Distribution-Free Assessment of Population Overlap in Observational Studies

Lihua Lei*[1], Alexander D'Amour†[2], Peng Ding‡[3], Avi Feller§[3], and Jasjeet Sekhon¶[4]

[1]Department of Statistics, Stanford University
[2]Google Brain, Cambridge
[3]Department of Statistics, University of California, Berkeley
[4]The Goldman School, University of California, Berkeley
[5]Department of Statistics and Data Science, Yale University

July 8, 2021

Abstract

Overlap in baseline covariates between treated and control groups, also known as positivity or common support, is a common assumption in observational causal inference. Assessing this assumption is often ad hoc, however, and can give misleading results. For example, the common practice of examining the empirical distribution of estimated propensity scores is heavily dependent on model specification and has poor uncertainty quantification. In this paper, we propose a formal statistical framework for assessing the extrema of the population propensity score; e.g., the propensity score lies in $[0.1, 0.9]$ almost surely. We develop a family of upper confidence bounds, which we term O-values, for this quantity. We show these bounds are valid in finite samples so long as the observations are independent and identically distributed, without requiring any further modeling assumptions on the data generating process. Finally, we demonstrate this approach using benchmark observational studies, showing how to build our proposed method into the observational causal inference workflow.

Keywords: overlap, observational study, propensity score, distribution-free, partial identification, uncertainty quantification

1 Introduction

Observational studies rely on covariate balance or, more formally, covariate overlap for credible estimation of causal effects. Given a binary treatment T and a set of covariates X, the overlap is measured by how extreme the propensity score $e(x) \triangleq \mathbb{P}(T = 1 \mid X)$ is over the covariate space. More formally, the *overlap condition* states that $0 < e(X) < 1$ almost surely, which is proved to be necessary for identification of the average treatment effect (ATE) without assumptions on the functional relationships between the outcome and covariates [Khan and Tamer, 2010]. Intuitively, if $e(x) = 0$ for a subset of covariate values with a non-negligible chance, the counterfactuals under the treatment can never be identified for those units, rendering the ATE on that subgroup unidentifiable. Unless untenable assumptions are imposed to extrapolate the missing counterfactual based on other units with $e(x) > 0$, the overall ATE cannot be identified.

*lihualei@stanford.edu
†alexdamour@google.com
‡pengdingpku@berkeley.edu
§afeller@berkeley.edu
¶jasjeet.sekhon@yale.edu

The above overlap condition is only sufficient to identification, which roughly amounts to consistent estimation with infinite samples. In practice where only a finite sample is available, stronger versions are needed. One of the most popular strengthened overlap condition is the strict overlap condition, which states the following:

Condition 1 (Strict Overlap)**.** *There exists some constant* $0 < \mathcal{O} \leq 0.5$ *such that*

$$\mathcal{O} \leq e(X) \leq 1 - \mathcal{O} \quad a.s.$$

Although it can be replaced by weaker conditions like finite moments of $1/e(X)$ for certain tasks [e.g. Chen et al., 2008, Hirshberg and Wager, 2017, Ma and Wang, 2019, Hong et al., 2020, Lei and Candès, 2020], the strict overlap condition remains widely-used for its interpretability in practice and convenience in deriving theoretical guarantees. The common practice of trimming extreme (estimated) propensity scores [e.g. Crump et al., 2009, Yang and Ding, 2017, Ju et al., 2019] is essentially trying to achieve the strict overlap condition if any violation is observed.

The strict overlap condition motivates a natural one-number summary $\mathcal{O}^*$ of the overlap for a given data generating process — the largest $\mathcal{O}$ for which Condition 1 is satisfied. We refer to it as the *population overlap slack*. By definition, the strict overlap condition with level $\mathcal{O}$ is equivalent to $\mathcal{O}^* \geq \mathcal{O}$.

One common approach to assess overlap in practice is the eyeball test based on distributions of estimated propensity scores in two groups, as suggested in the textbook by Imbens and Rubin [2015, Chapter 14]. Despite some rules of thumb, it is unclear whether this approach has any provable error control for testing the strict overlap condition, or how accurate it assesses $\mathcal{O}^*$ statistically. Intuitively, this approach would fail to provide a reliable assessment when the propensity score model is mis-specified.

A natural estimator for $\mathcal{O}^*$ is the plug-in estimator resting on the extreme value of estimated propensity scores. Specifically, given an estimate $\hat{e}(x)$, we can estimate $\mathcal{O}^*$ by the minimal value of $\min\{\hat{e}(x), 1 - \hat{e}(x)\}$ over the samples. Despite its simplicity, this approach lacks reliability from a statistical perspective since the estimand $\mathcal{O}^*$ is irregular and classical statistical justification does not apply to it without discretion. On the one hand, the propensity score model may be misspecified; even a slight deviation can change the extreme $\mathcal{O}^*$ drastically. On the other hand, even if the model is correctly specified, it is challenging to quantify the uncertainty of the estimated extreme propensity score due to the irregularity of the estimator. In Appendix A, we present examples with large sample sizes and simple correctly specified parametric models, in which the plug-in estimator may still drastically overestimate or underestimate the overlap.

Because of the irregularity of the parameter $\mathcal{O}^*$, it is unrealistic to seek a consistent estimator. Instead, we introduce the *O-value*. An O-value is an upper confidence bound of $\mathcal{O}^*$ in the sense that $\mathbb{P}(\mathcal{O}^* \leq \hat{\mathcal{O}}) \geq 1 - \alpha$. Given an O-value $\hat{\mathcal{O}}$, the induced test which rejects Condition 1 if $\hat{\mathcal{O}} < \mathcal{O}$ has valid type-I error control. Similar to the p-value, the O-value has a one-sided guarantee — a small O-value provides strong evidence of insufficient overlap while a large O-value does not justify sufficient overlap. Therefore, an O-value is a "detector" of overlap deficiency instead of an "estimator" of the amount of overlap. We put a "hat" on $\mathcal{O}$ just to highlight that an O-value is a statistic that is computed from data instead of an unknown population parameter.

Apparently 0.5 is a valid but futile O-value regardless of the model, just as 1 is a valid but useless p-value. In this paper, we propose four types of O-values — Difference-in-Means (DiM), Difference-in-Tails (DiT), Difference-in-Ranks (DiR) and Classification Error (CE) O-values — which are valid in finite samples without any modelling assumption, given independent and identically distributed (i.i.d.) observations $(T_i, X_i)_{i=1}^n$. As with the eyeball test, our method operates on the estimated propensity scores via arbitrary statistical or machine learning methods. However, instead of taking the extreme value directly, our method tackles with more estimable upper bounds of $\mathcal{O}^*$. In a nutshell, our test can take advantage of modern machine learning algorithms to estimate propensity scores; but rather than hoping for the best, it carefully wraps around them to protect against overfitting and produce trustable uncertainty quantification — however inaccurate the estimated propensity scores are, our proposed O-values are always valid.

Apparently, in the distribution-free setting as ours, it is impossible to derive a non-trivial O-value with two-sided worst-case guarantees in finite samples because one can always keep the extreme

propensity score on a sufficiently small subset of covariates that are barely observed in the data while raise the rest to 0.5. However, on synthetic and real datasets with moderate sample sizes, our O-values are observed to be powerful to detect lack of overlap.

In practice, an O-value can also be used as a tool to partially assess the quality of trimming or matching procedures. The motivation of both procedures is to select units with sufficient overlap. As a consequence, if the O-value computed on the post-trimming or post-matching dataset is smaller than expected, it suggests that more aggressive trimming or matching needs to be applied. Similarly, it can be viewed as a partial diagnostic tool for propensity score fitting in the sense that if the O-value is much smaller than the extreme of estimated propensity score, the model is invalidated and needs to be adjusted.

Although the population overlap slack $\mathcal{O}^*$ is the parameter of interest in this paper, we realize that it only provides a coarse summary of the extreme behaviors of the propensity score. For examples, it cannot distinguish the case with $\mathbb{P}(e(X) \in \{\mathcal{O}^*, 1-\mathcal{O}^*\}) = 1$ and the case with $\mathbb{P}(e(X) \in \{\mathcal{O}^*, 1-\mathcal{O}^*\}) = 0.01, \mathbb{P}(e(X) = 0.5) = 0.99$, although the latter case is clearly more "overlapped" than the former. We discuss alternative overlap measures at the end of the paper and show that part of our techniques can be extended to test those parameters. Nonetheless, we focus on $\mathcal{O}^*$ because it is already sufficiently challenging.

2 Preliminaries

2.1 Notation and assumption

Throughout this paper, we consider a binary treatment $T \in \{0, 1\}$ with $T = 1$ denoting the treatment group and $T = 0$ denoting the control group. Denote by X a generic covariate and by $e(x) = \mathbb{P}(T = 1 \mid X = x)$ the propensity score. The population overlap slack is defined as

$$\mathcal{O}^* \triangleq \sup\{\mathcal{O} : \mathcal{O} \leq e(X) \leq 1 - \mathcal{O},\ a.s.\}.$$

A simple continuity argument shows that

$$\mathcal{O}^* \leq e(X) \leq 1 - \mathcal{O}^*, \quad a.s. \tag{1}$$

Let $(T_1, X_1), \ldots, (T_n, X_n)$ denote the observations. In this paper, we rely merely on the following assumption.

Assumption 1. $(T_1, X_1), \ldots, (T_n, X_n) \overset{i.i.d.}{\sim} (T, X)$.

2.2 Variational representation of population overlap slack

In the distribution-free setting, we cannot expect to recover $\mathcal{O}^*$ by directly estimating the range of $e(X)$. Therefore, we need to find implications of (1) which yield estimable upper bounds of $\mathcal{O}^*$.

Since X is potentially mixed-typed and high-dimensional, the first step is to simplify X into a "standard form". Consider any function $s(\cdot)$ on the domain of X. Let $e_s(x) \triangleq \mathbb{P}(T = 1 \mid s(X) = s(x))$ and $\mathcal{O}_s^*$ denote the population overlap slack of $e_s(x)$, i.e.

$$\mathcal{O}_s^* \triangleq \sup\{\mathcal{O} : \mathcal{O} \leq e_s(X) \leq 1 - \mathcal{O},\ a.s.\}.$$

Then $\mathcal{O}_s^* \geq \mathcal{O}^*$ for any transformation $s(\cdot)$ because

$$e_s(x) = \mathbb{P}(T = 1 \mid s(X) = s(x)) = \mathbb{E}[e(X) \mid s(X) = s(x)] \in [\mathcal{O}^*, 1 - \mathcal{O}^*].$$

On the other hand, when $s(\cdot)$ is the true propensity score, $\mathcal{O}_s^* = \mathcal{O}^*$ because $e(X) = P(T = 1 \mid X) = P(T = 1 \mid e(X))$. This simple fact is a special case of Theorem 2 of Rosenbaum and Rubin [1983] which states that the propensity score is the coarsest balancing score. Putting two pieces together, we obtain the following variational representation of the population overlap slack.

Proposition 2.1. *Let $\mathcal{X}$ denote the domain of X and $\mathcal{F}$ be the set of all measurable functions $s : \mathcal{X} \mapsto [0, 1]$. Then*

$$\mathcal{O}^* = \inf_{s(\cdot)\in\mathcal{F}} \mathcal{O}^*_s,$$

where the equality holds when $s(\cdot) = e(\cdot)$.

Suppose we have a "magical" tool that can produce valid O-values when X is univariate and lies in $[0, 1]$. Then we can transform X into a single number $s(X)$ and treat $s(X)$ as the new covariate. The tool can provide $\hat{\mathcal{O}}$ which is an upper confidence bound of $\mathcal{O}^*_s$. As implied by Proposition 2.1, $\hat{\mathcal{O}}$ is a valid O-value. However, it is loose if $\mathcal{O}^*_s \gg \mathcal{O}^*$. By Proposition 2.1 again, we need to find a transformation $s(\cdot)$ that approximates the true propensity score.

2.3 Generic covariate standardization via data splitting

A natural strategy is to set $s(x) = \hat{e}(x)$ for some estimated propensity score $\hat{e}(\cdot)$. Noticing that Proposition 2.1 requires $s(\cdot)$ to be a deterministic function that does not depend on the data, we cannot estimate the propensity score using the full data and apply the O-value tool on $\hat{e}(X_i)$'s again. To avoid the double-dipping issue, we randomly split the data into two folds and train $\hat{e}(x)$ on one fold using any probabilistic classifier, including the logistic regression, random forest, gradient boosting, deep neural net, and so on. On the second fold, we transform the covariate X_i into $S_i \triangleq \hat{e}(X_i)$ and apply the O-value tool on the reduced dataset $\{(T_i, S_i) : i \in \text{fold } 2\}$. The resulting O-value is valid conditional on the fold 1 and thus valid unconditionally.

Contrasted with the heuristic approach to assess overlap, our method processes the estimated propensity score indirectly through the "magical" O-value tool for univariate covariates. Throughout the rest of the paper, we focus on the setting where

$$(T_1, S_1), \ldots, (T_n, S_n) \overset{i.i.d.}{\sim} (T, S), \quad \text{where } S \in [0, 1].$$

With a slight abuse of notation, we denote by $\mathcal{O}^*$ the induced population overlap slack $\sup\{\mathcal{O} : \mathcal{O} \le \mathbb{P}(T = 1 \mid S) \le 1 - \mathcal{O},\ a.s.\}$. The goal is to derive valid upper confidence bounds for $\mathcal{O}^*$. It should be kept in mind that an upper confidence bound of the induced population overlap slack is a valid O-value by Proposition 2.1.

2.4 Strict overlap as a density ratio condition

D'Amour et al. [2017] reveals the connection between the overlap and balance of covariate distribution. In particular, denote by P_0 and P_1 the covariate distributions under the treatment and control respectively, i,e.

$$P_1(S \in A) := P(S \in A \mid T = 1), \quad P_0(S \in A) := P(S \in A \mid T = 0).$$

Further let $\pi = \mathbb{P}(T = 1)$ denote the marginal intensity of treatment. Then by the Bayes formula,

$$\begin{aligned} &\mathcal{O}^* \le \mathbb{P}(T = 1 \mid S) \le 1 - \mathcal{O}^* \quad a.s. \quad \Longleftrightarrow \quad \frac{\mathcal{O}^*}{1 - \mathcal{O}^*} \le \frac{\mathbb{P}(T = 1 \mid S)}{\mathbb{P}(T = 0 \mid S)} \le \frac{1 - \mathcal{O}^*}{\mathcal{O}^*} \quad a.s. \\ &\Longleftrightarrow \quad b_{\min}(\mathcal{O}^*; \pi) \triangleq \frac{1 - \pi}{\pi} \frac{\mathcal{O}^*}{1 - \mathcal{O}^*} \le \frac{dP_1}{dP_0}(S) \le \frac{1 - \pi}{\pi} \frac{1 - \mathcal{O}^*}{\mathcal{O}^*} \triangleq b_{\max}(\mathcal{O}^*; \pi) \quad a.s.. \end{aligned} \tag{2}$$

As a consequence, (1) is equivalent to a density ratio condition.

2.5 A general strategy

Density ratio estimation is a widely-studied problem in the literature [e.g. Sugiyama et al., 2012]. However, in the distribution-free setting like ours, direct estimation of the likelihood ratio with reliable uncertainty quantification is hard even if S is univariate. Moreover, direct estimation of the extreme values of a density ratio is unstable and sensitive to hyperparameters like the bandwidth.

Instead, we seek implications of (2). Intuitively, (2) controls the "discrepancy" between P_1 and P_0. Since both P_1 and P_0 are probability measures, by (2), $b_{\min}(\mathcal{O}^*;\pi) \le 1 \le b_{\max}(\mathcal{O}^*;\pi)$. This yields a naive upper bound of $\mathcal{O}^*$ as $\min\{\pi, 1-\pi\}$. Furthermore, the equality holds if and only if $b_{\min}(\mathcal{O}^*;\pi) = 1$ or $b_{\max}(\mathcal{O}^*;\pi) = 1$, in which case $P_1 = P_0$ and thus the data is from a completely randomized experiment.

However, the naive bound completely ignores the imbalance between P_1 and P_0 and thus is not powerful. To derive tighter bounds, we need to take the estimated propensity scores into account. D'Amour et al. [2017] shows that the difference of means $|\mathbb{E}_{P_1}[S] - \mathbb{E}_{P_0}[S]|$ is bounded by a function of $\mathcal{O}^*$. Inverting their bound, we are able to obtain an upper bound of $\mathcal{O}^*$ as a function of the mean difference. As opposed to the density ratio, the mean difference is much easier to estimate, thereby enabling reliable uncertainty quantification in finite samples.

In general, given a measure of discrepancy $\Delta(P_1, P_0)$, our first step is to derive an $\mathcal{O}^*$-dependent upper bound $B_\Delta(\mathcal{O}^*)$ from the density ratio condition (2) in the sense that

$$\Delta(P_1, P_0) \le B_\Delta(\mathcal{O}^*). \tag{3}$$

For example, Theorem 1 of D'Amour et al. [2017] essentially considers a measure of discrepancy based on the normalized mean difference with

$$\Delta(P_1, P_0) = \frac{|\mathbb{E}_{P_1}[S] - \mathbb{E}_{P_0}[S]|}{\sqrt{\mathrm{Var}_{P_0}[S]}}, \quad B_\Delta(\mathcal{O}^*) = \sqrt{(1 - b_{\min}(\mathcal{O}^*;\pi))(b_{\max}(\mathcal{O}^*;\pi) - 1)}.$$

Typically, B_Δ is decreasing in $\mathcal{O}^*$ because larger overlap implies a smaller distributional discrepancy. In these cases, if we can find a $(1-\alpha)$ lower confidence bound $\hat{\Delta}^-(P_1, P_0)$ of $\Delta(P_1, P_0)$, then $\hat{\mathcal{O}} \triangleq B_\Delta^{-1}(\hat{\Delta}^-(P_1, P_0))$ is a valid O-value because

$$\mathbb{P}(\mathcal{O}^* \le \hat{\mathcal{O}}) = \mathbb{P}(\hat{\Delta}^-(P_1, P_0) \le \Delta(P_1, P_0)) \ge 1 - \alpha. \tag{4}$$

2.6 Contrast with two-sample testing-based methods for overlap

A closely related line of works treats the overlap assessment as a two-sample testing problem and tests the equality of the covariate distributions or the distribution of propensity scores under two groups [e.g. Gagnon-Bartsch et al., 2019, Kim et al., 2019, Chen and Small, 2016, Kim et al., 2016]. Those works essentially test the hypothesis that $dP_1(S)/dP_0(S) = 1$ almost surely. A simple algebra shows that $P_1 = P_0$ if and only if $e(X)$ is a constant almost surely. Despite being more statistically sound than the heuristic eyeball assessment, this type of methods is testing a overly stringent hypothesis that the observational study is a completely randomized experiment. As a result, rejecting this null hypothesis does not necessarily imply a lack of overlap — a non-constant propensity score lying in $[0.4, 0.6]$ is totally fine in terms of overlap. By contrast, the density ratio condition (2) is much less stringent and an violation of it with a small $\mathcal{O}^*$ does imply the lack of overlap because it is equivalent to the strict overlap condition that we aim to test.

3 O-values

In this section, we derive four types of O-values based on different discrepancy measures and different techniques to construct their lower confidence bounds. As mentioned in Section 2.6, our O-values are closely related to two-sample tests of distributional equality. It turns out that each type of O-value is analogous to a type of two-sample test. We summarize them in Table 1. The definitions of the symbols can be found in following subsections. The techniques to get $\hat{\Delta}_-(P_1, P_0)$ for all types of O-values are quite involved technically so we presented a looser but more interpretable version in this section and discuss the tighter version to Appendix C.

Type	$\Delta(P_1, P_0)$	Technique to get $\hat{\Delta}_-(P_1, P_0)$	Analogue
DiM	$\frac{\lvert\mu_1-\mu_0\rvert}{\sigma_0}, \frac{\lvert\mu_1-\mu_0\rvert}{\sigma_1}$	Hedged capital bound [Waudby-Smith and Ramdas, 2020] Maurer-Pontil inequality [Maurer and Pontil, 2009]	Student t-test
DiT	$\sup_{A\in\mathcal{A}} \frac{P_1(A)}{P_0(A)}, \sup_{A\in\mathcal{A}} \frac{P_0(A)}{P_1(A)}$	DKWM inequality [Massart, 1990b] Line-crossing probability [Dempster, 1959] Generalized Simes' inequality [Sarkar et al., 2008]	Kolmogorov-Smirnov test
DiR	$\mathbb{P}(S^{(1)} \geq S^{(0)})$ $(S^{(1)}, S^{(0)}) \sim P_1 \otimes P_0$	Hoeffding-Bentkus-Maurer inequality for U-statistics [Bates et al., 2021a]	Wilcoxon rank-sum test
CE	$\mathbb{P}(T = I(S > \eta))$	Same as DiT O-value	Classification-based test

Table 1: Overview of four types of O-values

For a moment, we pretend that π is known and $\pi \leq 0.5$ to avoid unnecessary complications. We will discuss this issue in Section 3.5. All technical proofs are relegated into Appendix B.

3.1 Difference-in-Means (DiM) O-value

D'Amour et al. [2017] show that the overlap implies balance in terms of the covariate means in two groups. Their bound is stated for general multivariate covariates and its proof is based on Rukhin [1993]'s bounds of f-divergences for families with bounded likelihood ratio. Here, the covariate S_i is univariate, the bound can be stated in terms of the normalized mean differences T_1 and T_0:

$$T_1 = \frac{|\mu_1 - \mu_0|}{\sigma_1}, \quad T_0 = \frac{|\mu_1 - \mu_0|}{\sigma_0},$$

where $\mu_1 = \mathbb{E}_{P_1}[S], \mu_0 = \mathbb{E}_{P_0}[S]$, and $\sigma_1^2 = \mathrm{Var}_{P_1}[S], \sigma_0^2 = \mathrm{Var}_{P_0}[S]$. When $\mu_1 = \mu_0$ and σ_1 (or σ_0) is zero, we set T_1 (or T_0) to be zero.

Theorem 3.1 (D'Amour et al. [2017], Theorem 1)**.**

$$T_0 \leq \sqrt{(1 - b_{\min}(\mathcal{O}^*;\pi))(b_{\max}(\mathcal{O}^*;\pi) - 1)}, \quad T_1 \leq \sqrt{(1 - b_{\max}(\mathcal{O}^*;\pi)^{-1})(b_{\min}(\mathcal{O}^*;\pi)^{-1} - 1)}.$$

Equivalently,

$$\mathcal{O}^* \leq \mathcal{O}^+_{\mathrm{DiM}}(T_0, T_1; \pi) \triangleq \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\pi(1-\pi)}{\max\{\pi T_0, (1-\pi)T_1\}^2 + 1}}$$

The proof is presented in Appendix B for completeness. According to the general strategy in Section 2.5, it remains to find simultaneous lower confidence bounds $\hat{T}_1^-$ and $\hat{T}_0^-$ such that

$$\mathbb{P}\left(\hat{T}_1^- \leq T_1, \hat{T}_0^- \leq T_0\right) \geq 1 - \alpha. \tag{5}$$

Intuitively, T_1 and T_0 can be estimated by replacing the means and variances with their empirical estimates. However, it is challenging to quantify the uncertainty in the distribution-free setting without resorting to asymptotics.

Here, we use the property that $S_i \in [0, 1]$ by construction. This simple property enables finite-sample valid uncertainty quantification via the well-known empirical Berstein inequality.

Proposition 3.1 (empirical Berstein inequality, [Maurer and Pontil, 2009]). *Let $Z_1, \dots, Z_n$ be i.i.d. with $Z_i \in [0,1]$. Further let $\hat{\mu}$ and $\hat{\sigma}^2$ be the empirical estimates of the mean and variance, i.e.*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \hat{\mu})^2.$$

Then with probability $1-\delta$,

$$|\hat{\mu} - \mu| \le \hat{\sigma}\sqrt{\frac{2\log\left(\frac{4}{\delta}\right)}{n}} + \frac{7\log\left(\frac{4}{\delta}\right)}{3(n-1)} \triangleq C_\mu(\hat{\sigma}, n; \delta), \quad \sigma - \hat{\sigma} \le \sqrt{\frac{2\log\left(\frac{4}{\delta}\right)}{n-1}} \triangleq C_\sigma(n;\delta).$$

Let $\hat{\mu}_1, \hat{\sigma}_1^2$ (resp. $\hat{\mu}_0, \hat{\sigma}_0^2$) be the empirical estimates of the mean and variance for $(S_i)_{T_i=1}$ (resp. $(S_i)_{T_i=0}$). Further let n_1 and n_0 denote the size of the treatment and control group. By setting $\delta = \alpha/4$, we have

$$|\hat{\mu}_t - \mu_t| \le C_\mu(\hat{\sigma}_t, n_t; \alpha/4), \quad \sigma_t \le \hat{\sigma}_t + C_\sigma(n_t; \alpha/4)$$

simultaneously for $t \in \{0,1\}$ with probability at least $1-\alpha$. This yields a lower confidence bound of T_t that satisfies (5) as

$$\hat{T}_t^- = \frac{\max\left\{0, |\hat{\mu}_1 - \hat{\mu}_0| - C_\mu(\hat{\sigma}_1, n_1; \alpha/4) - C_\mu(\hat{\sigma}_0, n_0; \alpha/4)\right\}}{\hat{\sigma}_t + C_\sigma(n_t; \alpha/4)}. \tag{6}$$

By (4) and Theorem 3.1, we obtain a closed-form expression of the DiM O-value.

Theorem 3.2. *Let $\hat{T}_1^-$ and $\hat{T}_0^-$ be defined in (6) and $\mathcal{O}_{\mathrm{DiM}}^+$ be defined in Theorem 3.1. Then $\hat{\mathcal{O}}_{\mathrm{DiM}} \triangleq \mathcal{O}_{\mathrm{DiM}}^+(T_0^-, T_1^-; \pi)$ is a valid O-value.*

As discussed in Section 1, $\hat{\mathcal{O}}$ induces a valid test for Condition 1 by rejecting it when $\hat{\mathcal{O}} < \mathcal{O}$. In the special case $\mathcal{O} = \min\{\pi, 1-\pi\}$, which corresponds to the completely randomized experiment, both bounds in Theorem 3.1 are zero. As a result, our test rejects Condition 1 with $\mathcal{O} = \pi$ if

$$\max\{\hat{T}_1^-, \hat{T}_0^-\} < 0 \iff |\hat{\mu}_1 - \hat{\mu}_0| < C_\mu(\hat{\sigma}_1, n_1; \alpha/4) + C_\mu(\hat{\sigma}_0, n_0; \alpha/4).$$

By contrast, the two-sample Student t-test, which works under the assumption of equal variances $\sigma_1^2 = \sigma_0^2 = \sigma^2$, rejects the hypothesis if

$$|\hat{\mu}_1 - \hat{\mu}_0| \le \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}\sqrt{\frac{(n_1-1)\hat{\sigma}_1^2 + (n_0-1)\hat{\sigma}_0^2}{n-2}}\, t_{\alpha/2}(n-2)$$

where $t_{\alpha/2}(n-2)$ is the $(\alpha/2)$-th upper quantile of the t-distribution with degree-of-freedom $n-2$. For illustration, consider $\alpha = 0.05$. When n_1 and n_0 are in the same order as n, $\hat{\sigma}_1, \hat{\sigma}_0 = \sigma + O_p(1/\sqrt{n})$, and the rejection region given by the DiM O-value is

$$|\hat{\mu}_1 - \hat{\mu}_0| < \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_0}}\right)\sigma\sqrt{2\log\left(\frac{16}{\alpha}\right)} + O\left(\frac{1}{n}\right) \approx \left(\frac{1}{\sqrt{n_1}} + \frac{1}{\sqrt{n_0}}\right) 3.4\sigma, \tag{7}$$

while the rejection region given by the two-sample Student t-test is

$$|\hat{\mu}_1 - \hat{\mu}_0| < \sqrt{\frac{1}{n_1} + \frac{1}{n_0}}\sigma t_{\alpha/2}(n-2) \approx \sqrt{\frac{1}{n_1} + \frac{1}{n_0}} 1.96\sigma.$$

Therefore, the test for $\mathcal{O}^* = \pi$ induced by the DiM O-value is very similar to the two-sample Student t-test. Nonetheless, our test is always valid in finite samples while the latter requires Gaussian variables and equal variances. As a byproduct, our DiM O-value provides an exact solution for non-parametric Behrens-Fisher problem [Behrens, 1929, Fisher, 1935].

Although the empirical Bernstein inequality provides a neat confidence region, the constants are unsatisfactory. In Appendix C.1, we derive a sharper DiM O-value based on the recently developed betting-based confidence interval for mean parameters [Waudby-Smith and Ramdas, 2020].

3.2 Difference-in-Tails (DiT) O-value

A direct implication of the density ratio condition (2) is that, for any event $A \subset [0,1]$,

$$b_{\min}(\mathcal{O}^*;\pi) \le \frac{P_1(A)}{P_0(A)} \le b_{\max}(\mathcal{O}^*;\pi). \tag{8}$$

When S_i is an estimated propensity score, we expect S_i to be larger for treated units and smaller for control units. As a result, (8) is most likely violated for tail events. In particular, we consider the collection of intervals $[0,x]$ and $[x,1]$ for any $x \in [0,1]$ and denote it by $\mathcal{A}$. Let

$$\nu_1 \triangleq \sup_{A\in\mathcal{A}} \frac{P_1(A)}{P_0(A)}, \quad \nu_0 \triangleq \sup_{A\in\mathcal{A}} \frac{P_0(A)}{P_1(A)}. \tag{9}$$

Then ν_1 and ν_0 can be viewed as discrepancies between P_1 and P_0 in tails. The $\mathcal{O}^*$-dependent bound can be directly derived from (8).

Theorem 3.3. *Let ν_1 and ν_0 be defined in* (9)*. Then*

$$\nu_1 \le b_{\max}(\mathcal{O}^*;\pi), \quad \nu_0 \le b_{\min}(\mathcal{O}^*;\pi)^{-1}.$$

Equivalently,

$$\mathcal{O}^* \le \mathcal{O}^+_{\mathrm{DiT}}(\nu_0,\nu_1;\pi) \triangleq \min\left\{\frac{\pi}{\pi+(1-\pi)\nu_0}, \frac{1-\pi}{1-\pi+\pi\nu_1}\right\}$$

Similar to DiM O-values, it remains to find simultaneous lower confidence bounds $\hat{\nu}_1^-$ and $\hat{\nu}_0^-$ such that

$$\mathbb{P}\left(\hat{\nu}_1^- \le \nu_1, \hat{\nu}_0^- \le \nu_0\right) \ge 1-\alpha. \tag{10}$$

To derive a confidence region for (ν_-,ν_+), we recall the well-known Dvoretzky-Kiefer-Wolfowitz-Massart (DKWM) inequality [Dvoretzky et al., 1956, Massart, 1990a].

Proposition 3.2 (DKWM inequality)**.** *Let $Z_1,\dots,Z_m$ be real-valued i.i.d. random variables with the cumulative distribution function (cdf) $F(z)$. Further let $F_m(z)$ be the empirical cdf. Then with probability $1-\delta$,*

$$\sup_{z\in\mathbb{R}} |F_m(z) - F(z)| \le \sqrt{\frac{\log(2/\delta)}{2m}}$$

Let $\hat{P}_1$ and $\hat{P}_0$ be the empirical measure of S_i's in the treatment and control group, respectively. Applying Proposition 3.2 to both P_1 and P_0, we obtain that

$$\max_{t\in\{0,1\}} \sup_{A\in\mathcal{A}} |\hat{P}_t(A) - P_t(A)| \le \sqrt{\frac{\log(4/\alpha)}{2n_t}} \quad \text{with probability } 1-\alpha.$$

On this event,

$$\nu_0 \ge \hat{\nu}_0^- \triangleq \sup_{A\in\mathcal{A}} \frac{\hat{P}_0(A) - \sqrt{\log(4/\alpha)/2n_0}}{\hat{P}_1(A) + \sqrt{\log(4/\alpha)/2n_1}}, \quad \nu_1 \ge \hat{\nu}_1^- \triangleq \sup_{A\in\mathcal{A}} \frac{\hat{P}_1(A) - \sqrt{\log(4/\alpha)/2n_1}}{\hat{P}_0(A) + \sqrt{\log(4/\alpha)/2n_0}}. \tag{11}$$

Therefore, $(\hat{\nu}_1^-, \hat{\nu}_0^-)$ satisfies (10). Inverting the bounds in Theorem 3.3, we obtain a closed-form expression of the DiT O-value.

Theorem 3.4. *Let $\hat{\nu}_1^-$ and $\hat{\nu}_0^-$ be defined in* (11) *and $\mathcal{O}^+_{\mathrm{DiT}}$ be defined in Theorem 3.3. Then $\hat{\mathcal{O}}_{\mathrm{DiT}} \triangleq \mathcal{O}^+(\hat{\nu}_0^-, \hat{\nu}_1^-;\pi)$ is a valid O-value.*

When applying the DiT O-value to test the strict overlap condition with $\mathcal{O} = \min\{\pi, 1-\pi\}$, or equivalently $P_1 = P_0$, the two-sample Kolmogorov-Smirnov's test uses the statistics $\sup_{x\in[0,1]} |\hat{F}_1(x) - \hat{F}_0(x)|$, where $\hat{F}_1$ and $\hat{F}_0$ are empirical cdfs of S_is for treated and control units. The test induced DiT O-value essentially uses the supremum of the log-difference of two empirical cdfs.

Although the DKWM inequality is the tightest bound for $|\hat{F}_m(x) - F(x)|$ in the literature, it is still not tight enough for DiT O-values because DiT O-values are useful only when there exists an event A that has relatively small probability under P_1 or P_0, for which the induced DKWM confidence bound can be trivial. In particular, when $\hat{F}_m(x) < \sqrt{\log(2/\delta)/2m}$, the lower confidence bound for $F(x)$ is 0. In Appendix C.3, we derive a sharper DiT O-value based on a hybrid of the DKWM inequality, Dempster's line-crossing probability [Dempster, 1959], and generalized Simes' inequality [Sarkar et al., 2008]. The techniques are proposed by Bates et al. [2021b] for outlier detection.

3.3 Difference-in-Ranks (DiR) O-value

Apart from the Student t-test and Kolmogorov-Smirnov test, the Mann-Whitney U-test, also known as the Wilcoxon rank-sum test, is another popular yet qualitatively different test for two-sample problems. Assuming no ties in $(S_i)_{i=1}^n$ (otherwise adding a tiny random noise to each S_i) and letting R_i denote the rank of S_i in ascending order, the Mann-Whitney U-statistic is defined as $\hat{U} = \sum_{T_i=1} R_i - n_1(n_1+1)/2$. When testing the equality null $P_1 = P_0$, the critical value can be found through permutations or by asymptotic normality theory.

However, in our case we do not have strict equality but a density ratio condition (2). Thus the standard critical value is invalid. To apply the idea to our problem, we use an equivalent formulation of the rank-sum statistic R. Since there is no tie,

$$R_i = 1 + \sum_{j=1}^{n} I(S_i > S_j). \tag{12}$$

As a consequence,

$$\hat{U} = \sum_{T_i=1, T_j=0} I(S_i > S_j). \tag{13}$$

It is well-known that [Hanley and McNeil, 1982] the rescaled Mann-Whitney U-statistic $\hat{U}/n_1 n_0$ is an unbiased estimator of the area under the receiver operating characteristic curve (AUROC) for thresholding classifiers $\{s \mapsto I(s \le c) : c \in [0,1]\}$. Intuitively, a larger overlap makes P_1 and P_0 less distinguishable, and thus drives the AUROC smaller. Let

$$p_* \triangleq \mathbb{P}\left(S^{(1)} > S^{(0)}\right), \quad \text{where } (S^{(1)}, S^{(0)}) \sim P_1 \otimes P_0, \tag{14}$$

where $\otimes$ denotes the Cartesian product. The following theorem quantifies the heuristics.

Theorem 3.5. *Let $S^{(1)}$ and $S^{(0)}$ be independent draws from P_1 and P_0 respectively. Assume that neither P_1 nor P_0 has a point mass (otherwise convolute them with a small random noise). Then*

$$\max\{p_*, 1-p_*\} \le \frac{1-\mathcal{O}^*}{2(1-2\mathcal{O}^*)}\left(2 - \frac{\mathcal{O}^*}{\pi(1-\pi)}\right).$$

When $\mathcal{O}^* = 1/2$, the bound is taken as $1/2$. Equivalently,

$$\mathcal{O}^* \le \mathcal{O}^+_{\mathrm{DiR}}(p_*;\pi) \triangleq \frac{1}{2} - \pi(1-\pi)|1-2p_*| - \sqrt{\frac{(1-2\pi)^2}{4} + \pi^2(1-\pi)^2(1-2p_*)^2}.$$

When $\mathcal{O}^* = \pi$, Theorem 3.5 implies that $\max\{p_*, 1-p_*\} \le 0.5$, and thus $p_* = 0.5$. This recovers the Mann-Whitney U-test whose critical value is based on this fact. Since (T_i, S_i)'s are i.i.d.,

$$\mathbb{E}[\hat{U}] = n(n-1)\mathbb{P}\left(S_i > S_j \mid T_i = 1, T_j = 0\right)\mathbb{P}(T_i = 1, T_j = 0) = n(n-1)p_*\pi(1-\pi). \tag{15}$$

It remains to construct lower and upper confidence bounds for $\mathbb{E}[\hat{U}]$.

Noting that the summands of $\hat{U}$ are dependent, we cannot apply the empirical Bernstein/Bennett inequality used in DiM O-values. Nonetheless, as suggested by its name, $\hat{U}$ can be formulated as a U-statistic of order 2 with a symmetric kernel:

$$\hat{U} = \sum_{i \neq j} T_i(1-T_j)I(S_i > S_j) = \sum_{i \neq j} \phi(Z_i, Z_j)$$

where $Z_i = (S_i, T_i)$ and

$$\phi(Z_i, Z_j) = \frac{T_i(1-T_j)I(S_i > S_j) + T_j(1-T_i)I(S_j > S_i)}{2}. \tag{16}$$

It is easy to see that $\phi(Z_i, Z_j) \in \{0, 0.5\}$ and thus ϕ is a bounded kernel. We can apply the empirical Bernstein inequality for U-statistics.

Proposition 3.3 (Peel et al. [2010], a slight modification[1] of Theorem 3 with $m = 2$)**.** *Let $Z_1, \dots, Z_n$ be i.i.d. and $\phi(z, z')$ be a symmetric bounded kernel taking values in $[0, 0.5]$. Let*

$$S = \frac{1}{n(n-1)} \sum_{i \neq j} \phi(Z_i, Z_j), \quad \hat{\sigma}^2 = \frac{1}{n(n-1)(n-2)(n-3)} \sum_{i_1 \neq i_2 \neq i_3 \neq i_4} (\phi(Z_{i_1}, Z_{i_2}) - \phi(Z_{i_3}, Z_{i_4}))^2.$$

Then with probability $1 - \delta$,

$$|S - \mathbb{E}[S]| \leq \hat{\sigma} \sqrt{\frac{4 \log\left(\frac{4}{\delta}\right)}{n}} + \frac{5 \log\left(\frac{4}{\delta}\right)}{n} \triangleq C_U(\hat{\sigma}, n; \delta).$$

Although $\hat{\sigma}^2$ involves $O(n^4)$ terms, we show that it is easy to compute in our case.

Lemma 3.1. $\hat{\sigma}^2 = \dfrac{(n-1)(n-4)\hat{U} - 2\hat{U}^2 + 2\sum_{i=1}^n \tilde{R}_i^2}{n(n-1)(n-2)(n-3)}$ *where $\tilde{R}_i$ is defined as*

$$\tilde{R}_i = \begin{cases} \sum_{T_j=0} I(S_i > S_j) & (\text{if } T_i = 1) \\ \sum_{T_j=1} I(S_j > S_i) & (\text{if } T_i = 0) \end{cases}. \tag{17}$$

Combining (15) and Proposition 3.3, we obtain simultaneous lower and upper confidence bounds of p_* as

$$\hat{p}^- \triangleq \max\left\{ \frac{\hat{U}}{n(n-1)\pi(1-\pi)} - \frac{C_U(\hat{\sigma}, n; \alpha)}{\pi(1-\pi)}, 0 \right\}, \ \hat{p}^+ \triangleq \max\left\{ \frac{\hat{U}}{n(n-1)\pi(1-\pi)} + \frac{C_U(\hat{\sigma}, n; \alpha)}{\pi(1-\pi)}, 1 \right\}. \tag{18}$$

This induces a valid lower confidence bound for $\max\{p_*, 1-p_*\}$ as $\max\{\hat{p}^-, 1-\hat{p}^+\}$.

Theorem 3.6. *Let $\hat{p}^-$ and $\hat{p}^+$ are defined in* (18) *and $\mathcal{O}^+_{\mathrm{DiR}}$ be defined in Theorem 3.5. Then $\hat{\mathcal{O}}_{\mathrm{DiR}} \triangleq \mathcal{O}^+_{\mathrm{DiR}}(\max\{\hat{p}^-, 1-\hat{p}^+\}; \pi)$ is a valid O-value.*

Under the null hypothesis $H_0 : \mathcal{O} = \pi$, $p_* = 0.5$ as shown by Theorem 3.5, and the Mann-Whitney U-test rejects the null if

$$\left| \frac{\hat{U}}{n_1 n_0} - \frac{1}{2} \right| \leq \sqrt{\mathrm{Var}\left(\frac{\hat{U}}{n_1 n_0} \right)} z_{\alpha/2} = \sqrt{\frac{(n+1)}{12 n_1 n_0}} z_{\alpha/2} \approx \sqrt{\frac{1}{12\pi(1-\pi)n}} z_{\alpha/2}, \tag{19}$$

where $z_{\alpha/2}$ is the $\alpha/2$-th quantile of the standard normal distribution.

If we use the induced test by the DiR O-value, H_0 is rejected if

$$\hat{p}^- \leq \frac{\pi(1-\pi)}{2} \leq \hat{p}^+ \iff \left| \frac{\hat{U}}{n(n-1)\pi(1-\pi)} - \frac{1}{2} \right| \leq \frac{C_U(\hat{\sigma}, n; \alpha)}{\pi(1-\pi)}.$$

[1]The original result is stated for kernels bounded by $[0, 1]$. This version is obtained by a simple rescaling argument.

Since $n_1/n \approx \pi, n_0/n \approx 1-\pi$, this rejection region is approximately

$$\left|\frac{\hat{U}}{n_1 n_0} - \frac{1}{2}\right| \leq \frac{\hat{\sigma}}{\pi(1-\pi)}\sqrt{\frac{4\log\left(\frac{4}{\alpha}\right)}{n}}. \tag{20}$$

By (19) and Chebyshev's inequality,

$$\frac{\hat{U}}{n_1 n_0} = \mathbb{E}\left[\frac{\hat{U}}{n_1 n_0}\right] + O_{\mathbb{P}}\left(\sqrt{\mathrm{Var}\left[\frac{\hat{U}}{n_1 n_0}\right]}\right) = \frac{1}{2} + O_{\mathbb{P}}\left(\sqrt{\frac{1}{n}}\right).$$

and by Lemma 3.1,

$$\hat{\sigma}^2 \approx \frac{\hat{U}}{n^2} - \frac{2\hat{U}^2}{n^4} \approx \frac{\hat{U}\pi(1-\pi)}{n_1 n_0} - 2\left(\frac{\hat{U}\pi(1-\pi)}{n_1 n_0}\right)^2 \approx \frac{\pi(1-\pi)(1-\pi(1-\pi))}{2},$$

where the last term in Lemma 3.1 is ignored because $\sum_{i=1}^n \tilde{R}_i^2 \leq n^3 = o(n^4)$. As a result, the rejection region (20) is approximately

$$\left|\frac{\hat{U}}{n_1 n_0} - \frac{1}{2}\right| \leq \sqrt{\frac{2(1-\pi(1-\pi))\log\left(\frac{4}{\alpha}\right)}{\pi(1-\pi)n}}.$$

When $\alpha = 0.05$ and $\pi = 0.5$, the rejection threshold of $\sqrt{n}|\hat{U}/n_1 n_0 - 1/2|$ is 1.132 for the Mann-Whitney U-test and 5.128 for the DiR O-value induced test. Again, the constants of empirical Bernstein's inequality are unsatisfactory. In Appendix C.2, we will derive a sharper DiR O-value based on the Hoeffding-Bentkus-Maurer inequality [Bates et al., 2021a].

3.4 Classification-error (CE) O-value

In the special case where $\mathcal{O}^* = \pi$, $e(X) \equiv \pi$ almost surely, T is independent of X. As a result, the permutation test with any test statistic is valid in finite samples if T is permuted. In particular, the classification permutation test [Gagnon-Bartsch et al., 2019] uses the classification accuracy as the test statistic and rejects the null if it is significantly smaller than the typical values of the accuracy obtained with T randomly permuted.

In our case, the permutation test is no longer valid because the independence between T and X may fail to hold if $\mathcal{O}^* < \pi$. Nonetheless, D'Amour et al. [2017] derive another important implication of the strict overlap condition, which states that no classifier can achieve an expected accuracy higher than $1 - \mathcal{O}^*$ if X is used as covariates to predict T. The proof is based on the simple fact that the Bayes optimal classifier is $I(e(x) \geq 0.5)$ [Devroye et al., 2013], for which the expected classification error is $\mathbb{E}[\min\{e(X), 1-e(X)\}]$, which is lower bounded by $\mathcal{O}^*$. Intuitively, the optimal classification accuracy measures the discrepancy between P_1 and P_0.

Given the estimated propensity scores, we can classify T by a threshold learner $I(S > \eta)$ for some η. The above result shows that the expected classification error is at least $\mathcal{O}^*$.

Theorem 3.7. *Let $\mathcal{E}(\eta) \triangleq \mathbb{P}[T = I(S > \eta)]$. Then $\mathcal{E}(\eta) \leq 1 - \mathcal{O}^*$ for any $\eta \in [0,1]$. Equivalently, $\mathcal{O}^* \leq \mathcal{O}^+_{\mathrm{CE}} \triangleq 1 - \mathcal{E}_{\max}$, where $\mathcal{E}_{\max} = 1 - \sup_{\eta\in[0,1]} \mathcal{E}(\eta)$.*

Recalling that S is the estimated propensity score, the above class of threshold learners is quite general because most classifiers essentially threshold an estimate of $\mathbb{P}(T = 1 \mid X = x)$. The threshold can be chosen by minimizing $\hat{\mathcal{E}}(\eta)$, the classification error on the validation set with threshold η. A lower confidence bound for $\mathcal{E}_{\max}$ can be obtained based on $\hat{\mathcal{E}}_{\max} \triangleq \max_\eta \hat{\mathcal{E}}(\eta)$ by controlling the maximum gap between $\hat{\mathcal{E}}(\eta)$ and $\mathcal{E}(\eta)$. Since the class of threshold learners has a Vapnik-Chervonenkis dimension 1, we can use the empirical process theory to bound the discrepancy. In particular, we consider the following version by Vapnik [1995] which has tighter constants than most of the textbook versions.

Proposition 3.4 (adapted from Vapnik [1995], Section 3.4, (3.15)). $\mathbb{P}(\hat{\mathcal{E}}^-_{\max} \le \mathcal{E}_{\max}) \ge 1-\alpha$ *where*

$$\hat{\mathcal{E}}^-_{\max} \triangleq \hat{\mathcal{E}}_{\max} - \sqrt{\frac{\log(2n+1) + \log\left(\frac{4}{\alpha}\right)}{n}}.$$

Proposition 3.4 yields a valid O-value $\hat{\mathcal{O}} = 1 - \hat{\mathcal{E}}^-_{\max}$, which we refer to as the CE O-value. It is attempting to remove the logarithmic factor using the chaining argument [Talagrand, 2006], resulting in a bound $C\sqrt{\log(1/\alpha)/n}$. However, we are not aware of any version of this kind with a satisfactory constant C that outperforms Proposition 3.4 for reasonable sample sizes, say $n \le 10^{10}$. This makes a case where the constant matters more than the convergence rate. Indeed, we discuss a sharper CE O-value using the techniques for the DiT O-value in Appendix C.4.

3.5 Nuisance parameter and derandomization

In previous subsections, we assume π is known while in practice it is unknown in general. Among the four types of O-values we discussed, only the CE O-value does not depend on π while the DiM, DiT, and DiR O-values depend on π in complicate ways. Let $\hat{\mathcal{O}}(\pi;\alpha)$ denote a generic valid O-value that depends on π with level α. If we can find a $(1-\gamma)$ confidence interval $[\hat{\pi}^-, \hat{\pi}^+]$ for π for a $\gamma < \alpha$, then

$$\hat{\mathcal{O}} \triangleq \sup_{\pi \in [\hat{\pi}^-, \hat{\pi}^+]} \hat{\mathcal{O}}(\pi; \alpha - \gamma) \tag{21}$$

is a valid O-value with level α via a simple union bound. In our implementation, we use the Clopper-Pearson bound with $\gamma = 0.1\alpha$ because the interval is not sensitive to γ.

The data splitting step or the usage of a randomized algorithm to fit propensity scores, e.g. random forest, introduces extra unwanted randomness into O-values. To derandomize O-values, we can repeat the procedure for B times with level $\alpha/2$ and then report the median. The following proposition shows that the resulting O-value remains valid in finite samples.

Proposition 3.5. *Let $\hat{\mathcal{O}}^{(1)}, \ldots, \hat{\mathcal{O}}^{(B)}$ are B O-values with level $q\alpha$ for some $q \in (0,1)$, which can be arbitrarily dependent. Define $\hat{\mathcal{O}}$ be qth lower quantile of $\{\hat{\mathcal{O}}^{(1)}, \ldots, \hat{\mathcal{O}}^{(B)}\}$, i.e.*

$$\hat{\mathcal{O}} \triangleq \inf\left\{ \mathcal{O} : \frac{1}{B}\sum_{b=1}^{B} I(\hat{\mathcal{O}}^{(b)} \le \mathcal{O}) \ge q \right\}.$$

Then $\hat{\mathcal{O}}$ is a valid O-value with level α.

Conditional on the data, the O-values in each run are i.i.d.. Therefore, when the $B \to \infty$, the median converges to a deterministic quantity that only depends on the observations and the resulting O-value is purely deterministic.

3.6 Approximate O-values in large samples

The O-values discussed in Section 3 are intrinsically conservative for two reasons. First they estimate upper bounds of $\mathcal{O}^*$, e.g. $\mathcal{O}^+_{\mathrm{DiM}}(T_0, T_1; \pi)$ defined in Theorem 3.2, and the gap could be potentially large. Second, the O-value incorporates the finite sample uncertainty by underestimating the determinants (e.g. T_0, T_1) of the upper bound. Since the aforementioned O-values are valid in finite samples uniformly for all propensity score estimators, the second source of conservatism is non-negligible in order to handle small sample sizes and poor estimators.

Given an upper bound of $\mathcal{O}^*$, we may derive a less conservative O-value by being more optimistic in the finite sample uncertainty if the sample size is large and the propensity score estimator is relatively well-behaved. In particular, we consider the case where $\hat{e}(X)$ is uniformly consistent in the sense that there exists a function $e^*(x)$, which is not necessarily the true propensity score $e(x)$, such that

$$\sup_x |\hat{e}(x) - e^*(x)| \xrightarrow{p} 0 \quad \text{as the size of the training set goes to infinity,} \tag{22}$$

where $\xrightarrow{p}$ denotes the convergence in probability. This condition is much weaker than the consistency. When $\hat{e}(x) = f(x;\hat{\theta})$ for some parametric family $f(x;\theta)$ where $\hat{\theta} = \arg\min_\theta (1/n)\sum_{i=1}^n \ell(T_i, f(X_i,\theta))$ for some loss function ℓ, the classical quasi-likelihood theory [White, 1982] shows that $\hat{\theta}$ converges to $\theta^* = \arg\min_\theta \mathbb{E}[\ell(T, f(X,\theta))]$ in probability under mild regularity conditions. The condition (22) is then satisfied if f is uniformly continuous in θ, even if the parametric model is misspecified. For general nonparametric estimators, (22) continues to hold with e^* identified as the "projection" onto the specified model, provided that the model class is relative simple; see Pollard [1990] for more details.

Under (22), the naive eyeball test still does not work since $e^*(x)$ may be drastically different from $e(x)$. Nonetheless, it ensures that $S_i \approx e^*(X_i)$ uniformly for all i with high probability. If $S_i = e^*(X_i)$ for any deterministic function e^*, $\mathcal{O}^+_{\mathrm{DiM}}$ (Theorem 3.1), $\mathcal{O}^+_{\mathrm{DiT}}$ (Theorem 3.3), $\mathcal{O}^+_{\mathrm{DiR}}$ (Theorem 3.5), and $\mathcal{O}^+_{\mathrm{CE}}$ (Theorem 3.7) are still legitimate upper bounds of $\mathcal{O}^*$ since the theorems do not depend on the specification of e^*. When $S_i \approx e^*(X_i)$, with some mild regularity conditions, we can show that the plug-in estimates of the upper bounds are approximate O-values in the sense that

$$\mathbb{P}(\mathcal{O}^* - \epsilon \le \hat{\mathcal{O}}) \to 1$$

for any $\epsilon > 0$ as the total sample size tends to infinity.

Theorem 3.8. *Let $\hat{\pi} = n_1/n$. Assume that $\pi \in (0,1)$ and $\hat{e}(x)$ satisfies* (22).

- *(Approximate DiM O-value) Let $\zeta > 0$ be any constant. Assume that $\mathrm{Var}[e^*(X) \mid T = 1], \mathrm{Var}[e^*(X) \mid T = 0] > 0$ or $\mathbb{E}[e^*(X) \mid T = 1] \neq \mathbb{E}[e^*(X) \mid T = 0]$. Then $\hat{\mathcal{O}}^{\approx}_{\mathrm{DiM}}$ is an approximate O-value where*

$$\hat{\mathcal{O}}^{\approx}_{\mathrm{DiM}} \triangleq \mathcal{O}^+_{\mathrm{DiM}}(\hat{T}_0, \hat{T}_1; \hat{\pi}), \quad \hat{T}_1 = \frac{|\hat{\mu}_1 - \hat{\mu}_0|}{\hat{\sigma}_1}, \quad \hat{T}_0 = \frac{|\hat{\mu}_1 - \hat{\mu}_0|}{\hat{\sigma}_0}.$$

- *(Approximate DiR O-value) Let $\hat{U}$ be defined as in* (13). *If the distribution of $e^*(X)$ has no point mass, then $\hat{\mathcal{O}}^{\approx}_{\mathrm{DiR}}$ is an approximate O-value where*

$$\hat{\mathcal{O}}^{\approx}_{\mathrm{DiR}} \triangleq \mathcal{O}^+_{\mathrm{DiR}}(\hat{U}; \hat{\pi}).$$

- *(Approximate CE O-value) Let $\hat{\mathcal{E}}(\eta)$ be defined as in Theorem 3.7. $\mathcal{E}^*_t(\eta) = \mathbb{P}(T = I(e^*(X) > \eta), T = t)$. If $\mathcal{E}^*_t(\eta)$ has a uniformly bounded derivative for both $t = 0, 1$, then $\hat{\mathcal{O}}^{\approx}_{\mathrm{CE}}$ is an approximate O-value where*

$$\hat{\mathcal{O}}^{\approx}_{\mathrm{CE}} \triangleq 1 - \sup_{\eta \in [0,1]} \hat{\mathcal{E}}(\eta).$$

Theorem 3.8 does not include an approximate DiT O-value because $\mathcal{O}^+_{\mathrm{DiT}}$ is characterized by density ratios where the denominators can approach zero no matter how large the training set is. Put another way, unlike the other upper bounds, $\mathcal{O}^+_{\mathrm{DiT}}$ is an irregular estimator, and thus the plug-in estimator is unreliable.

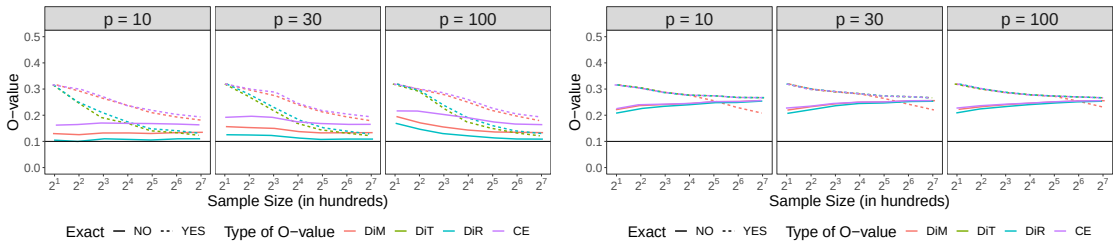
4 Empirical Demonstration

4.1 An illustrative simulation study

We compare different O-values on an illustrative synthetic dataset. In particular, we consider a simple data generating process where $X_i \overset{i.i.d.}{\sim} N(0, I_p)$ with $p \in \{10, 30, 100\}$. The propensity score function is designed as

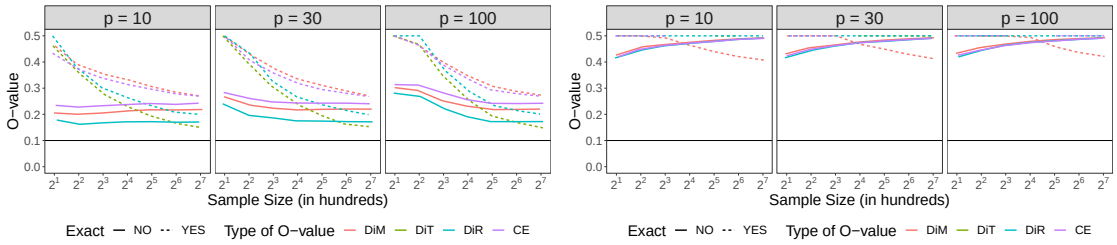
$$e(x) = f(x^T\beta), \quad f(y) = \begin{cases} 0.1 & (y < c_1) \\ 0.9 & (y > c_2) \end{cases}, \ f(y) \text{ is linear in } [c_1, c_2],$$

where $\beta_1 = \cdots = \beta_{10} = 1, \beta_{11} = \ldots = \beta_p = 0$, c_1 is chosen such that $\mathbb{P}[e(X) = 0.1] \in \{0.2, 0.8\}$, and c_2 is chosen such that $\mathbb{P}[e(X) = 0.9] = 0.2$. Clearly, the population overlap slack $\mathcal{O}^* = 0.1$.



(a) Gradient Boosting (b) Random Forest

Figure 1: Comparison of refined and approximate O-values when $\mathbb{P}[e(X) = 0.1] = 0.8$. Each line plots the 95% quantile across 50 replicates.



(a) Gradient Boosting (b) Random Forest

Figure 2: Comparison of refined and approximate O-values when $\mathbb{P}[e(X) = 0.1] = 0.2$. Each line plots the 95% quantile across 50 replicates.

The case with $\mathbb{P}[e(X) = 0.1] = 0.2$ is harder than the case with $\mathbb{P}[e(X) = 0.1] = 0.8$, because $\mathbb{P}[e(X) \in \{\mathcal{O}^*, 1-\mathcal{O}^*\}] = 0.4$ in the former while $\mathbb{P}[e(X) \in \{\mathcal{O}^*, 1-\mathcal{O}^*\}] = 1$ in the latter. For each sample size $n \in 100 \times \{2^1, 2^2, \ldots, 2^6\}$, we generate 50 independent datasets $\{(T_i, X_i) : i = 1, \ldots, n\}$.

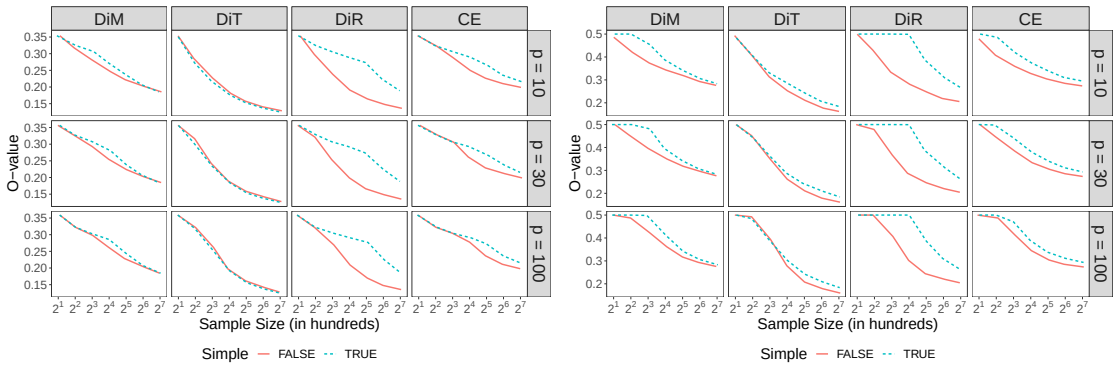
For each dataset, we compute the DiM, DiT, DiR, and CE O-values with $\alpha = 0.05$. We apply the gradient boosting and random forest as the learner to estimate propensity scores. For comparison, we compute the simple versions (Section 3), the refined versions (Appendix C), and the approximate versions (Section 3.6) of O-values. The O-values are implemented in the `R ovalue` package, available at `https://github.com/lihualei71/ovalue`. The programs to reproduce the simulation results are available at `https://github.com/lihualei71/ovaluePaper`.

Figure 1 and Figure 2 present the 95% quantile of the refined and approximate O-values, which is supposed to be at least $\mathcal{O}^* = 0.1$ when the O-value is valid. Unsurprisingly, all exact O-values are valid as guaranteed by our theory. The approximate O-values are also valid in this example even with a sample size 200, and they are generally tighter than the exact ones. As expected, the O-values are tighter when a larger fraction of propensity scores are taking the extreme values. In all settings, the gradient boosting yields substantially more powerful O-values than the random forest. With the gradient boosting, the DiT O-value is the best exact one and the DiR O-value is the best approximate one.

Figure 3 displays the mean of simple and refined versions of all O-values with gradient boosting. Clearly, the refined techniques yield tighter O-values, especially when the sample size is moderate.

4.2 Real data analysis

O-values can be computed whenever an estimate of the propensity score function is available. Therefore, we can evaluate them on almost any real-world dataset. For demonstration, we compute the O-values on the Lalonde data, initially collected by LaLonde [1986] and later reanalyzed by Dehejia and Wahba [2002]. The dataset is used to evaluate the effect of National Supported Work Demonstration on earnings. It includes a treated group and a control group from a randomized experiment,



(a) $\mathbb{P}[e(X) = 0.1] = 0.8$ (b) $\mathbb{P}[e(X) = 0.1] = 0.2$

Figure 3: Comparison between simple and refined (exact) O-value with gradient boosting. Each line plots the average over 50 replicates.

together with three external control groups from the Current Population Survery (CPS) and three external control groups from the Panel Study of Income Dynamics (PSID). Here, we use the data curated by Dehejia and Wahba [2002], which involves 185 treated units and 8 covariates: age, education, Black (1 if black, 0 otherwise), Hispanic (1 if Hispanic, 0 otherwise), married (1 if married, 0 otherwise), nodegree (1 if no degree, 0 otherwise), earnings in 1974 and 1975. The sample sizes of the control groups are summarized in Table 1. For each control group, Table 1 presents the exact DiT O-value and the approximate DiR O-value with $\alpha = 0.05$ and gradient boosting as the learner to estimate propensity scores. Other types of O-values with other propensity score estimators are presented in Appendix D. To reduce the algorithmic randomness, we report the median O-value from 1000 independent data splits.

	CPS			PSID			RCT		
	n_0	DiT	DiR (app.)	n_0	DiT	DiR (app.)	n_0	DiT	DiR (app.)
Raw	15992	0.0026	0.0006	2490	0.0176	0.0036	260	0.4831	0.3865
V2	2369	0.0213	0.0079	253	0.2336	0.0750			
V3	429	0.1428	0.0691	128	0.3133	0.1002			

Table 2: Sample sizes, exact DiT O-values, and approximate DiR O-values for each control group in Lalonde data. Both O-values use the gradient boosting to estimate propensity scores.

Note that the population overlap slack is not scale-free in that $\mathcal{O}^* \leq \pi$. As a result, when the treated and control groups are highly imbalanced in sizes, even a completely randomized experiment would have a small population overlap slack. Hong et al. [2020] showed that the convergence rate for the usual ATE estimator is $\sqrt{n\mathcal{O}^*}$ without additional restrictions on the outcome. Therefore, $n\mathcal{O}^*$ can be viewed as an effective sample size. If the data had been drawn from a completely randomized experiment, the effective sample size would have been $n \min\{\pi, 1-\pi\}$. Therefore, we can consider the normalized population overlap slack $\mathcal{O}^* / \min\{\pi, 1-\pi\}$, which measures the efficiency relative to the most overlapped case, i.e., the completely randomized experiment. A natural estimator of the relative efficiency is $n\hat{\mathcal{O}} / \min\{n_1, n_0\}$. Here, we compute estimates of the efficiency loss $\hat{L} = \max\{0, 1 - n\hat{\mathcal{O}} / \min\{n_1, n_0\}\}$. Table 3 summarizes the results for exact DiT O-values and approximate DiR O-values. Since $\hat{\mathcal{O}}$ is an upper confidence bound of $\mathcal{O}^*$, $\hat{L}$ is an optimistic assessment of the efficiency loss caused by lack of overlap.

Clearly, the external control groups suffer from substantial lacks of overlap. The numbers show that both the exact and approximate O-values are able to detect a fair amount of nonoverlap, though they are obtained via a partial identification approach, though the approximate O-values are substantially more powerful. On the other hand, despite the smaller sample size, the randomized

	CPS		PSID		RCT	
	DiT	DiR (app.)	DiT	DiR (app.)	DiT	DiR (app.)
Raw	76.9%	94.7%	74.7%	94.8%	0%	6.9%
V2	70.8%	89.4%	44.7%	82.2%		
V3	52.5%	77.2%	22.8%	75.1%		

Table 3: Estimated efficiency loss yielded by the exact DiT O-values and approximate DiR O-values. The other details are the same as in Table 2.

experiment has the highest effective sample size. Unsurprisingly, the DiT O-value detects no efficiency loss in this case. By contrast, the DiR O-value detects a small efficiency loss. It could either be finite sample errors or potential undocumented noncompliance that makes the experiment less randomized than desired.

5 Other Measures of Overlap

5.1 Population overlap slack for ATT and ATC

The population overlap slack introduced in Section 2 is tied to ATE. In many applications, the inferential target is the average treatment effect on the treated (ATT), especially when the treated group and the control group differ substantially in sizes. ATT is easier to infer than ATE because it requires a weaker overlap condition. Specifically, the strict overlap condition for ATT is stated as $e(X) \le 1 - \mathcal{O}$ almost surely. Thus, a natural population overlap slack for ATT is defined as follows:

$$\mathcal{O}^*_{\text{ATT}} = \sup\{\mathcal{O} : e(X) \le 1 - \mathcal{O},\ a.s.\}.$$

Analogous to (2), we can show a one-sided likelihood ratio condition,

$$0 \le \frac{dP_1}{dP_0}(S) \le \frac{1-\pi}{\pi}\frac{1-\mathcal{O}^*_{\text{ATT}}}{\mathcal{O}^*_{\text{ATT}}} \quad a.s.. \tag{23}$$

This can be viewed as the likelihood condition for ATE with the same $b_{\max}(\mathcal{O}^*;\pi)$ and $b_{\min}(\mathcal{O}^*;\pi) = 0$. For DiM O-values, the upper bound for T_1 in Theorem 3.1 becomes vacuous because $b_{\min}(\mathcal{O}^*;\pi)^{-1} = \infty$. However, the bound for T_0 remains meaningful.

By symmetry, we can also define the population overlap slack for the average treatment effect on the controls (ATC):

$$\mathcal{O}^*_{\text{ATC}} = \sup\{\mathcal{O} : e(X) \ge \mathcal{O},\ a.s.\}.$$

Theorem 5.1 below shows that the upper bounds $\mathcal{O}^+_{\text{DiM}}$, $\mathcal{O}^+_{\text{DiT}}$ and $\mathcal{O}^+_{\text{DiR}}$ can be extended to $\mathcal{O}^*_{\text{ATT}}$ and $\mathcal{O}^*_{\text{ATC}}$.

Theorem 5.1. *With the same notation as in Theorem 3.1, 3.3, and 3.5,*

- *(Analogue of $\mathcal{O}^+_{\text{DiM}}$ for ATT and ATC)*

$$\mathcal{O}^*_{ATT} \le \frac{1-\pi}{1+\pi T_0^2}, \quad \mathcal{O}^*_{ATC} \le \frac{\pi}{1+(1-\pi)T_1^2}.$$

- *(Analogue of $\mathcal{O}^+_{\text{DiT}}$ for ATT and ATC)*

$$\mathcal{O}^*_{ATT} \le \frac{1-\pi}{1-\pi+\pi\nu_1}, \quad \mathcal{O}^*_{ATC} \le \frac{\pi}{\pi+(1-\pi)\nu_0}.$$

- *(Analogue of $\mathcal{O}^+_{\text{DiR}}$ for ATT and ATC)*

$$\mathcal{O}^*_{ATT} \le \frac{2(1-\pi)(1-p_*)}{2(1-\pi)(1-p_*)+\pi}, \quad \mathcal{O}^*_{ATC} \le \frac{2\pi p_*}{2\pi p_* + 1 - \pi}.$$

Remark 1. *Although $\mathcal{O}^* = \min\{\mathcal{O}^*_{ATT}, \mathcal{O}^*_{ATC}\}$, the DiM and DiR bound for $\mathcal{O}^*$ are tighter than the minimum of the bounds for $\mathcal{O}^*_{ATT}$, $\mathcal{O}^*_{ATC}$, reflecting that using the two-sided density ratio bound is more effective than using two one-sided density ratio bounds separately.*

Based on Theorem 5.1, we can construct approximate O-values as in Theorem 3.8 and exact O-values using the same confidence bounds for T_0, T_1, ν_0, ν_1, and p_* as discussed in Appendix C. These O-values for ATT and ATC are also implemented in the `ovalue` package.

Similar to the O-values for ATE, we can normalize the O-values for ATT and ATC. For completely randomized experiments, the population overlap slacks for ATT and ATC are $1 - \pi$ and π, respectively. Therefore, we can normalize the O-values for ATT and ATC by $1 - \hat{\pi}$ and $\hat{\pi}$, respectively, where $\hat{\pi} = n_1/n$. Again, one minus the normalized O-value can be interpreted as the efficiency loss relative to a completely randomized experiment.

For Lalonde data, ATT is more often the inferential target than ATE. Table 4 presents the estimated efficiency loss for ATT, yielded by exact DiT O-value and approximate DiR O-value as in Table 3. As with ATE, the O-values show strong evidence on the lack of overlap when external control groups are used. Other types of O-values for ATT and ATC are presented in Appendix D.

	CPS		PSID		RCT	
	DiT	DiR (app.)	DiT	DiR (app.)	DiT	DiR (app.)
Raw	36.3%	17.2%	59.3%	57%	0%	4.7%
V2	44.6%	39.4%	43.5%	69.4%		
V3	33.9%	55.2%	21.7%	68.1%		

Table 4: Estimated efficiency loss yielded by the exact DiT O-values and approximate DiR O-values for ATT. The other details are the same as in Table 3.

5.2 Quantile overlap slack

As alluded to earlier in Section 1, the population overlap slack is overly conservative when only a small fraction of propensity scores are in the tails. When the potential outcomes are bounded, dropping a handful of units with extreme propensity scores would only introduce a tolerable bias to the ATE estimator. If the remaining samples satisfy a decent strict overlap condition, the inference is still approximately valid. Therefore, a less conservative measure of the population overlap is the quantile of the propensity score distribution. For ATE, we can define the quantile overlap slack as

$$\mathcal{O}^*_\gamma = \text{Quantile}_\gamma(\min\{e(X), 1 - e(X)\}).$$

Then the population overlap slack is essentially $\mathcal{O}^*_1$. As argued above, $\mathcal{O}^*_{0.95}$ might be more informative than $\mathcal{O}^*_1$.

A natural upper bound for $\mathcal{O}^*_\gamma$ can be derived based on $\mathcal{O}^+_{\text{CE}}$. As discussed in Section 3.4, the expected classification error of the Bayes optimal classifier is $\mathbb{E}[\min\{e(X), 1-e(X)\}]$. By definition, the integrand is lower bounded by $\mathcal{O}^*_\gamma$ when $\min\{e(X), 1-e(X)\}$ lies below its γ quantile. Therefore,

$$\mathbb{E}[\min\{e(X), 1 - e(X)\}] \geq \gamma \mathcal{O}^*_\gamma.$$

Analogous to Theorem 3.7, we can derive the following estimable upper bound for $\mathcal{O}^*_\gamma$.

Theorem 5.2. *With the same notation as in Theorem 3.7, for any $\gamma \in (0, 1]$,*

$$\mathcal{O}^*_\gamma \leq \gamma^{-1}(1 - \mathcal{E}_{\max}).$$

Theorem 5.2 implies that inflating a valid (approximate) CE O-value by a factor of $(\gamma^{-1} - 1)$ yields a valid (approximate) O-value for the γ-th quantile overlap slack. When $\gamma = 0.95$, the inflation factor is small.

The upper bounds for DiM, DiT, and DiR O-values can also be modified for the quantile overlap slack, though with substantially more complex adjustments. We leave them to future work.

For a completely randomized experiment, the quantile overlap slack is $\min\{\pi, 1-\pi\}$. Therefore, we can apply the same normalization as the O-values for population overlap slack for ATE. Table 5 presents the estimated efficiency loss yielded by exact and approximate CE O-values for $\mathcal{O}^*_{0.95}$. We observe that the exact O-values are much less powerful than those in Table 3. This is not surprising because CE O-values are not powerful for $\mathcal{O}^*_1$. Nevertheless, the approximate O-values still show a substantial lack of overlap even after dropping 5% units, especially when the control group is from PSID. The O-values with other estimator of propensity scores can be found in Appendix D.

	CPS		PSID		RCT	
	CE	CE (app.)	CE	CE (app.)	CE	CE (app.)
Raw	0%	19.5%	15.5%	56.1%	0%	0%
V2	0%	32%	35%	65.1%		
V3	9.8%	46.9%	18.5%	58%		

Table 5: Estimated efficiency loss yielded by the exact and approximate CE O-values for quantile population slack. The other details are the same as in Table 3.

5.3 Measure of overlap under additional outcome restrictions

The population overlap slack and quantile overlap slack are designed for extrapolation-free inference, that is, inference without modelling assumptions on the outcomes. When an outcome model is assumed, the treatment effect for a group without treated units can be inferred by extrapolation from a similar group with sufficient overlap. Therefore, the overlap condition can be weakened when extrapolation is possible. D'Amour et al. [2017] provide a detailed discussion on the tradeoff between outcome models and ovelap.

Here, we discuss the partially linear model, which is widely studied in the literature [e.g. Robinson, 1988, Chernozhukov et al., 2016]. The model assumes that the observed outcome Y follows

$$Y = T\tau + m(X) + \epsilon$$

where τ is the constant treatment effect, $m(X)$ is a unknown function of X, and ϵ is the stochastic error. Under this model, the overlap is measured by

$$\mathcal{O}^*_{\mathrm{PL}} \triangleq \mathbb{E}[\mathrm{Var}(T \mid X)].$$

Indeed, the semiparametric efficiency bound for ATE under this model is inversely proportional to $\mathcal{O}^{*2}_{\mathrm{PL}}$ [Bhattacharya and Zhao, 1997].

For a binary treatment, $\mathrm{Var}(T \mid X) = e(X)(1-e(X))$ and thus

$$\mathcal{O}^*_{\mathrm{PL}} = \mathbb{E}[e(X)(1-e(X))].$$

The following Theorem presents two estimable upper bounds of $\mathcal{O}^*_{\mathrm{PL}}$.

Theorem 5.3. *(1) Quasi-Variance (QVar) bound:* $\mathcal{O}^*_{PL} \le \mathbb{E}[(T-S)^2]$.

(2) With the same notation as in Theorem 3.7, $\mathcal{O}^*_{PL} \le 1 - \mathcal{E}_{\max}$.

For Quasi-Variance upper bound, we can apply the hedged capital bound detailed in Appendix C.1 to derive a lower confidence bound for $\mathbb{E}[(T-S)^2]$, which yields an upper confidence bound for $\mathcal{O}^*_{\mathrm{PL}}$. The second upper bound is identical to $\mathcal{O}^+_{\mathrm{CE}}$, implying that a CE O-value is valid upper confidence bound for $\mathcal{O}^*_{\mathrm{PL}}$. For both bounds, we can construct approximate O-values using the plug-in estimates for $\mathbb{E}[(T-S)^2]$ and $\mathcal{E}_{\max}$.

Clearly, $\mathcal{O}^*_{\text{PL}} = \pi(1-\pi)$ for a completely randomized experiment. Therefore, we can normalize the O-value for $\mathcal{O}^*_{\text{PL}}$ by $\hat{\pi}(1-\hat{\pi})$. Table 6 presents the estimated efficiency loss yielded by exact and approximate QVar O-values. Again, these O-values show strong evidence on the lack of overlap with observational control groups even under a partial linear models. Other types of O-values for $\mathcal{O}^*_{\text{PL}}$ are presented in Appendix D.

	CPS		PSID		RCT	
	QVar	QVar (app.)	QVar	QVar (app.)	QVar	QVar (app.)
Raw	56.5%	35.8%	59.6%	62.9%	0%	0%
V2	48.7%	45.8%	52.6%	54.5%		
V3	39.3%	46.6%	54.3%	47.4%		

Table 6: Estimated efficiency loss yielded by the exact and approximate QVar O-values for partial linear models. The other details are the same as in Table 3.

6 Conclusion

In this paper, we introduce the population overlap slack and propose four different types of O-values — the Difference-in-Means (DiM), Difference-in-Tails (DiT), Difference-in-Ranks (DiR), and Classification-error (CE) O-values. These O-values are all valid upper confidence bounds for the population overlap slack in finite samples, without any distributional assumption other than operating on i.i.d. samples. The O-values are built on estimated propensity scores but the validity does not impose any requirement on the estimation (e.g., consistency). As a consequence, O-values can wrap aroung any black-box machine learning algorithm to estimate propensity scores without hurting the validity. A general strategy to construct a valid O-value proceeds as (1) choosing a measure of discrepancy between the distributions of estimated propensity scores in the treated and control groups, (2) deriving an estimable bound as a function of the population overlap slack in theory, and (3) inverting the function at a lower confidence bound for the discrepancy.

There are at least two ways to further improve O-values — tighter upper bounds for the population overlap slack, and tighter concentration inequalities for the discrepancy measures. The former is related to partial identification, though our strategies to derive the bounds are relatively non-standard. It would be interesting to explore whether our problem resembles other partial identification problems. The latter is related to many areas outside causal inference. For example, the hedged capital bound used for DiM O-values was initially developed for sequential inference, and the generalized Simes' inequality used for DiT O-values was initially developed for multiple testing.

On the other hand, if the finite sample validity is not of primary concern, approximate O-values might be more appropriate because they are likely less conservative. In our paper, we propose the zero-order approximation by completely ignoring the uncertainty of the discrepancy estimates. It is also reasonable to consider first-order approximations by replacing the concentration inequalities, which are valid in finite samples, by large sample distributional approximations such as the central limit theorem.

References

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I Jordan. Distribution-free, risk-controlling prediction sets. *arXiv preprint arXiv:2101.02703*, 2021a.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *arXiv preprint arXiv:2104.08279*, 2021b.

WU Behrens. Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. *Landwirtschaftliche Jahrbücher*, 68:807–837, 1929.

Vidmantas Bentkus et al. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004.

PK Bhattacharya and Peng-Liang Zhao. Semiparametric inference in a partial linear model. *The annals of statistics*, pages 244–262, 1997.

Hao Chen and Dylan S Small. New multivariate tests for assessing covariate balance in matched observational studies. *arXiv preprint arXiv:1609.03686*, 2016.

Xiaohong Chen, Han Hong, Alessandro Tarozzi, et al. Semiparametric efficiency in gmm models with auxiliary data. *The Annals of Statistics*, 36(2):808–843, 2008.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/Debiased Machine Learning for Treatment and Causal Parameters. 2016. ISSN 13684221. doi: 10.1920/wp.cem.2016.4916. URL `http://arxiv.org/abs/1608.00060`.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009. ISSN 00063444. doi: 10.1093/biomet/asn055.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*, 2017.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

AP Dempster. Generalized d_n^+ statistics. *Ann. Math. Stat.*, 30(2):593–597, 1959.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Aryeh Dvoretzky, Jack Kiefer, Jacob Wolfowitz, et al. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Ronald A Fisher. The fiducial argument in statistical inference. *Annals of eugenics*, 6(4):391–398, 1935.

Johann Gagnon-Bartsch, Yotam Shem-Tov, et al. The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *The Annals of Applied Statistics*, 13(3):1464–1483, 2019.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Han Hong, Michael P Leung, and Jessie Li. Inference on finite-population treatment effects under limited overlap. *The Econometrics Journal*, 23(1):32–47, 2020.

Guido W Imbens and Donald B Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/9781139025751.

Cheng Ju, Joshua Schwab, and Mark J van der Laan. On adaptive propensity score truncation in causal inference. *Statistical methods in medical research*, 28(6):1741–1760, 2019.

S Khan and E Tamer. Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica*, 78(6):2021–2042, 2010. ISSN 0012-9682. doi: 10.3982/ECTA7372.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*, 2016.

Ilmun Kim, Ann B Lee, Jing Lei, et al. Global and local two-sample tests via regression. *Electronic Journal of Statistics*, 13(2):5253–5305, 2019.

Arun Kumar Kuchibhotla and Qinqing Zheng. Near-optimal confidence sequences for bounded random variables. *arXiv preprint arXiv:2006.05022*, 2020.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, pages 1–10, 2019.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990a.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990b.

Andreas Maurer. Concentration inequalities for functions of independent variables. *Random Structures & Algorithms*, 29(2):121–138, 2006.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Thomas Peel, Sandrine Anthoine, and Liva Ralaivola. Empirical bernstein inequalities for u-statistics. In *Advances in Neural Information Processing Systems*, pages 1903–1911, 2010.

David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

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

PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. URL `http://biomet.oxfordjournals.org/content/70/1/41.short`.

Andrew L Rukhin. Lower Bound on the Error Probability for Families with Bounded Likelihood Ratios. *Proceedings of the American Mathematical Society*, 119(4):1307, dec 1993. ISSN 00029939. doi: 10.2307/2159993. URL `http://www.jstor.org/stable/2159993?origin=crossref`.

Andrew L Rukhin et al. Information-type divergence when the likelihood ratios are bounded. *Applicationes Mathematicae*, 24(4):415–423, 1997.

Sanat K Sarkar et al. Generalizing simes' test and hochberg's stepup procedure. *Annals of Statistics*, 36(1):337–363, 2008.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media, 2006.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1995.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

Shu Yang and Peng Ding. Asymptotic causal inference with observational studies trimmed by the estimated propensity scores. *arXiv preprint arXiv:1704.00666*, 2017.

APPENDIX

A Failure of Plug-in Estimators for Population Overlap Slack

The most straightforward estimator of the population overlap slack is the plug-in estimator

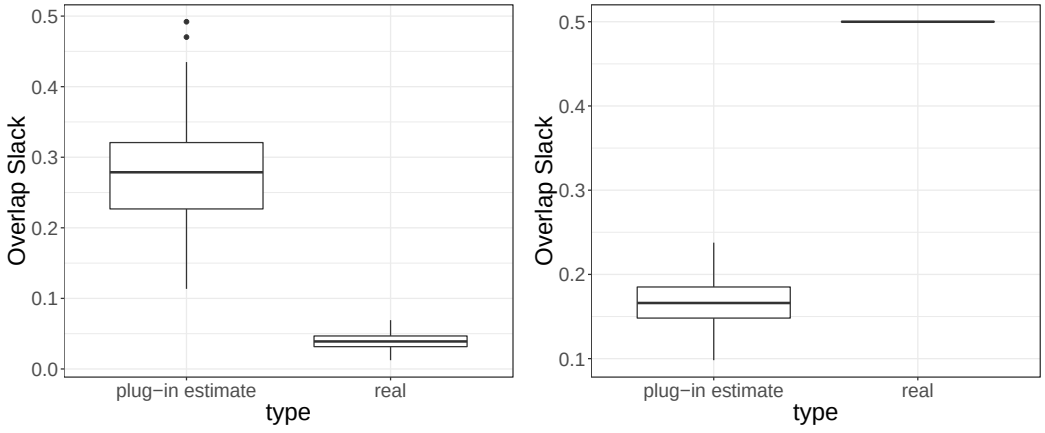
$$\hat{\mathcal{O}}_{\text{plugin}} \triangleq \min_i \min\{\hat{e}(X_i), 1 - \hat{e}(X_i)\}.$$

Although this approach permits flexible methods in estimation, there is typically no statistical guarantee like whether the assessment tends to be conservative or anti-conservative, unless strong assumptions are imposed. For instance, if a parametric method is used, the minimal requirement is the correct model specification; if a non-parametric method is used, the propensity score must satisfy some smoothness assumptions, depending on the method, which are not easy to justify in practice. Even if the regularity conditions are satisfied, the consistency is typically measured by pointwise convergence or convergence in some metrics, typified by the mean square errors $1/n \sum_{i=1}^{n} (\hat{e}(x_i) - e(x_i))^2$. It is unclear how accurate the approximation of the extreme values of $e(x)$ is based on $\hat{e}(x)$ in finite samples, as the former is clearly an irregular parameter. We use a toy simulation study to illustrate that the plug-in method can be both severely conservative or severely anti-conservative even for seemingly easy problems. In particular, we consider two data generating process:

C1se 1 $n = 1000, p = 50$ and $e(x) \equiv 0.5$ for all x. We fit a logistic regression to obtain $\hat{e}(x)$;

C2se 2 $n = 500, p = 1000$ and $\log(e(x)/(1 - e(x))) = x^T\beta$ where $\beta_1 = \ldots = \beta_{10} = 10/\sqrt{p}, \beta_{11} = \ldots = \beta_{1000} = 0$. We fit an L_1 penalized logistic regression to obtain $\hat{e}(x)$, with penalty level selected by 10-fold cross-validation.

In both cases, X is generated with i.i.d. standard normal entries. For each case, we calculate their population overlap slack and $\hat{\mathcal{O}}_{\text{plugin}}$. Figure 4 shows the boxplots of the real and estimated overlap slack on 1000 independent datasets.



(a) Case 1 (b) Case 2

Figure 4: Box-plots of real and estimated overlap slack over 1000 replicates.

Although the models are correctly specified for both cases and the estimate $\hat{e}(x)$ is supposed to "work well" as suggested by the theory, the plug-in estimates of $\mathcal{O}^*$ can be either extremely anti-conservative or extremely conservative in spite of the reasonably large sample size. This illustrates that failure to take finite-sample uncertainty into consideration may yield inadequate results for this problem even in large samples due to the irregularity of the estimand. In a nutshell, the assessment given by the plug-in method can be statistically invalid due to model misspecification or poor finite-sample performance.

B Technical Proofs

We start with a useful lemma that will be used for multiple times.

Lemma B.1. *[Generalization of Theorem 2.1 in Rukhin et al. [1997]] Let X be a random variable. Given any convex function f, non-decreasing function w, and non-negative function h that satisfies $\mathbb{E}[h(X)] = 1$ and $a_{\min} \le h(X) \le a_{\max}$ almost surely for some $0 \le a_{\min} \le a_{\max} < \infty$, let*

$$h^*(x) = \begin{cases} a_{\min} & (x < c) \\ a_{\max} & (x \ge c) \end{cases} \quad \text{if } f(a_{\max}) \ge f(a_{\min}), \quad h^*(x) = \begin{cases} a_{\max} & (x < c) \\ a_{\min} & (x \ge c) \end{cases} \quad \text{otherwise},$$

where c is chosen such that $\mathbb{E}[h^*(X)] = 1$. Then

$$\mathbb{E}[w(X)f(h(X))] \le \mathbb{E}[w(X)f(h^*(X))].$$

Proof. By Jensen's inequality,

$$\begin{aligned} f(h(x)) &= f\left(\frac{h(x) - a_{\min}}{a_{\max} - a_{\min}} a_{\max} + \frac{a_{\max} - h(x)}{a_{\max} - a_{\min}} a_{\min}\right) \\ &\le \frac{h(x) - a_{\min}}{a_{\max} - a_{\min}} f(a_{\max}) + \frac{a_{\max} - h(x)}{a_{\max} - a_{\min}} f(a_{\min}) \\ &= \frac{f(a_{\max}) - f(a_{\min})}{a_{\max} - a_{\min}} h(x) + \frac{a_{\max} f(a_{\min}) - a_{\min} f(a_{\max})}{a_{\max} - a_{\min}} \end{aligned}$$

Since $h^*(x)$ only takes two values $a_{\max}$ and $a_{\min}$, it is easy to see that

$$f(h^*(x)) = \frac{f(a_{\max}) - f(a_{\min})}{a_{\max} - a_{\min}} h^*(x) + \frac{a_{\max} f(a_{\min}) - a_{\min} f(a_{\max})}{a_{\max} - a_{\min}}.$$

If $f(a_{\max}) \ge f(a_{\min})$, it remains to prove that

$$\mathbb{E}[w(X)h(X)] \le \mathbb{E}[w(X)h^*(X)] \Longrightarrow \mathbb{E}[w(X)(h^*(X) - h(X))] \ge 0$$

Since $h(X) \in [a_{\min}, a_{\max}]$ almost surely, $h^*(x) - h(x) \ge 0$ if $x \ge c$ and $h^*(x) - h(x) \le 0$ if $x < c$. Due to the monotonicity and non-negativity of $w(x)$,

$$\begin{aligned} \mathbb{E}[w(X)(h^*(X) - h(X))] &= \mathbb{E}[w(X)(h^*(X) - h(X))I(X \ge c)] + \mathbb{E}[w(X)(h^*(X) - h(X))I(X < c)] \\ &\ge \mathbb{E}[w(c)(h^*(X) - h(X))I(X \ge c)] + \mathbb{E}[w(c)(h^*(X) - h(X))I(X < c)] \\ &= w(c)\mathbb{E}[h^*(X) - h(X)] = 0, \end{aligned}$$

where the last equality uses the fact that $\mathbb{E}[h^*(X)] = \mathbb{E}[h(X)] = 1$.

If $f(a_{\max}) < f(a_{\min})$, it remains to prove that

$$\mathbb{E}[w(X)h(X)] \ge \mathbb{E}[w(X)h^*(X)] \Longrightarrow \mathbb{E}[w(X)(h^*(X) - h(X))] \le 0.$$

This can be proved similarly as above. □

Proof of Theorem 3.1. If $\mathcal{O}^* = 1/2$, we must have $P_1 = P_0$ and thus $\mu_1 = \mu_0$. In this case it is easy to prove the theorem. Throughout the rest of the proof, we assume $\mathcal{O}^* < 1/2$. By definition,

$$\mu_1 - \mu_0 = \mathbb{E}_{P_1}[S] - \mathbb{E}_{P_0}[S] = \mathbb{E}_{P_0}\left[S\left(\frac{dP_1}{dP_0} - 1\right)\right].$$

Since $\mathbb{E}_{P_0}[dP_1/dP_0 - 1] = \mathbb{E}_{P_1}[S] - \mathbb{E}_{P_0}[S] = 0$,

$$\mu_1 - \mu_0 = \mathbb{E}_{P_0}\left[(S - \mathbb{E}_{P_0}[S])\left(\frac{dP_1}{dP_0} - 1\right)\right].$$

By Cauchy-Schwarz inequality,

$$|\mu_1 - \mu_0| \le \sqrt{\mathbb{E}_{P_0}[(S - \mathbb{E}_{P_0}[S])^2]}\sqrt{\mathbb{E}_{P_0}\left(\frac{dP_1}{dP_0} - 1\right)^2} = \sigma_0\sqrt{\mathbb{E}_{P_0}\left(\frac{dP_1}{dP_0} - 1\right)^2}. \tag{24}$$

Applying Lemma B.1 with $X \sim P_0$, $w(x) \equiv 1$, $h(x) = dP_1(x)/dP_0(x)$ which satisfies $\mathbb{E}_{P_0}[h(X)] = 1$, $a_{\min} = b_{\min}(\mathcal{O}^*;\pi)$, $a_{\max} = b_{\max}(\mathcal{O}^*;\pi)$ and $f(y) = (y-1)^2$ which is convex, if $f(a_{\max}) \ge f(a_{\min})$, we obtain that

$$\mathbb{E}_{P_0}\left(\frac{dP_1}{dP_0} - 1\right)^2 \le (b_{\max}(\mathcal{O}^*;\pi) - 1)^2\mathbb{P}(X \ge c) + (1 - b_{\min}(\mathcal{O}^*;\pi))^2\mathbb{P}(X < c)$$

where c satisfies that

$$1 = \mathbb{E}[b_{\max}(\mathcal{O}^*;\pi)I(X \ge c) + b_{\min}(\mathcal{O}^*;\pi)I(X < c)] = b_{\max}(\mathcal{O}^*;\pi)\mathbb{P}(X \ge c) + b_{\min}(\mathcal{O}^*;\pi)\mathbb{P}(X < c).$$

Using the fact that $\mathbb{P}(X \ge c) + \mathbb{P}(X < c) = 1$, we obtain that

$$\mathbb{P}(X \ge c) = \frac{1 - b_{\min}(\mathcal{O}^*;\pi)}{b_{\max}(\mathcal{O}^*;\pi) - b_{\min}(\mathcal{O}^*;\pi)}.$$

By (24),

$$T_0 \le \sqrt{\mathbb{E}_{P_0}\left(\frac{dP_1}{dP_0} - 1\right)^2} \le \sqrt{(1 - b_{\min}(\mathcal{O}^*;\pi))(b_{\max}(\mathcal{O}^*;\pi) - 1)}.$$

Similarly we can prove the bound for T_1. If $f(a_{\max}) < f(a_{\min})$, we can show the above bound using the same arguments.

Now we prove that $\mathcal{O}^* \le \mathcal{O}^+_{\mathrm{DiM}}(T_0, T_1;\pi)$. By definitions of $b_{\min}(\mathcal{O}^*;\pi)$ and $b_{\max}(\mathcal{O}^*;\pi)$,

$$\begin{aligned}
T_0^2 &\le \left(1 - \frac{1-\pi}{\pi}\frac{\mathcal{O}^*}{1-\mathcal{O}^*}\right)\left(\frac{1-\pi}{\pi}\frac{1-\mathcal{O}^*}{\mathcal{O}^*} - 1\right) \\
&= \frac{1-\pi}{\pi}\left(\frac{1-\mathcal{O}^*}{\mathcal{O}^*} + \frac{\mathcal{O}^*}{1-\mathcal{O}^*}\right) - \frac{(1-\pi)^2}{\pi^2} - 1 \\
&= \frac{1-\pi}{\pi}\left(\frac{1}{\mathcal{O}^*(1-\mathcal{O}^*)} - 2\right) - \frac{(1-\pi)^2}{\pi^2} - 1 \\
&= \frac{1-\pi}{\pi\mathcal{O}^*(1-\mathcal{O}^*)} - \left(\frac{1-\pi}{\pi} + 1\right)^2 \\
&= \frac{1}{\pi^2}\left(\frac{\pi(1-\pi)}{\mathcal{O}^*(1-\mathcal{O}^*)} - 1\right).
\end{aligned}$$

This implies that

$$\mathcal{O}^*(1-\mathcal{O}^*) \le \frac{\pi(1-\pi)}{(\pi T_0)^2 + 1} \Longrightarrow \left(\frac{1}{2} - \mathcal{O}^*\right)^2 \ge \frac{1}{4} - \frac{\pi(1-\pi)}{(\pi T_0)^2 + 1}.$$

Since $\mathcal{O}^* \le 1/2$, we conclude that

$$\mathcal{O}^* \le \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\pi(1-\pi)}{(\pi T_0)^2 + 1}}.$$

Similarly, we can prove that

$$\mathcal{O}^* \le \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{\pi(1-\pi)}{((1-\pi) T_1)^2 + 1}}.$$

The proof is then completed. □

Proof of Theorem 3.3. The bounds of ν_1 and ν_0 are direct consequence of (8). As a result,

$$\nu_1 \leq \frac{1-\pi}{\pi}\frac{1-\mathcal{O}^*}{\mathcal{O}^*} \Longrightarrow \mathcal{O}^* \leq \frac{1-\pi}{1-\pi+\pi\nu_1}.$$

Similarly,

$$\mathcal{O}^* \leq \frac{\pi}{\pi+(1-\pi)\nu_0}.$$

□

Proof of Theorem 3.5. If $\mathcal{O}^* = 1/2$, we must have $P_1 = P_0$, $\pi = 1/2$ and $p_* = 1/2$. In this case it is easy to prove the theorem. Throughout the rest of the proof, we assume $\mathcal{O}^* < 1/2$. Let F_1 and F_0 denote the cdfs of P_1 and P_0. Since P_1 and P_0 do not have point mass, $p_* = \mathbb{P}(S^{(1)} \geq S^{(0)})$. Then

$$p_* = \mathbb{P}(S^{(1)} \geq S^{(0)}) = \int_0^1 dP_1(x)\left(\int_0^x dP_0(y)\right) = \int_0^1 F_0(x)dP_1(x) = \int_0^1 F_0(x)\frac{dP_1(x)}{dP_0(x)}dP_0(x).$$

Applying Lemma B.1 with $X \sim P_0$, $w(x) \equiv F_0(x)$ which is non-decreasing, $h(x) = dP_1(x)/dP_0(x)$ which satisfies $\mathbb{E}_{P_0}[h(X)] = 1$, $a_{\min} = b_{\min}(\mathcal{O}^*;\pi)$, $a_{\max} = b_{\max}(\mathcal{O}^*;\pi)$ and $f(y) = y$ which is convex, since $f(a_{\max}) = a_{\max} \geq a_{\min} = f(a_{\min})$, we have that

$$\begin{aligned} p_* &= \mathbb{E}[F_0(X)h(X)] \leq \mathbb{E}\left[F_0(X)\left(b_{\max}(\mathcal{O}^*;\pi)I(X \geq c) + b_{\min}(\mathcal{O}^*;\pi)I(X < c)\right)\right] \\ &= b_{\max}(\mathcal{O}^*;\pi)\mathbb{E}[F_0(X)I(X \geq c)] + b_{\min}(\mathcal{O}^*;\pi)\mathbb{E}[F_0(X)I(X < c)], \end{aligned} \tag{25}$$

where, as shown in the proof of (3.1), c satisfies that

$$\mathbb{P}(X \geq c) = \frac{1-b_{\min}(\mathcal{O}^*;\pi)}{b_{\max}(\mathcal{O}^*;\pi) - b_{\min}(\mathcal{O}^*;\pi)}.$$

Since P_0 does not have point mass,

$$F_0(c) = \mathbb{P}(X < c) = 1 - F_0(c) = \frac{b_{\max}(\mathcal{O}^*;\pi) - 1}{b_{\max}(\mathcal{O}^*;\pi) - b_{\min}(\mathcal{O}^*;\pi)} \tag{26}$$

Since F_0 is non-decreasing,

$$\mathbb{E}[F_0(X)I(X \geq c)] \leq \mathbb{E}[F_0(X)I(F_0(X) \geq F_0(c))].$$

Let $c' = \inf\{y : F_0(y) \geq F_0(c)\}$. Then

$$\mathbb{E}[F_0(X)I(F_0(X) \geq F_0(c))] \leq \mathbb{E}[F_0(X)I(X \geq c')].$$

Due to the right-continuity of F_0, $F_0(c') = F_0(c)$, implying that $\mathbb{P}(X \in (c', c]) = 0$. Since F_0 has no point mass, $\mathbb{P}(X \in [c', c]) = 0$, implying that

$$\mathbb{E}[F_0(X)I(X \geq c')] = \mathbb{E}[F_0(X)I(X \geq c)].$$

Therefore,

$$\mathbb{E}[F_0(X)I(X \geq c)] \leq \mathbb{E}[F_0(X)I(F_0(X) \geq F_0(c))].$$

As a result,

$$\mathbb{E}[F_0(X)I(X \geq c)] = \mathbb{E}[F_0(X)I(F_0(X) \geq F_0(c))] = \int_{F_0(c)}^1 F_0(x)dF_0(x) = \frac{1}{2}\left(1 - F_0(c)^2\right).$$

This entails that

$$\mathbb{E}[F_0(X)I(X < c)] = \mathbb{E}[F_0(X)] - \mathbb{E}[F_0(X)I(X \geq c)] = \frac{1}{2}F_0(c)^2.$$

By (25) and (26),

$$
\begin{aligned}
p_* &\le \frac{b_{\max}(\mathcal{O}^*;\pi)}{2} - \frac{(b_{\max}(\mathcal{O}^*;\pi)-1)^2}{2(b_{\max}(\mathcal{O}^*;\pi)-b_{\min}(\mathcal{O}^*;\pi))} = \frac{2b_{\max}(\mathcal{O}^*;\pi)-1-b_{\max}(\mathcal{O}^*;\pi)b_{\min}(\mathcal{O}^*;\pi)}{2(b_{\max}(\mathcal{O}^*;\pi)-b_{\min}(\mathcal{O}^*;\pi))} \\
&= \frac{2\frac{1-\pi}{\pi}\frac{1-\mathcal{O}^*}{\mathcal{O}^*} - \left(\frac{1-\pi}{\pi}\right)^2 - 1}{2\frac{1-\pi}{\pi}\left(\frac{1-\mathcal{O}^*}{\mathcal{O}^*} - \frac{\mathcal{O}^*}{1-\mathcal{O}^*}\right)} = \frac{\mathcal{O}^*(1-\mathcal{O}^*)}{2(1-2\mathcal{O}^*)}\left(\frac{2(1-\mathcal{O}^*)}{\mathcal{O}^*} - \frac{1-\pi}{\pi} - \frac{\pi}{1-\pi}\right) \\
&= \frac{\mathcal{O}^*(1-\mathcal{O}^*)}{2(1-2\mathcal{O}^*)}\left(\frac{2}{\mathcal{O}^*} - \frac{1}{\pi(1-\pi)}\right) = \frac{1-\mathcal{O}^*}{2(1-2\mathcal{O}^*)}\left(2 - \frac{\mathcal{O}^*}{\pi(1-\pi)}\right).
\end{aligned}
$$

Similarly, by replacing $b_{\max}$ by $b_{\min}^{-1}$ and $b_{\min}$ by $b_{\max}^{-1}$, we can prove that the above bound is also an upper bound of $1-p_* = \mathbb{P}(S^{(1)} \le S^{(0)})$. Therefore,

$$
\max\{p_*, 1-p_*\} \le \frac{1-\mathcal{O}^*}{2(1-2\mathcal{O}^*)}\left(2 - \frac{\mathcal{O}^*}{\pi(1-\pi)}\right).
$$

Let $q = \max\{p_*, 1-p_*\}$ for notational convenience. The above inequality implies that

$$
\begin{aligned}
& 2\pi(1-\pi)q(1-2\mathcal{O}^*) \le (1-\mathcal{O}^*)\left(2\pi(1-\pi)-\mathcal{O}^*\right) \\
\Longrightarrow\ & \mathcal{O}^{*2} - \left(1+2\pi(1-\pi)(1-2q)\right)\mathcal{O}^* + 2\pi(1-\pi)(1-q) \ge 0 \\
\Longrightarrow\ & \left(\mathcal{O}^* - \frac{1}{2} - \pi(1-\pi)(1-2q)\right)^2 \ge \left(\frac{1}{2} + \pi(1-\pi)(1-2q)\right)^2 - 2\pi(1-\pi)(1-q) \\
\Longrightarrow\ & \left(\mathcal{O}^* - \frac{1}{2} - \pi(1-\pi)(1-2q)\right)^2 \ge \frac{(1-2\pi)^2}{4} + \left(\pi(1-\pi)(1-2q)\right)^2 \\
\Longrightarrow\ & \mathcal{O}^* \le \frac{1}{2} + \pi(1-\pi)(1-2q) - \sqrt{\frac{(1-2\pi)^2}{4} + \left(\pi(1-\pi)(1-2q)\right)^2} \\
& \quad \text{or } \mathcal{O}^* \ge \frac{1}{2} + \pi(1-\pi)(1-2q) + \sqrt{\frac{(1-2\pi)^2}{4} + \left(\pi(1-\pi)(1-2q)\right)^2}.
\end{aligned}
$$

Note that $1-2q = -|1-2p_*|$. It remains to prove that the second bound never holds. In fact, since $\pi \le 1/2$,

$$
\frac{1}{2} + \pi(1-\pi)(1-2q) + \sqrt{\frac{(1-2\pi)^2}{4} + \left(\pi(1-\pi)(1-2q)\right)^2} \ge \frac{1}{2} + \pi(1-\pi)(1-2q) + \left|\pi(1-\pi)(1-2q)\right| \ge \frac{1}{2}.
$$

The bound never holds since $\mathcal{O}^* < 1/2$ as assumed at the beginning of the proof. □

Proof of Lemma 3.1. Since $\hat{\sigma}^2$ is permutation-invariant, we can assume without loss of generality that $T_1 = \cdots = T_{n_1} = 1$ and $T_{n_1+1} = \cdots = T_n = 0$. In this case, $T_i(1-T_j) \ne 0$ only if $i < j$. For any (i_1, i_2, i_3, i_4), the summand $(\phi(z_{i_1}, z_{i_2}) - \phi(z_{i_3}, z_{i_4}))^2$ remains invariant if i_1 and i_2 are interchanged

or i_3 and i_4 are interchanged. As a result,

$$\begin{aligned}
&n(n-1)(n-2)(n-3)\hat{\sigma}^2 \\
&= \sum_{i_1\neq i_2\neq i_3\neq i_4} (\phi(z_{i_1},z_{i_2})-\phi(z_{i_3},z_{i_4}))^2 = \sum_{\substack{i_1<i_2,i_3<i_4\\ i_1\neq i_2\neq i_3\neq i_4}} 4(\phi(z_{i_1},z_{i_2})-\phi(z_{i_3},z_{i_4}))^2 \\
&= \sum_{\substack{i_1<i_2,i_3<i_4\\ i_1\neq i_2\neq i_3\neq i_4}} (T_{i_1}(1-T_{i_2})I(S_{i_1}>S_{i_2})-T_{i_3}(1-T_{i_4})I(S_{i_3}>S_{i_4}))^2 \\
&= \sum_{\substack{i_1<i_2,i_3<i_4\\ i_1\neq i_2\neq i_3\neq i_4}} (T_{i_1}(1-T_{i_2})I(S_{i_1}>S_{i_2})+T_{i_3}(1-T_{i_4})I(S_{i_3}>S_{i_4})) \\
&\quad -2\sum_{\substack{i_1<i_2,i_3<i_4\\ i_1\neq i_2\neq i_3\neq i_4}} T_{i_1}(1-T_{i_2})T_{i_3}(1-T_{i_4})I(S_{i_1}>S_{i_2},S_{i_3}>S_{i_4}) \\
&= 2\sum_{\substack{i_1<i_2,i_3<i_4\\ i_1\neq i_2\neq i_3\neq i_4}} T_{i_1}(1-T_{i_2})I(S_{i_1}>S_{i_2}) \\
&\quad -2\sum_{\substack{i_1<i_2,i_3<i_4\\ i_1\neq i_2\neq i_3\neq i_4}} T_{i_1}(1-T_{i_2})T_{i_3}(1-T_{i_4})I(S_{i_1}>S_{i_2},S_{i_3}>S_{i_4}) \\
&\triangleq \Sigma_1-\Sigma_2.
\end{aligned}$$

Fix any $i_1<i_2$, there are $(n-2)(n-3)/2$ pairs of (i_3,i_4) with $i_3<i_4$ that are distinct from (i_1,i_2). As a result,

$$\Sigma_1=(n-2)(n-3)\sum_{i_1<i_2} T_{i_1}(1-T_{i_2})I(S_{i_1}>S_{i_2})=(n-2)(n-3)\hat{U}.$$

On the other hand,

$$\begin{aligned}
\Sigma_2 &= 2\sum_{\substack{i_1\neq i_3\\ i_1,i_3\leq n_1}}\sum_{\substack{i_2\neq i_4\\ i_2,i_4>n_1}} I(S_{i_1}>S_{i_2},S_{i_3}>S_{i_4}) \\
&= 2\sum_{i_1,i_3\leq n_1}\sum_{i_2,i_4>n_1} I(S_{i_1}>S_{i_2},S_{i_3}>S_{i_4}) - 2\sum_{i_1\leq n_1}\sum_{i_2,i_4>n_1} I(S_{i_1}>S_{i_2},S_{i_1}>S_{i_4}) \\
&\quad -2\sum_{i_1,i_3\leq n_1}\sum_{i_2>n_1} I(S_{i_1}>S_{i_2},S_{i_3}>S_{i_2})+2\sum_{i_1\leq n_1}\sum_{i_2>n_1} I(S_{i_1}>S_{i_2}) \\
&\triangleq 2\Sigma_2^{(1)}-2\Sigma_2^{(2)}-2\Sigma_2^{(3)}+2\Sigma_2^{(4)}.
\end{aligned}$$

By definition, $\Sigma_2^{(4)}=\hat{U}$, and

$$\begin{aligned}
\Sigma_2^{(1)} &= \sum_{i_1\leq n_1}\sum_{i_2>n_1}\sum_{i_3\leq n_1}\sum_{i_4>n_1} I(S_{i_1}>S_{i_2},S_{i_3}>S_{i_4}) \\
&= \left(\sum_{i_1\leq n_1,i_2>n_1} I(S_{i_1}>S_{i_2})\right)\left(\sum_{i_3\leq n_1,i_4>n_1} I(S_{i_3}>S_{i_4})\right)=\hat{U}^2.
\end{aligned}$$

To derive $\Sigma_2^{(2)}$ and $\Sigma_2^{(3)}$, we recall the definition of $\tilde{R}_i$ given by (17). Then

$$\begin{aligned}
\Sigma_2^{(2)} &= \sum_{i_1\leq n_1}\left\{\left(\sum_{i_2>n_1} I(S_{i_1}>S_{i_2})\right)\left(\sum_{i_4>n_1} I(S_{i_1}>S_{i_4})\right)\right\} \\
&= \sum_{i_1\leq n_1}\left(\sum_{i_2>n_1} I(S_{i_1}>S_{i_2})\right)^2=\sum_{i\leq n_1}\tilde{R}_i^2,
\end{aligned}$$

and

$$\Sigma_2^{(2)} = \sum_{i_2>n_1} \left\{ \left(\sum_{i_1 \le n_1} I(S_{i_1} > S_{i_2}) \right) \left(\sum_{i_3 \le n_1} I(S_{i_3} > S_{i_2}) \right) \right\}$$
$$= \sum_{i_2>n_1} \left(\sum_{i_1 \le n_1} I(S_{i_1} > S_{i_2}) \right)^2 = \sum_{i>n_1} \tilde{R}_i^2.$$

Putting the pieces together, we conclude that

$$\Sigma_2 = 2\hat{U} + 2\hat{U}^2 - 2\sum_{i=1}^{n} \tilde{R}_i^2.$$

Therefore,

$$n(n-1)(n-2)(n-3)\hat{\sigma}^2 = \Sigma_1 - \Sigma_2 = (n-1)(n-4)\hat{U} - 2\hat{U}^2 + 2\sum_{i=1}^{n} \tilde{R}_i^2.$$

□

Proof of Proposition 3.4. Let A_η denote a subset of $\mathbb{R}^2$ such that $I(T_i = I(S_i > \eta)) = I(Z_i \in A_\eta)$ where $Z_i = (T_i, S_i)$. For instance, $A(\eta)$ can be chosen as $(\{1\} \times (\eta, 1]) \cup (\{0\} \times [0, \eta])$. Note that $I(t = I(s > \eta)) \in [0, 1]$ for any (t, s). By equation (3.15) of Vapnik [1995], Then,

$$\sup_{\eta \in [0,1]} \left(\mathcal{E}(\eta) - \hat{\mathcal{E}}(\eta) \right) \le \sqrt{\frac{\log \Delta(2n) + \log\left(\frac{4}{\alpha}\right)}{n}}$$

where $\Delta(m)$ is the shattering number, defined as

$$\Delta(m) = \max_{z_1,\ldots,z_m} \text{Card} \left\{ (I(z_1 \in A(\eta)), \ldots, I(z_m \in A(\eta)) : \eta \in [0, 1] \right\}, \tag{27}$$

and Card denotes the cardinality of a set. It is easy to see that

$$\Delta(m) \le \max_{s_1,\ldots,s_m \in [0,1]} \text{Card} \left\{ (I(s_1 > \eta), \ldots, I(s_m > \eta) : \eta \in [0, 1] \right\} = m + 1.$$

The proof is then completed by setting $m = 2n$.

□

Proof of Proposition 3.5. By definition, $\mathbb{P}(\hat{\mathcal{O}}^{(b)} \le \mathcal{O}^*) \le q\alpha$. On the event $\hat{\mathcal{O}} \le \mathcal{O}^*$,

$$q \le \frac{1}{B}\sum_{b=1}^{B} I(\hat{\mathcal{O}}^{(b)} \le \hat{\mathcal{O}}) \le \frac{1}{B}\sum_{b=1}^{B} I(\hat{\mathcal{O}}^{(b)} \le \mathcal{O}^*)$$

where the first inequality is obtained from the definition of $\hat{\mathcal{O}}$. Taking expectation over both sides,

$$q\mathbb{P}(\hat{\mathcal{O}} \le \mathcal{O}^*) \le \mathbb{E}\left[I(\hat{\mathcal{O}} \le \mathcal{O}^*) \frac{1}{B}\sum_{b=1}^{B} I(\hat{\mathcal{O}}^{(b)} \le \mathcal{O}^*) \right] \le \mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} I(\hat{\mathcal{O}}^{(b)} \le \mathcal{O}^*) \right] \le q\alpha.$$

The proof is then completed. □

Proof of Theorem 3.8. Let N denote the total sample size, N_1 denote the size of the treatment group and N_0 denote the size of the control group. Further let $S_i^* = e^*(X_i)$. Since $\pi \in (0, 1)$, we have $\hat{\pi} \xrightarrow{p} \pi$. Furthermore, the condition (22) implies that

$$\mathbb{P}\left(\max_i |S_i - S_i^*| \ge \rho_N \right) \le \rho_N, \quad \text{for some sequence } \rho_N \to 0. \tag{28}$$

For notational convenience, we let $\mathcal{V}$ denote event that $\max_i |S_i - S_i^*| \le \rho_N$.

- (Approximate DiM O-value) Let $\mu_t^* = \mathbb{E}[e^*(X) \mid T = t]$ and $\sigma_t^{2*} = \mathrm{Var}[e^*(X) \mid T = t]$. Further let $T_t^* = |\mu_1^* - \mu_0^*| / \max\{\sigma_t^*, \zeta\}$. By Theorem 3.1,

$$\mathcal{O}^+_{\mathrm{DiM}}(T_0^*, T_1^*; \pi) \geq \mathcal{O}^*.$$

Since $\hat{\pi} \xrightarrow{p} \pi$ and $\mathcal{O}^+_{\mathrm{DiM}}$ is continuous in all arguments, it remains to prove $\hat{T}_t \xrightarrow{p} T_t^*$ for both $t = 0, 1$. By (28),

$$\left| \hat{\mu}_1 - \frac{1}{N_1} \sum_{T_i = 1} S_i^* \right| \xrightarrow{p} 0.$$

By the law of large numbers,

$$\frac{1}{N_1} \sum_{T_i = 1} S_i^* = \frac{1}{\hat{\pi}} \frac{1}{N} \sum_{i=1}^N S_i^* I(T_i = 1) \xrightarrow{p} \frac{1}{\pi} \mathbb{E}[S_i^* I(T_i = 1)] = \mu_1^*.$$

On the other hand, Since $S_i, S_i^* \in [0, 1]$,

$$|S_i^2 - S_i^{*2}| = |S_i - S_i^*||S_i + S_i^*| \leq 2|S_i - S_i^*|.$$

Then (28) implies that

$$\left| \frac{1}{N_1 - 1} \sum_{T_i = 1} S_i^2 - \frac{1}{N_1 - 1} \sum_{T_i = 1} S_i^{*2} \right| \xrightarrow{p} 0.$$

Similar to the mean, by the law of large numbers, we have

$$\frac{1}{N_1 - 1} \sum_{T_i = 1} S_i^{*2} \xrightarrow{p} \mathbb{E}[e^*(X) \mid T = 1]^2.$$

As a result,

$$\hat{\sigma}_1^2 = \frac{1}{N_1 - 1} \sum_{i=1}^N S_i^2 - \frac{N_1}{N_1 - 1} \left(\frac{1}{N_1} \sum_{i=1}^N S_i \right)^2 \xrightarrow{p} \sigma_1^{*2}.$$

Similarly, we have $\hat{\mu}_0 \xrightarrow{p} \mu_0^*$ and $\hat{\sigma}_0^2 \xrightarrow{p} \sigma_0^{*2}$.

Therefore, $|\hat{\mu}_1 - \hat{\mu}_0| \xrightarrow{p} |\mu_1^* - \mu_0^*|$ and $1/\max\{\hat{\sigma}_t, \zeta\} \xrightarrow{p} 1/\max\{\sigma_t^*, \zeta\}$ for any $\zeta \geq 0$. When $\zeta > 0$ or $\hat{\sigma}_1^*, \hat{\sigma}_0^* > 0$, $1/\max\{\sigma_t^*, \zeta\} < \infty$ and by Slusky's lemma, $\hat{T}_t \xrightarrow{p} T_t^*$. When $\zeta = 0$ and $\hat{\sigma}_1^* = 0$ (or $\hat{\sigma}_0^* = 0$), $1/\max\{\sigma_1^*, \zeta\} = \infty$. If $|\mu_1^* - \mu_0^*| > 0$, $\hat{T}_1 \xrightarrow{p} \infty = T_1^*$ (or $\hat{T}_0 \xrightarrow{p} \infty = T_0^*$). The proof is then completed.

- (Approximate DiR O-value) Let $Z_i^* = (T_i, S_i^*)$ and $\hat{U}^* = \frac{1}{N(N-1)} \sum_{i \neq j} \phi(Z_i^*, Z_j^*)$. Then By Theorem 3.5, we have

$$\mathcal{O}^* \leq \mathcal{O}^+_{\mathrm{DiR}}(\mathbb{E}[\hat{U}^*]; \pi).$$

It remains to prove that $|\hat{U} - \hat{U}^*| \xrightarrow{p} 0$. Since $|\hat{U} - \hat{U}^*| \leq 2$, it is left to prove that $|\hat{U} - \hat{U}^*| I_{\mathcal{V}} \xrightarrow{p} 0$. On the event $\mathcal{V}$, $\phi(Z_i, Z_j) = \phi(Z_i^*, Z_j^*)$ if $|S_i^* - S_j^*| > 2\rho_N$. As a result,

$$\begin{aligned} |\hat{U} - \hat{U}^*| I_{\mathcal{V}} &\leq \frac{1}{N(N-1)} \sum_{i \neq j} \phi(Z_i^*, Z_j^*) I(|S_i^* - S_j^*| \leq 2\rho_N) I_{\mathcal{V}} \\ &\leq \frac{1}{N(N-1)} \sum_{i \neq j} I(|S_i^* - S_j^*| \leq 2\rho_N) \triangleq \tilde{U}. \end{aligned}$$

By Hoeffding inequality for U-statistics (Hoeffding [1963], equation 5.7),

$$\mathbb{P}\left(|\tilde{U} - \mathbb{E}[\tilde{U}]| \geq t \right) \leq 2 \exp\left\{ -2 \left\lfloor \frac{N}{2} \right\rfloor t^2 \right\}.$$

As a result,

$$|\tilde{U} - \mathbb{E}[\tilde{U}]| \xrightarrow{p} 0.$$

By definition, $\mathbb{E}[\tilde{U}] = \mathbb{P}(|S_i^* - S_j^*| \leq \rho_N)$. By the dominated convergence theorem,

$$\mathbb{P}(|S_i^* - S_j^*| \leq \rho_N) \to \mathbb{P}(|S_i^* - S_j^*| = 0) = \mathbb{P}\left(S_i^* = S_j^*\right).$$

Since the distribution of $e^*(X_i)$ has no point mass, $\mathbb{P}(S_i^* = S_j^*) = 0$. The proof is then completed by putting the pieces together.

- (Approximate CE O-value) Let $\mathcal{E}^*(\eta) = \mathbb{P}(T = I(e^*(X) > \eta))$. By Theorem 3.7, $\mathcal{O}^* \leq 1 - \sup_{\eta\in[0,1]} \mathcal{E}^*(\eta)$. It remains to prove that

$$\sup_{\eta\in[0,1]} \hat{\mathcal{E}}(\eta) \xrightarrow{p} \sup_{\eta\in[0,1]} \mathcal{E}^*(\eta).$$

Let

$$\hat{\mathcal{E}}_t(\eta) = \frac{1}{N}\sum_{T_i=t} I(T_i = I(S_i > \eta)), \quad \hat{\mathcal{E}}_t^*(\eta) = \frac{1}{N}\sum_{T_i=t} I(T_i = I(S_i^* > \eta)).$$

On the event $\mathcal{V}$,

$$\begin{aligned}
\hat{\mathcal{E}}_1(\eta) &\leq \hat{\mathcal{E}}_1^*(\max\{0, \eta - \rho_N\}) = \mathcal{E}_1^*(\max\{0, \eta - \rho_N\}) + \sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_1^*(\eta) - \mathbb{E}[\hat{\mathcal{E}}_1^*(\eta)]| \\
&\leq \mathcal{E}_1^*(\max\{0, \eta - \rho_N\}) + \sup_{|\eta_1-\eta_2|\leq\rho_n} |\mathcal{E}_1^*(\eta_1) - \mathcal{E}_1^*(\eta_2)| + \sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_1^*(\eta) - \mathbb{E}[\hat{\mathcal{E}}_1^*(\eta)]| \\
&\triangleq \mathcal{E}_1^*(\max\{0, \eta - \rho_N\}) + \Delta_{t1} + \Delta_{t2}.
\end{aligned}$$

Similarly, on the event $\mathcal{V}$,

$$\hat{\mathcal{E}}_1(\eta) \geq \hat{\mathcal{E}}_1^*(\eta) - \Delta_{t1} - \Delta_{t2}.$$

As a consequence,

$$\sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_1(\eta) - \pi\mathcal{E}_1^*(\eta)| I_{\mathcal{V}} \leq \Delta_{t1} + \Delta_{t2} \Longrightarrow \sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_1(\eta) - \pi\mathcal{E}_1^*(\eta)| \leq \Delta_{t1} + \Delta_{t2} + I_{\mathcal{V}}.$$

By (28), $I_{\mathcal{V}} \xrightarrow{p} 0$. Since $\mathcal{E}_t^*(\eta)$ has a bounded derivative, $\Delta_{t1} \to 0$. Finally, the inequality that Proposition 3.4 adapts to implies that for any $\alpha > 0$,

$$\mathbb{P}\left(\Delta_{t2} \geq \sqrt{\frac{\log(2N+1) + \log(4/\alpha)}{N}}\right) \leq \alpha.$$

Therefore, $\Delta_{t2} \xrightarrow{p} 0$. Putting the pieces together, we conclude that

$$\sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_1(\eta) - \mathcal{E}_1^*(\eta)| \xrightarrow{p} 0.$$

Similarly, we can prove that

$$\sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_0(\eta) - \mathcal{E}_0^*(\eta)| \xrightarrow{p} 0.$$

Since $\hat{\mathcal{E}}(\eta) = \hat{\mathcal{E}}_1(\eta) + \hat{\mathcal{E}}_0(\eta)$ and $\mathcal{E}^*(\eta) = \mathcal{E}_1^*(\eta) + \mathcal{E}_0^*(\eta)$,

$$\sup_{\eta\in[0,1]} |\hat{\mathcal{E}}(\eta) - \mathcal{E}^*(\eta)| \leq \sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_1(\eta) - \mathcal{E}_1^*(\eta)| + \sup_{\eta\in[0,1]} |\hat{\mathcal{E}}_0(\eta) - \mathcal{E}_0^*(\eta)| \xrightarrow{p} 0.$$

□

Proof of Theorem 5.1. We only present the proof for ATT. The results for ATC can be proved by symmetry. By definition,

$$b_{\min,\mathrm{ATT}}(\mathcal{O}^*;\pi) \triangleq 0 \le \frac{dP_1}{dP_0}(S) \le \frac{1-\pi}{\pi}\frac{1-\mathcal{O}^*_{\mathrm{ATT}}}{\mathcal{O}^*_{\mathrm{ATT}}} \triangleq b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi).$$

- The proof of Theorem 3.1 implies that

$$T_0 \le \sqrt{(1-b_{\min,\mathrm{ATT}}(\mathcal{O}^*;\pi))(b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)-1)} = \sqrt{\frac{1-\pi}{\pi}\frac{1-\mathcal{O}^*_{\mathrm{ATT}}}{\mathcal{O}^*_{\mathrm{ATT}}}-1}.$$

  Squaring both sides implies that

$$\frac{\pi}{1-\pi}(1+T_0^2) \le \frac{1-\mathcal{O}^*_{\mathrm{ATT}}}{\mathcal{O}^*_{\mathrm{ATT}}} \Longrightarrow \mathcal{O}^*_{\mathrm{ATT}} \le \frac{1-\pi}{1+\pi T_0^2}.$$

- It is obvious that $\nu_1 \le b_{\max}(\mathcal{O}^*;\pi)$, which implies the bound for $\mathcal{O}^*_{\mathrm{ATT}}$.
- Recalling (25) and (26) in the proof of Theorem 3.5,

$$\begin{aligned} p_* &\le \frac{b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)}{2} - \frac{(b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)-1)^2}{2(b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)-b_{\min,\mathrm{ATT}}(\mathcal{O}^*;\pi))} \\ &= \frac{b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)}{2} - \frac{(b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)-1)^2}{2b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)} \\ &= 1 - \frac{1}{2b_{\max,\mathrm{ATT}}(\mathcal{O}^*;\pi)} = 1 - \frac{\pi\mathcal{O}^*_{\mathrm{ATT}}}{2(1-\pi)(1-\mathcal{O}^*_{\mathrm{ATT}})}. \end{aligned}$$

  This implies the desired result.

□

Proof of Theorem 5.3. (1) By definition,

$$\mathcal{O}^+_{\mathrm{PL}} = \mathbb{E}[\mathrm{Var}(T \mid X)] = \mathbb{E}[(T-e(X))^2] \le \mathbb{E}[(T-\hat{e}(X))^2].$$

The proof is completed by noting that $S = \hat{e}(X)$.

(2) Recall that $1-\mathcal{E}_{\max} = \mathbb{E}[\min\{e(X), 1-e(X)\}]$. The result is proved by the simple fact that

$$e(X)(1-e(X)) \le \min\{e(X), 1-e(X)\}.$$

□

C Refined Concentration Inequalities For Tighter O-values

C.1 Tighter DiM O-values

Since S_i is an estimated propensity score, we expect that $\mu_1 > \mu_0$. This motivates us to consider one-sided confidence intervals for μ_1 and μ_0. In particular, (T_1^-, T_0^-) is a valid lower confidence envelope of (T_1, T_0) if

$$T_1^- = \frac{\max\{\hat{\mu}_1^- - \hat{\mu}_0^+, 0\}}{\hat{\sigma}_1^+}, \quad T_0^- = \frac{\max\{\hat{\mu}_1^- - \hat{\mu}_0^+, 0\}}{\hat{\sigma}_0^+}$$

where $\hat{\mu}_0^+, \hat{\sigma}_1^+, \hat{\sigma}_0^+$ are $(1-\alpha/4)$ upper confidence bounds of μ_0, σ_1, and σ_0, respectively, and $\hat{\mu}_1^-$ is a $(1-\alpha/4)$ lower confidence bound of μ_1.

For the mean parameters, it is known that the Hoeffding's and empirical Bernstein's inequalities are loose in constants [e.g. Waudby-Smith and Ramdas, 2020, Kuchibhotla and Zheng, 2020, Bates et al., 2021a]. For the refined DiM O-value, we apply the following tighter bound.

Proposition C.1 (Hedged Captial bound, Waudby-Smith and Ramdas [2020], Theorem 3)**.** *Let $Z_1, \ldots, Z_n \in [0,1]$ be i.i.d. random variables with $\mathbb{E}[Z_1] = \mu$. For each $i = 1, \ldots, n$,*

$$\hat{\mu}_i = \frac{1/2 + \sum_{j=1}^{i} Z_j}{1+i}, \ \hat{\sigma}_i^2 = \frac{1/4 + \sum_{j=1}^{i}(Z_j - \hat{\mu}_j)^2}{1+i}, \ \nu_i = \min\left\{1, \sqrt{\frac{2\log(1/\alpha)}{n\hat{\sigma}_{i-1}^2}}\right\}.$$

Further let

$$\mathcal{K}_t^{\pm}(m) = \prod_{i=1}^{t}\{1 \pm \nu_i(Z_i - m)\}.$$

Then

$$\mathbb{P}\left(\max_{t \le n} \mathcal{K}_t^{\pm}(\mu) > \frac{1}{\alpha}\right) \le \alpha.$$

By Proposition C.1, we can construct $\hat{\mu}_1^-$ and $\hat{\mu}_0^+$ as

$$\hat{\mu}_1^- = \sup\left\{m : \max_{t \le n} \mathcal{K}_{t,1}^+(m) \le \frac{4}{\alpha}\right\}, \quad \hat{\mu}_0^+ = \inf\left\{m : \max_{t \le n} \mathcal{K}_{t,0}^-(m) \le \frac{4}{\alpha}\right\},$$

where $\mathcal{K}_{t,1}$ and $\mathcal{K}_{t,0}$ are constructed from the estimated propensity scores in the treated and control groups, respectively.

For the standard deviations, we prove three lower-tail inequalities for the empirical variance estimate.

Proposition C.2. *Let $Z_1, \ldots, Z_n \in [0,1]$ be i.i.d. random variables with $\mathrm{Var}[Z_1] = \sigma^2$ and $\hat{\sigma}$ be the standard error, i.e., $\hat{\sigma}^2 = (1/(n-1))\sum_{i=1}^{n}(Z_i - \bar{Z})^2$, where $\bar{Z} = (1/n)\sum_{i=1}^{n} Z_i$. Further let $m = \lfloor n/2 \rfloor$. Then*

1. *(Improved Hoeffding inequality for U-statistics, modified from Section 5 of Hoeffding [1963])*

$$\mathbb{P}(\hat{\sigma}^2 \le x) \le \exp\left\{-mh\left(x \wedge \sigma^2; \sigma^2\right)\right\},$$

   where $h(a;b) = a\log\{a/b\} + (1-a)\log\{(1-a)/(1-b)\}$.

2. *(Bentkus inequality for U-statistics, modified from Bentkus et al. [2004])*

$$\mathbb{P}\left(\hat{\sigma}^2 \le x\right) \le e\mathbb{P}\left(\mathrm{Bin}\left(m; \sigma^2\right) \le \lceil mx \rceil\right).$$

3. *(Maurer-Pontil inequality, Equation (5) of Maurer and Pontil [2009])*

$$\mathbb{P}\left(\hat{\sigma}^2 \le x\right) \le \exp\left\{-\frac{(n-1)((\sigma^2 - x) \vee 0)^2}{2\sigma^2}\right\}.$$

Proposition C.2 implies that

$$\mathbb{P}\left(\hat{\sigma}^2 \le x\right) \le F(x; \sigma^2),$$

where

$$F(x;\sigma^2) = \min\left\{\exp\left\{-mh\left(x \wedge \sigma^2; \sigma^2\right)\right\}, e\mathbb{P}\left(\mathrm{Bin}\left(m; \sigma^2\right) \le \lceil mx \rceil\right), \exp\left\{-\frac{(n-1)((\sigma^2 - x) \vee 0)^2}{2\sigma^2}\right\}\right\}.$$

Let $F^{-1}(y;\sigma^2)$ be the solution of $F(x;\sigma^2) = y$ given σ^2, and $f(\sigma^2) = F^{-1}(\alpha;\sigma^2)$. Clearly, f is an increasing function. Then

$$\mathbb{P}\left[\hat{\sigma}^2 \le f(\sigma^2)\right] \le \alpha \Longrightarrow \mathbb{P}\left[f^{-1}(\hat{\sigma}^2) \le \sigma^2\right] \le \alpha.$$

As a consequence, $\hat{\sigma}^+ \triangleq \sqrt{f^{-1}(\hat{\sigma}^2)}$ is a valid upper confidence bound of σ. To simplify the computation, note that

$$\hat{\sigma}^+ = \sqrt{f^{-1}(\hat{\sigma}^2)} \Longleftrightarrow \hat{\sigma}^2 = f(\hat{\sigma}^{+2}) = F^{-1}(\alpha; \hat{\sigma}^{+2}) \Longleftrightarrow F(\hat{\sigma}^2; \hat{\sigma}^{+2}) = \alpha.$$

Thus, $\hat{\sigma}^+$ can be reformulated as

$$\hat{\sigma}^+ = \sup\{\sigma : F(\hat{\sigma}^2; \sigma^2) > \alpha\}. \tag{29}$$

Therefore, we can construct $\hat{\sigma}_1^+$ and $\hat{\sigma}_0^+$ based on (29) with α replaced by $\alpha/4$.

Proof of Proposition C.2. 1. Note that $\hat{\sigma}^2$ can be written as a U-statistic

$$\hat{\sigma}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} (Z_i - Z_j)^2.$$

Let $\pi : [n] \mapsto [n]$ be a uniform random permutation. Further let $\tilde{Z}_{k,\pi} = 1 - (Z_{\pi(2k-1)} - Z_{\pi(2k)})^2$. Then $\tilde{Z}_{k,\pi} \in [0,1]$ with $\mathbb{E}[\tilde{Z}_k] = 1 - \sigma^2$ and $\tilde{Z}_{1,\pi}, \ldots, \tilde{Z}_{m,\pi}$ are i.i.d.. Let

$$W_\pi = \frac{1}{m} \sum_{k=1}^{m} \tilde{Z}_{k,\pi}. \tag{30}$$

It is easy to see that

$$1 - \hat{\sigma}^2 = \mathbb{E}_\pi[W_\pi],$$

where $\mathbb{E}_\pi$ denotes the expectation over the randomness of π. By Lemma 1 of Hoeffding [1963],

$$\mathbb{E}\left[e^{\lambda \tilde{Z}_{k,\pi}}\right] \leq \sigma^2 + (1-\sigma^2)e^\lambda.$$

This implies that for any π,

$$\mathbb{E}\left[e^{\lambda m W_\pi}\right] \leq \left(\sigma^2 + (1-\sigma^2)e^\lambda\right)^m.$$

Since $z \mapsto e^{\lambda z}$ is convex, by Jensen's inequality,

$$\mathbb{E}\left[e^{\lambda m(1-\hat{\sigma}^2)}\right] = \mathbb{E}\left[e^{\lambda m \mathbb{E}_\pi[W_\pi]}\right] \leq \mathbb{E}_\pi \mathbb{E}\left[e^{\lambda m W_\pi}\right] \leq \left(\sigma^2 + (1-\sigma^2)e^\lambda\right)^m.$$

By Markov's inequality, for any $\lambda \geq 0$,

$$\mathbb{P}\left(\hat{\sigma}^2 \leq x\right) = \mathbb{P}\left(1 - \hat{\sigma}^2 \geq 1 - x\right) \leq \left(e^{-\lambda(1-x)}\left(\sigma^2 + (1-\sigma^2)e^\lambda\right)\right)^m.$$

Let $\lambda = \log((1-x)\sigma^2/x(1-\sigma^2))$. For any $x \leq \sigma^2$, $\lambda \geq 0$. The tail inequality can be proved via some algebra.

2. By equation (3.1) of Bentkus et al. [2004], for any π,

$$\mathbb{P}\left(\hat{\sigma}^2 \geq 1 - x\right) \leq \inf_{t < 1-x} \frac{\mathbb{E}(\hat{\sigma}^2 - t)_+}{1 - x - t} = \inf_{t < 1-x} \frac{\mathbb{E}(\mathbb{E}_\pi[W_\pi] - t)_+}{1 - x - t}.$$

Since $z \mapsto (z-t)_+$ is convex, by Jensen's inequality,

$$\mathbb{E}(\mathbb{E}_\pi[W_\pi] - t)_+ \leq \mathbb{E}_\pi \mathbb{E}(W_\pi - t)_+.$$

By Lemma 4.3 of Bentkus et al. [2004],

$$\mathbb{E}(W_\pi - t)_+ \leq \mathbb{E}\left(\mathrm{Bin}\left(m, 1-\sigma^2\right)/m - t\right)_+.$$

By Lemma 4.2 of Bentkus et al. [2004], if $m(1-x)$ is an integer,

$$\inf_{t<1-x} \frac{\mathbb{E}(\mathrm{Bin}(m, 1-\sigma^2)/m - t)_+}{1-x-t} \leq e\mathbb{P}(\mathrm{Bin}(m, 1-\sigma^2) \geq m(1-x)) = e\mathbb{P}(\mathrm{Bin}(m, \sigma^2) \leq mx).$$

The proof is completed by putting pieces together.

□

C.2 Tighter DiT O-values

By Theorem 3.3, DiT O-values rely on confidence envelopes of the distributions P_1 and P_0. Although the DKWM inequality has a tight constant for the ℓ_∞ perturbation bound $\sup|\hat{P}_t(A) - P_t(A)|$, the resulting confidence bound is uninformative when

$$\hat{P}_t(A) < \sqrt{\log(4/\alpha)/2n_t} \quad \text{or} \quad \hat{P}_t(A) > 1 - \sqrt{\log(4/\alpha)/2n_t}.$$

Since $\mathcal{O}^+_{\text{DiT}}$ in Theorem 3.3 is likely attained at events with extreme probabilities, the DKWM inequality ma be conservative.

To improve the bounds for rare events, Bates et al. [2021b] develop an upper confidence band of the CDF based on the generalized Simes' inequality [Sarkar et al., 2008].

Proposition C.3 (Bates et al. [2021b], Theorem 4 and Proposition 3)**.** *Let $b_0(\delta) = 0, b_{m+1}(\delta) = 1$, and*

$$b_{m+1-i}(\delta) = 1 - \delta^{1/k}\left(\frac{i\cdots(i-k+1)}{m\cdots(m-k+1)}\right)^{1/k}, \quad i = 1,\ldots,m.$$

Further let $h_{\textit{Simes}}(\cdot;\delta) : [0,1] \mapsto [0,1]$ be a piece-wise constant function such that

$$h_{\textit{Simes}}(t;\delta) = b_{\lceil (m+1)t\rceil}(\delta),\ t \in [0,1].$$

With the same notation as in Proposition 3.2,

$$\mathbb{P}\left[F(z) \le h_{\textit{Simes}}(F_m(z);\delta),\ \forall z \in \mathbb{R}\right] \ge 1-\delta.$$

Bates et al. [2021b] recommend $k = m/2$ for which

$$h_{\text{Simes}}\left(\frac{1}{m};\delta\right) = 1 - \delta^{2/m} = 1 - \exp\left\{-\frac{2}{m}\log\left(\frac{1}{\delta}\right)\right\} \approx \frac{2\log(1/\delta)}{m}.$$

When $\delta = 0.01, m = 1000$, the Simes' upper confidence bound is 0.0092 while the DKWM bound is 0.049, which is more than five times larger. On the other hand, Proposition C.3 is loose when $\hat{F}_m(z)$ is large. For example, the bound is 1 when $\hat{F}_m(z) \ge k/(m+1) \approx 0.5$. To take the best of the worlds, we consider a hybrid upper confidence band by applying the Bonferroni correction on Simes' and DKWM inequalities, i.e.,

$$\hat{P}^+_t([0,x];\delta) = \min\left\{h_{\text{Simes}}\left(\hat{F}_t(x);\frac{\delta}{2}\right), \hat{F}_t(x) + \sqrt{\frac{\log(2/\delta)}{2m}}\right\}. \tag{31}$$

The generalized Simes' inequality only provides an upper confidence bound. To obtain a tighter lower confidence bound for rare events, we invoke Dempster's line crossing probability identity.

Proposition C.4 (Dempster [1959], equation (10))**.** *With the same setting as in Proposition C.3,*

$$\mathbb{P}\left[\hat{F}_m(z) \le b + \frac{1-b}{1-a}F(z),\ \forall z \in \mathbb{R}\right] = 1 - \Delta_{\text{Dempster}}(a,b;m),$$

for any $a, b \in (0,1)$, where

$$\Delta_{\text{Dempster}}(a,b;m) \triangleq a\sum_{j=0}^{\lfloor m(1-b)\rfloor}\frac{m!}{j!(m-j)!}\left(a + \frac{1-a}{1-b}\frac{j}{m}\right)^{j-1}\left(1 - a - \frac{1-a}{1-b}\frac{j}{m}\right)^{m-j}.$$

For any $b, \delta \in (0,1)$, let $a(b;\delta)$ be the solution of

$$\Delta_{\text{Dempster}}(a,b;m) = \delta.$$

The Proposition C.4 implies that

$$\mathbb{P}\left[F(z) \geq \frac{1-a(b;\delta)}{1-b}\max\{0, \hat{F}_m(z)-b\},\ \forall z \in \mathbb{R}\right] = 1-\delta.$$

Here, we set $b = 5/m$ so that the bound is informative whenever $\hat{F}_m(z) > 5/m$. For example, when $\delta = 0.01, m = 1000$, $a(5/m;\delta) = 0.37$. Then the resulting lower confidence band is

$$\frac{1-a(5/m;\delta)}{1-5/m}\cdot\max\left\{0, \hat{F}_m(z)-\frac{5}{m}\right\} = 0.632\cdot\max\left\{0, \hat{F}_m(z)-0.005\right\}.$$

The DKWM bound becomes non-zero only when $\hat{F}_t(z) > \sqrt{\log(2/\delta)/2m} = 0.051$, at which the above bound is 0.029.

As with the Simes' upper confidence band, the Dempster's lower confidence band is loose when $\hat{F}_t(z)$ is large. For example, when $\delta = 0.01, m = 1000, \hat{F}_t(z) = 0.25$, the DKWM bound is 0.198 and the Dempster's bound is 0.155. Therefore, we apply the Bonferroni correction to obtain the best of the worlds, i.e.,

$$\hat{P}_t^-([0,x];\delta) = \min\left\{\frac{1-a(5/m;\delta)}{1-5/m}\cdot\max\left\{0, \hat{F}_m(z)-\frac{5}{m}\right\}, \hat{F}_t(x)-\sqrt{\frac{\log(2/\delta)}{2m}}\right\}. \tag{32}$$

Finally, to obtain confidence envelopes for $P_t([x,1])$, we replace S_i by $1-S_i$ and apply (31) and (32). To guarantee the simultaneity, we set $\delta = \alpha/8$ because eight confidence bands are constructed.

C.3 Tighter DiR O-values

Bates et al. [2021a] develop a substantially tighter concentration inequality for U-statistics that blend three different bounds [Hoeffding, 1963, Bentkus et al., 2004, Maurer, 2006].

Proposition C.5 (Hoeffding–Bentkus–Maurer inequality for bounded U-statistics of order two)**.** *Let $Z_1, \ldots, Z_n$ be i.i.d. and $\phi(z, z')$ be a symmetric bounded kernel taking values in $[0,1]$. Let*

$$S = \frac{1}{n(n-1)}\sum_{i\neq j}\phi(Z_i, Z_j), \quad \mu = \mathbb{E}[S].$$

Let $m = \lfloor n/2 \rfloor$ and $h(a;b)$ be defined as in Proposition C.2. Then, for any $t \in (0,\mu)$,

$$P(S \leq t) \leq \min\Bigg(\exp\left\{-mh(t;\mu)\right\}, eP\left(\mathrm{Bin}(m;\mu) \leq \lceil mt \rceil\right) \inf_{\nu>0}\exp\left\{-\frac{n\nu}{2}\left(\frac{\mu}{1+2G(\nu)}-t\right)\right\}\Bigg), \tag{33}$$

where $G(\nu) = (e^\nu - \nu - 1)/\nu$.

Note that the first two terms are the same as the Proposition C.2. Let $H(t;\mu)$ denote the RHS of (33). Similar to the upper confidence bound $\hat{\sigma}^+$ in Appendix C.1, (33) implies an upper confidence bound for S as

$$\hat{\mu}^+(\delta) = \sup\left\{\mu : H(S;\mu) > \alpha\right\}. \tag{34}$$

For a reasonably good propensity score estimator, p_* is expected to be above 1/2. In this case, we can safely replace bounded $\max\{p^*, 1-p^*\}$ from below by a lower confidence bound of p^*, or equivalently one minus an upper confidence bound of $1-p^*$. Let $\tilde{U} = 2(1-\hat{U})$. Then $\tilde{U}$ is a U-statistics of order 2 with a kernel bounded by 1 and

$$\mathbb{E}[\tilde{U}] = 2n(n-1)\pi(1-\pi)(1-p_*).$$

Applying (34) on $\tilde{U}/n(n-1)$ with α, we can obtain an upper confidence bound $\tilde{\mu}^+$ for $2\pi(1-\pi)(1-p_*)$ in the sense that

$$\mathbb{P}\left(2\pi(1-\pi)(1-p_*) \le \tilde{\mu}^+\right) \ge 1-\alpha.$$

This implies

$$\mathbb{P}\left(\pi(1-\pi)(2p_*-1) \ge \pi(1-\pi) - \tilde{\mu}^+\right) \ge 1-\alpha.$$

As a result, with probability at least $1-\alpha$,

$$\pi(1-\pi)|1-2p_*| \ge \max\left\{\pi(1-\pi) - \tilde{\mu}^+, 0\right\}.$$

Recall that

$$\mathcal{O}^+_{\mathrm{DiR}} = \frac{1}{2} - \pi(1-\pi)|1-2p_*| - \sqrt{\frac{(1-2\pi)^2}{4} + \pi^2(1-\pi)^2(1-2p_*)^2}.$$

For any $a \in \mathbb{R}$, the mapping $y \mapsto -y - \sqrt{a^2+y^2}$ has derivative $-1 - y/\sqrt{a^2+y^2} < 0$. Thus, $\mathcal{O}^+_{\mathrm{DiR}}$ is decreasing in $\pi(1-\pi)|1-2p_*|$. Therefore, replacing $\pi(1-\pi)|1-2p_*|$ by $\max\{\pi(1-\pi) - \tilde{\mu}^+, 0\}$ yields a valid DiR O-value.

C.4 Tighter CE O-values

Proposition 3.4 holds for any empirical process with bounded envelopes and VC dimension 1. Nevertheless, we can sharpen the bound by utilizing the special structure of the classification error. Throughout this section, we assume that S is a continuous; otherwise we perturb S by a tiny random noise (e.g., $\mathrm{Unif}([0, 10^{-9}])$). Then the classification error can be decomposed as

$$\begin{aligned}\mathcal{E}(\eta) = \mathbb{P}\left[T \neq I(S \ge \eta)\right] &= \mathbb{P}(T=1)\mathbb{P}(S < \eta \mid T=1) + \mathbb{P}(T=0)\mathbb{P}(S > \eta \mid T=0) \\ &= \pi P_1([0,\eta]) + (1-\pi)P_0([\eta,1]). \end{aligned} \tag{35}$$

Then we can derive upper confidence bands $\hat{P}_1^+(\cdot;\delta)$ and $\hat{P}_0^+(\cdot;\delta)$ for $P_1(\cdot)$ and $P_0(\cdot)$, respectively, using the same techniques for DiT O-values in Section C.2. The uniformity implies that

$$\mathbb{P}\left(\mathcal{E}(\eta) \le \pi\hat{P}_1^+\left([0,\eta];\frac{\alpha}{2}\right) + (1-\pi)\hat{P}_0^+\left([\eta,1];\frac{\alpha}{2}\right),\ \forall \eta \in [0,1]\right) \ge 1-\alpha.$$

This induces a valid upper confidence bound for $\mathcal{O}^*$:

$$1 - \sup_{\eta\in[0,1]}\left\{\pi\hat{P}_1^+\left([0,\eta];\frac{\alpha}{2}\right) + (1-\pi)\hat{P}_0^+\left([\eta,1];\frac{\alpha}{2}\right)\right\}.$$

D Additional Numerical Results for Lalonde data

In this section, we compute all types of O-values discussed in the main text on Lalonde data. This includes (1) exact DiM, DiT, DiR, and CE O-values and approximate DiM, DiR, and CE O-values for population overlap slacks for ATE, ATT, and ATC, (2) exact and approximate CE O-values for quantile overlap slack $\mathcal{O}^*_\gamma$ with $\gamma = 0.95$, and (3) exact and approximate CE and QVar O-values for $\mathcal{O}^*_{\mathrm{PL}}$. For all O-values, we consider three estimators of propensity scores: Gradient Boosting, Random Forest, and Logistic Regression. We compute the efficiency loss on 50 independent splits and present the boxplots in Figure 5 - 9.

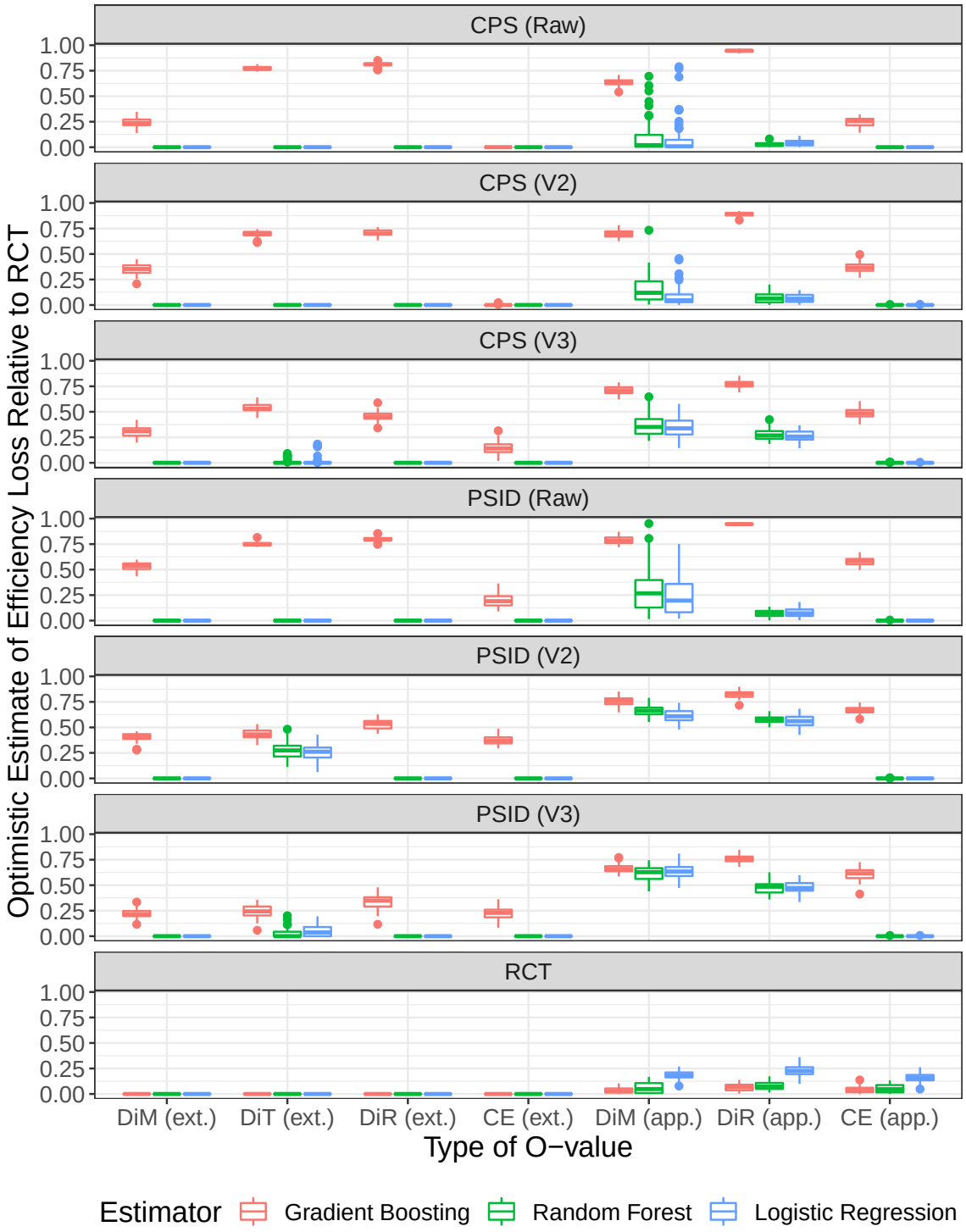


Figure 5: Boxplots of normalized O-values across 50 data splits for population overlap slacks for ATE. The normalized O-value is defined as $\max\{0, 1 - \hat{\mathcal{O}}/\min\{\hat{\pi}, 1 - \hat{\pi}\}\}$. "ext." and "app." are short for "exact" and "approximate", respectively. Each panel corresponds to a control group.

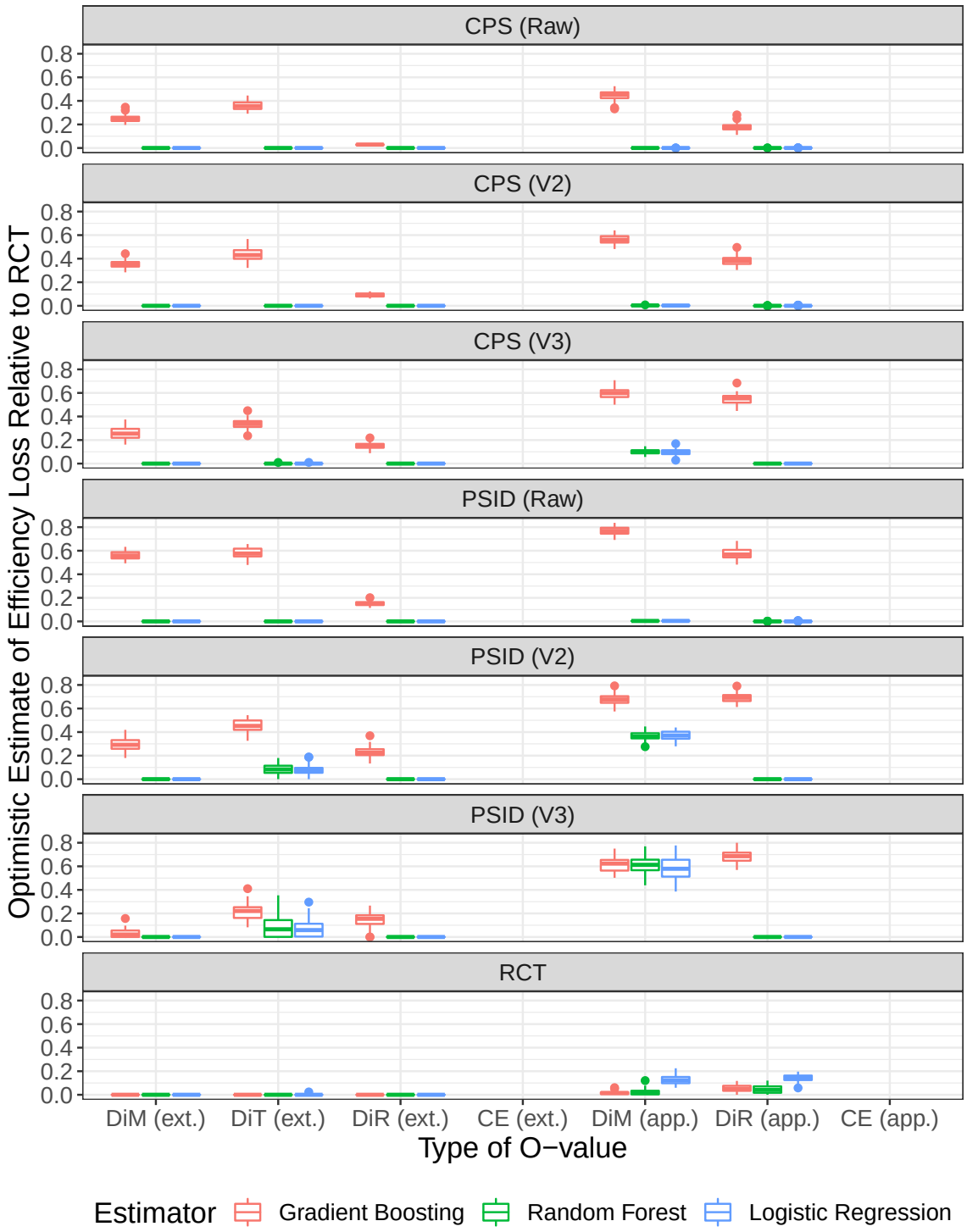


Figure 6: Boxplots of normalized O-values across 50 data splits for population overlap slacks for ATE. The normalized O-value is defined as $\max\{0, \hat{\mathcal{O}}/(1-\hat{\pi})\}$. Other details are same as Figure 5.

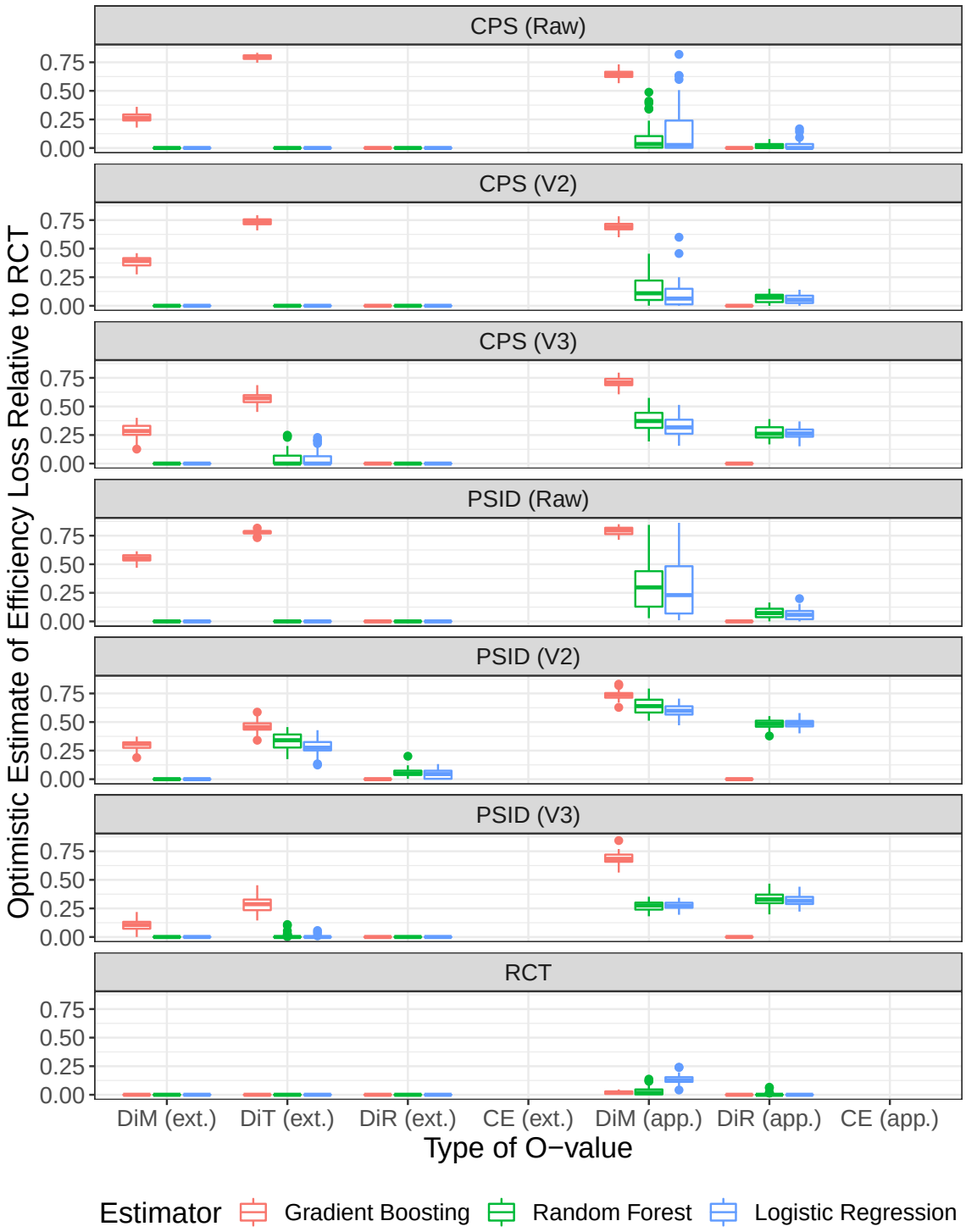


Figure 7: Boxplots of normalized O-values across 50 data splits for population overlap slacks for ATC. The normalized O-value is defined as $\max\{0, \hat{\mathcal{O}}/\hat{\pi}\}$. Other details are same as Figure 5.

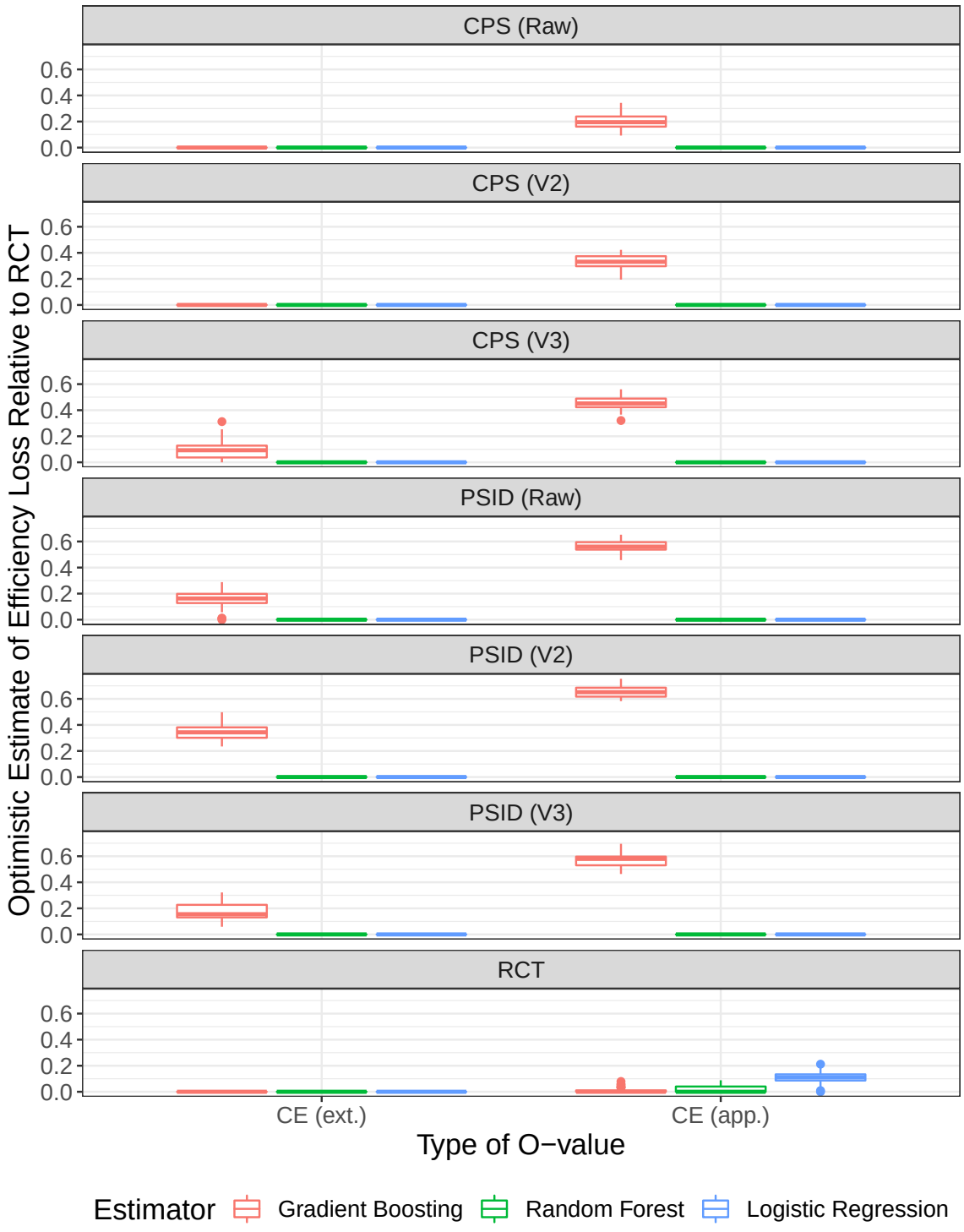


Figure 8: Boxplots of normalized O-values across 50 data splits for quantile overlap slacks $\mathcal{O}^*_{0.95}$. The normalized O-value is defined as $\max\{0, \hat{\mathcal{O}}/\min\{\hat{\pi}, 1-\hat{\pi}\}\}$. Other details are same as Figure 5.

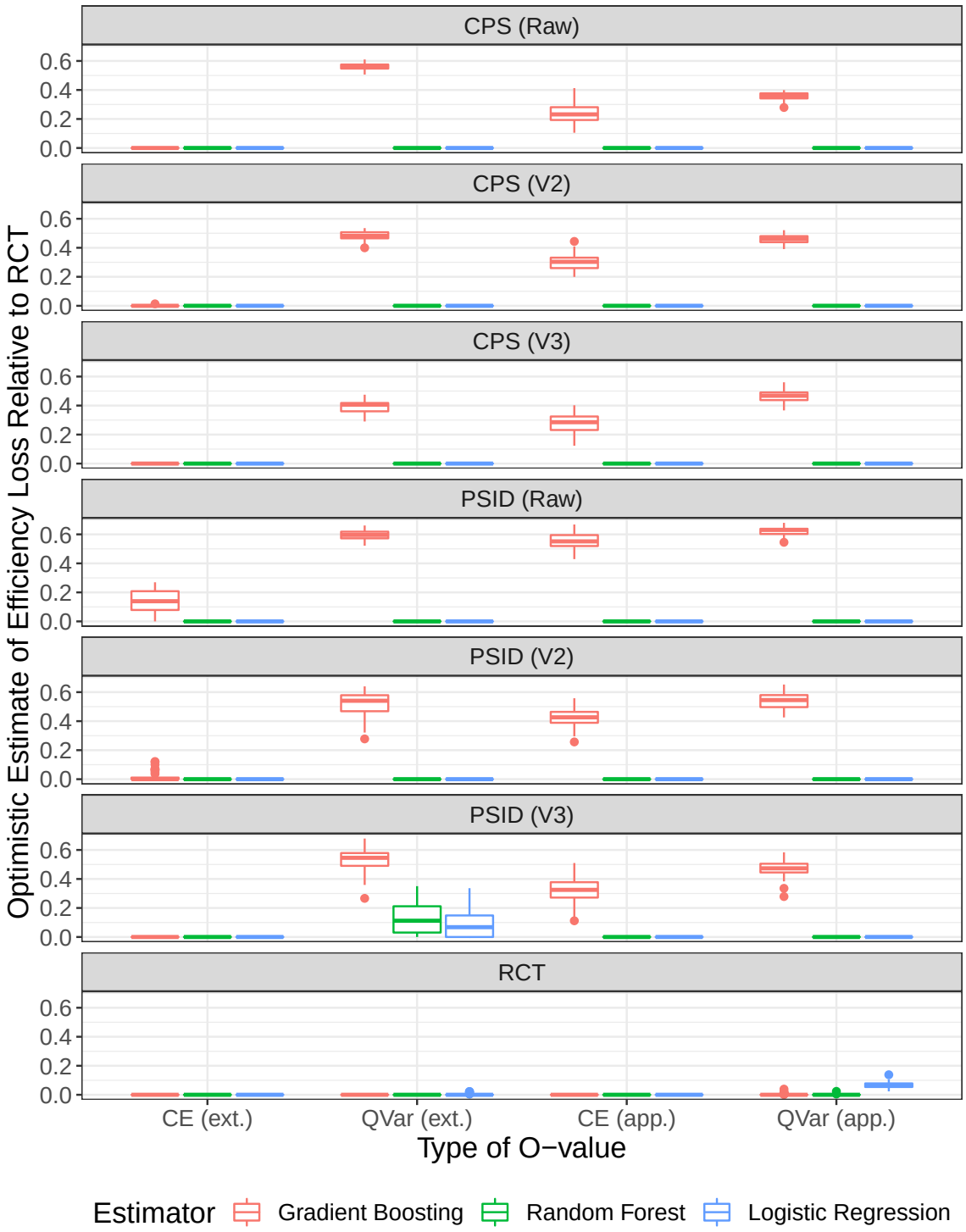


Figure 9: Boxplots of normalized O-values across 50 data splits for quantile overlap slacks $\mathcal{O}^*_{0.95}$. The normalized O-value is defined as $\max\{0, \hat{\mathcal{O}}/\hat{\pi}(1-\hat{\pi})\}$. Other details are same as Figure 5.