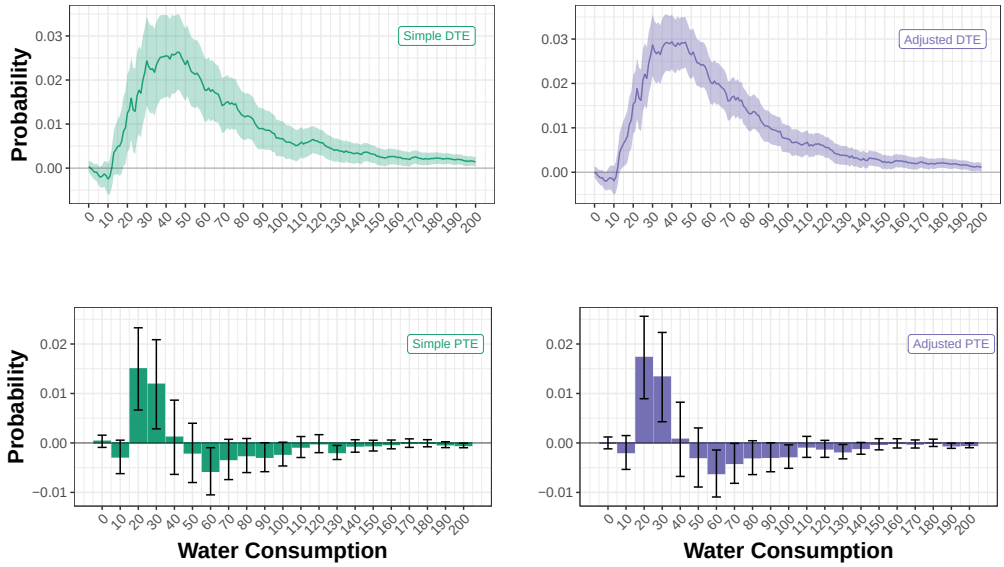

Figure 2. Distributional Treatment Effect (DTE) and Probability Treatment Effect (PTE) of a nudge (Strong Social Norm vs. Control) on water consumption (in thousands of gallons). The top left figure represents the simple DTE; the top right figure depicts the regression-adjusted DTE, computed for $y \in \{0, 1, 2, \ldots, 200\}$. The bottom left figure represents the simple PTE; the bottom right figure represents the regression-adjusted PTE, computed for $y \in \{0, 10, 20, \ldots, 200\}$ and $h = 10$. The regression adjustment is implemented via gradient boosting with 10-fold cross-fitting. The shaded areas and error bars represent the 95% pointwise confidence intervals. $n = 78,500$.

performs linear adjustment in all cases. Specifically, for $n = 5000$, RMSE reduction over the simple estimator is around 40-50% for the ML adjusted estimator, while it is between roughly 15-35% for the linearly adjusted estimator. This is as expected because our data generating process consists of variables that are irrelevant to the outcome, and so ML adjustment better captures the relationship between the outcome and covariates compared to the linear model. Improved prediction quality for $\widehat{\gamma}_y$ results in more variance reduction for the DTE.

6.2. Nudges to reduce water consumption

We reanalyze data from a randomized experiment conducted by Ferraro & Price (2013b) in 2007, to examine the effect of norm-based messages or nudges on water usage in Cobb County, Atlanta, Georgia. Three different interventions to reduce water usage were implemented and compared to the control group (no nudge). List et al. (2022) re-estimate the regression-adjusted ATE for the intervention called "strong social norm" that combines prosocial appeal and social comparison (the strongest intervention) relative to the control group. We extend their analysis by estimating the regression-adjusted DTE and PTE of this intervention over the control group ($W \in \{0, 1\}$) using the same pre-treatment covariates X, which is monthly water consumption in the year prior to the experiment. Thus, the dimension of the covariate space $\mathcal{X}$ is $d_x = 12$. The outcome variable Y is level of water consumption from June to September of 2007. The unit of our outcome variable is in thousands of gallons and is discretely distributed. Note that although the measure in gallons appears to be approximately continuous in practice, the presence of subtle discreteness can create problems for both theoretical and practical statistical inference. Thus, the QTE is not applicable here. The results for the other two treatments - "technical advice" and "weak social norm" - relative to the control group are summarized in Appendix E.2.

Figure 2 plots the DTE and the PTE of the intervention compared to the control group. We compute the DTE for $y \in \{0, 1, 2, \ldots, 200\}$. Figure 2 top left represents the simple estimate of the DTE, whereas the top right figure depicts the regression-adjusted estimate of the DTE. For regression adjustment, we estimate the conditional distribution functions $\widehat{\gamma}_y$ via gradient boosting using 10-fold cross-fitting. The shaded areas represent the 95% pointwise confidence intervals. We can see that the regression-adjusted DTE has substantially smaller variance and hence tighter confidence intervals compared to the simple DTE, especially between 15 and 110. Based on the regression-adjusted DTE results, we see that the DTE is increasing up until 40-50 and starts declining after that. We can draw conclusions about how the outcome distributions differ under treatment and control, keeping in mind this is differences in cumulative distributions.

More intuitive measure of this distribution change is the PTE. Figure 2 bottom left represents the simple PTE, while the bottom right represents the regression-adjusted PTE, with the 95% pointwise confidence intervals. We show the PTE in increments of $h = 10$ for $y \in \{0, 10, 20, \ldots, 200\}$. We see that the treatment is effective in that it reduces the probability of high water consumption and increases the probability of low water consumption. Specifically, the results from the regression-adjusted PTE indicate increase in water usage in the range of 20-40; decrease in water usage in the range of 60-110 (with a minor exception that within the range (80, 90], which is represented by point 80 in the graph, the confidence interval exceeds 0 by only a little). The variance reduction is especially helpful at the range (70, 80]. The probability change for the range (70,80] is not significant under simple estimates, while it is significantly negative under regression adjustment.

7. Conclusion

We provide a novel regression adjustment method to estimate various measures of distributional treatment effects to capture heterogeneity. Our framework accommodates high-dimensional setup with many pre-treatment covariates and offers flexible modeling by incorporating machine learning techniques for the regression adjustment.

Some limitations of our method are as follows. Firstly, we consider a setting where we have an experimental data with perfect compliance and no interference (no network or peer effects). While suitable for some applications, these assumptions may prove restrictive in other contexts. Secondly, our approach relies on the presence of pre-treatment covariates highly predictive of the outcome. While we enhance variance reduction compared to linear regression by employing flexible machine learning methods to improve prediction quality, substantial variance reduction may not occur if the covariates lack high-quality information. Thirdly, we focus on a setup where experimental data is already collected, neglecting opportunities to incorporate variance reduction strategies at the design stage of the experiment. These limitations suggest avenues for future research on distributional analysis that incorporates these concerns.

Acknowledgements

We extend our gratitude to the four anonymous reviewers and the program chairs for their insightful comments and discussions, which significantly enhanced the quality of this paper. We also appreciate the feedback provided by Hiroki Yanagisawa and Shuting Wu. Additionally, Oka acknowledges the financial support from JSPS KAKENHI Grant Number 24K04821.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abadie, A. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292, 2002.

Abadie, A., Angrist, J., and Imbens, G. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, 70(1): 91–117, 2002.

Alaa, A. M. and Van Der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Andrews, D. W. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pp. 43–72, 1994a.

Andrews, D. W. Empirical process methods in econometrics. *Handbook of econometrics*, 4:2247–2294, 1994b.

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

Athey, S. and Imbens, G. W. Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497, 2006.

Bakshy, E., Eckles, D., and Bernstein, M. S. Designing and deploying online field experiments. In *Proceedings of the 23rd international conference on World wide web*, pp. 283–292, 2014.

Belloni, A., Chernozhukov, V., Fernandez-Val, I., and Hansen, C. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

Berk, R., Pitkin, E., Brown, L., Buja, A., George, E., and Zhao, L. Covariance adjustments for the analysis of randomized field experiments. *Evaluation review*, 37 (3-4):170–196, 2013.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Bitler, M. P., Gelbach, J. B., and Hoynes, H. W. What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, 96(4):988–1012, 2006.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S., and Yu, B. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

Callaway, B. and Li, T. Quantile treatment effects in difference in differences models with panel data. *Quantitative Economics*, 10(4):1579–1618, 2019.

Callaway, B., Li, T., and Oka, T. Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413, 2018.

Chen, T. and Guestrin, C. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. ACM.

Chernozhukov, V. and Hansen, C. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

Chernozhukov, V., Fernández-Val, I., and Melly, B. Inference on counterfactual distributions. *Econometrica*, 81 (6):2205–2268, 2013.

Chernozhukov, V., Chetverikov, D., and Kato, K. Gaussian approximation of suprema of empirical processes. *The Annals of Statistics*, 42(4):1564, 2014.

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

Chernozhukov, V., Fernandez-Val, I., Melly, B., and Wüthrich, K. Generic inference on quantile and quantile effect functions for discrete outcomes. *Journal of the American Statistical Association*, 2019.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022.

Chiang, H. D., Matsushita, Y., and Otsu, T. Regression adjustment in randomized controlled trials with many covariates. *arXiv preprint arXiv:2302.00469*, 2023.

Cochran, W. G. *Sampling techniques*. john wiley & sons, 1977.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pp. 123–132, 2013.

Ding, P., Feller, A., and Miratrix, L. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.

Doksum, K. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The annals of statistics*, pp. 267–277, 1974.

Ferraro, P. and Price, M. Replication data for: Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. Technical report, 2013a. URL https://doi.org/10.7910/DVN1/22633.

Ferraro, P. J. and Price, M. K. Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1): 64–73, 2013b.

Firpo, S. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276, 2007.

Fisher, R. A. *Statistical methods for research workers*. Oliver and Boyd, 1932.

Fisher, R. A. *The Design of Experiments*. Oliver and Boyd, 1935.

Freedman, D. A. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008a.

Freedman, D. A. On regression adjustments in experiments with several treatments. *Annals of Applied Statistics*, 2: 176–96, 2008b.

Friedman, J., Tibshirani, R., and Hastie, T. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01.

Ge, Q., Huang, X., Fang, S., Guo, S., Liu, Y., Lin, W., and Xiong, M. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in genetics*, 11:585804, 2020.

Giné, E. and Zinn, J. Some limit theorems for empirical processes. *The Annals of Probability*, pp. 929–989, 1984.

Guo, X., Zhang, Y., Wang, J., and Long, M. Estimating heterogeneous treatment effects: mutual information bounds and learning algorithms. In *International Conference on Machine Learning*, pp. 12108–12121. PMLR, 2023.

Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., and Goldman, M. Machine learning for variance reduction in online experiments. *Advances in Neural Information Processing Systems*, 34:8637–8648, 2021.

Heckman, J. J., Smith, J., and Clements, N. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4):487–535, 1997.

Ichimura, H. and Newey, W. K. The influence function of semiparametric estimators. *Quantitative Economics*, 13 (1):29–61, 2022.

Imai, K. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, 2005.

Imai, K. and Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 2013.

Imbens, G. W. and Rubin, D. B. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997.

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

Jiang, L., Phillips, P. C., Tao, Y., and Zhang, Y. Regression-adjusted estimation of quantile treatment effects under covariate-adaptive randomizations. *Journal of Econometrics*, 234(2):758–776, 2023.

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

Kallus, N. and Oprescu, M. Robust and agnostic learning of conditional distributional treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pp. 6037–6060. PMLR, 2023.

Kallus, N., Mao, X., and Uehara, M. Localized debiased machine learning: Efficient inference on quantile treatment effects and beyond. *Journal of Machine Learning Research*, 25(16):1–59, 2024.

Klaassen, C. A. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, 15(4):1548–1562, 1987.

Koenker, R. *Quantile regression*, volume 38. Cambridge university press, 2005.

Koenker, R., Chernozhukov, V., He, X., and Peng, L. *Handbook of quantile regression*. CRC press, 2017.

Kohavi, R., Tang, D., and Xu, Y. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

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

Lehmann, E. L. and D'Abrera, H. J. *Nonparametrics: statistical methods based on ranks.* Holden-day, 1975.

Lei, L. and Ding, P. Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828, 2021.

Lin, W. Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

List, J. A., Muir, I., and Sun, G. K. Using machine learning for efficient flexible regression adjustment in economic experiments. Technical report, National Bureau of Economic Research, 2022.

Negi, A. and Wooldridge, J. M. Revisiting regression adjustment in experiments with heterogeneous treatment effects. *Econometric Reviews*, 40(5):504–534, 2021.

Newey, W. K. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pp. 1349–1382, 1994.

Neyman, J. Optimal asymptotic tests of composite hypotheses. *Probability and statsitics*, pp. 213–234, 1959.

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

Oka, T., Yasui, S., Hayakawa, Y., and Byambadalai, U. Regression adjustment for estimating distributional treatment effects in randomized controlled trials. *arXiv preprint*, 2024.

Park, J., Shalit, U., Schölkopf, B., and Muandet, K. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *International Conference on Machine Learning*, pp. 8401–8412. PMLR, 2021.

Poyarkov, A., Drutsa, A., Khalyavin, A., Gusev, G., and Serdyukov, P. Boosted decision tree regression adjustment for variance reduction in online controlled experiments. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 235–244, 2016.

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

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

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pp. 931–954, 1988.

Rosenbaum, P. R. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002.

Rosenblum, M. and Van Der Laan, M. J. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics*, 6(1), 2010.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

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

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Sverdrup, E. and Cui, Y. Proximal causal learning of conditional average treatment effects. In *International Conference on Machine Learning*, pp. 33285–33298. PMLR, 2023.

Tang, D., Agarwal, A., O'Brien, D., and Meyer, M. Overlapping experiment infrastructure: More, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 17–26, 2010.

Tay, J. K., Narasimhan, B., and Hastie, T. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31, 2023. doi: 10.18637/jss.v106.i01.

Tsiatis, A. A., Davidian, M., Zhang, M., and Lu, X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in medicine*, 27(23):4658–4677, 2008.

van der Laan, L., Ulloa-Pérez, E., Carone, M., and Luedtke, A. Causal isotonic calibration for heterogeneous treatment effects. *arXiv preprint arXiv:2302.14011*, 2023.

van der Vaart, A. and Wellner, J. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

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

Wager, S., Du, W., Taylor, J., and Tibshirani, R. J. High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(45):12673–12678, 2016.

Xie, H. and Aurisset, J. Improving the sensitivity of online controlled experiments: Case studies at netflix. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 645–654, 2016.

Yang, L. and Tsiatis, A. A. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

Zhou, T., Carson IV, W. E., and Carlson, D. Estimating potential outcome distributions with collaborating causal networks. *Transactions on machine learning research*, 2022, 2022.

Appendix

The Appendix is structured as follows. Section A summarizes the notation in the main text and introduces additional notations used in the Appendix. Section B provides key tools and proofs for the claims appeared in Section 4. Section C describes the multiplier bootstrap procedure for inference. Section D provides key tools and proofs for the claims presented in Section 5. Finally, Section E contains more detailed information and additional results from the experiments.

A. Summary of Notation

Table 1. Summary of Notation

Notation in the main text	
X	Pre-treatment covariates
W	Treatment variable
Y	Outcome variable
$Y(w)$	Potential outcome for treatment group w
$\{Z_i\}_{i=1}^n$	$\{(X_i, W_i, Y_i)\}_{i=1}^n$, observed data
π_w	Treatment assignment probability for treatment group w
n_w	Number of observations in treatment group w
$F_{Y(w)}(y)$	$E[1\!\!1_{\{Y(w)\leqslant y\}}]$, potential outcome distribution function
θ_y	$(F_{Y(1)}(y), \ldots, F_{Y(K)}(y))^\top$, vector of potential outcome distribution functions
$\gamma_y^{(w)}(x)$	$E[1\!\!1_{\{Y\leqslant y\}}\|W = w, X = x]$, conditional distribution function
γ_y	$(\gamma_y^{(1)}, \ldots, \gamma_y^{(K)})^\top$, vector of conditional distribution functions
$\psi_y^{(w)}(Z; \theta_y, \gamma_y)$	moment function defined in (2)
$\psi_y(Z)$	$\psi_y(Z; \theta_y, \gamma_y) := \left(\psi_y^{(1)}(Z; \theta_y, \gamma_y), \ldots, \psi_y^{(K)}(Z; \theta_y, \gamma_y)\right)^\top$, vector of moment functions
$\succeq$	positive semi-definiteness
$\|a\|$	$\sqrt{a^\top a}$, Euclidean norm of a vector $a = (a_1, \ldots, a_p)^\top \in \mathbb{R}^p$
$\ell^\infty(\mathcal{Y})$	space of uniformly bounded functions mapping an arbitrary index set $\mathcal{Y}$ to the real line
$UC(\mathcal{Y})$	space of uniformly continuous functions mapping an arbitrary index set $\mathcal{Y}$ to the real line
$\mathbb{G}_n f$	$\sqrt{n}\sum_{i=1}^n (f(Z_i) - \int f(z)dP(z))$, empirical process
$\mathbb{G}_P$	P-Brownian bridge
$Z_{n,P}$	$(\mathbb{G}_n \psi_y)_{y\in\mathcal{Y}}$
Z_P	$(\mathbb{G}_P \psi_y)_{y\in\mathcal{Y}}$
T_P	$\phi'_{\theta_P^0}(Z_P)$, where ϕ'_θ is derivative map of functional ϕ
$\mathcal{P}_n$	set of probability measures, that is weakly increasing in n
$\rightsquigarrow$	convergence in distribution or law
$\rightsquigarrow_B$	weak convergence of the bootstrap law in probability
Additional notation in the Appendix	
$N(\epsilon, \mathcal{F}, \|\cdot\|)$	ϵ-covering number of the class of functions $\mathcal{F}$ with respect to the norm $\|\cdot\|$
$a \vee b$	$\max\{a, b\}$ for real numbers a and b
$a \wedge b$	$\min\{a, b\}$ for real numbers a and b
$[K]$	$\{1, \ldots, K\}$ for a positive integer K
$x_n \lesssim y_n$	for sequences x_n and y_n in $\mathbb{R}$, $x_n \leqslant Ay_n$ for a constant A that does not depend on n
$\|\cdot\|_{P,q}$	$L^q(P)$ norm
$BL_1(\mathbb{D})$	space of functions mapping $\mathbb{D}$ to $[0, 1]$ with Lipschitz norm at most 1
$X_n = O_P(a_n)$	$\lim_{K\to\infty}\lim_{n\to\infty} P(\|X_n\| > Ka_n) = 0$ for a sequence of positive constants a_n
$X_n = o_P(a_n)$	$\sup_{K>0}\lim_{n\to\infty} P(\|X_n\| > Ka_n) = 0$ for a sequence of positive constants a_n

B. Key Tools and Proofs for Section 4

B.1. Proofs of Lemmas in Section 4

Proof of Lemma 4.1. Let $y \in \mathcal{Y}$ held constant. It is sufficient to show that each element in the vector of moment conditions in (3) holds. From the definition of $\psi_y^{(w)}(Z;\theta_y,\gamma_y)$ in (2), we can show, for each $w \in \mathcal{W}$,

$$\begin{aligned}
E\big[\psi_y^{(w)}(Z;\theta_y,\gamma_y)\big] &= E\Big[\frac{\mathbb{1}_{\{W=w\}}\cdot(\mathbb{1}_{\{Y\leqslant y\}}-\gamma_y^{(w)}(X))}{\pi_w}+\gamma_y^{(w)}(X)-F_{Y(w)}(y)\Big]\\
&= E\Big[\frac{\mathbb{1}_{\{W=w\}}\cdot(\mathbb{1}_{\{Y(w)\leqslant y\}}-\gamma_y^{(w)}(X))}{\pi_w}+\gamma_y^{(w)}(X)-F_{Y(w)}(y)\Big]\\
&= E\Big[E[\mathbb{1}_{\{W=w\}}]\cdot\frac{(\mathbb{1}_{\{Y(w)\leqslant y\}}-\gamma_y^{(w)}(X))}{\pi_w}+\gamma_y^{(w)}(X)-F_{Y(w)}(y)\Big]\\
&= E\Big[\pi_w\cdot\frac{(\mathbb{1}_{\{Y(w)\leqslant y\}}-\gamma_y^{(w)}(X))}{\pi_w}+\gamma_y^{(w)}(X)-F_{Y(w)}(y)\Big]\\
&= E\Big[\mathbb{1}_{\{Y(w)\leqslant y\}}-\gamma_y^{(w)}(X)+\gamma_y^{(w)}(X)-F_{Y(w)}(y)\Big]\\
&= 0.
\end{aligned}$$

Here, the first equality is due to the definition in (2) and the second equality comes from the definition of potential outcomes $Y = Y(W)$. The third equality holds because of the independence assumption in Assumption 3.1. The fourth equality comes from the fact that $E[\mathbb{1}_{\{W=w\}}] = P(W = w) = \pi_w$. Finally, all terms cancel out to be zero since $E[\mathbb{1}_{\{Y(w)\leqslant y\}}] = F_{Y(w)}(y)$ by definition. □

Proof of Lemma 4.2. For each $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, with probability approaching 1, we have

$$\begin{aligned}
\frac{\partial}{\partial t}E\big[\psi_y^{(w)}(Z;\theta_y,t)\big] &= E\Big[\frac{\partial}{\partial t}\Big(\frac{\mathbb{1}_{\{W=w\}}\cdot(\mathbb{1}_{\{Y\leqslant y\}}-t)}{\pi_w}+t-F_{Y(w)}(y)\Big)\Big]\\
&= E\Big[-\frac{\mathbb{1}_{\{W=w\}}}{\pi_w}+1\Big]\\
&= -\frac{E[\{\mathbb{1}_{\{W=w\}}]}{\pi_w}+1\\
&= 0.
\end{aligned}$$

Thus, the desired conclusion follows. □

Proof of Lemma 4.3. For each $y \in \mathcal{Y}$ and $w \in \mathcal{W}$, we substitute the estimate $\widehat{\gamma}_y^{(w)}(X_i)$ for $\gamma_y^{(w)}(X_i)$ and n_w/n for π_w in equation (3). Then, $\widehat{\theta}_y := \big(\widehat{\mathbb{F}}_{Y(1)}(y),\ldots,\widehat{\mathbb{F}}_{Y(K)}(y)\big)^\top$ solves the sample counterpart of equation (3) for θ_y, so that

$$\widehat{\mathbb{F}}_{Y(w)}(y) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\mathbb{1}_{\{W_i=w\}}\cdot(\mathbb{1}_{\{Y_i\leqslant y\}}-\widehat{\gamma}_y^{(w)}(X_i))}{n_w/n}+\widehat{\gamma}_y^{(w)}(X_i)\Big].$$

Rearranging the terms in the above equation, we obtain

$$\widehat{\mathbb{F}}_{Y(w)}(y) = \frac{1}{n_w}\sum_{i:W_i=w}\big(\mathbb{1}_{\{Y_i\leqslant y\}}-\widehat{\gamma}_y^{(w)}(X_i)\big)+\frac{1}{n}\sum_{i=1}^{n}\widehat{\gamma}_y^{(w)}(X_i),$$

which is the regression-adjusted estimator given in equation (1). □

B.2. Proof of Theorem 4.4

For the sake of proof completeness, we first present a variant of Lagrange's identity and Bergström's inequality in the below lemma, which is useful to prove the efficiency gain of the regression adjustment.

Lemma B.1. *For any $(a_1, \dots, a_K) \in \mathbb{R}^K$ and $(b_1, \dots, b_K) \in \mathbb{R}^K$ with $b_k > 0$ for all $k = 1, \dots, K$, we can show that*

$$\sum_{k=1}^{K} \frac{a_k^2}{b_k} - \frac{\left(\sum_{k=1}^{K} a_k\right)^2}{\sum_{k=1}^{K} b_k} = \frac{1}{\sum_{k=1}^{K} b_k} \cdot \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \frac{(a_k b_\ell - a_\ell b_k)^2}{b_k b_\ell},$$

which implies Bergström's inequality, given by

$$\sum_{k=1}^{K} \frac{a_k^2}{b_k} \geqslant \frac{\left(\sum_{k=1}^{K} a_k\right)^2}{\sum_{k=1}^{K} b_k}.$$

Proof. Lagrange's identity is that, for any $(c_1, \dots, c_K) \in \mathbb{R}^K$ and $(d_1, \dots, d_K) \in \mathbb{R}^K$,

$$\left(\sum_{k=1}^{K} c_k^2\right)\left(\sum_{k=1}^{K} d_k^2\right) - \left(\sum_{k=1}^{K} c_k d_k\right)^2 = \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} (c_k d_\ell - c_\ell d_k)^2. \tag{4}$$

Fix arbitrary $(a_1, \dots, a_K) \in \mathbb{R}^K$ and $(b_1, \dots, b_K) \in \mathbb{R}^K$ with $b_k > 0$ for all $k = 1, \dots, K$. Then, taking $c_k = a_k/\sqrt{b_k}$ and $d_k = \sqrt{b_k}$ for all $k = 1, \dots, K$ in (4), we can show that

$$\begin{aligned}
\left(\sum_{k=1}^{K} \frac{a_k^2}{b_k}\right)\left(\sum_{k=1}^{K} b_k^2\right) - \left(\sum_{k=1}^{K} a_k\right)^2 &= \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \left(\frac{a_k}{\sqrt{b_k}}\sqrt{b_\ell} - \frac{a_\ell}{\sqrt{b_\ell}}\sqrt{b_k}\right)^2 \\
&= \frac{1}{2} \sum_{k=1}^{K} \sum_{\substack{\ell=1 \\ \ell \neq k}}^{K} \frac{(a_k b_\ell - a_\ell b_k)^2}{b_k b_\ell},
\end{aligned}$$

which leads to the desired equality. Also, the last expression in the math display above is non-negative, which leads to Bergström's inequality. □

To prove Theorem 4.4, we introduce additional notation. Define the empirical probability measures of X as

$$\widehat{\mathbb{P}}_X := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \text{ and } \widehat{\mathbb{P}}_X^{(w)} := \frac{1}{n_w} \sum_{i=1}^{n} 1\!\!1_{\{W_i = w\}} \cdot \delta_{X_i},$$

for all observations and observations in the treatment group $w \in \mathcal{W}$, respectively. Here, δ_x is the measure that assigns mass 1 at $x \in \mathcal{X}$ and thus $\widehat{\mathbb{P}}_X$ and $\widehat{\mathbb{P}}_X^{(w)}$ can be interpreted as the random discrete probability measures, which put mass $1/n$ and $1/n_w$ at each of the n and n_w points $\{X_i\}_{i=1}^n$ and $\{X_i : W_i = w\}_{i=1}^n$, respectively. Given a real-valued function $f : \mathcal{X} \to \mathbb{R}$, we denote by

$$\widehat{\mathbb{P}}_X f = \int f \widehat{\mathbb{P}}_X = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \quad \text{and} \quad \widehat{\mathbb{P}}_X^{(w)} f = \int f \widehat{\mathbb{P}}_X^{(w)} = \frac{1}{n_w} \sum_{i=1}^{n} 1\!\!1_{\{W_i = w\}} \cdot f(X_i).$$

Given that the true conditional distribution $\gamma_y^{(w)}(X) \equiv F_{Y(w)|X}(y|X)$, the infeasible version of regression-adjusted distribution function for treatment $w \in \mathcal{W}$ is written as

$$\widetilde{\mathbb{F}}_{Y(w)}(y) = \widehat{\mathbb{F}}_{Y(w)}^{simple}(y) - (\widehat{\mathbb{P}}_X^{(w)} - \widehat{\mathbb{P}}_X)\gamma_y^{(w)}.$$

Proof of Theorem 4.4. **Part (a)** Choose any arbitrary $w \in \mathcal{W}$ and $y \in \mathcal{Y}$. Applying the quadratic expansion for $\widetilde{\mathbb{F}}_{Y(w)}(y) = \widehat{\mathbb{F}}^{simple}_{Y(w)}(y) - (\widehat{\mathbb{P}}^{(w)}_X - \widehat{\mathbb{P}}_X)\gamma^{(w)}_y$, we can show that

$$\mathrm{Var}\big(\widetilde{\mathbb{F}}_{Y(w)}(y)\big) = \mathrm{Var}\big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y)\big) - 2\mathrm{Cov}\Big(\widehat{\mathbb{F}}^{simple}_{Y(1)}(y), (\widehat{\mathbb{P}}^{(w)}_X - \widehat{\mathbb{P}}_X)\gamma^{(w)}_y\Big) + \mathrm{Var}\Big((\widehat{\mathbb{P}}^{(w)}_X - \widehat{\mathbb{P}}_X)\gamma^{(w)}_y\Big). \tag{5}$$

We can write $\widehat{\mathbb{P}}_X = \sum_{w' \in \mathcal{W}} \widehat{\pi}_{w'} \widehat{\mathbb{P}}^{(w')}_X$. It is assumed that observations are a random sample and $n_{w'}/n = \pi_{w'} + o(1)$ for every $w' \in \mathcal{W}$ as $n \to \infty$. Furthermore, all unconditional and conditional functions are bounded. By applying the dominated convergence theorem, we can show

$$\begin{aligned} n\mathrm{Cov}\Big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y), (\widehat{\mathbb{P}}^{(w)}_X - \widehat{\mathbb{P}}_X)\gamma^{(w)}_y\Big) &= \mathrm{Cov}\Big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y), (1-\widehat{\pi}_w)\widehat{\mathbb{P}}^{(w)}_X \gamma^{(w)}_y(X)\Big) \\ &= \frac{1-\pi_w}{\pi_w}\mathrm{Cov}\big(1\!\!1_{\{Y(w)\leqslant y\}}, \gamma^{(w)}_y(X)\big) + o(1). \end{aligned} \tag{6}$$

Similarly, we can show that

$$\begin{aligned} n\mathrm{Var}\big((\widehat{\mathbb{P}}^{(w)}_X - \widehat{\mathbb{P}}_X)\gamma^{(w)}_y\big) &= n\mathrm{Var}\big((1-\widehat{\pi}_w)\widehat{\mathbb{P}}^{(w)}_X\gamma^{(w)}_y\big) + n\sum_{w':w'\neq w}\mathrm{Var}\big(\widehat{\pi}_{w'}\widehat{\mathbb{P}}^{(w')}_X\gamma^{(w)}_y\big) \\ &= \frac{(1-\pi_w)^2}{\pi_w}\mathrm{Var}\big(\gamma^{(w)}_y(X)\big) + \sum_{w':w'\neq w}\frac{\pi^2_{w'}}{\pi_{w'}}\mathrm{Var}\big(\gamma^{(w)}_y(X)\big) + o(1) \\ &= \left(\frac{(1-\pi_w)^2}{\pi_w} + \sum_{w':w'\neq w}\pi_{w'}\right)\mathrm{Var}\big(\gamma^{(w)}_y(X)\big) + o(1) \\ &= \frac{1-\pi_w}{\pi_w}\mathrm{Var}\big(\gamma^{(w)}_y(X)\big) + o(1). \end{aligned} \tag{7}$$

It follows from (5)-(7) that

$$n\big\{\mathrm{Var}\big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y)\big) - \mathrm{Var}\big(\widetilde{\mathbb{F}}_{Y(w)}(y)\big)\big\} = \frac{1-\pi_w}{\pi_w}\big\{2\mathrm{Cov}\big(1\!\!1_{\{Y(w)\leqslant y\}}, \gamma^{(w)}_y(X)\big) - \mathrm{Var}\big(\gamma^{(w)}_y(X)\big)\big\} + o(1). \tag{8}$$

An application of the law of iterated expectation yields

$$\mathrm{Cov}\big(1\!\!1_{\{Y(w)\leqslant y\}}, \gamma^{(w)}_y(X)\big) = \mathrm{Var}\big(E[1\!\!1_{\{Y(w)\leqslant y\}}|X]\big),$$

which together with (8) shows

$$n\big\{\mathrm{Var}\big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y)\big) - \mathrm{Var}\big(\widetilde{\mathbb{F}}_{Y(w)}(y)\big)\big\} = \frac{1-\pi_w}{\pi_w}\mathrm{Var}\big(\gamma^{(w)}_y(X)\big) + o(1).$$

Since $\pi_w \in (0,1)$ and $\mathrm{Var}\big(\gamma^{(w)}_y(X)\big) \geqslant 0$, it follows that $\mathrm{Var}\big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y)\big) \geqslant \mathrm{Var}\big(\widetilde{\mathbb{F}}_{Y(w)}(y)\big) + o(n^{-1})$. Here, the equality hold only when $F_{Y(w)|X}(y) = F_{Y(w)}(y)$ or X has no predictive power for the event $1\!\!1_{\{Y(w)\leqslant y\}}$.

Part (b) Choose any arbitrary $y \in \mathcal{Y}$. First, we shall show that, for any $w, w' \in \mathcal{W}$,

$$n\mathrm{Cov}\big(\widehat{\mathbb{F}}_{Y(w)}(y), \widetilde{\mathbb{F}}_{Y(w')}(y)\big) = \mathrm{Cov}\big(\gamma^{(w)}_y(X), \gamma^{(w')}_y(X)\big). \tag{9}$$

Fix any two distinct treatment statuses $w, w' \in \mathcal{W}$. We can write $\widetilde{\mathbb{F}}_{Y(w)}(y) = \big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y) - \widehat{\mathbb{P}}^{(w)}_X\gamma^{(w)}_y\big) + \widehat{\mathbb{P}}_X\gamma^{(w)}_y$ and also $\widehat{\mathbb{P}}_X\gamma^{(w)}_y = \sum_{v\in\mathcal{W}}\widehat{\pi}_v\widehat{\mathbb{P}}^{(v)}_X\gamma^{(w)}_y$. Given random sample and the bi-linear property of the covariance function, we can show that

$$\begin{aligned} \mathrm{Cov}\Big(\widetilde{\mathbb{F}}_{Y(w)}(y), \widetilde{\mathbb{F}}_{Y(w')}(y)\Big) =& \mathrm{Cov}\Big(\widehat{\mathbb{F}}^{simple}_{Y(w)}(y) - \widehat{\mathbb{P}}^{(w)}_X\gamma^{(w)}_y, \widehat{\pi}_{w'}\widehat{\mathbb{P}}^{(w')}_X\gamma^{(w')}_y\Big) \\ &+ \mathrm{Cov}\big(\widehat{\pi}_w\widehat{\mathbb{P}}^{(w)}_X\gamma^{(w)}_y, \widehat{\mathbb{F}}^{simple}_{Y(w')} - \widehat{\mathbb{P}}^{(w')}_X\gamma^{(w')}_y\big) \\ &+ \mathrm{Cov}\big(\widehat{\mathbb{P}}_X\gamma^{(w)}_y, \widehat{\mathbb{P}}_X\gamma^{(w')}_y\big), \end{aligned}$$

where it can be shown that the first and second terms on the right-hand side are equal zero, due to the fact that $E\big[\widehat{\mathbb{F}}^{simple}_{Y(w)}(y) - \widehat{\mathbb{P}}^{(w)}_X \gamma^{(w)}_y | X_1, \ldots, X_n\big] = 0$. Furthermore, under the random sample assumption, we can show that $\mathrm{Cov}\big(\widehat{\mathbb{P}}_X \gamma^{(w)}_y, \widehat{\mathbb{P}}_X \gamma^{(w')}_y\big) = n^{-1}\mathrm{Cov}\big(\gamma^{(w)}_y(X), \gamma^{(w')}_y(X)\big)$. Thus, we can prove the equality in (9).

Next, we compare the variance-covariance matrices of the simple and regression-adjusted estimators. By applying the result from part (a) of this theorem and the one in (9), we are able to show that

$$
\begin{aligned}
&n\big\{\mathrm{Var}\big(\widehat{\theta}^{simple}_y\big) - \mathrm{Var}\big(\widetilde{\theta}_y\big)\big\} \\
&= \begin{bmatrix} \frac{1-\pi_1}{\pi_1}\mathrm{Var}\big(\gamma^{(1)}_y(X)\big), & -\mathrm{Cov}\big(\gamma^{(1)}_y(X), \gamma^{(2)}_y(X)\big), & \ldots, & -\mathrm{Cov}\big(\gamma^{(1)}_y(X), \gamma^{(K)}_y(X)\big) \\ -\mathrm{Cov}\big(\gamma^{(2)}_y(X), \gamma^{(1)}_y(X)\big), & \frac{1-\pi_2}{\pi_2}\mathrm{Var}\big(\gamma^{(2)}_y(X)\big), & \ldots, & -\mathrm{Cov}\big(F_{Y(2)|X}, F_{Y(K)|X}\big) \\ \vdots & \vdots & \ddots & \vdots \\ -\mathrm{Cov}\big(\gamma^{(K)}_y(X), \gamma^{(1)}_y(X)\big), & -\mathrm{Cov}\big(\gamma^{(K)}_y(X), \gamma^{(2)}_y(X)\big), & \ldots, & \frac{1-\pi_K}{\pi_K}\mathrm{Var}\big(\gamma^{(K)}_y(X)\big) \end{bmatrix} + o(1),
\end{aligned}
$$

which can be written as

$$
n\big\{\mathrm{Var}\big(\widehat{\theta}^{simple}_y\big) - \mathrm{Var}\big(\widetilde{\theta}_y\big)\big\} = E\Big[\big(\gamma_y(X) - E[\gamma_y(X)]\big)A\big(\gamma_y(X) - E[\gamma_y(X)]\big)^\top\Big] + o(1),
$$

where $\gamma_y(X) = [\gamma^{(1)}_y(X), \ldots, \gamma^{(K)}_y(X)]^\top$ and

$$
A := \begin{bmatrix} \pi_1^{-1} - 1, & -1, & \ldots, & -1 \\ -1, & \pi_2^{-1} - 1, & \ldots, & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1, & -1, & \ldots, & \pi_K^{-1} - 1 \end{bmatrix}.
$$

The variant of Lagrange's identity in Lemma B.1 with $\sum_{w\in\mathcal{W}} \pi_w = 1$ shows that, for an arbitrary vector $v := (v_1, \ldots, v_K)^\top \in \mathbb{R}^k$,

$$
\begin{aligned}
&v^\top\big(\gamma_y(X) - E[\gamma_y(X)]\big)A\big(\gamma_y(X) - E[\gamma_y(X)]\big)^\top v \\
&= \sum_{w\in\mathcal{W}} \frac{v_w^2\big(\gamma^{(w)}_y(X) - E[\gamma^{(w)}_y(X)]\big)^2}{\pi_w} - \Big(\sum_{w\in\mathcal{W}} v_w\big(\gamma^{(w)}_y(X) - E[\gamma^{(w)}_y(X)]\big)\Big)^2 \\
&= \frac{1}{2}\sum_{w\in\mathcal{W}}\sum_{\substack{w'\in\mathcal{W} \\ w'\neq w}} \frac{\big\{v_w\big(\gamma^{(w)}_y(X) - E[\gamma^{(w)}_y(X)]\big)\pi_{w'} - v_{w'}\big(\gamma^{(w')}_y(X) - E[\gamma^{(w')}_y(X)]\big)\pi_w\big\}^2}{\pi_w\pi_{w'}}.
\end{aligned}
$$

It follows that

$$
v^\top\big\{\mathrm{Var}\big(\widehat{\theta}^{simple}_y\big) - \mathrm{Var}\big(\widetilde{\theta}_y\big)\big\}v = \frac{1}{2}\sum_{w\in\mathcal{W}}\sum_{\substack{w'\in\mathcal{W} \\ w'\neq w}} \frac{\mathrm{Var}\Big(v_w\gamma^{(w)}_y(X)\pi_{w'} - v_{w'}\gamma^{(w')}_y(X)\pi_w\Big)}{\pi_w\pi_{w'}} + o(n^{-1}).
$$

The above equality implies the desired positive semi-definiteness result, because $\mathrm{Var}\big(v_w\gamma^{(w)}_y(X)\pi_{w'} - v_{w'}\gamma^{(w')}_y(X)\pi_w\big) \geqslant 0$ for any $w, w' \in \mathcal{W}$ with $w \neq w'$.

Furthermore, the positive definite result holds when $\mathrm{Var}\big(v_w\gamma^{(w)}_y(X)\pi_{w'} - v_{w'}\gamma^{(w')}_y(X)\pi_w\big) > 0$ for any $v \in \mathbb{R}^k$ with $v \neq 0$ and for any $w, w' \in \mathcal{W}$ with $w \neq w'$. Because $v \in \mathbb{R}^k$ is chosen arbitrarily except $v \neq 0$ and $\pi_w \in (0, 1)$ for all $w \in \mathcal{W}$, the condition for the positive definiteness can be written as $\mathrm{Var}\big(\gamma^{(w)}_y(X) - r \cdot \gamma^{(w')}_y(X)\big) > 0$ for any $r \in \mathbb{R}$ and for any $w, w' \in \mathcal{W}$ with $w \neq w'$. □

C. Multiplier Bootstrap Procedure

We can obtain pointwise and uniform confidence bands for distributional parameters using multiplier bootstrap following, for example, Chernozhukov et al. (2013) and Belloni et al. (2017). We outline the procedure to obtain uniform confidence bands in Algorithm 2. The algorithm can be altered slightly to generate pointwise confidence bands as explained in Remark C.1.

Algorithm 2 Multiplier bootstrap procedure to obtain uniform confidence bands

Input: Data $\{(X_i, W_i, Y_i)\}_{i=1}^n$; point estimates $\widehat{\theta}_y$; influence functions $\widehat{\psi}_y(Z_i) := \psi_y(Z_i; \widehat{\theta}_y, \widehat{\gamma}_y)$

(1) Draw multipliers $\{\xi_i\}_{i=1}^n = \{m_{1,i}/\sqrt{2} + ((m_{2,i})^2 - 1)/2\}_{i=1}^n$ independently from the data $\{Z_i\}_{i=1}^n$, where $m_{1,i}$ and $m_{2,i}$ are i.i.d. draws from two independent standard normal random variables.

(2) For each $y \in \mathcal{Y}$, obtain the bootstrap draws $\phi^b(\widehat{\theta}_y)$ of $\phi(\widehat{\theta}_y)$ as

$$\phi^b(\widehat{\theta}_y) = \phi(\widehat{\theta}_y^b) \text{ where } \widehat{\theta}_y^b = \widehat{\theta}_y + \frac{1}{n}\sum_{i=1}^n \xi_i \widehat{\psi}_y(Z_i).$$

(3) Repeat (1)-(2) B times and index the bootstrap draws by $b = 1, \ldots, B$.

(4) Obtain bootstrap standard error estimates for $\phi(\widehat{\theta}_y)$ for each $y \in \mathcal{Y}$ as

$$\widehat{\Sigma}(y) = \frac{q_{0.75}(y) - q_{0.25}(y)}{z_{0.75} - z_{0.25}},$$

where $q_p(y)$ is the p-th quantile of $\{\phi^b(\widehat{\theta}_y) : 1 \leqslant b \leqslant B\}$ and z_p is the p-th quantile of the standard normal distribution.

(5) Construct the bootstrap draw of the Kolmogorov-Smirnov maximal t-statistic as

$$t_{\max}^b = \max_{y \in \mathcal{Y}} \frac{|\phi^b(\widehat{\theta}_y) - \phi(\widehat{\theta}_y)|}{\widehat{\Sigma}(y)},$$

(6) Obtain the bootstrap estimators of the critical values as

$$\widehat{t}_{1-\alpha} = (1-\alpha) - \text{quantile of } \{t_{\max}^b : 1 \leqslant b \leqslant B\}.$$

(7) Construct $(1-\alpha) \times 100\%$ uniform confidence band for $(\phi(\theta_y))_{y \in \mathcal{Y}}$ as

$$I^{1-\alpha} = \{[\phi(\widehat{\theta}_y) \pm \widehat{t}_{1-\alpha} \times \widehat{\Sigma}(y)] : y \in \mathcal{Y}\}.$$

Result: Uniform confidence band $I^{1-\alpha}$ for $(\phi(\theta_y))_{y \in \mathcal{Y}}$

Remark C.1. To obtain pointwise confidence intervals using bootstrap, we skip Steps (5)-(6) and use $z_{1-\alpha/2}$ as a critical value instead of $\widehat{t}_{1-\alpha}$.

Remark C.2. Note that the multiplier bootstrap method is computationally efficient since it does not involve recomputing the nuisance estimates $\widehat{\gamma}_y$ and the influence functions $\widehat{\psi}_y$.

D. Key Tools and Proofs for Section 5

D.1. Additional Definitions: Empirical Processes

We first introduce some basic definitions related to empirical processes in this section. The following definitions are taken from van der Vaart & Wellner (1996).

Definition D.1 (Brownian bridge)**.** $\mathbb{G}_P$ is called a *P-Brownian bridge* on $\mathcal{F}$ if it is a mean-zero Gaussian process with covariance function $E[\mathbb{G}_P f \mathbb{G}_P g] = P(fg) - P(f)P(g)$.

Definition D.2 (Covering numbers and entropies)**.** The *covering number* $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimal number of balls $\{g : \|g - f\| < \epsilon\}$ of radius ϵ needed to cover the set $\mathcal{F}$. The centers of the balls need not belong to $\mathcal{F}$, but they should have finite norms. The *entropy* is the logarithm of the covering number.

Definition D.3 (Envelope function)**.** An *envelope function* of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \leqslant F(x)$ for every x and f.

D.2. Key tools

In this subsection, we introduce some key tools to derive our asymptotic results. Let $\{Z_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. copies of the random element Z taking values in the measure space $(\mathcal{Z}, \mathcal{A}_{\mathcal{Z}})$ according to the probability law P on that space. Let $\mathcal{F}_P = \{f_{t,P} : t \in T\}$ be a set of suitably measurable functions $z \longmapsto f_{t,P}(z)$ mapping $\mathcal{Z}$ to $\mathbb{R}$, equipped with a measurable envelope $F_P : \mathcal{Z} \longmapsto \mathbb{R}$. The class is indexed by $P \in \mathcal{P}$ and $t \in T$, where T is a fixed, totally bounded semi-metric space equipped with a semi-metric d_T. Let $N(\epsilon, \mathcal{F}_P, \|\cdot\|_{Q,2})$ denote the ϵ-covering number of the class of functions $\mathcal{F}_P$ with respect to the $L^2(Q)$ seminorm $\|\cdot\|_{Q,2}$ for Q a finitely-discrete measure on $(\mathcal{Z}, \mathcal{A}_{\mathcal{Z}})$.

Theorem D.4 (**Uniform in P Donsker Property**)**.** *Suppose that for $q > 2$*

$$\sup_{P \in \mathcal{P}} \|F_P\|_{P,q} \leqslant C \text{ and } \lim_{\delta \searrow 0} \sup_{P \in \mathcal{P}} \sup_{d_T(t,\bar{t}) \leqslant \delta} \|f_{t,P} - f_{\bar{t},P}\|_{P,2} = 0. \tag{10}$$

Furthermore, suppose that

$$\lim_{\delta \searrow 0} \sup_{P \in \mathcal{P}} \int_0^{\delta} \sup_{Q} \sqrt{\log N(\epsilon \|F_P\|_{Q,2}, \mathcal{F}_P, \|\cdot\|_{Q,2})} d\epsilon = 0. \tag{11}$$

Let $\mathbb{G}_P$ denote the P-Brownian Bridge, and consider

$$Z_{n,P} := (Z_{n,P}(t))_{t \in T} := (\mathbb{G}_n(f_{t,P}))_{t \in T}, \;\; Z_P := (Z_P(t))_{t \in T} := (\mathbb{G}_P(f_{t,P}))_{t \in T}.$$

(a) Then, $Z_{n,P} \rightsquigarrow Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely

$$\sup_{P \in \mathcal{P}} \sup_{h \in BL_1(\ell^\infty(T))} |\mathrm{E}_P^* h(Z_{n,P}) - \mathrm{E}_P h(Z_P)| \to 0.$$

(b) The process $Z_{n,P}$ is stochastically equicontinuous uniformly in $P \in \mathcal{P}$, i.e., for every $\varepsilon > 0$,

$$\lim_{\delta \searrow 0} \limsup_{n \to \infty} \sup_{P \in \mathcal{P}} \mathrm{P}_P^* \left(\sup_{d_T(t,\bar{t}) \leqslant \delta} |Z_{n,P}(t) - Z_{n,P}(\bar{t})| > \varepsilon \right) = 0.$$

(c) The limit process Z_P has the following continuity properties:

$$\sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{t \in T} |Z_P(t)| < \infty, \;\; \lim_{\delta \searrow 0} \sup_{P \in \mathcal{P}} \mathrm{E}_P \sup_{d_T(t,\bar{t}) \leqslant \delta} |Z_P(t) - Z_P(\bar{t})| = 0.$$

(d) The paths $t \longmapsto Z_P(t)$ are a.s. uniformly continuous on (T, d_T) under each $P \in \mathcal{P}$.

Proof. See the proof of Theorem B.1 in Belloni et al. (2017). □

Uniform in P Validity of Multiplier Bootstrap Let $(\xi_i)_{i=1}^n$ be i.i.d. multipliers whose distribution does not depend on P, such that $\mathrm{E}\xi = 0$, $\mathrm{E}\xi^2 = 1$, and $\mathrm{E}|\xi|^q \leqslant C$ for $q > 2$. Consider the multiplier empirical process:

$$Z_{n,P}^* := (Z_{n,P}^*(t))_{t \in T} := (\mathbb{G}_n(\xi f_{t,P}))_{t \in T} := \left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i f_{t,P}(Z_i) \right)_{t \in T}.$$

Here $\mathbb{G}_n$ is taken to be an extended empirical processes defined by the empirical measure that assigns mass $1/n$ to each point (Z_i, ξ_i) for $i = 1, ..., n$. Let $Z_P = (Z_P(t))_{t \in T} = (\mathbb{G}_P(f_{t,P}))_{t \in T}$ as defined in Theorem D.4.

Theorem D.5 (Uniform in P Validity of Multiplier Bootstrap). *Assume the conditions of Theorem D.4 hold.*

(a) Then, the following unconditional convergence takes place, $Z^*_{n,P} \rightsquigarrow Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely

$$\sup_{P\in\mathcal{P}} \sup_{h\in BL_1(\ell^\infty(T))} |\mathrm{E}^*_P h(Z^*_{n,P}) - \mathrm{E}_P h(Z_P)| \to 0.$$

(b) The following conditional convergence takes place, $Z^*_{n,P} \rightsquigarrow_B Z_P$ in $\ell^\infty(T)$ uniformly in $P \in \mathcal{P}$, namely uniformly in $P \in \mathcal{P}$

$$\sup_{h\in BL_1(\ell^\infty(T))} |\mathrm{E}_{B_n} h(Z^*_{n,P}) - \mathrm{E}_P h(Z_P)| = o^*_P(1),$$

where E_{B_n} denotes the expectation over the multiplier weights $(\xi_i)_{i=1}^n$ holding the data $(Z_i)_{i=1}^n$ fixed.

Proof. See Theorem B.2 of Belloni et al. (2017) for the proof. □

Definition D.6 (Uniform Hadamard Tangential Differentiability). Consider a map $\phi : \mathbb{D}_\phi \longmapsto \mathbb{E}$, where the domain of the map $\mathbb{D}_\phi$ is a subset of a normed space $\mathbb{D}$ and the range is a subset of the normed space $\mathbb{E}$. Let $\mathbb{D}_0$ be a normed space, with $\mathbb{D}_0 \subset \mathbb{D}$, and $\mathbb{D}_\rho$ be a compact metric space, a subset of $\mathbb{D}_\phi$. The map $\phi : \mathbb{D}_\phi \longmapsto \mathbb{E}$ is called *Hadamard-differentiable uniformly* in $\rho \in \mathbb{D}_\rho$ tangentially to $\mathbb{D}_0$ with derivative map $h \longmapsto \phi'_\rho(h)$, if

$$\left|\frac{\phi(\rho_n + t_n h_n) - \phi(\rho_n)}{t_n} - \phi'_\rho(h)\right| \to 0, \quad \left|\phi'_{\rho_n}(h_n) - \phi'_\rho(h)\right| \to 0, \quad n \to \infty,$$

for all convergent sequences $\rho_n \to \rho$ in $\mathbb{D}_\rho$, $t_n \to 0$ in $\mathbb{R}$, and $h_n \to h \in \mathbb{D}_0$ in $\mathbb{D}$ such that $\rho_n + t_n h_n \in \mathbb{D}_\phi$ for every n. As a part of the definition, we require that the derivative map $h \longmapsto \phi'_\rho(h)$ from $\mathbb{D}_0$ to $\mathbb{E}$ is linear for each $\rho \in \mathbb{D}_\rho$.

Theorem D.7 (Functional delta-method uniformly in $P \in \mathcal{P}$). *Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \longmapsto \mathbb{E}$ be Hadamard-differentiable uniformly in $\rho \in \mathbb{D}_\rho \subset \mathbb{D}_\phi$ tangentially to $\mathbb{D}_0$ with derivative map ϕ'_ρ. Let $\widehat{\rho}_{n,P}$ be a sequence of stochastic processes taking values in $\mathbb{D}_\phi$, where each $\widehat{\rho}_{n,P}$ is an estimator of the parameter $\rho_P \in \mathbb{D}_\rho$. Suppose there exists a sequence of constants $r_n \to \infty$ such that $Z_{n,P} = r_n(\widehat{\rho}_{n,P} - \rho_P) \rightsquigarrow Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. The limit process Z_P is separable and takes its values in $\mathbb{D}_0$ for all $P \in \mathcal{P} = \cup_{n\geqslant n_0}\mathcal{P}_n$, where n_0 is fixed. Moreover, the set of stochastic processes $\{Z_P : P \in \mathcal{P}\}$ is relatively compact in the topology of weak convergence in $\mathbb{D}_0$, that is, every sequence in this set can be split into weakly convergent subsequences. Then, $r_n\left(\phi(\widehat{\rho}_{n,P}) - \phi(\rho_P)\right) \rightsquigarrow \phi'_{\rho_P}(Z_P)$ in $\mathbb{E}$ uniformly in $P \in \mathcal{P}_n$. If $(\rho, h) \longmapsto \phi'_\rho(h)$ is defined and continuous on the whole of $\mathbb{D}_\rho \times \mathbb{D}$, then the sequence $r_n\left(\phi(\widehat{\rho}_{n,P}) - \phi(\rho_P)\right) - \phi'_{\rho_P}\left(r_n(\widehat{\rho}_{n,P} - \rho_P)\right)$ converges to zero in outer probability uniformly in $P \in \mathcal{P}_n$. Moreover, the set of stochastic processes $\{\phi'_{\rho_P}(Z_P) : P \in \mathcal{P}\}$ is relatively compact in the topology of weak convergence in $\mathbb{E}$.*

Proof. See Theorem B.3 of Belloni et al. (2017) for the proof. □

Let $D_{n,P} = (W_{i,P})_{i=1}^n$ denote the data vector and $B_n = (\xi_i)_{i=1}^n$ be a vector of random variables used to generate bootstrap. Consider sequences of stochastic processes $\widehat{\rho}_{n,P} = \widehat{\rho}_{n,P}(D_{n,P})$, where $Z_{n,P} = r_n(\widehat{\rho}_{n,P} - \rho_P) \rightsquigarrow Z_P$ in the normed space $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. Also consider the bootstrap stochastic process $Z^*_{n,P} = Z_{n,P}(D_{n,P}, B_n)$ in $\mathbb{D}$, where $Z_{n,P}$ is a measurable function of B_n for each value of D_n. Suppose that $Z^*_{n,P}$ converges conditionally given D_n in distribution to Z_P uniformly in $P \in \mathcal{P}_n$, namely that

$$\sup_{h\in BL_1(\mathbb{D})} |\mathrm{E}_{B_n}[h(Z^*_{n,P})] - \mathrm{E}_P h(Z_P)| = o^*_P(1),$$

uniformly in $P \in \mathcal{P}_n$, where E_{B_n} denotes the expectation computed with respect to the law of B_n holding the data $D_{n,P}$ fixed and $BL_1(\mathbb{D})$ denotes the space of functions mapping $\mathbb{D}$ to $[0, 1]$ with Lipschitz norm at most 1. This is denoted as $Z^*_{n,P} \rightsquigarrow_B Z_P$ uniformly in $P \in \mathcal{P}_n$. Finally, let $\widehat{\rho}^*_{n,P} = \widehat{\rho}_{n,P} + Z^*_{n,P}/r_n$ denote the bootstrap or simulation draw of $\widehat{\rho}_{n,P}$.

Theorem D.8 (Uniform in P functional delta-method for bootstrap and other simulation methods). *Assume the conditions of Theorem D.7 hold. Let $\widehat{\rho}_{n,P}$ and $\widehat{\rho}^*_{n,P}$ be maps as indicated previously taking values in $\mathbb{D}_\phi$ such that $r_n(\widehat{\rho}_{n,P} - \rho_P) \rightsquigarrow Z_P$ and $r_n(\widehat{\rho}^*_{n,P} - \widehat{\rho}_{n,P}) \rightsquigarrow_B Z_P$ in $\mathbb{D}$ uniformly in $P \in \mathcal{P}_n$. Then, $X^*_{n,P} = r_n(\phi(\widehat{\rho}^*_{n,P}) - \phi(\widehat{\rho}_{n,P})) \rightsquigarrow_B X_P = \phi'_{\rho_P}(Z_P)$ uniformly in $P \in \mathcal{P}_n$.*

Proof. See Theorem B.4 of Belloni et al. (2017) for the proof. □

Lemma D.9 (**Maximal Inequality (Chernozhukov et al., 2014)**). *Suppose that $F \geqslant \sup_{f\in\mathcal{F}}|f|$ is a measurable envelope with $\|F\|_{P,q} < \infty$ for some $q \geqslant 2$. Let $M = \max_{i\leqslant n} F(Z_i)$ and $\sigma^2 > 0$ be any positive constant such that $\sup_{f\in\mathcal{F}}\|f\|_{P,2}^2 \leqslant \sigma^2 \leqslant \|F\|_{P,2}^2$. Suppose that there exist constants $a \geqslant e$ and $v \geqslant 1$ such that $\log\sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, \|\cdot\|_{Q,2}) \leqslant v(\log a + \log(1/\epsilon)),\ 0 < \epsilon \leqslant 1$. Then*

$$\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] \leqslant C\left(\sqrt{v\sigma^2\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{\mathrm{P}_P,2}}{\sqrt{n}}\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)\right),$$

where C is an absolute constant. Moreover, for every $t \geqslant 1$, with probability $> 1 - t^{-q/2}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant (1+\alpha)\mathrm{E}_P[\|\mathbb{G}_n\|_{\mathcal{F}}] + C(q)\Big[(\sigma + n^{-1/2}\|M\|_{\mathrm{P}_P,q})\sqrt{t} + \alpha^{-1}n^{-1/2}\|M\|_{\mathrm{P}_P,2}t\Big],\quad \forall\alpha > 0,$$

where $C(q) > 0$ is a constant depending only on q. In particular, setting $a \geqslant n$ and $t = \log n$, with probability $> 1 - c(\log n)^{-1}$,

$$\|\mathbb{G}_n\|_{\mathcal{F}} \leqslant C(q,c)\left(\sigma\sqrt{v\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)} + \frac{v\|M\|_{\mathrm{P}_P,q}}{\sqrt{n}}\log\left(\frac{a\|F\|_{P,2}}{\sigma}\right)\right), \tag{12}$$

where $\|M\|_{\mathrm{P}_P,q} \leqslant n^{1/q}\|F\|_{P,q}$ and $C(q,c) > 0$ is a constant depending only on q and c.

Proof. See Chernozhukov et al. (2014) for the proof. □

Lemma D.10 (**Algebra for Covering Entropies**). *(1) Let $\mathbb{F}$ be a VC subgraph class with a finite VC index k or any other class whose entropy is bounded above by that of such a VC subgraph class, then the covering entropy of $\mathcal{F}$ obeys:*

$$\sup_Q \log N(\epsilon\|F\|_{Q,2}, \mathbb{F}, \|\cdot\|_{Q,2}) \lesssim 1 + k\log(1/\epsilon) \vee 0$$

(2) For any measurable classes of functions $\mathbb{F}$ and $\mathbb{F}'$ mapping $\mathcal{Z}$ to $\mathbb{R}$,

$$\begin{aligned}
&\log N(\epsilon\|F+F'\|_{Q,2}, \mathbb{F}+\mathbb{F}', \|\cdot\|_{Q,2}) \leqslant \log N\left(\tfrac{\epsilon}{2}\|F\|_{Q,2}, \mathbb{F}, \|\cdot\|_{Q,2}\right) + \log N\left(\tfrac{\epsilon}{2}\|F'\|_{Q,2}, \mathbb{F}', \|\cdot\|_{Q,2}\right),\\
&\log N(\epsilon\|F\cdot F'\|_{Q,2}, \mathbb{F}\cdot\mathbb{F}', \|\cdot\|_{Q,2}) \leqslant \log N\left(\tfrac{\epsilon}{2}\|F\|_{Q,2}, \mathbb{F}, \|\cdot\|_{Q,2}\right) + \log N\left(\tfrac{\epsilon}{2}\|F'\|_{Q,2}, \mathbb{F}', \|\cdot\|_{Q,2}\right),\\
&N(\epsilon\|F\vee F'\|_{Q,2}, \mathbb{F}\cup\mathbb{F}', \|\cdot\|_{Q,2}) \leqslant N\left(\epsilon\|F\|_{Q,2}, \mathbb{F}, \|\cdot\|_{Q,2}\right) + N\left(\epsilon\|F'\|_{Q,2}, \mathbb{F}', \|\cdot\|_{Q,2}\right).
\end{aligned}$$

(3) Given a measurable class $\mathcal{F}$ mapping $\mathcal{Z}$ to $\mathbb{R}$ and a random variable ξ taking values in $\mathbb{R}$,

$$\log\sup_Q N(\epsilon\||\xi|F\|_{Q,2}, \xi\mathbb{F}, \|\cdot\|_{Q,2}) \leqslant \log\sup_Q N\left(\epsilon/2\|F\|_{Q,2}, \mathbb{F}, \|\cdot\|_{Q,2}\right)$$

(4) Given measurable classes $\mathbb{F}_j$ and envelopes F_j, $j = 1,\dots,k$, mapping $\mathcal{Z}$ to $\mathbb{R}$, a function $\phi : \mathbb{R}^k \to \mathbb{R}$ such that for $f_j, g_j \in \mathbb{F}_j$, $|\phi(f_1,\dots,f_k) - \phi(g_1,\dots,g_k)| \leqslant \sum_{j=1}^k L_j(x)|f_j(x) - g_j(x)|$, $L_j(x) \geqslant 0$, and fixed functions $\bar{f}_j \in \mathbb{F}_j$, the class of functions $\mathcal{L} = \{\phi(f_1,\dots,f_k) - \phi(\bar{f}_1,\dots,\bar{f}_k) : f_j \in \mathcal{F}_j, j = 1,\dots,k\}$ satisfies

$$\log\sup_Q N(\epsilon\|\sum_{j=1}^k L_jF_j\|_{Q,2}, \mathcal{L}, \|\cdot\|_{Q,2}) \leqslant \sum_{j=1}^k \log\sup_Q N\left(\tfrac{\epsilon}{k}\|F_j\|_{Q,2}, \mathbb{F}_j, \|\cdot\|_{Q,2}\right).$$

Proof. See Andrews (1994b) for the proofs of (1) and (2). (3) follows from (2). See Lemma K.1 of Belloni et al. (2017) for the proof of (4). □

D.3. Regularity conditions for Theorem 5.1

In this subsection, we lay out the regularity conditions for Theorem 5.1. In what follows, let δ, c_0, c, and C denote some positive constants. Let $\Delta_n \searrow 0$, $\delta_n \searrow 0$, and $\tau_n \searrow 0$ be sequences of constants approaching zero from above at a speed at most polynomial in n.

Assumption D.11 (Moment condition problem)**.** Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold: (i) The true parameter value θ_y obeys (3) and is interior relative to $\Theta_y \subset \Theta \subset \mathbb{R}^K$, namely there is a ball of radius δ centered at θ_y contained in Θ_y for all $y \in \mathcal{Y}$, and Θ is compact.

(ii) For $\nu := (\nu_k)_{k=1}^{2K} = (\theta, t)$, each $w \in \mathcal{W}$ and $y \in \mathcal{Y}$, the map $\Theta_y \times \Gamma_y \ni \nu \longmapsto \mathrm{E}_P[\psi_y^{(w)}(Z;\nu)]$ is twice continuously differentiable a.s. with derivatives obeying the integrability conditions specified in Assumption D.12.

(iii) The following identifiability condition holds: $\|\mathrm{E}_P[\psi_y(Z,\theta,\gamma_y)]\| \geqslant 2^{-1}(\|(\theta-\theta_y)\| \wedge c_0)$ for all $\theta \in \Theta_y$.

Assumption D.12 (Entropy and smoothness)**.** The set $(\mathcal{Y}, d_{\mathcal{Y}})$ is a semi-metric space such that $\log N(\epsilon, \mathcal{Y}, d_{\mathcal{Y}}) \leqslant C\log(\mathrm{e}/\epsilon) \vee 0$. Let $\alpha \in [1,2]$, and let α_1 and α_2 be some positive constants. Uniformly for all $n \geqslant n_0$ and $P \in \mathcal{P}_n$, the following conditions hold:

(i) The set of functions $\mathcal{F}_0 = \{\psi_y^{(w)}(Z;\theta_y,\gamma_y) : w \in \mathcal{W}, y \in \mathcal{Y}\}$, viewed as functions of Z is suitably measurable; has an envelope function $F_0(Z) = \sup_{w\in\mathcal{W}, y\in\mathcal{Y}, \nu\in\theta_y\times\Gamma_y} |\psi_y^{(w)}(Z;\nu)|$ that is measurable with respect to Z and obeys $\|F_0\|_{P,q} \leqslant C$, where $q \geqslant 4$ is a fixed constant; and has a uniform covering entropy obeying $\sup_Q \log N(\epsilon\|F_0\|_{Q,2}, \mathcal{F}_0, \|\cdot\|_{Q,2}) \leqslant C\log(\mathrm{e}/\epsilon) \vee 0$.

(ii) For all $w \in \mathcal{W}$ and $k, r \in [2K]$, and $\psi_y^{(w)}(Z) := \psi_y^{(w)}(Z;\theta_y,\gamma_y)$,

(a) $\sup_{y\in\mathcal{Y},(\nu,\bar\nu)\in(\theta_y\times\Gamma_y)^2} \mathrm{E}_P[(\psi_y^{(w)}(Z;\nu) - \psi_y^{(w)}(Z;\bar\nu))^2]/\|\nu-\bar\nu\|^\alpha \leqslant C$, P-a.s.,

(b) $\sup_{d_{\mathcal{Y}}(y,\bar y)\leqslant\delta} \mathrm{E}_P[(\psi_y^{(w)}(Z) - \psi_{\bar y}^{(w)}(Z))^2] \leqslant C\delta^{\alpha_1}$,

(c) $\mathrm{E}_P \sup_{y\in\mathcal{Y},\nu\in\Theta_y\times\Gamma_y} |\partial_{\nu_r}\mathrm{E}_P\left[\psi_y^{(w)}(Z;\nu)\right]|^2 \leqslant C$,

(d) $\sup_{y\in\mathcal{Y},\nu\in\Theta_y\times\Gamma_y} |\partial_{\nu_k}\partial_{\nu_r}\mathrm{E}_P[\psi_y^{(w)}(Z;\nu)]| \leqslant C$, P-a.s.

Assumption D.13 (Estimation of nuisance functions)**.** The following conditions hold for each $n \geqslant n_0$ and all $P \in \mathcal{P}_n$. The estimated functions $\widehat\gamma_y = (\widehat\gamma_y^{(w)})_{w=1}^K \in \mathcal{G}_{yn}$ with probability at least $1-\Delta_n$, where $\mathcal{G}_{yn}$ is the set of measurable maps $x \longmapsto \gamma = (\gamma^{(w)})_{w=1}^K(x) \in \Gamma_y(x)$ such that

$$\|\gamma^{(w)} - \gamma_y^{(w)}\|_{P,2} \leqslant \tau_n, \quad \tau_n^2\sqrt{n} \leqslant \delta_n,$$

and whose complexity does not grow too quickly in the sense that $\mathcal{F}_1 = \{\psi_y^{(w)}(Z;\theta,\gamma) : w \in \mathcal{W}, y \in \mathcal{Y}, \theta \in \Theta_y, \gamma \in \mathcal{G}_{yn}\}$ is suitably measurable and its uniform covering entropy obeys

$$\sup_Q \log N(\epsilon\|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leqslant \log(e/\epsilon) \vee 0,$$

where $F_1(Z)$ is an envelope for $\mathcal{F}_1$ which is measurable with respect to Z and satisfies $F_1(Z) \leqslant F_0(Z)$ for F_0 defined in Assumption D.12.

D.4. Proofs for Section 5

In this subsection, we state the proofs for Theorems 5.1 and 5.2.

Proof of Theorem 5.1. **Step 0.** In the proof $a \lesssim b$ means that $a \leqslant Ab$, where the constant A depends on the constants in Assumptions D.12, but not on n once $n \geqslant n_0$, and not on $P \in \mathcal{P}_n$. In Step 1, we consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $P = P_n$ throughout the proof, suppressing the index n. Since the argument is asymptotic, we can assume that $n \geqslant n_0$ in what follows.

Also, let

$$B(Z) \quad := \quad \max_{w\in\mathcal{W},k\in[2K]} \sup_{\nu\in\Theta_y\times\Gamma_y, y\in\mathcal{Y}} \left|\partial_{\nu_k}\mathrm{E}_P[\psi_y^{(w)}(Z;\nu)]\right|. \tag{13}$$

STEP 1. (A Preliminary Rate Result). In this step, we claim that with probability $1 - o(1)$,

$$\sup_{y \in \mathcal{Y}} \|\widehat{\theta}_y - \theta_y\| \lesssim \tau_n.$$

Since $\widehat{\theta}_y$ is defined as a solution to the sample moment condition, we have

$$\|\mathbb{E}_n \psi_y(Z, \widehat{\theta}_y, \widehat{\gamma}_y)\| \leqslant \inf_{\theta \in \Theta_y} \|\mathbb{E}_n \psi_y(Z, \theta, \widehat{\gamma}_y)\| + \epsilon_n \text{ for each } y \in \mathcal{Y},$$

where $\epsilon_n = o(n^{-1/2})$. This implies via triangle inequality that uniformly in $y \in \mathcal{Y}$ with probability $1 - o(1)$

$$\left\| P[\psi_y(Z; \widehat{\theta}_y, \gamma_y)] \right\| \leqslant \epsilon_n + 2I_1 + 2I_2 \lesssim \tau_n, \tag{14}$$

for I_1 and I_2 defined in Step 2 below. The $\lesssim$ bound in (14) follows from Step 2 and from the assumption $\epsilon_n = o(n^{-1/2})$. Since by Assumption D.11 (iii), $2^{-1}(\|(\widehat{\theta}_y - \theta_y)\| \wedge c_0)$ does not exceed the left side of (14), we conclude that $\sup_{y \in \mathcal{Y}} \|\widehat{\theta}_y - \theta_y\| \lesssim \tau_n$.

STEP 2. (Define and bound I_1 and I_2) We claim that with probability $1 - o(1)$:

$$\begin{aligned} I_1 &:= \sup_{\theta \in \Theta_y, y \in \mathcal{Y}} \left\| \mathbb{E}_n \psi_y(Z; \theta, \widehat{\gamma}_y) - \mathbb{E}_n \psi_y(Z; \theta, \gamma_y) \right\| \lesssim \tau_n, \\ I_2 &:= \sup_{\theta \in \Theta_y, y \in \mathcal{Y}} \left\| \mathbb{E}_n \psi_y(Z; \theta, \gamma_y) - P\psi_y(Z; \theta, \gamma_y) \right\| \lesssim \tau_n. \end{aligned}$$

To establish this, we can bound $I_1 \leqslant 2I_{1a} + I_{1b}$ and $I_2 \leqslant I_{1a}$, where with probability $1 - o(1)$,

$$\begin{aligned} I_{1a} &:= \sup_{\theta \in \Theta_y, y \in \mathcal{Y}, \gamma \in \mathcal{G}_{yn} \cup \{\gamma_y\}} \left\| \mathbb{E}_n \psi_y(Z; \theta, \gamma) - P\psi_y(Z; \theta, \gamma) \right\| \lesssim \tau_n, \\ I_{1b} &:= \sup_{\theta \in \Theta_y, y \in \mathcal{Y}, \gamma \in \mathcal{G}_{yn} \cup \{\gamma_y\}} \left\| P\psi_y(Z; \theta, \gamma) - P\psi_y(Z; \theta, \gamma_y) \right\| \lesssim \tau_n. \end{aligned}$$

These bounds in turn hold by the following arguments.

In order to bound I_{1b}, we employ Taylor's expansion and the triangle inequality. For $\bar{\gamma}(X, y, w, \theta)$ denoting a point on a line connecting vectors $\gamma(X)$ and $\gamma_y(X)$, and t_m denoting the mth element of the vector t,

$$\begin{aligned} I_{1b} &\leqslant \sum_{w=1}^{K} \sum_{m=1}^{K} \sup_{\theta \in \Theta_y, y \in \mathcal{Y}, \gamma \in \mathcal{G}_{yn}} \left| P\left[\partial_{t_m} P\left[\psi_y^{(w)}(Z, \theta, \bar{\gamma}(X, y, w, \theta)) \right] (\gamma^{(m)}(X) - \gamma_y^{(m)}(X)) \right] \right| \\ &\leqslant K \cdot K \cdot \|B\|_{P,2} \max_{y \in \mathcal{Y}, \gamma \in \mathcal{G}_{yn}, w \in \mathcal{W}} \|\gamma^{(w)} - \gamma_y^{(w)}\|_{P,2}, \end{aligned}$$

where the last inequality holds by the definition of $B(Z)$ given earlier and Hölder's inequality. By Assumption D.12(ii)(c), $\|B\|_{P,2} \leqslant C$, and by Assumption D.13, $\sup_{y \in \mathcal{Y}, \gamma \in \mathcal{G}_{yn}, w \in \mathcal{W}} \|\gamma^{(w)} - \gamma_y^{(w)}\|_{P,2} \lesssim \tau_n$, hence we conclude that $I_{1b} \lesssim \tau_n$ since K is fixed.

In order to bound I_{1a}, we employ the maximal inequality of Lemma D.9 to the class

$$\mathcal{F}_1 = \{\psi_y^{(w)}(Z, \theta, \gamma) : w \in \mathcal{W}, y \in \mathcal{Y}, \theta \in \Theta_y, \gamma \in \mathcal{G}_{yn} \cup \{\gamma_y\}\},$$

defined in Assumption D.13 and equipped with an envelope $F_1 \leqslant F_0$, to conclude that with probability $1 - o(1)$,

$$I_{1a} \lesssim \tau_n.$$

Here we use that $\log \sup_Q N(\epsilon \|F_1\|_{Q,2}, \mathcal{F}_1, \|\cdot\|_{Q,2}) \leqslant \log(e/\epsilon) \vee 0$ by Assumption D.13; $\|F_0\|_{P,q} \leqslant C$ and $\sup_{f \in \mathcal{F}_1} \|f\|_{P,2}^2 \leqslant \sigma^2 \leqslant \|F_0\|_{P,2}^2$ for $c \leqslant \sigma \leqslant C$ by Assumption D.12(i).

STEP 3. (Linearization) By definition, we have

$$\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\widehat{\theta}_y,\widehat{\gamma}_y)\| \leqslant \inf_{\theta\in\Theta_y}\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\theta,\widehat{\gamma}_y)\| + \sqrt{n}\cdot\epsilon_n.$$

By Taylor's theorem, for all $y \in \mathcal{Y}$,

$$\begin{aligned}\sqrt{n}\mathbb{E}_n\psi_y(Z;\widehat{\theta}_y,\widehat{\gamma}_y) &= \sqrt{n}\mathbb{E}_n\psi_y(Z;\theta_y,\gamma_y)\\ &- \sqrt{n}(\widehat{\theta}_y-\theta_y) + \mathrm{D}_{u,0}(\widehat{\gamma}_y-\gamma_y) + II_1(y) + II_2(y),\end{aligned}$$

where the terms $II_1(y)$ and $II_2(y)$ are defined in Step 4 and $\mathrm{D}_{y,0}(\widehat{\gamma}_y-\gamma_y)$ is treated in the next paragraph. Then, by the triangle inequality, for all $y \in \mathcal{Y}$ and Steps 4 and 5, we have

$$\begin{aligned}&\left\|\sqrt{n}\mathbb{E}_n\psi_y(Z;\theta_y,\gamma_y) - \sqrt{n}(\widehat{\theta}_y-\theta_y) + \mathrm{D}_{y,0}(\widehat{\gamma}_y-\gamma_y)\right\|\\ &\leqslant \epsilon_n\sqrt{n} + \sup_{y\in\mathcal{Y}}\left(\inf_{\theta\in\Theta_y}\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\theta,\widehat{\gamma}_y)\| + \|II_1(y)\| + \|II_2(y)\|\right) = o_P(1),\end{aligned}$$

where the $o_P(1)$ bound follows from Step 4, $\epsilon_n\sqrt{n} = o(1)$ by assumption, and Step 5.

Moreover, by the orthogonality condition:

$$\mathrm{D}_{y,0}(\widehat{\gamma}_y-\gamma_y) := \left(\sum_{m=1}^{K}\sqrt{n}P\left[\partial_{t_m}P[\psi_y^{(w)}(Z;\theta_y,\gamma_y)](\widehat{\gamma}^{(m)}(X)-\gamma_y^{(m)}(X))\right]\right)_{w=1}^{K} = 0.$$

Conclude using Assumption D.11 (iii) that

$$\sup_{y\in\mathcal{Y}}\left\|-\sqrt{n}\mathbb{E}_n\psi_y(Z;\theta_y,\gamma_y) + \sqrt{n}(\widehat{\theta}_y-\theta_y)\right\| \leqslant o_P(1).$$

Furthermore, the empirical process $(\sqrt{n}\mathbb{E}_n\psi_y(Z;\theta_y,\gamma_y))_{y\in\mathcal{Y}}$ is equivalent to an empirical process $\mathbb{G}_n$ indexed by $\mathcal{F}_P := \left\{\psi_y^{(w)} : w\in\mathcal{W}, y\in\mathcal{Y}\right\}$, where $\psi_y^{(w)}$ is the w-th element of $\psi_y(Z;\theta_y,\gamma_y)$ and we make explicit the dependence of $\mathcal{F}_P$ on P.

The conditions on $\mathcal{F}_0$ in Assumption D.12(ii) imply that $\mathcal{F}_P$ has a uniformly well-behaved uniform covering entropy by Lemma D.10, namely

$$\sup_{P\in\mathcal{P}=\cup_{n\geqslant n_0}\mathcal{P}_n}\log\sup_Q N(\epsilon\|CF_0\|_{Q,2},\mathcal{F}_P,\|\cdot\|_{Q,2}) \lesssim \log(e/\epsilon)\vee 0,$$

where $F_P = CF_0$ is an envelope for $\mathcal{F}_P$ since $\sup_{f\in\mathcal{F}_P}|f| \lesssim CF_0$ by Assumption D.12 (i). The class $\mathcal{F}_P$ is therefore Donsker uniformly in P because $\sup_{P\in\mathcal{P}}\|F_P\|_{P,q} \leqslant C\sup_{P\in\mathcal{P}}\|F_0\|_{P,q}$ is bounded by Assumption D.12 (ii), and $\sup_{P\in\mathcal{P}}\|\psi_y-\psi_{\bar{y}}\|_{P,2}\to 0$ as $d_{\mathcal{Y}}(y,\bar{y})\to 0$ by Assumption D.12 (ii) (b). Application of Theorem D.4 gives the results of the theorem.

STEP 4. (Define and Bound $II_1(y)$ and $II_2(y)$). Let $II_1(y) := (II_{1w}(y))_{w=1}^K$ and $II_2(y) = (II_{2w}(y))_{w=1}^K$, where

$$\begin{aligned}II_{1w}(y) &:= \sum_{r,k=1}^{2K}\sqrt{n}P\left[\partial_{\nu_k}\partial_{\nu_r}P[\psi_y^{(w)}(Z,\bar{\nu}_y(X,w))]\{\widehat{\nu}_{yr}(X)-\nu_{yr}(X)\}\{\widehat{\nu}_{yk}(X)-\nu_{yk}(X)\}\right],\\ II_{2w}(y) &:= \mathbb{G}_n(\psi_y^{(w)}(Z,\widehat{\theta}_y,\widehat{\gamma}_y) - \psi_y^{(w)}(Z,\theta_y,\gamma_y)),\end{aligned}$$

$\nu_y(X) := (\nu_{yk}(X))_{k=1}^{2K} := (\theta_y^\top,\gamma_y(X)^\top)^\top$, $\widehat{\nu}_y(X) := (\widehat{\nu}_{yk}(X))_{k=1}^{d_\nu} := (\widehat{\theta}_y^\top,\widehat{\gamma}_y^\top)^\top$, and $\bar{\nu}_y(X,w)$ is a vector on the line connecting $\nu_y(X)$ and $\widehat{\nu}_y(X)$.

First, by Assumptions D.12(ii)(d) and D.13, the claim of Step 1, and the Hölder inequality,

$$\begin{aligned}\max_{w\in\mathcal{W}}\sup_{y\in\mathcal{Y}}|II_{1w}(y)| &\leqslant \sup_{y\in\mathcal{Y}}\sum_{r,k=1}^{2K}\sqrt{n}P\left[C|\widehat{\nu}_{yr}(X)-\nu_{yr}(X)||\widehat{\nu}_{yk}(X)-\nu_{yk}(X)|\right]\\ &\leqslant C\sqrt{n}K^2\max_{k\in[2K]}\sup_{y\in\mathcal{Y}}\|\widehat{\nu}_{yk}-\nu_{yk}\|_{P,2}^2 \lesssim_P \sqrt{n}\tau_n^2 = o(1).\end{aligned}$$

Second, we have that with probability $1-o(1)$, $\max_{w\in\mathcal{W}}\sup_{y\in\mathcal{Y}}|II_{2w}(y)|\lesssim\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|$, where, for $\Theta_{yn}:=\{\theta\in\Theta_y:\|\theta-\theta_y\|\leqslant C\tau_n\}$,

$$\mathcal{F}_2=\left\{\psi_y^{(w)}(Z;\theta,\gamma)-\psi_y^{(w)}(Z;\theta_y,\gamma_y):w\in\mathcal{W},y\in\mathcal{Y},\gamma\in\mathcal{G}_{yn},\theta\in\Theta_{yn}\right\}.$$

Application of Lemma D.9 with an envelope $F_2\lesssim F_0$ gives that with probability $1-o(1)$

$$\sup_{f\in\mathcal{F}_2}|\mathbb{G}_n(f)|\lesssim\tau_n^{\alpha/2}+n^{-1/2}n^{\frac{1}{q}},\tag{15}$$

since $\sup_{f\in\mathcal{F}_2}|f|\leqslant 2\sup_{f\in\mathcal{F}_1}|f|\leqslant 2F_0$ by Assumption D.13; $\|F_0\|_{P,q}\leqslant C$ by Assumption D.12(i); $\log\sup_Q N(\epsilon\|F_2\|_{Q,2},\mathcal{F}_2,\|\cdot\|_{Q,2})\lesssim(1+\log(e/\epsilon))\vee 0$ by Lemma D.10 because $\mathcal{F}_2=\mathcal{F}_1-\mathcal{F}_0$ for the $\mathcal{F}_0$ and $\mathcal{F}_1$ defined in Assumptions D.12(i) and D.13; and σ can be chosen so that $\sup_{f\in\mathcal{F}_2}\|f\|_{P,2}\leqslant\sigma\lesssim\tau_n^{\alpha/2}$. Indeed,

$$\begin{aligned}\sup_{f\in\mathcal{F}_2}\|f\|_{P,2}^2 &\leqslant \sup_{w\in\mathcal{W},y\in\mathcal{Y},\nu\in\Theta_{yn}\times\mathcal{G}_{yn}} P\left(P[(\psi_y^{(w)}(Z;\nu(X))-\psi_y^{(w)}(Z,\nu_y(X)))^2]\right)\\ &\leqslant \sup_{y\in\mathcal{Y},\nu\in\Theta_{yn}\times\mathcal{G}_{yn}} P\left(C\|\nu(X)-\nu_y(X)\|^\alpha\right)\\ &= \sup_{y\in\mathcal{Y},\nu\in\Theta_{yn}\times\mathcal{G}_{yn}} C\|\nu-\nu_y\|_{P,\alpha}^\alpha\leqslant\sup_{y\in\mathcal{Y},\nu\in\Theta_{yn}\times\mathcal{G}_{yn}} C\|\nu-\nu_u\|_{P,2}^\alpha\lesssim\tau_n^\alpha,\end{aligned}$$

where the first inequality follows by the law of iterated expectations; the second inequality follows by Assumption D.12(ii)(a); and the last inequality follows from $\alpha\in[1,2]$ by Assumption D.12, the monotonicity of the norm $\|\cdot\|_{P,\alpha}$ in $\alpha\in[1,\infty]$, and Assumption D.13. Conclude that with probability $1-o(1)$

$$\max_{w\in\mathcal{W}}\sup_{y\in\mathcal{Y}}|II_{2w}(y)|\lesssim\tau_n^{\alpha/2}+n^{-1/2}n^{\frac{1}{q}}=o(1).\tag{16}$$

STEP 5. In this step we show that $\sup_{y\in\mathcal{Y}}\inf_{\theta\in\Theta_y}\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\theta,\widehat{\gamma}_y)\|=o_P(1)$. We have that with probability $1-o(1)$

$$\inf_{\theta\in\Theta_y}\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\theta,\widehat{\gamma}_y)\|\leqslant\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\bar{\theta}_y,\widehat{\gamma}_y)\|,$$

where $\bar{\theta}_y=\theta_y+\mathbb{E}_n\psi_y(Z,\theta_y,\gamma_y)$, since $\bar{\theta}_y\in\Theta_y$ for all $y\in\mathcal{Y}$ with probability $1-o(1)$, and, in fact, $\sup_{y\in\mathcal{Y}}\|\bar{\theta}_y-\theta_y\|=O_P(1/\sqrt{n})$ by the last paragraph of Step 3.

Then, arguing similarly to Step 3 and 4, we can conclude that uniformly in $y\in\mathcal{Y}$:

$$\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\bar{\theta}_y,\widehat{\gamma}_y)\|\leqslant\sqrt{n}\|\mathbb{E}_n\psi_y(Z;\theta_y,\gamma_y)-(\bar{\theta}_y-\theta_y)+\mathrm{D}_{y,0}(\widehat{\gamma}_y-\gamma_y)\|+o_P(1)$$

where the first term on the right side is zero by definition of $\bar{\theta}_y$ and $\mathrm{D}_{y,0}(\widehat{\gamma}_y-\gamma_y)=0$. □

Proof of Theorem 5.2. STEP 0. In the proof $a\lesssim b$ means that $a\leqslant Ab$, where the constant A depends on the constants in Assumptions D.11– D.13, but not on n once $n\geqslant n_0$, and not on $P\in\mathcal{P}_n$. In Step 1, we consider a sequence P_n in $\mathcal{P}_n$, but for simplicity, we write $P=P_n$ throughout the proof, suppressing the index n. Since the argument is asymptotic, we can assume that $n\geqslant n_0$ in what follows.

We first show that

$$\widehat{Z}_{n,P}^*\rightsquigarrow_B Z_P\text{ in }\ell^\infty(\mathcal{Y})^K,\text{ uniformly in }P\in\mathcal{P}_n.$$

In other words, we first show that the multiplier bootstrap provides a valid approximation to the large sample law of $\sqrt{n}(\widehat{\theta}_y-\theta_y)_{y\in\mathcal{Y}}$. Let $\mathbb{P}_n$ denote the measure that puts mass n^{-1} at the points (ξ_i,Z_i) for $i=1,...,n$. Let $\mathbb{E}_n$ denote the expectation with respect to this measure, so that $\mathbb{E}_nf=n^{-1}\sum_{i=1}^n f(\xi_i,Z_i)$, and $\mathbb{G}_n$ denote the corresponding empirical process $\sqrt{n}(\mathbb{E}_n-P)$, i.e.

$$\mathbb{G}_nf=\sqrt{n}(\mathbb{E}_nf-Pf)=\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(f(\xi_i,Z_i)-\int f(s,z)dP_\xi(s)dP(z)\right).$$

Recall that we define the bootstrap draw as:

$$Z^*_{n,P} := \sqrt{n}(\widehat{\theta}^* - \widehat{\theta}) = \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i \widehat{\psi}_y(Z_i)\right)_{y\in\mathcal{Y}} = \left(\mathbb{G}_n \xi \widehat{\psi}_y\right)_{y\in\mathcal{Y}},$$

where $\widehat{\psi}_y(Z) = \psi_y(Z, \widehat{\theta}_y, \widehat{\gamma}_y)$.

STEP 1.(Linearization) In this step we establish that

$$\zeta^*_{n,P} := Z^*_{n,P} - G^*_{n,P} = o_P(1) \quad \text{in } \mathbb{D} = \ell^\infty(\mathcal{Y})^K, \tag{17}$$

where $G^*_{n,P} := (\mathbb{G}_n \xi \bar{\psi}_y)_{y\in\mathcal{Y}}$, and $\bar{\psi}_y(Z) = \psi_y(Z; \theta_y, \gamma_y)$.

To show (17), we note that with probability $1-\delta_n$, $\widehat{\gamma}_y \in \mathcal{G}_{yn}$, $\widehat{\theta}_y \in \Theta_{yn} = \{\theta \in \Theta_y : \|\theta - \theta_y\| \leqslant C\tau_n\}$, so that $\|\zeta^*_{n,P}\|_{\mathbb{D}} \lesssim \sup_{f\in\mathcal{F}_3} |\mathbb{G}_n[\xi f]|$, where

$$\mathcal{F}_3 = \left\{\tilde{\psi}^{(w)}_y(\bar{\theta}_y, \bar{\gamma}_y) - \bar{\psi}^{(w)}_y : w \in \mathcal{W}, y \in \mathcal{Y}, \bar{\theta}_y \in \Theta_{yn}, \bar{\gamma}_y \in \mathcal{G}_{yn}\right\},$$

where $\tilde{\psi}^{(w)}_y(\bar{\theta}_y, \bar{\gamma}_y))$ is the j-th element of $\psi_y(Z; \bar{\theta}_y, \bar{\gamma}_y(X))$, and $\bar{\psi}^{(w)}_y$ is the j-th element of $\psi_y(Z; \theta_y, \gamma_y(X))$. By the arguments similar to those employed in the proof of the previous theorem, $\mathcal{F}_3$ obeys

$$\log \sup_Q N(\epsilon \|F_3\|_{Q,2}, \mathcal{F}_3, \|\cdot\|_{Q,2}) \lesssim (1 + \log(e/\epsilon)) \vee 0,$$

for an envelope $F_3 \lesssim F_0$. By Lemma D.10, multiplication of this class by ξ does not change the entropy bound modulo an absolute constant, namely

$$\log \sup_Q N(\epsilon \||\xi| F_3\|_{Q,2}, \xi\mathcal{F}_3, \|\cdot\|_{Q,2}) \lesssim (1 + \log(e/\epsilon)) \vee 0.$$

Also $\mathrm{E}[\exp(|\xi|)] < \infty$ implies $(\mathrm{E}[\max_{i\leqslant n} |\xi_i|^2])^{1/2} \lesssim \log n$, so that, using independence of $(\xi_i)_{i=1}^n$ from $(Z_i)_{i=1}^n$ and Assumption D.12(i),

$$\|\max_{i\leqslant n} \xi_i F_0(Z_i)\|_{\mathrm{P}_P,2} \leqslant \|\max_{i\leqslant n} \xi_i\|_{\mathrm{P}_P,2} \|\max_{i\leqslant n} F_0(Z_i)\|_{\mathrm{P}_P,2} \lesssim n^{1/q} \log n.$$

Applying Lemma D.9,

$$\sup_{f\in\xi\mathcal{F}_3} |\mathbb{G}_n(f)| = O_P\left(\tau_n^{\alpha/2} + \frac{n^{1/q}\log n}{\sqrt{n}}\right) = o_P(1),$$

for $\sup_{f\in\xi\mathcal{F}_3} \|f\|_{P,2} = \sup_{f\in\mathcal{F}_3} \|f\|_{P,2} \lesssim \sigma_n \lesssim \tau_n^{\alpha/2}$, where the details of calculations are similar to those in the proof of Theorem 5.1. Indeed, with probability $1 - o(\delta_n)$,

$$\begin{aligned}
\sup_{f\in\mathcal{F}_3} \|f\|^2_{P,2} &\lesssim \sup_{w\in\mathcal{W}, y\in\mathcal{Y}, \nu\in\Theta_{yn}\times\mathcal{G}_{yn}} P\left(P[(\psi^{(w)}_y(Z, \nu(X)) - \psi^{(w)}_y(Z, \nu_y(X)))^2]\right) \\
&\lesssim \sup_{y\in\mathcal{Y}, \nu\in\Theta_{yn}\times\mathcal{G}_{yn}} \|\nu - \nu_y\|^\alpha_{P,\alpha} \\
&\lesssim \sup_{y\in\mathcal{Y}, \nu\in\Theta_{yn}\times\mathcal{G}_{yn}} \|\nu - \nu_y\|^\alpha_{P,2} \\
&\lesssim \tau_n^\alpha,
\end{aligned}$$

where the first inequality follows from the triangle inequality and the law of iterated expectations; the second inequality follows by Assumption D.12(ii)(a) and Assumption D.12(i); the third inequality follows from $\alpha \in [1, 2]$ by Assumption D.12, the monotonicity of the norm $\|\cdot\|_{P,\alpha}$ in $\alpha \in [1, \infty]$, and Assumption D.13; and the last inequality follows from $\|\nu - \nu_y\|_{P,2} \lesssim \tau_n$ by the definition of Θ_{yn} and $\mathcal{G}_{yn}$. The equation (17) follows.

STEP 2. Here we are claiming that $Z^*_{n,P} \rightsquigarrow_B Z_P$ in $\mathbb{D} = \ell^\infty(\mathcal{Y})^K$, under any sequence $P = P_n \in \mathcal{P}_n$, were $Z_P = (\mathbb{G}_P \bar{\psi}_y)_{y \in \mathcal{Y}}$. By the triangle inequality and Step 1,

$$\sup_{h \in BL_1(\mathbb{D})} \left| \mathrm{E}_{B_n} h(Z^*_{n,P}) - \mathrm{E}_P h(Z_P) \right| \leqslant \sup_{h \in BL_1(\mathbb{D})} \left| \mathrm{E}_{B_n} h(G^*_{n,P}) - \mathrm{E}_P h(Z_P) \right| + \mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2),$$

where the first term is $o^*_P(1)$, since $G^*_{n,P} \rightsquigarrow_B Z_P$ by Theorem D.5, and the second term is $o_P(1)$ because $\|\zeta^*_{n,P}\|_{\mathbb{D}} = o_P(1)$ implies that $\mathrm{E}_P(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2) = \mathrm{E}_P \mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2) \to 0$, which in turn implies that $\mathrm{E}_{B_n}(\|\zeta^*_{n,P}\|_{\mathbb{D}} \wedge 2) = o_P(1)$ by the Markov inequality. □

E. Additional Experimental Details and Results

All experiments are carried out using R version 4.3.1 on a MacBook Pro with Apple M2 Max chip and 64GB memory. The code is available at https://github.com/CyberAgentAILab/dte-ml-adjustment. Additionally, we are in the process of developing a Python package that implements our proposed method.

E.1. Simulation Study

E.1.1. DATA GENERATING PROCESS (DGP)

We fix the number of covariates d_x as $d_x = 100$ and the sample size n to be in $\{500, 1000, 5000\}$. For each $i = 1, \ldots, n$, we generate $X_i = (X_{1i}, \ldots, X_{100i})$ from $U_{100}((0,1)^{100})$, a multivariate uniform distribution on $(0,1)$. Binary treatment variable D_i follows Bernoulli distribution with success probability of $\rho = 0.5$. A continuous outcome variable Y_i is then generated from the outcome equation $Y_i = f(X_i, D_i) + U_i$, where the error term $U_i \sim N(0,1)$. We consider the functional form of

$$f(X_i, D_i) = D_i + \sum_{j=1}^{100} \beta_j X_{ij} + \sum_{j=1}^{100} \gamma_j X_{ij}^2 \tag{18}$$

so that the outcome is nonlinear in covariates. We set

$$\beta_j = \begin{cases} 1 \text{ for } j \in \{1, \ldots, 50\} \\ 0 \text{ for } j \in \{51, \ldots, 100\}, \end{cases}$$

and

$$\gamma_j = \begin{cases} 1 \text{ for } j \in \{1, \ldots, 50\} \\ 0 \text{ for } j \in \{51, \ldots, 100\}. \end{cases}$$

In other words, the first 50 variables are relevant but the other 50 variables are irrelevant to the outcome variable.

E.1.2. EVALUATION METRICS

We evaluate the performance of our estimators using

1. Bias ratio, computed as $100\% \times \frac{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\Delta}_{y,r} - \Delta_y)}{\Delta_y}$
2. Root mean squared error (RMSE), computed as $RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\widehat{\Delta}_{y,r} - \Delta_y)^2}$
3. RMSE reduction, computed as $100\% \times (1 - \frac{RMSE_{adjusted}}{RMSE_{simple}})$

where Δ_y is the true distributional parameter (e.g., DTE, QTE) at threshold y, $\widehat{\Delta}_{y,r}$ is the estimate from the replication r, and R is the number of replications. We consider $R = 1000$ in our experiments. To approximate the true distributions, we generate a dataset with 1,000,000 observations and calculate the distributional parameter at each y. In the simulations, we consider 9 values of threshold y at the quantiles $\{0.1, 0.2, \ldots, 0.9\}$ of the true outcome distribution.

E.1.3. Implementation

In the simulation study, we implement logistic LASSO with *cv.glmnet* function in `glmnet` package in R (Friedman et al., 2010; Tay et al., 2023).

E.1.4. Relevance of covariates and RMSE reduction

We demonstrate the relationship between the predictive power of covariates on the outcome and the reduction in RMSE using a simple experiment. For this purpose, we explore a series of data generating processes where the relevance of covariates ranges from high to low. Specifically, we construct a slowly decaying sequence of coefficients $\kappa_s = 2 \times s^{-1}$ for $s = 1, \ldots, 10$. Then, in our outcome equation (18), we set the coefficients as follows:

$$\beta_j = \begin{cases} \kappa_s \text{ for } j \in \{1, \ldots, 50\} \\ 0 \text{ for } j \in \{51, \ldots, 100\}, \end{cases}$$

and

$$\gamma_j = \begin{cases} \kappa_s \text{ for } j \in \{1, \ldots, 50\} \\ 0 \text{ for } j \in \{51, \ldots, 100\}. \end{cases}$$

Note that we consider a decaying sequence: $s = 1$ corresponds to the case with the highest relevance, and the relevance diminishes as we increase s up to $s = 10$. Figure 3 illustrates how covariates with higher relevance result in a greater reduction in RMSE across all quantiles when sample size is $n = 1000$.

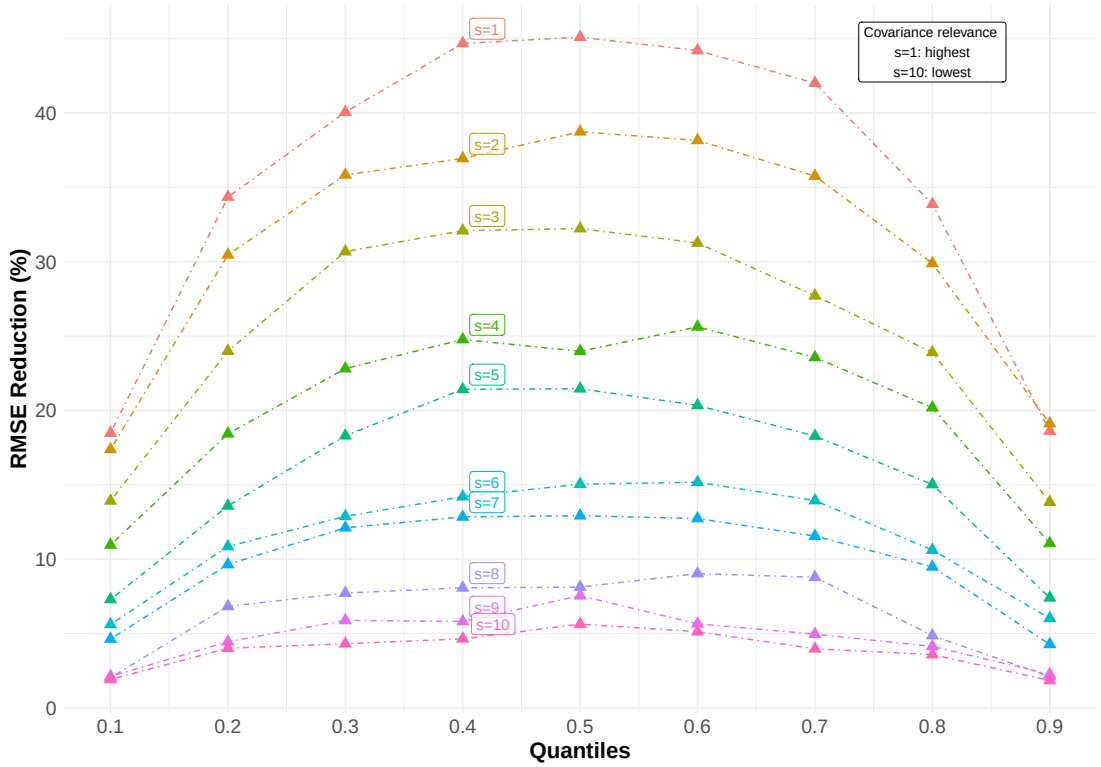

Figure 3. RMSE reduction in % of the ML adjusted estimator compared to the simple DTE estimator, under various data generating processes indexed by $s = 1, \ldots, 10$, calculated over 1000 simulations. $s = 1$: highest relevance of covariates, diminishing relevance of covariates as s increases up to $s = 10$. The simple estimator is derived from empirical distribution functions, while the ML adjusted estimator is obtained using LASSO with 5-fold cross-fitting. $n = 1000$.

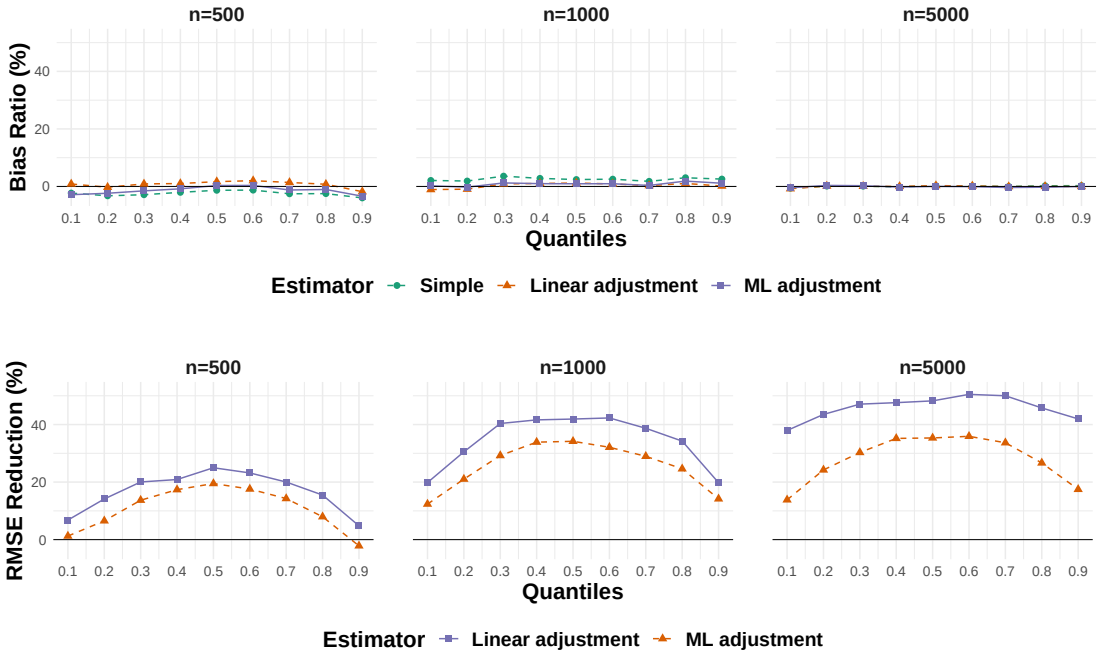

Figure 4. Bias (top figure), as a % of true value, of different QTE estimators and RMSE reduction in % (bottom figure) of adjusted estimators compared to simple QTE estimator, under sample sizes $\{500, 1000, 5000\}$, calculated over 1000 simulations. The simple estimator is calculated from empirical distribution functions. The regression-adjusted estimators (linear adjustment and ML adjustment based on LASSO) are implemented using 5-fold cross-fitting.

E.1.5. Quantile Treatment Effect (QTE)

We also consider simple and regression-adjusted QTE estimators. In the setup introduced in Section E.1.1, the outcome variable is continuous and hence the QTE is well-defined. The true value of QTE is constant and equals 1 at all quantiles. The top figure of Figure 4 plots the bias as a % of the true value of the QTE. The bottom figure of Figure 4 plots the RMSE reduction in % terms for the linear and ML adjustment, compared to the simple estimator. We confirm the bias is small for all QTE estimators. Even when sample size is small ($n = 500$), the bias is at most 4%. As for the RMSE, the results are similar to that for the DTE explained in Section 6.1. The variance reduction is around 13%-35% for linearly adjusted estimator and is around 37%-50% for the ML adjusted estimator when sample size is large ($n = 5000$).

E.2. Nudges to reduce water consumption

E.2.1. Data and implementation

The dataset from the randomized experiment can be downloaded at https://doi.org/10.7910/DVN1/22633 (Ferraro & Price, 2013a).

In our analysis, we implement gradient boosting with `xgboost` package in R (Chen & Guestrin, 2016).

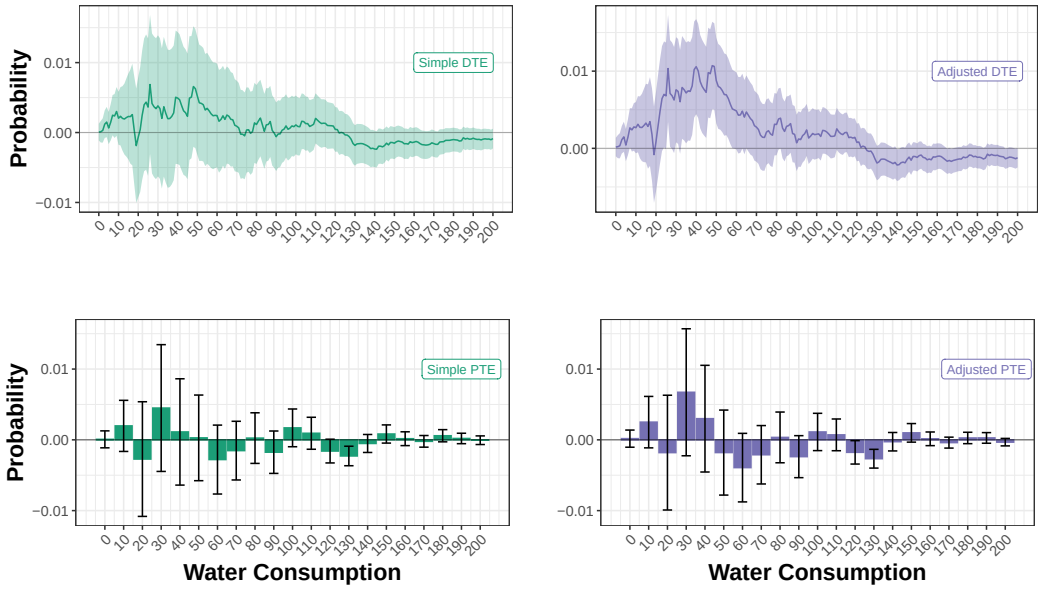

Figure 5. Technical Advice (T1) vs. Control. Distributional Treatment Effect (DTE) and Probability Treatment Effect (PTE) on water consumption (in thousands of gallons). The top left figure represents the simple DTE; the top right figure depicts the regression-adjusted DTE, computed for $y \in \{0, 1, 2, \ldots, 200\}$. The bottom left figure represents the simple PTE; the bottom right figure represents the regression-adjusted PTE, computed for $y \in \{0, 10, 20, \ldots, 200\}$ and $h = 10$. The regression adjustment is implemented via gradient boosting with 10-fold cross-fitting. The shaded areas and error bars represent the 95% pointwise confidence intervals. $n = 78,478$.

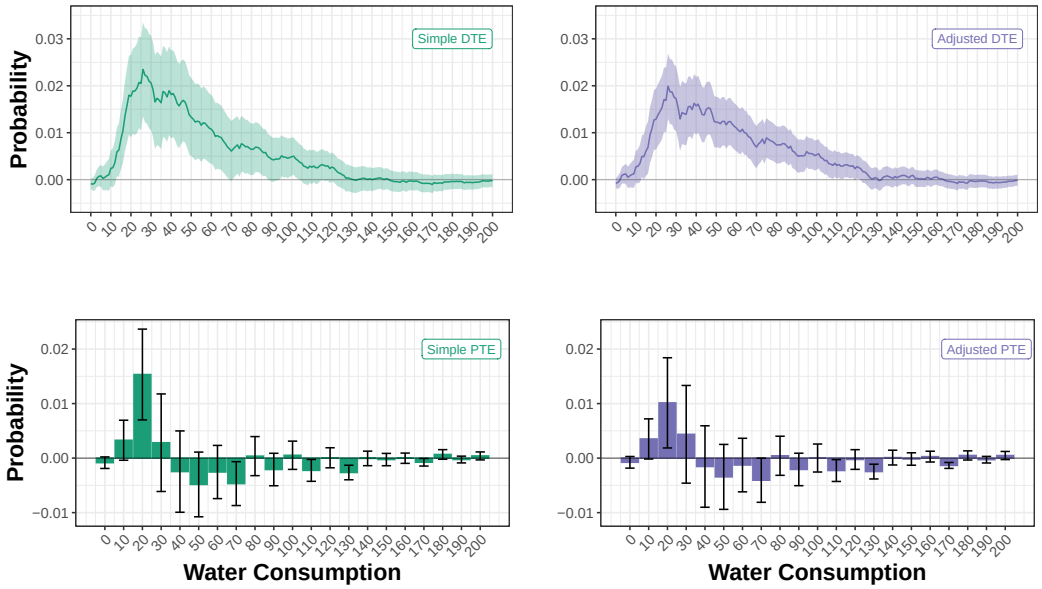

Figure 6. Weak Social Norm (T2) vs. Control. Distributional Treatment Effect (DTE) and Probability Treatment Effect (PTE) on water consumption (in thousands of gallons). The top left figure represents the simple DTE; the top right figure depicts the regression-adjusted DTE, computed for $y \in \{0, 1, 2, \ldots, 200\}$. The bottom left figure represents the simple PTE; the bottom right figure represents the regression-adjusted PTE, computed for $y \in \{0, 10, 20, \ldots, 200\}$ and $h = 10$. The regression adjustment is implemented via gradient boosting with 10-fold cross-fitting. The shaded areas and error bars represent the 95% pointwise confidence intervals. $n = 78,468$.

E.2.2. RESULTS WITH MULTIPLE TREATMENTS

The randomized experiment considered four treatment groups: technical advice (T1), weak social norm (T2), strong social norm (T3), and a control group. Technical advice (T1) involved providing residents with information on ways to reduce water use. The weak social norm (T2) treatment combined technical advice with an appeal to prosocial preferences. The strong social norm (T3) treatment further included social comparisons along with the elements of T2. On average, all three treatments resulted in a reduction in water use compared to the control group, with the strong social norm (T3) showing the largest effect and the technical advice (T1) showing the smallest effect. See Ferraro & Price (2013b) for more details about the experimental design and average treatment effect analysis.

We extended the analysis by examining the entire distribution of water use. Figure 5 displays the DTE and PTE of technical advice (T1) compared to the control group. The PTE results indicate a reduction in water use in the range of (120, 140] under T1. Although regression adjustment results in tighter confidence intervals for the DTE and PTE, the overall conclusions remain the same.

Figure 6 presents the DTE and PTE of the weak social norm (T2) compared to the control group. Under T2, water use increased in the range of (20, 30] but decreased in the ranges of (110, 120], (130, 140], and (only slightly in) (170, 180]. Similar to T1, regression adjustment leads to tighter confidence intervals for the DTE and PTE without altering the primary conclusions for this treatment.