

# Improving Precision of RCT-Based CATE Estimation using Data Borrowing with Double Calibration

Amir Asiaee<sup>1</sup>

AMIR.ASIAEETAHERI@VUMC.ORG

Chiara Di Gravio<sup>2</sup>

C.DI-GRAVIO@IMPERIAL.AC.UK

Cole Beck<sup>1</sup>

COLE.BECK@VUMC.ORG

Yuting Mei<sup>1</sup>

YUTINGMEI.VU@GMAIL.COM

Samhita Pal<sup>1</sup>

SAMHITA.PAL@VUMC.ORG

Jared D. Huling<sup>3</sup>

HULING@UMN.EDU

<sup>1</sup>Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, U.S.A.

<sup>2</sup>Department of Epidemiology and Biostatistics, Imperial College London, London SW7 2AZ, U.K.

<sup>3</sup>Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, Minnesota, U.S.A.

## Abstract

Understanding how treatment effects vary across patient characteristics is essential for personalized medicine, yet randomized controlled trials (RCTs) are often underpowered to detect heterogeneous treatment effects (HTEs). We propose a framework that improves the efficiency of conditional average treatment effect (CATE) estimation in RCTs by leveraging large observational studies (OS) while preserving the unbiasedness of RCT estimates. By framing CATE estimation as a supervised learning problem, we show that estimation variance is minimized using the counterfactual mean outcome (CMO) as an augmentation function. We derive finite-sample error bounds and establish conditions under which OS data improves CMO estimation, and thus CATE efficiency, even in the presence of confounding in the OS or outcome distribution shifts between populations. We introduce R-OSCAR (Robust Observational Studies for CMO-Augmented RCT), a two-stage estimator that calibrates OS outcome predictions to the RCT population and corrects residual biases through regularized regression. Simulations show that R-OSCAR can reduce the RCT sample size needed for HTE detection by up to 75%, maintaining robustness to model misspecification. Application to the Tennessee STAR study confirms these efficiency gains. Our framework offers a principled approach to integrating observational and experimental data using tools from statistical learning and transfer learning.

**Keywords:** causal inference, heterogeneous treatment effect, conditional average treatment effect, data integration; observational studies; outcome augmentation; personalized medicine; randomized controlled trial; transfer learning

## 1 Introduction

Understanding how treatment effects vary with patient characteristics is essential for tailoring interventions (Kosorok and Laber, 2019). While randomized controlled trials (**RCTs**) are the gold standard for causal inference and estimating heterogeneous treatment effects (**HTEs**), they are often constrained by cost, time, and limited sample sizes, making it difficult to estimate HTEs across diverse covariates. In contrast, large-scale observational studies (**OS**) offer richer data and broader population coverage, but raise confounding concerns, making causal inference reliant on untestable assumptions such as unconfoundedness. As a result, recent work has focused on integrating RCT

and OS data to mitigate bias in OS, improve the efficiency of RCT estimates, or enhance RCT finding’s generalizability.

Most work on RCT–OS integration has focused on generalizing findings from RCTs to broader OS populations ( $\mathcal{R} \rightarrow \mathcal{O}$ ); see Colnet et al. (2024) and Degtiar and Rose (2023). Our focus differs: we aim to improve estimation efficiency within the RCT population using OS data ( $\mathcal{O} \rightarrow \mathcal{R}$ ).

Data integration to improve RCT precision has a long history, starting with Pocock (1976), who proposed augmenting control arms with data from previous RCTs or OS, treating these as *historical controls*. Approaches fall into two categories: static borrowing (Li and Song, 2020; Liu et al., 2022), which treats external data as part of the RCT, and dynamic borrowing (Ibrahim and Chen, 2000; Neuenschwander et al., 2010; Kaizer et al., 2018; Kotalik et al., 2021), which weights external data based on its similarity or agreement with the current RCT data.

While prior work has mostly focused on borrowing control arms to reduce variance of the treatment effect, less attention has been given to leveraging both treatment arms from OS for improving CATE estimation. Moreover, existing methods often assume CATEs are transportable between populations, a strong assumption that can fail when unmeasured systematic differences exist between populations (Hernán and Robins, 2020). New methods that relax this transportability assumption are needed that reliably leverage both treatment arms of OS data when integrating with RCTs for more efficient CATE estimation.

**Our contribution.** We propose a framework that efficiently leverages OS data to improve CATE estimation in RCTs, without assuming CATE transportability. Our key contributions include:

1. developing an  $\mathcal{O} \rightarrow \mathcal{R}$  framework that borrows both treatment arms from OS, not just controls;
2. relaxing the assumption of CATE or outcome mean exchangeability across populations;
3. using learning theory to derive non-asymptotic error bounds, showing how CATE estimation depends on an augmentation functional we call the Counterfactual Mean Outcome (CMO);
4. identifying conditions under which borrowing improves CMO and thus CATE estimation;
5. providing a practical implementation via regularized regressions in settings with linear outcome models and sparse population differences;
6. and validating our method in extensive simulations and real-world data, showing improved efficiency and power for HTE detection.

Our approach has potentials to enhance personalized treatment strategies by making more efficient use of available data while preserving the unbiasedness guarantees of RCTs. The remainder of this introduction sets up the necessary notation and presents an overview of our results.

## 1.1 Mathematical notation

Let  $Y^{\text{obs}}$  be the observed outcome,  $X \in \mathcal{X}^s$  the observed covariates, and  $A \in \mathcal{A} = \{a_1, \dots, a_T\}$  the treatment in population  $s \in \{r, o\}$ , where  $r$  denotes an RCT and  $o$  an OS. Let  $Y(a)$  denote the potential outcome under treatment  $a$ , and assume consistency:  $Y = Y(a)$  when  $A = a$ . Then the joint density decomposes as:

$$\begin{aligned} \mathbb{P}^s(Y^{\text{obs}}, X = x, A = a) &= \mathbb{P}^s(Y(a), X = x, A = a) \\ &= \mathbb{P}^s(Y(a) | X = x, A = a) \mathbb{P}^s(A = a | X = x) \mathbb{P}^s(X = x) \\ &= \mathbb{P}(Y(a) | X = x, A = a, S = s) \mathbb{P}(A = a | X = x, S = s) \mathbb{P}(X = x | S = s). \end{aligned} \tag{1}$$

The last line articulates dependence on the study source  $s$ . When  $\mathbb{P}^r(Y(a), X, A) = \mathbb{P}^o(Y(a), X, A)$ , both datasets share the same distribution. Otherwise, any of the three terms in (1) may differ. In machine learning,  $\mathbb{P}^r(X) \neq \mathbb{P}^o(X)$  is called covariate shift, while  $\mathbb{P}^r(Y(a)|X) \neq \mathbb{P}^o(Y(a)|X)$  is known as outcome (concept or label) shift (Kouw and Loog, 2018).

A summary of key notation appears in Table 1. We assume binary treatment  $A \in \{-1, +1\}$ . Our goal is to use OS data ( $s = o$ ) to improve the efficiency of CATE estimation in the RCT ( $s = r$ ), where  $\tau^r(x) = \mathbb{E}^r[Y(+1) - Y(-1) | X = x]$ . We denote the OS dataset as  $\{(X_i^o, A_i^o, Y_i^o)\}_{i=1}^{n_o}$  and the RCT dataset as  $\{(X_i^r, A_i^r, Y_i^r)\}_{i=1}^{n_r}$ , with  $n_r \ll n_o$  typically. A key quantity in our framework is the *counterfactual mean outcome* (CMO),  $\mu^s(x)$ , defined as the average of  $\mu_{+1}^s(x)$  and  $\mu_{-1}^s(x)$  under swapped treatment assignment probabilities. Intuitively, the CMO represents the expected outcome at covariates  $x$  under a hypothetical reversal of the observed treatment assignment.

Table 1: Notation and definitions used throughout the manuscript.

Symbol	Description
$A$	Treatment indicator; $A \in \{-1, 1\}$ where 1 is for treatment
$Y$	Observed outcome of interest
$S$	Study indicator; $S \in \{r, o\}$ with $r$ : RCT, $o$ : observational
$p$	Number of measured covariates in both studies
$X^s$	Covariates observed in study $S = s$ ; $X^s \in \mathcal{X}^s \subseteq \mathbb{R}^p$
$n^s$	Sample size of study $S = s$
$\tau^s$	Average treatment effect (ATE) in population $s$ ; $\tau^s \equiv \mathbb{E}^s[Y(+1) - Y(-1)]$
$\mu_a^s(x)$	Conditional mean potential outcome in study $s$ ; $\mu_a^s(x) \equiv \mathbb{E}^s[Y(a)   X = x]$
$\mu^s(x, a)$	Regression function (outcome mean) in study $s$ ; $\mu^s(x, a) \equiv \mathbb{E}^s[Y   X = x, A = a]$
$\tau_a^s(x)$	Conditional average treatment effect (CATE); $\tau_a^s(x) \equiv \mathbb{E}^s[Y(+1) - Y(-1)   X = x]$
$\pi_a^s(x)$	Treatment assignment probability in study $s$ ; $\pi_a^s(x) \equiv \mathbb{P}^s(A = a   X = x)$
$\mu^s(x)$	Counterfactual mean outcome (CMO) in population $s$ ; $\mu^s(x) \equiv \sum_a \pi_a^s(x) \mu_a^s(x)$
$\delta_a(x)$	Mean discrepancy between populations $r$ and $o$ ; $\delta_a(x) \equiv \mu_a^r(x) - \mu_a^o(x, a)$
$\Delta_2(f, g)$	$L^2$ distance (RMSE) between $f$ and ground truth $g$ ; $\Delta_2^s(f, g) \equiv \mathbb{E}_X^s[(f(X) - g(X))^2]$
$\Delta_\infty(f, g)$	$L^\infty$ distance between function $f$ and ground truth $g$ ; $\Delta_\infty(f, g) \equiv \sup_X  f(X) - g(X) $
$\Delta_2^2(\mathcal{F}, g)$	Approximation error; $\Delta_2^2(\mathcal{F}, g) \equiv \inf_{f \in \mathcal{F}} \mathbb{E}_X^s[(f(X) - g(X))^2]$

## 1.2 Technical overview of our framework and results

Our goal is to *improve the efficiency* of CATE estimation in an RCT population by leveraging OS data. We frame CATE estimation as a supervised learning task: given observed triplets  $(X, A, Y)$  and an *augmentation function*  $m(X)$ , we define the pseudo-outcome  $\tau_m(X, A, Y) \equiv \frac{A(Y - m(X))}{\pi_A(X)}$ . We estimate  $\hat{\tau}(x)$  as the conditional mean of pseudo-outcomes by minimizing squared loss over a function class  $\mathcal{F}$ . Specifically, the CATE for the RCT is estimated as:

$$\hat{\tau}^r(\cdot) = \arg \min_{f \in \mathcal{F}} \frac{1}{n^r} \sum_{i=1}^{n^r} (\tau_m(X_i^r, A_i^r, Y_i^r) - f(X_i^r))^2. \quad (2)$$

Crucially, the estimator’s risk depends on the choice of  $m$ . Our key insight is that the risk-minimizing choice is the *counterfactual mean outcome* (CMO):  $\mu^r(x) \equiv \pi_{-1}^r(x) \mu_{+1}^r(x) + \pi_{+1}^r(x) \mu_{-1}^r(x)$ .

Under standard assumptions, our analysis yields two key results. **First**, with high probability:

$$\Delta_2^2(\hat{\tau}^r, \tau^r) \lesssim (1 + \Delta_2(m, \mu^r)) \mathcal{R}_{n^r}(\mathcal{F}),$$

where  $\Delta_2(\cdot, \cdot)$  is the  $L^2$  error and  $\mathcal{R}_{n^r}(\mathcal{F})$  is the Rademacher complexity. This shows that a more accurate CMO (as  $m$ ) reduces CATE estimation error. **Second**, the prediction risk is bounded as:

$$R(\hat{\tau}^r) \equiv \mathbb{E}^r [\tau_m(X, A, Y) - \hat{\tau}^r(X)]^2 \lesssim \bar{\sigma}^2 + \Delta_2^2(m, \mu^r) + \Delta_2^2(\hat{\tau}^r, \tau^r),$$

capturing irreducible noise ( $\bar{\sigma}^2$ ), CMO, and CATE mean square errors (**MSE**). These bounds motivate improving CMO estimates to reduce both prediction and estimation error.

Since the CMO is a weighted average of  $\mu_{+1}^r(x)$  and  $\mu_{-1}^r(x)$ , improving its estimation requires accurate estimation of these per-arm outcomes. Intuitively, estimating  $\mu_a^r(x)$  using only RCT data yields unbiased but high-variance estimates, while using OS data can reduce variance at the cost of bias. We propose methods that combine RCT and OS data to reduce the overall error in estimating  $\mu_a^r(x)$ , and thus the CMO, without introducing bias in the CATE.

To relax the CATE and outcome mean exchangeability assumptions, often required for data integration, we adopt an *outcome-shift* framework in which conditional outcome models differ across populations. Let  $\mu^o(x, a) \equiv \mathbb{E}^o[Y \mid X = x, A = a]$  and  $\mu_a^r(x) \equiv \mathbb{E}^r[Y(a) \mid X = x]$ . We define the discrepancy  $\delta_a(x) \equiv \mu_a^r(x) - \mu^o(x, a)$ , and assume  $\delta_a(x)$  belongs to a structured class (e.g., sparse or low-dimensional). This formulation captures the relationship between the two populations without requiring full exchangeability and aligns with the concept-shift framework in transfer learning (Zhuang et al., 2020). It allows us to improve estimates of  $\mu_a^r(x)$ , and thus the CMO, by appropriately correcting for systematic differences between OS and RCT data.

We propose, a doubly-calibrated OS-augmented procedure for estimating CATE in the RCT, R-OSCAR. Algorithm 1 implements R-OSCAR through four steps: estimate outcome models  $\hat{\mu}^o(\cdot, a)$  (line 3) using the OS, estimate their discrepancies  $\hat{\delta}_a(\cdot)$  from the RCT data to calibrate for population differences (line 4), construct the CMO  $m(\cdot)$  (line 6), then estimate the CATE discrepancy  $\hat{\delta}(\cdot)$  using pseudo-outcomes.

---

**Algorithm 1** R-OSCAR: Robust OS for CMO-Augmented RCT

---

- 1: **Input:** OS data  $\{(X_i^o, A_i^o, Y_i^o)\}_{i=1}^{n^o}$ , RCT data  $\{(X_i^r, A_i^r, Y_i^r)\}_{i=1}^{n^r}$
  - 2: **for**  $a \in \{-1, +1\}$  **do**
  - 3:     **OS outcome modeling:**  $\hat{\mu}^o(\cdot, a) = \operatorname{argmin}_f \frac{1}{n_a^o} \sum_{i: A_i^o=a} [Y_i^o - f(X_i^o)]^2 + R_a^o(f)$
  - 4:     **Outcome calibration to RCT:**  $\hat{\delta}_a(\cdot) = \operatorname{argmin}_d \frac{1}{n_a^r} \sum_{i: A_i^r=a} [Y_i^r - \hat{\mu}^o(X_i^r, a) - d(X_i^r)]^2 + R_a^r(d)$
  - 5: **end for**
  - 6: **CMO construction:**  $m(\cdot) = \sum_a \pi_a^r(\cdot) [\hat{\mu}^o(\cdot, a) + \hat{\delta}_a(\cdot)]$
  - 7: **CATE calibration:**  
 $\hat{\delta}(\cdot) = \operatorname{argmin}_d \frac{1}{n^r} \sum_{i=1}^{n^r} [\tau_m(X_i^r, A_i^r, Y_i^r) - \sum_a a(\hat{\mu}^o(X_i^r, a) + \hat{\delta}_a(X_i^r)) - d(X_i^r)]^2 + R(d)$
  - 8: **Output:**  $\hat{\tau}^r(\cdot) = \sum_a a[\hat{\mu}^o(\cdot, a) + \hat{\delta}_a(\cdot)] + \hat{\delta}(\cdot)$
- 

All  $R$  terms represent suitable function regularizers (e.g., ridge or lasso penalties). Without regularization in outcome calibration, estimated discrepancies may ignore information from the OS, so the penalties  $R_a^r$  ensure effective “knowledge transfer.” Separate arm regularization can introduce

regularization bias, where independently regularized outcome functions yield poor CATE estimates when contrasted. The CATE calibration step corrects this bias while its penalty term helps borrow information from the preliminary CATE estimate. For ease of analysis, our theoretical results use constrained formulations of all steps, while practical implementations employ regularized versions that maintain equivalent solution sets and theoretical guarantees.

For theoretical validity, outcome and CATE calibration require independent RCT samples. In practice, we use  $K$ -fold sample splitting: partition RCT data into  $K$  folds, use  $K - 1$  folds for outcome calibration and one fold for CATE calibration in each iteration. The final estimator averages across folds:  $\hat{\tau}^r(x) = \sum_a a[\hat{\mu}^o(x, a) + \frac{1}{K} \sum_{k=1}^K \hat{\delta}_a^k(x)] + \frac{1}{K} \sum_{k=1}^K \hat{\delta}^k(x)$ . Empirically, sample-split results closely match the full-data approach.

### 1.3 Related Work

#### 1.3.1 CATE ESTIMATION

The field of heterogeneous treatment effect estimation has seen significant methodological development in recent years, particularly through the integration of machine learning techniques (Jacob, 2021). Current methodological approaches in CATE estimation can be broadly categorized into two groups: meta-learners that leverage off-the-shelf machine learning methods (Künzel et al., 2019) and specialized algorithms specifically designed for causal inference.

Meta-learners represent a flexible framework that adapts standard machine learning algorithms for estimating required functions such as outcome means and propensities for causal inference tasks. These include the *S-learner*, which models a single function including the treatment as a feature, and the *T-learner*, which estimates separate conditional mean functions for treated and control groups. In our notation, the T-learner estimator becomes  $\sum_a a\hat{\mu}_a^r(x)$ . However, separate regularized estimation of outcome models may lead to different regularization patterns, potentially introducing artifacts in the final CATE estimates known as regularization bias (Wager, 2024). In contrast, the estimator in (7) provides greater robustness by directly targeting CATE estimation, making it less susceptible to such estimation artifacts. Other meta-learners include the *doubly-robust (DR) learner*, which combines regression adjustment with inverse probability weighting (Kennedy, 2023); the *R-learner*, which employs orthogonalization techniques (Nie and Wager, 2021); and the *X-learner*, which addresses treatment effect heterogeneity in imbalanced experimental settings (Künzel et al., 2019). These approaches leverage various machine learning algorithms such as random forests, gradient boosting, neural networks, and LASSO regression to tailor estimation to specific dataset characteristics.

In parallel, specialized algorithms have been developed to directly estimate heterogeneous treatment effects, including causal forests (Athey et al., 2019), causal BART (Hahn et al., 2020), and causal boosting (Powers et al., 2018). These methods modify traditional machine learning approaches to specifically target treatment effect estimation, often incorporating sample-splitting and cross-fitting techniques to improve robustness and reduce overfitting. As the field continues to evolve, research increasingly focuses on understanding the statistical properties of these estimators, developing appropriate inference techniques, and addressing challenges in high-dimensional settings (Fan et al., 2022).

A closely related concept is that of *individualized treatment rules (ITRs)*, which aim to recommend the optimal treatment for each individual based on their covariates (Qian and Murphy, 2011). For binary treatments, ITRs are often derived by thresholding the estimated CATE: treatment is

assigned when the estimated CATE is positive. Recent advances have also focused on directly estimating ITRs using methods such as outcome-weighted learning (Zhao et al., 2012), policy learning (Athey and Wager, 2017), and doubly robust policy evaluation (Dudik et al., 2011). Generalizing ITRs to target populations under covariate shift has recently been addressed by Chen et al. (2024). These methods are typically evaluated using the value function, which quantifies the expected outcome under the learned policy. While conceptually distinct, ITR estimation is closely tied to CATE estimation, as accurate CATEs support more effective individualized decision-making.

### 1.3.2 RCT AND OS DATA INTEGRATION ASSUMPTIONS

As mentioned earlier, most work in data integration follows the  $\mathcal{R} \rightarrow \mathcal{O}$  direction, commonly known as generalization or transportability. While the exact definitions and their corresponding assumptions vary between potential outcome (Dahabreh et al., 2020) and structural causal model (Bareinboim and Pearl, 2016) frameworks, we broadly refer to these as generalizability methods where the goal is to generalize the average treatment effect from RCT to OS. Although our work studies the  $\mathcal{O} \rightarrow \mathcal{R}$  direction, understanding the assumptions of generalizability methods remains relevant. These methods primarily address shifts in covariate distribution ( $\mathbb{P}^r(x) \neq \mathbb{P}^o(x)$ ) while assuming some aspect of the outcome model remains invariant across populations. This invariance typically relies on one or more of the following assumptions in descending order of strength (Colnet et al., 2024): **1)** CATE invariance ( $\tau^r(x) = \tau^o(x)$ ), **2)** mean exchangeability ( $\forall a : \mu_a^r(x) = \mu_a^o(x)$ ), or **3)** trial participation ignorability ( $Y(a), Y(a') \perp S \mid X$ ). The strongest assumption, trial participation ignorability, is equivalent to the “no outcome shift” assumption, which states that  $\mathbb{P}^r(Y(a) \mid X = x) = \mathbb{P}^o(Y(a) \mid X = x)$ .

These invariance assumptions have been criticized as unrealistic in practical applications. Hernán and Robins (2020) identifies three primary ways the trial participation ignorability assumption can be violated: shifts in the distribution of unobserved effect modifiers across populations, inconsistencies in treatment versions between populations, and altered interference patterns of units across different populations. We address these concerns by explicitly modeling and accounting for outcome shifts between the RCT and OS populations rather than assuming mean exchangeability.

### 1.3.3 DATA INTEGRATION FOR CATE ESTIMATION IN RCT

Recent work has explored combining OS and RCT data to improve CATE estimation for the RCT population. These approaches can be categorized into two main methodological frameworks, each with distinct assumptions and strategies.

**Weighted Combination Approaches.** The first category focuses on combining CATE estimates from both data sources through optimal weighting schemes. The approach of Cheng and Cai (2021) uses a weighted combination of preliminary CATE estimates from both RCT and OS datasets, where the weight is determined adaptively based on the mean squared error of the combined estimator relative to a proxy derived from out-of-fold RCT estimates. This allows efficient borrowing of information from the observational data while controlling for potential bias. A related line of work by Oberst et al. (2022) studies the optimal convex combination of an unbiased and a possibly biased estimator in a general setting, providing sharp MSE-based bias thresholds to guide integration. While not specific to CATE, their framework is directly applicable to combining RCT and OS estimators for causal inference.

**Confounding Function Approaches.** The second category, which is closest in spirit to our approach, relaxes strong CATE transportability assumption by explicitly modeling the bias between populations. This category comprises only two recent works: (Wu and Yang, 2022; Yang et al., 2025), which integrate OS and RCT data to improve CATE estimation efficiency for the RCT population. Like our work, they relax stringent transportability conditions, though they focus on relaxing the CATE invariance assumption by introducing a confounding function that captures hidden biases specific to the OS data. In contrast, our approach relaxes the mean exchangeability assumption. Building on semiparametric modeling techniques, Wu and Yang (2022) propose orthogonalized loss functions, while Yang et al. (2025) derive semiparametric efficient scores and construct estimators that achieve the efficiency bound.

#### 1.3.4 RELATIONSHIP TO OTHER LEARNING PARADIGMS

Transfer learning can be viewed as the predictive analogue of data integration for causal effect estimation, where the goal is to use source datasets to improve prediction accuracy—rather than estimate causal effects—in a target population. This area is well-studied in the machine learning literature (Weiss et al., 2016; Zhuang et al., 2020). Within transfer learning, domain adaptation methods have received the most attention. These approaches primarily aim to address covariate shift, where the input distribution differs between source and target domains (Ben-David et al., 2010), while outcome shift has been comparatively less explored (Kouw and Loog, 2018).

While transfer learning generally assumes a sequential relationship between source and target tasks, multi-task learning instead seeks to jointly train models across multiple related tasks, enabling the sharing of structure and inductive bias (Zhang and Yang, 2021). The tasks in transfer and multi-task learning methods may involve predicting related but distinct outcomes (Li et al., 2022; Zhili et al., 2020), predicting the same outcome across different populations (Suresh et al., 2019; Steingrimsson et al., 2023) or treatment groups (Wang et al., 2022; Strauch and Asiaee, 2024), or addressing temporal distributional changes (Nguyen et al., 2020; Helli et al., 2024).

Our framework enables the integration of transfer learning techniques to estimate conditional mean outcomes in the RCT population, using OS data as the source domain. These transferred outcome models are then used to construct CMO estimates in the RCT population, which can be safely incorporated into CATE estimation without introducing bias. Our outcome discrepancy formulation provides a principled way to implement this idea, drawing on prior work in high-dimensional multi-task learning under sparse outcome shift assumptions (Gross and Tibshirani, 2016; Asiaee et al., 2018).

To the best of our knowledge, our work is the first to explicitly connect transfer learning theory to CATE estimation across populations. The proposed method offers a flexible framework that accommodates a wide range of transfer learning strategies for outcome modeling, enhancing the efficiency of CATE estimation while preserving consistency guarantees.

The remainder of the manuscript is organized as follows. Section 2 introduces our theoretical framework, formulating CATE estimation as a supervised learning problem and establishing how its efficiency depends on the MSE between the augmentation function ( $m$ ) and the true CMO ( $\mu$ ) (Section 2.1). We then show how to leverage OS data to improve CMO estimation (Section 2.2), including a special case based on sparse shifts in linear outcome models (Section 2.3). Section 3 presents simulation studies, and Section 4 validates our method on real-world data. Appendix A contain all proofs.

## 2 A general framework for improving efficiency of CATE estimation

This section presents our general framework for improving the efficiency of CATE estimation by incorporating OS data. We begin by showing that CATE estimation can be formulated as supervised learning with a pseudo-outcome whose variance is minimized when using the true CMO as the augmentation function (Section 2.1). We then propose strategies for estimating the CMO using both RCT and OS data under outcome-shift assumptions (Section 2.2). Finally, we illustrate the framework using linear models with sparse differences across populations (Section 2.3).

### 2.1 Improved CATE estimation via augmentation: general prediction and estimation risk bounds

Here, we reformulate CATE estimation as an empirical risk minimization problem, first demonstrating how efficient point-wise CATE estimation is achievable through a careful choice of the augmentation function. We then show how this formulation facilitates the systematic reduction of prediction risk and estimation error of the CATE. For clarity, in this section, we present our results without study population indicators, as the results hold under appropriate assumptions in any study population. While these assumptions (except A1) are typically satisfied by design in RCTs, they become explicit requirements for observational studies. Importantly, since our ultimate goal is to estimate CATE for the RCT population, these assumptions are only required to hold for the RCT data; their validity in the observational study is not necessary.

#### 2.1.1 MORE EFFICIENT POINTWISE CATE ESTIMATION VIA AUGMENTATION

The following proposition formalizes the supervised learning perspective for CATE estimation by showing how it can be framed as a risk minimization problem using transformed outcomes.

**Proposition 1** *Consider any study population where the following hold:*

- (A1) (SUTVA)  $Y_i = Y_i(A_i)$
- (A2) (Conditional Ignorability)  $(Y(-1), Y(+1)) \perp A \mid X$
- (A3) (Weak Positivity)  $\forall X, A : 0 < \pi_A(X) < 1$

*Let  $m(X)$  be any function of  $X$ . For any given sample  $(X, A, Y)$ , define the following transformation:  $\tau_m(X, A, Y) \equiv \frac{A(Y - m(X))}{\pi_A(X)}$ . Then,  $\tau_m(X, A, Y)$  is an unbiased estimator of the CATE at  $X$ , that is:  $\mathbb{E}_{A,Y}(\tau_m(x, A, Y) \mid X = x) = \tau(x)$ . Moreover, for any  $x$ , the true CATE  $\tau(x)$  minimizes the following conditional risk:*

$$\tau(x) = \underset{f}{\operatorname{argmin}} \mathbb{E}_{Y,A} \left[ (\tau_m(x, A, Y) - f(x))^2 \mid X = x \right], \quad (3)$$

The special case of the transformation  $\tau_0(X, A, Y) \equiv 2AY$  with equal treatment assignment probability  $\pi_A(X) = 1/2$  and augmentation function  $m(X) = 0$  was introduced by Tian et al. (2014), where it was referred to as the “modified outcome.” Subsequently, Athey and Imbens (2015) extended this idea to general propensity functions  $\pi_A(X)$ , still with  $m(X) = 0$ , calling it the “CATE-generating transformation.”, we further generalize this transformation by allowing an arbitrary augmentation function  $m(X)$ . We denote this general form by  $\tau_m(X, A, Y)$  and refer to it as the “pseudo-outcome,” emphasizing its use within supervised learning frameworks.



Since  $m(X)$  is arbitrary, we have the flexibility to choose it in a way that minimizes the variance of the CATE estimator  $\tau_m(X, A, Y)$  at  $X = x$ . The following theorem demonstrates that minimum conditional variance for  $\tau_m(X, A, Y)$  is achieved by choosing  $m(X)$  to be the CMO,  $\mu(X)$ .

**Theorem 2** *Assuming assumptions (A1)-(A3) of Proposition 1 hold. Then, setting  $m(X) := \mu(X)$  in (3) results in the minimum variance CATE estimator. In other words, the solution of the following is the CMO:  $\mu(x) = \operatorname{argmin}_m \mathbb{V}(\tau_m(x, A, Y) | X = x)$ .*

One can show (see Proposition 11 in Appendix A) that the mean potential outcome decomposes into the sum of the CMO and the CATE. This decomposition offers an intuitive interpretation of Theorem 2: subtracting the CMO from outcome yields the most efficient estimator of the CATE. The function  $m(X)$  can be viewed as either a *nuisance function* that removes baseline variation without introducing bias, or as an *augmentation function* that improves efficiency when properly chosen. We adopt the term **augmentation function**, following the augmented inverse probability weighting (AIPW) literature (Kurz, 2022), as it better captures the specific role of  $m(X)$  in enhancing CATE estimation efficiency.

### 2.1.2 BOUNDING AND MINIMIZING THE EXPECTED CATE PREDICTION ERROR

Use of the true CMO is optimal for variance reduction with respect to the population loss function. In practice an empirical loss function is minimized using the observed data. Furthermore, the true CMO is never available and must be estimated from finite samples, introducing an additional layer of estimation error. This raises critical questions about the practical implications of outcome transformation: how does the selected  $m(X)$  impact CATE estimation across all covariates  $x$  when the corresponding plug-in estimator of (3) is used? and how does the error in estimating the CMO influence the efficiency of the resulting CATE estimates?

To address these questions, we leverage the standard concept of risk, or expected squared prediction error, from supervised learning. In a typical supervised learning task, the risk of a function  $f$  is defined and decomposes as:  $R(f) \equiv \mathbb{E}_{X,Y}[(Y - f(X))^2] = \mathbb{E}_X \mathbb{V}(Y | X) + \Delta_2^2(f, g)$ , where  $(X, Y)$  is the predictor-outcome pair,  $\Delta_2^2(f, g) = \mathbb{E}_X[f(X) - g(X)]^2$ , and  $g(X) \equiv \mathbb{E}[Y | X]$  is the Bayes estimator, minimizing the risk in the MSE sense. This decomposition reveals two components: 1) The **irreducible error**,  $\mathbb{E}_X \mathbb{V}(Y | X)$ , representing the variance of  $Y$  given  $X$ . 2) The **estimation error**,  $\Delta_2^2(f, g) = \mathbb{E}_X[f(X) - g(X)]^2$ , quantifying the MSE of  $f$  with respect to  $g$ . For homoskedastic additive noise, the risk simplifies to  $R(f) = \sigma^2 + \Delta_2^2(f, g)$ , where  $\sigma^2$  represents the irreducible noise variance in the model  $Y = g(X) + \epsilon$  with  $\mathbb{E}[\epsilon | X] = 0$ .

Analogously, in the supervised learning formulation of CATE estimation, the transformed outcome  $\tau_m(X, A, Y)$  serves as the analog of  $Y$ , and the true CATE  $\tau(X) = \mathbb{E}[\tau_m | X]$  plays the role of  $g(X)$ . Then for the risk of any CATE estimator we have:

$$R_m(\hat{\tau}) = \mathbb{E}_{X,A,Y}[\tau_m(X, A, Y) - \hat{\tau}(X)]^2 = \mathbb{E}_X \mathbb{V}(\tau_m | X) + \Delta_2^2(\hat{\tau}, \tau). \quad (4)$$

**Bounding and minimizing irreducible error.** Here, unlike a typical supervised learning setting, the **irreducible error**  $\mathbb{E}_X \mathbb{V}(\tau_m | X)$  depends on the choice of  $m(X)$ . As shown in Theorem 2, this term is minimized when  $m(X) = \mu(X)$ , the true CMO. The “irreducible error” is a misnomer in this context because it can be reduced by improving the choice of  $m(X)$ . However, given its established usage in regression, we retain the term here for consistency. Proposition 3 summarizes our discussion.

**Proposition 3 (Risk Reduction with the Optimal Augmentation Function)** *The variance of the transformed outcome  $\tau_m$  determines the irreducible error in CATE estimation. Selecting the augmentation function  $m(X)$  to minimize  $\mathbb{E}_X \mathbb{V}(\tau_m | X)$  reduces the risk of ANY estimator  $\hat{\tau}(X)$ . The optimal augmentation function,  $m(X) = \mu(X)$ , achieves the smallest possible irreducible error.*

Proposition 3 underscores the importance of quantifying the relationship between the quality of  $m(X)$  and the variance reduction achieved. In the following theorem, we establish bounds on the mean irreducible variance  $\mathbb{E}_X \mathbb{V}(\tau_m | X)$  as a function of the MSE of  $m(X)$  relative to the true CMO  $\mu(X)$ ,  $\Delta_2^2(m, \mu)$ . These bounds provide a practical criterion for determining when an estimated CMO, subject to estimation error and possible misspecification, improves the risk over simpler alternatives, such as setting  $m(X) = 0$ .

**Theorem 4** *Under assumptions (A1)-(A3) of Proposition 1 and:*

(A4) (Strong Positivity)  $\exists \rho \in (0, 1/2]$  such that  $\rho \leq \pi_a(X) \leq 1 - \rho$  for  $a \in \{-1, 1\}$ ;  
Let  $\mu$  be the true CMO and  $m$  be any function used to generate pseudo-outcome  $\tau_m$  introduced in Proposition 1. Define  $\Delta_2^2(m, \mu) = \mathbb{E}_X [(m(X) - \mu(X))^2]$  as the mean squared error between  $m(X)$  and  $\mu(X)$ . The excess irreducible error from using  $m$  instead of  $\mu$  is bounded as:

$$\frac{\Delta_2^2(m, \mu)}{(1 - \rho)^2} \leq \mathbb{E}[\mathbb{V}(\tau_m | X)] - \mathbb{E}[\mathbb{V}(\tau_\mu | X)] \leq \frac{\Delta_2^2(m, \mu)}{\rho^2}, \quad (5)$$

where  $\mathbb{V}(\tau_\mu | X)$  is the variance of the oracle estimator using the true CMO.

For the special case of  $\pi_{-1}^r = \pi_{+1}^r = 1/2$ , the inequalities simplify to:  $\mathbb{E}[\mathbb{V}(\tau_m | X)] = \mathbb{E}[\mathbb{V}(\tau_\mu | X)] + 4\Delta_2^2(m, \mu)$ . Note that  $\mathbb{V}(\tau_\mu | X = x)$  represents the minimum achievable variance at  $X = x$  according to Theorem 2 and is therefore truly irreducible. By combining (4) and (5), we establish how the MSE between  $m(x)$  and the CMO  $\mu(x)$  controls the prediction risk:

**Corollary 5** *Let  $\bar{\sigma}^2 \equiv \sup_{x \in \mathcal{X}} \mathbb{V}(\tau_\mu | X = x)$  denote the supremum of the conditional variance of the optimal transformed outcome. Then, under the assumptions of Theorem 4, the prediction risk admits the following bound:  $R_m(\hat{\tau}) \leq \bar{\sigma}^2 + \frac{\Delta_2^2(m, \mu)}{\rho^2} + \Delta_2^2(\hat{\tau}, \tau)$ .*

This risk decomposition shows that, beyond the irreducible variance term  $\bar{\sigma}^2$  and the CATE estimation error  $\Delta_2^2(\hat{\tau}, \tau)$ , the prediction risk is further governed by  $\Delta_2^2(m, \mu)$ , the quality of the CMO estimate. In the next section, we demonstrate that this same term also influences the upper bound on  $\Delta_2^2(\hat{\tau}, \tau)$ , highlighting the central role of accurate CMO estimation.

**Bounding and minimizing the CATE estimation error.** Now we turn to the second term of the risk (4), which is the **CATE estimation error**. Here we focus on the estimation error of the empirical risk minimizer (ERM), i.e., the plug-in estimator corresponding to the population-level objective introduced in (3). We restrict  $f$  to a function class  $\mathcal{F}$  to ensure tractability and avoid overfitting in finite sample settings. For instance, in Section 2.3, we consider the class of linear functions with bounded  $\ell_1$  norm coefficients, which is the commonly-used lasso-penalized linear model. We show that, similar to the irreducible error part of the prediction risk, the CATE's non-asymptotic estimation error bound also relies on the choice of the augmentation function. Using tools from statistical learning theory, Theorem 6 quantifies how the choice of both the function class  $\mathcal{F}$  and  $m(X)$  influence the risk of the ERM estimator, establishing that accurately estimating the CMO and using it as  $m(X)$  can significantly reduce the CATE estimation error.

**Theorem 6 (Non-asymptotic MSE bound for an empirical risk minimizer)** *Suppose Assumptions (A1)–(A4) hold, and let  $\mathcal{F}$  be a function class for approximating the CATE  $\tau(x)$ . Consider a given augmentation function  $m(x)$  used to form the pseudo-outcome  $\tau_m(X_i, A_i, Y_i)$  from an i.i.d. dataset  $\{(X_i, A_i, Y_i)\}_{i=1}^n$ . Define the empirical risk minimizer as*

$$\hat{\tau}_n(\cdot) = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (\tau_m(X_i, A_i, Y_i) - f(X_i))^2.$$

*Then, with probability at least  $1 - 2\varepsilon$ , the mean squared error of  $\hat{\tau}_n$  satisfies*

$$\Delta_2^2(\hat{\tau}_n, \tau) \leq \Delta_2^2(\mathcal{F}, \tau) + 2C(m, \varepsilon) \mathcal{R}_n(\mathcal{F}) + C^2(m, \varepsilon) \sqrt{\frac{\log(2/\varepsilon)}{n}}, \quad (6)$$

*where the quantities are defined as follows:*

- $\Delta_2^2(\mathcal{F}, \tau) = \inf_{f \in \mathcal{F}} \mathbb{E}_X [(f(X) - \tau(X))^2]$  is the approximation error of  $\mathcal{F}$ , i.e., the minimal achievable error within the function class  $\mathcal{F}$ , capturing the inherent bias due to model misspecification,
- $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\epsilon, X} [\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i)]$  is the Rademacher complexity of  $\mathcal{F}$  based on  $n$  samples and  $\epsilon_i \stackrel{i.i.d.}{\sim} \text{Unif}\{-1, 1\}$ ,
- $B$  is a constant such that  $|\tau(x)| \leq B$  and  $|f(x)| \leq B$  for all  $f \in \mathcal{F}$ ,
- $C(m, \varepsilon) = \left(2B + \frac{\bar{\sigma}}{\sqrt{\varepsilon}} + \frac{\Delta_\infty(m, \mu)}{\rho}\right)$  depends on the uniform distance of  $m(x)$  from  $\mu(x)$ ,  $\Delta_\infty(m, \mu)$ .

Rademacher complexity is a well-studied, unifying measure of function class capacity that leads to tight generalization bounds. Tables 5–8 in Appendix A.6 collect sharp bounds for many classes used in statistical learning, illustrating how structural constraints such as sparsity, smoothness, or norm constraints govern learnability. For instance, for the class of linear  $s$ -sparse predictors  $\mathcal{F}_s = \{x \mapsto w^\top x : x \in \mathbb{R}^p, \|w\|_0 \leq s, \|x\|_2 \leq 1\}$ , the empirical process theory (Bartlett and Mendelson, 2006; Koltchinskii, 2011) gives the complexity bound  $\mathcal{R}_n(\mathcal{F}_s) \lesssim \sqrt{\frac{s \log(p/s)}{n}}$ , showing that the burden of high dimensionality enters only through a benign  $\log(p/s)$  factor.

These Rademacher bounds can be converted into minimax squared-error rates by the standard *localization argument* (Bartlett and Mendelson, 2006). Localization requires that the class  $\mathcal{F}$  be *star-shaped about the oracle*  $f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E} \ell(f)$ ; that is, for every  $f \in \mathcal{F}$  and  $\lambda \in [0, 1]$ , the convex combination  $\lambda f^* + (1 - \lambda)f$  also belongs to  $\mathcal{F}$ . Since every norm ball (e.g.  $\ell_1$ ,  $\ell_2$ , RKHS, Hölder, or Sobolev) is convex, this condition is automatically satisfied for all classes listed in the tables. Consequently, a bound  $\mathcal{R}_n(\mathcal{F}) \lesssim \varepsilon_n$  implies the minimax rate  $\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathbb{E} \|\hat{f} - f\|^2 \lesssim \varepsilon_n^2$ . This connection unifies the Rademacher-complexity viewpoint with classical minimax theory (Bartlett and Mendelson, 2006; Wainwright, 2019). For example, the bound  $(1/n)^{\alpha/(2\alpha+p)}$  for Hölder-smooth functions yields the familiar nonparametric rate  $(1/n)^{2\alpha/(2\alpha+p)}$  (Tsybakov, 2008), while the bound  $\sqrt{s \log(p/s)/n}$  for  $s$ -sparse linear functions produces the optimal high-dimensional linear-regression rate  $s \log(p/s)/n$  (Wainwright, 2019).

Theorem 6 shows that the excess MSE of  $\hat{\tau}_n$  decomposes into three familiar pieces: approximation, estimation, and concentration components. Both stochastic terms are multiplied by the factor  $C(m, \varepsilon) = 2B + \frac{\bar{\sigma}}{\sqrt{\varepsilon}} + \frac{\Delta_\infty(m, \mu)}{\rho}$ , so the quality of the augmentation  $m$  directly modulates the sample

complexity. If the oracle augmentation  $m = \mu$  were available, then  $\Delta_\infty(m, \mu) = 0$  and we recover the classical Rademacher bound. In practice, however, we plug in an estimate  $\hat{\mu}$ , and the uniform error  $\Delta_\infty(\hat{\mu}, \mu)$  inflates both the estimation and concentration terms. Thus, accurate estimation of the CMO is indispensable for fully realizing the benefits of low-complexity CATE classes. The next section quantifies how errors in the per-arm mean functions drive  $\Delta_\infty(\hat{\mu}, \mu)$  and, consequently, the overall precision of CATE estimation.

### 2.1.3 CONTROLLING THE CMO ERROR VIA TREATMENT-ARM ERROR BOUNDS

The preceding analysis has demonstrated that both prediction and estimation risks are governed by the accuracy of our CMO estimate, measured through MSE (Theorem 4) and uniform distance (Theorem 6), respectively. In what follows, we establish that controlling the mean squared estimation error for each treatment arm separately suffices for reliable control of the uniform and MSE estimation error of the CMO jointly. While uniform convergence generally implies convergence in MSE, the converse does not always hold. However, under mild conditions, convergence in MSE *does* imply uniform convergence. We show below how the per-arm rate of convergence in MSE simultaneously controls both the uniform and convergence in MSE of the CMO.

**Proposition 7 (MSE and Uniform Convergence of the CMO via Separate-Arm Estimation)** *Let  $\mathcal{D}' = \{(X_j, A_j, Y_j)\}_{j=1}^{n'}$  be an i.i.d. dataset of size  $n'$ , reserved for estimating the CMO,  $\mu$ . Partition  $\mathcal{D}'$  by treatment arm  $A_j \in \{+1, -1\}$  into two subsets  $\mathcal{D}'_{+1} = \{(X_j, Y_j) : A_j = +1\}$ ,  $\mathcal{D}'_{-1} = \{(X_j, Y_j) : A_j = -1\}$ , with sizes  $n'_{+1} = |\mathcal{D}'_{+1}|$  and  $n'_{-1} = |\mathcal{D}'_{-1}|$ , so that  $n'_{+1} + n'_{-1} = n'$ . Suppose there exists a constant  $\eta > 0$  such that  $n'_{+1} \geq \eta n'$  and  $n'_{-1} \geq \eta n'$ .*

*For each arm  $a \in \{+1, -1\}$ , let  $\hat{\mu}_a$  be an estimator of the outcome mean of treatment  $a$ ,  $\mu_a : \mathcal{X} \rightarrow \mathbb{R}$ , and assume that the per-arm MSE satisfies  $\Delta_2^2(\hat{\mu}_a, \mu_a) = \mathcal{O}_p(r^2(n'_a))$ . Define the true and estimated CMO functions as  $\mu(x) \equiv \pi_{-1}(x) \mu_{+1}(x) + \pi_{+1}(x) \mu_{-1}(x)$  and  $\hat{\mu}(x) \equiv \pi_{-1}(x) \hat{\mu}_{+1}(x) + \pi_{+1}(x) \hat{\mu}_{-1}(x)$ , respectively. Then, the following hold:*

- (i) *The MSE of the CMO estimator satisfies  $\Delta_2^2(\hat{\mu}, \mu) := \mathbb{E}_X[(\hat{\mu}(X) - \mu(X))^2] = \mathcal{O}_p(r^2(n'))$ .*
- (ii) *If, in addition, each  $\mu_a$  is  $L$ -Lipschitz and the domain  $\mathcal{X} \subset \mathbb{R}^d$  is compact, then the uniform error of the CMO estimator satisfies  $\Delta_\infty(\hat{\mu}, \mu) := \sup_{x \in \mathcal{X}} |\hat{\mu}(x) - \mu(x)| = \mathcal{O}_p(r(n'))$ .*

This proposition reveals a fundamental connection between per-arm estimation rates and the overall CMO convergence, enabling us to leverage standard statistical learning results for individual outcome estimation to control both prediction risk and estimation error of CATE. The following corollary further elucidates this relationship.

**Corollary 8 (Simplified Risk Bounds)** *Under the conditions of Proposition 7, the following simplified risk bounds hold:*

1. *With high probability, the MSE of the CATE estimator satisfies (up to a standard additive concentration term)  $\Delta_2^2(\hat{\tau}_n, \tau) \lesssim \Delta_2^2(\mathcal{F}, \tau) + (1 + \sum_a \Delta_2(\hat{\mu}_a, \mu_a)) \cdot \mathcal{R}_n(\mathcal{F})$ .*
2. *The overall prediction risk satisfies  $R(\hat{\tau}_n) \lesssim \sigma^2 + \sum_a \Delta_2^2(\hat{\mu}_a, \mu_a) + \Delta_2^2(\hat{\tau}, \tau)$ .*

We have now completed presenting our general CATE estimation framework. So far, the discussion has been limited to a single data source, where we partitioned the dataset into two parts:

one for estimating CATE under assumed conditions of Proposition 1 using the variational form in (3), and another for estimating per-arm outcome means to construct an optimal augmentation function for minimizing the risk of the estimated CATE. In the subsequent sections, we explore various CATE estimators for *the RCT population*, focusing on strategies to enhance per-arm outcome mean estimation by leveraging large auxiliary OS data to construct the augmentation function.

## 2.2 Enhancing CATE estimation efficiency via CMO estimation from OS and RCT data

We now develop approaches to integrate OS with RCT data for improved CATE estimation. Our methods focus on enhancing CMO estimation, which directly improves CATE estimates via our established risk bounds. Importantly, in this approach, leveraging potentially confounded OS data for CMO estimation does not introduce bias into the RCT-derived CATE estimate.

By design, all the assumptions required for Proposition 1 and Theorems 2 and 4 are satisfied in the RCT setting, with the exception of SUTVA (A1). Specifically, randomization ensures conditional ignorability (A2), while both weak (A3) and strong (A4) positivity are guaranteed by the trial protocol. Therefore, under the additional assumptions of SUTVA and correct specification of the CATE model, we establish the consistency of the ERM-based CATE estimator for the RCT setting, as formalized in the following proposition.

**Proposition 9 (Consistency of ERM in RCTs)** *Assume SUTVA and (A5) (Correct specification of the CATE model): that  $\tau^r \in \mathcal{F}$ . Then, the empirical risk minimizer*

$$\hat{\tau}^r(\cdot) = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n^r} \sum_{i=1}^{n^r} \left[ \frac{A_i^r (Y_i^r - m(X_i^r))}{\pi_{A_i^r}^r(X_i^r)} - f(X_i^r) \right]^2 \quad (7)$$

*is consistent for the true CATE  $\tau^r(\cdot)$  as  $n^r \rightarrow \infty$ .*

While  $\mathcal{F}$  could theoretically include all measurable functions, in practice, we must control its complexity through constraints or regularization.

### 2.2.1 BASELINE: NAIVE CATE ESTIMATOR

We refer to the most basic version of estimator (7), obtained by setting  $m(X) := 0$ , as the *Naive* CATE estimator. This approach corresponds to the “transformed outcome” method explored in (Athey and Imbens, 2015). However, using  $m(X) = 0$  is known to result in an inefficient estimator with high variance (Athey and Imbens, 2016).

### 2.2.2 BASELINE: RCT-AUGMENTED CATE ESTIMATOR

To reduce the variance of the CATE estimate in (7), as guided by Section 2.1, we use part of the RCT data to estimate the CMO:

$$m(x) := \hat{\mu}^r(x) = \sum_a \pi_{-a}^r(x) \hat{\mu}_a^r(x), \quad \hat{\mu}_a^r(x) \in \mathcal{M}_a^r, \quad (8)$$

where  $\hat{\mu}_a^r(x)$  denotes the estimated regression function from RCT data, constrained to lie in the function class  $\mathcal{M}_a^r$ . The function class  $\mathcal{M}_a^r$  can be specified, for instance, as the set of linear

functions with  $\ell_1$  norm of coefficients bounded. We refer to the resulting CATE estimator as **RACER** (RCT Augmented by CMO Estimated from RCT). While RACER achieves improved efficiency over the naive estimator, its performance can be constrained by the high variance of the estimated outcome means due to RCT's limited sample size.

### 2.2.3 PROPOSED: OS-AUGMENTED CATE ESTIMATOR WITH RCT CALIBRATION

To address the high variance of estimated outcome means due to the sample size limitations inherent in RCT data, we propose utilizing OS data to estimate the CMO. A direct but potentially suboptimal approach would set the augmentation function to  $m(x) := \sum_a \pi_{-a}^r(x) \hat{\mu}^o(x, a)$ , where  $\hat{\mu}^o(x, a)$  is an outcome mean estimate derived from OS data and applied to covariates from the RCT population. Ideally, if both ignorability in the OS population ( $\mu^o(x, a) = \mu_a^o(x)$ ) and mean exchangeability ( $\forall a : \mu_a^r(x) = \mu_a^o(x)$ ) hold, then  $\hat{\mu}^o(x, a)$  would be a valid substitute for  $\hat{\mu}_a^r(x)$ , enabling unbiased estimation of the RCT's potential outcome means from the OS.

However, in practice, OS data may suffer from unmeasured confounding and/or outcome shifts are often observed across populations, implying that  $\forall a, x : \mu_a^r(x) \neq \mu^o(x, a)$ . To account for this bias, we introduce an outcome mean discrepancy and shifted outcome mean functions defined as:

$$\text{outcome mean discrepancy:} \quad \delta_a(x) \equiv \mu_a^r(x) - \mu^o(x, a), \text{ and} \quad (9)$$

$$\text{shifted outcome mean:} \quad \tilde{\mu}_a^o(x; d_a) \equiv \hat{\mu}^o(x, a) + d_a(x), \quad \hat{\mu}^o(x, a) \in \mathcal{M}_a^o, \quad (10)$$

where  $\delta_a(x)$  represents the **true discrepancy** that will be estimated in the following steps and  $\tilde{\mu}_a^o(x; d_a)$  is defined for an arbitrary discrepancy function  $d_a(x)$  with  $\hat{\mu}^o(x, a)$  being estimated from the OS data. Here,  $\mathcal{M}_a^o$  denote the function classes used for estimating the OS outcome means. The discrepancy functions are then estimated (using RCT data) so the overall estimated shifted outcome means best predict the observed outcomes in the RCT data. A key issue to note is that if the discrepancies  $\delta_a(x)$  are estimated without constraining their function class, minimal to no information would be borrowed from the OS. Therefore, we constrain  $d_a(x)$  to lie in the function class  $\mathcal{D}_a^j$  so that the estimated shifted outcome mean is shrunk towards the preliminary estimates  $\hat{\mu}^o(x, a)$  of the CMO from the OS. In summary, the discrepancy functions are estimated by solving the following constrained optimization problem:

$$(\hat{\delta}_{-1}^j, \hat{\delta}_{+1}^j) = \underset{d_{-1} \in \mathcal{D}_{-1}^j, d_{+1} \in \mathcal{D}_{+1}^j}{\operatorname{argmin}} \frac{1}{n^r} \sum_{i=1}^{n^r} \left[ \frac{A_i^r}{\pi_{A_i^r}(X_i^r)} (Y_i^r - \tilde{\mu}^o(X_i^r; d_{+1}, d_{-1})) - \tilde{\tau}^o(X_i^r; d_{+1}, d_{-1}) \right]^2. \quad (11)$$

Here, we used CMO and CATE *template functions* using the shifted outcome means from (10) as  $\tilde{\mu}^o(x; d_{+1}, d_{-1}) \equiv \sum_a \pi_{-a}^r(x) \tilde{\mu}_a^o(x; d_a)$  and  $\tilde{\tau}^o(x; d_{+1}, d_{-1}) \equiv \sum_a a \tilde{\mu}_a^o(x; d_a)$ . These template functions serve as placeholders in the objective of (11), where the discrepancy functions  $d_a$  are optimized. The superscript  $j$  on the estimated discrepancies  $\hat{\delta}_a^j$  indicates that these functions are estimated jointly, in the sense that both the calibrated CMO and the resulting CATE depend on the same discrepancy functions optimized within the joint risk-minimization framework.

Once the optimal discrepancies are obtained, the final *calibrated CMO* is  $\hat{\mu}^{o,j}(x) \equiv \tilde{\mu}^o(x; \hat{\delta}_{+1}^j, \hat{\delta}_{-1}^j) = \sum_a \pi_{-a}^r(x) [\hat{\mu}^o(x, a) + \hat{\delta}_a^j(x)]$  and our proposed **OSCAR** (Observational Studies for CMO-Augmented RCT) estimator for the CATE will be (Table 2):

Table 2: **Summary of four CATE estimators for the RCT population within our general framework.** The naive estimator serves as a baseline and does not utilize any augmentation functions. RACER uses RCT data alone to estimate outcome means, forming the counterfactual mean outcome (CMO) used as the augmentation function in CATE estimation. The proposed methods, OSCAR and R-OSCAR, leverage large observational study (OS) samples to estimate outcome means. These are calibrated to match the RCT’s outcome means, yielding an estimator for the RCT’s CMO. OSCAR performs joint estimation of CMO and CATE. R-OSCAR follows a two-stage procedure: it calibrates the outcome means to estimate the CMO, then revises the CATE using a correction term  $\hat{\delta}(x)$ .

Selected Augmentation Function $m(x)$	Functional Form of the Corresponding CATE Estimator
0	$\hat{\tau}_{\text{Naive}}(x) \in \mathcal{F}$
$\hat{\mu}^r(x) = \sum_a \pi_{-a}^r(x) \hat{\mu}_a^r(x)$	$\hat{\tau}_{\text{RACER}}(x) \in \mathcal{F}, \hat{\mu}_a^r(x) \in \mathcal{M}_a^r$
$\hat{\mu}^{o,j}(x) = \sum_a \pi_{-a}^o(x) [\hat{\mu}^o(x, a) + \hat{\delta}_a^j(x)]$	$\hat{\tau}_{\text{OSCAR}}(x) = \sum_a a \left[ \hat{\mu}^o(x, a) + \hat{\delta}_a^j(x) \right]$ $\hat{\mu}^o(x, a) \in \mathcal{M}_a^o, \hat{\delta}_a^j(x) \in \mathcal{D}_a^j$
$\hat{\mu}^{o,t}(x) = \sum_a \pi_{-a}^o(x) [\hat{\mu}^o(x, a) + \hat{\delta}_a^t(x)]$	$\hat{\tau}_{\text{R-OSCAR}}(x) = \sum_a a \left[ \hat{\mu}^o(x, a) + \hat{\delta}_a^t(x) \right] + \hat{\delta}(x)$ $\hat{\mu}^o(x, a) \in \mathcal{M}_a^o, \hat{\delta}_a^t(x) \in \mathcal{D}_a^t, \hat{\delta} \in \mathcal{D}$

$$\hat{\tau}_{\text{OSCAR}}(x) \equiv \tilde{\tau}^o(x; \hat{\delta}_{+1}^j, \hat{\delta}_{-1}^j) = \sum_a a \left[ \hat{\mu}^o(x, a) + \hat{\delta}_a^j(x) \right]. \quad (12)$$

As an example of function class constraints in (11), suppose  $d_a(x)$  is modeled as a linear function of covariates. Then, the constraint sets  $\mathcal{D}_a^j$  may be defined as an  $\ell_1$  or  $\ell_2$  norm ball on the coefficients, corresponding to LASSO or Ridge regularization, respectively.

#### 2.2.4 PROPOSED: DOUBLY-CALIBRATED OS-AUGMENTED CATE ESTIMATOR

The OSCAR estimator may suffer from two potential limitations: model misspecification in either the OS outcome regression means or mean discrepancy functions and regularization bias from the separate estimation of treatment arm outcomes (Wager, 2024). To address these issues, we propose a robust variant called **R-OSCAR** (Robust OSCAR) that employs a two-stage calibration process that decouples the CMO estimation from CATE estimation. First, it calibrates the OS regression means to the RCT data to reduce their biases and uses these calibrated estimates to construct both the CMO and a preliminary CATE estimate. Second, it applies an additional calibration step to the preliminary CATE estimate to mitigate regularization and/or misspecification bias.

Specifically, we first calibrate the learned OS outcome means  $\hat{\mu}^o(x, a)$  to RCT by estimating the discrepancy functions for each treatment arm:

$$\hat{\delta}_a^t = \underset{d_a \in \mathcal{D}_a^t}{\operatorname{argmin}} \frac{1}{n_a^r} \sum_{i: A_i^r = a} \|Y_i^r - [\hat{\mu}^o(X_i^r, a) + d_a(X_i^r)]\|_2^2, \quad (13)$$

where  $n_a^r$  denotes the number of RCT samples in treatment arm  $a$  and  $\hat{\delta}_a^t$  is constrained to a function class  $\mathcal{D}_a^t$  and superscript  $t$  emphasizes that this estimate is for the two-stage calibration process in contrast to the joint estimation in (11). This step is inspired by the calibration of computer experiments (Kennedy and O’Hagan, 2001; Dai and Chien, 2018), which involves using observational data to refine the parameters of a computer model, making it better match real-world behavior.

In the next step, we construct the calibrated CMO as (Table 2):

$$\hat{\mu}^{o,t}(x) \equiv \tilde{\mu}^o(x; \hat{\delta}_{+1}^t, \hat{\delta}_{-1}^t) = \sum_a \pi_{-a}^r(x) \left[ \hat{\mu}^o(x, a) + \hat{\delta}_a^t(x) \right].$$

Then, we estimate an additional CATE discrepancy function  $\delta(x)$  to refine the preliminary CATE estimate further:

$$\hat{\delta} = \underset{d \in \mathcal{D}}{\operatorname{argmin}} \frac{1}{n^r} \sum_{i=1}^{n^r} \left[ \frac{A_i^r}{\pi_{A_i^r}^r(X_i^r)} \left( Y_i^r - \tilde{\mu}^o(x; \hat{\delta}_{+1}^t, \hat{\delta}_{-1}^t) \right) - \left[ \tilde{\tau}^o(X_i^r; \hat{\delta}_{+1}^t, \hat{\delta}_{-1}^t) + d(X_i^r) \right] \right]^2, \quad (14)$$

where  $\tilde{\tau}^o(x; \hat{\delta}_{+1}^t, \hat{\delta}_{-1}^t) = \sum_a a \left[ \hat{\mu}^o(x, a) + \hat{\delta}_a^t(x) \right]$  is the preliminary CATE estimate and  $\hat{\delta}$  is constrained to a function class  $\mathcal{D}$ . The final R-OSCAR estimator combines these components:

$$\hat{\tau}_{\text{R-OSCAR}}(x) \equiv \tilde{\tau}^o(x; \hat{\delta}_{+1}^t, \hat{\delta}_{-1}^t) + \hat{\delta}(x) = \sum_a a \left[ \hat{\mu}^o(x, a) + \hat{\delta}_a^t(x) \right] + \hat{\delta}(x). \quad (15)$$

Compared to (12), R-OSCAR estimates the per-arm discrepancy functions separately in the first calibration step (13), and incorporates an additional CATE calibration term  $\hat{\delta}(x)$  for robustness.

### 2.2.5 CATE IDENTIFIABILITY AND ESTIMATOR CONSISTENCY

**Identifiability.** Our framework leverages the key strength of RCTs: both conditional ignorability (A2) and positivity (A3–A4) hold by design. As a result, the identifiability of the CATE requires only the SUTVA (A1), making the assumption set minimal (Proposition 1).

**Estimator Consistency.** The consistency result in Proposition 9 requires correct specification of the CATE model (A5). In both OSCAR and R-OSCAR, CATE estimation involves combining estimated outcome means and discrepancy functions. Thus, consistency of these components must be connected to the overall consistency of the CATE estimator. We now examine the consistency behavior of OSCAR and R-OSCAR separately.

1. **OSCAR Estimator.** When both the observational outcome means  $\mu^o(x, a)$  and the discrepancy functions  $\delta_a(x)$  are correctly specified, their sum  $\hat{\mu}_a^o(x) + \hat{\delta}_a^j(x)$  consistently estimates the true RCT outcome mean  $\mu_a^r(x)$ . Consequently, their contrast  $\hat{\tau}_{\text{OSCAR}}(x)$  is consistent for the RCT's CATE. However, if either  $\mu^o(x, a)$  or  $\delta_a(x)$  is misspecified, then both CMO and CATE may be misspecified. While CMO misspecification alone affects only efficiency, misspecification of the CATE leads to bias and inconsistency in  $\hat{\tau}_{\text{OSCAR}}(x)$ .
2. **R-OSCAR Estimator.** To mitigate OSCAR's reliance on correct specification of multiple functions, R-OSCAR introduces a separate CATE discrepancy function. This reduces the number of functions that must be correctly specified for consistent CATE estimation from four to one. In this framework, misspecification of  $\mu^o(x, a)$  and  $\delta_a^r(x)$  results in variance inflation through CMO misspecification, but does not compromise CATE consistency if the CATE estimation step is correctly specified. Here, two estimation strategies are possible:

- *Direct CATE Estimation.* One approach is to directly estimate CATE using a model class known to contain the true CATE function, as assumed in Proposition 9. This mirrors standard CATE estimation from RCTs, with the distinction that our method would incorporate calibrated OS outcome functions  $\hat{\mu}_a^o(x) + \hat{\delta}_a^t(x)$  for augmentation.



- *CATE Calibration (Our Approach).* Our preferred strategy is CATE calibration, in which we first form a preliminary CATE estimate using calibrated outcome means  $\hat{\mu}_a^o(x) + \hat{\delta}_a^t(x)$ , then refine this using a CATE discrepancy function learned from RCT data. Provided this discrepancy function is correctly specified, the resulting estimator is consistent (Table 2).

In both approaches, CATE consistency depends on the correct specification of a single function: either the CATE itself or the CATE discrepancy function.

**Theoretical Advantages of R-OSCAR.** An additional benefit of R-OSCAR is its compatibility with sharper theoretical analysis. The prediction and estimation error bounds in Section 2.1 require that the CMO and CATE be estimated using independent datasets. This assumption fails for OSCAR, where CMO and CATE are jointly learned, complicating rigorous analysis. In contrast, while R-OSCAR uses RCT data for both the outcome mean calibration and CATE calibration steps, theoretical guarantees can still be obtained through sample splitting and cross-fitting (Zeng et al., 2023; Chernozhukov et al., 2018). Specifically, the RCT dataset is divided into  $K$  folds: one fold is used for calibration (Eq. (19)) and the remaining  $K - 1$  folds for CATE estimation (Eq. (20)). The final CATE estimator aggregates results across folds as

$$\hat{\tau}^r(x) = \sum_a a \left[ \hat{\mu}^o(x, a) + \frac{1}{K} \sum_{k=1}^K \hat{\delta}_a^k(x) \right] + \frac{1}{K} \sum_{k=1}^K \hat{\delta}_a^k(x).$$

#### 2.2.6 THEORETICAL JUSTIFICATION OF BORROWING STRENGTH FROM AN OS

As shown in Section 2.1, improving the MSE of each treatment arm’s outcome mean reduces both prediction and estimation errors. Here, we justify how borrowing from OS data via our proposed method can lower this MSE relative to relying solely on RCT data. Specifically, we compare outcome mean estimation under RACER (using partial RCT data) and R-OSCAR (with two-stage calibration), focusing on per-arm MSE.

For clarity, assume correct model specification: each true function lies within the function class used to estimate it (e.g.,  $\mu_a^r(x) \in \mathcal{M}_a^r$ ). Under standard assumptions, the high-probability non-asymptotic MSE of an estimator  $\hat{g}_n$  of  $g \in \mathcal{G}$  with  $n$  samples satisfies:  $\Delta_2^2(\hat{g}_n, g) \leq \mathcal{O}_p\left(\frac{c(\mathcal{G})}{n^\eta}\right)$ , where  $c(\mathcal{G})$  reflects function class complexity and  $\eta$  the convergence rate. Such bounds typically arise from localized Rademacher complexity arguments. For example, in LASSO regression with sparsity  $s$  and  $p$  covariates,  $c(\mathcal{G}) \approx s \log(p)$ ; in kernel methods,  $c(\mathcal{G}) \approx B^2 \text{tr}(K)$  for RKHS norm bound  $B$  and kernel matrix  $K$ . The rate  $\eta$  depends on the function class:  $\eta = 1$  for parametric models, and  $\eta = \frac{2\alpha}{2\alpha+p}$  for nonparametric Hölder classes with smoothness  $\alpha$  in  $p$  dimensions.

Assuming such bounds hold, we now state the following proposition:

**Proposition 10 (RACER vs. R-OSCAR)** *Assume standard regularity conditions and that the following non-asymptotic MSE bounds hold for per-arm outcome mean estimation:*

1. **Direct Estimation (RACER):** Using RCT data only,  $\Delta_2^2(\hat{\mu}_a^r, \mu_a^r) \leq \mathcal{O}_p\left(\frac{c(\mathcal{M}_a^r)}{n_a^r \eta_a^r}\right)$ .
2. **Two-Stage Calibration (R-OSCAR):** Using OS data with RCT-based calibration,  $\Delta_2^2(\hat{\mu}_a^{o,t}, \mu_a^r) \leq \mathcal{O}_p\left(\frac{c(\mathcal{M}_a^o)}{n_a^o \eta_a^o}\right) + \mathcal{O}_p\left(\frac{c(\mathcal{D}_a^t)}{n_a^r \eta_a^t}\right)$ , where  $\hat{\mu}_a^{o,t}(x) = \hat{\mu}_a^o(x, a) + \hat{\delta}_a^t(x)$ .

Then, borrowing from OS data will lead to CATE estimation improvement when the following inequality holds for some constants  $C_1, C_2, C_3 > 0$ :  $\forall a$  :  $C_1 \frac{c(\mathcal{M}_a^o)}{n_a^o \eta_a^o} + C_2 \frac{c(\mathcal{D}_a^t)}{n_a^r \eta_a^r} < C_3 \frac{c(\mathcal{M}_a^r)}{n_a^r \eta_a^r}$ .

The result follows directly: if the inequality holds, then the per-arm outcome means are more accurately estimated under R-OSCAR than under RACER,  $\Delta_2^2(\hat{\mu}_a^{o,t}, \mu_a^r) < \Delta_2^2(\hat{\mu}_a^r, \mu_a^r)$ , leading to a more accurate estimate of the CMO and, in turn, improved CATE estimation.

To interpret the proposition, suppose  $C_1 = C_2 = C_3$  and that the RCT and OS model classes have comparable complexity, i.e.,  $c(\mathcal{M}_a^o) \approx c(\mathcal{M}_a^r)$ , with identical convergence rates  $\eta_a^o = \eta_a^r = \eta$ . The inequality then reduces to:  $\frac{c(\mathcal{M}_a^r)}{n_a^o \eta} + \frac{c(\mathcal{D}_a^t)}{n_a^r \eta} < \frac{c(\mathcal{M}_a^r)}{n_a^r \eta}$ . This condition is satisfied when (i)  $n_a^o \gg n_a^r$ , i.e., the OS model benefits from a much larger sample size, and (ii)  $c(\mathcal{D}_a^t) < c(\mathcal{M}_a^r)$ , meaning the discrepancy function is simpler than the full outcome model. That is, by estimating  $\hat{\mu}^o(x, a)$  from abundant OS data and only calibrating a simpler discrepancy  $\hat{\delta}_a^t(x)$  using the limited RCT data, R-OSCAR achieves lower error than directly estimating  $\mu_a^r(x)$  from the RCT alone.

Crucially, this does not require  $\mathcal{M}_a^o \subseteq \mathcal{M}_a^r$ , the OS model class can be more complex (e.g., a neural network) so long as the OS sample size is large enough to keep its error term small, i.e.,  $c(\mathcal{M}_a^o)/n_a^o \lesssim c(\mathcal{M}_a^r)/n_a^r$ . If the discrepancy function belongs to a much simpler class  $\mathcal{D}_a^t$  such that  $c(\mathcal{D}_a^t) \ll c(\mathcal{M}_a^r)$ , the setup resembles transfer learning: a large source dataset establishes a strong baseline, and only a lightweight correction is learned in the target domain.

**Example 1 (Sparse Linear Model)** Consider a setting where:  $\mu_a^r(x) = x^\top \beta_a^r$ ,  $\mu^o(x, a) = x^\top \beta_a^o$ ,  $\delta_a^t(x) = x^\top (\beta_a^r - \beta_a^o)$ . Assume  $\beta_a^r, \beta_a^o \in \mathbb{R}^p$ , and their difference  $\beta_a^r - \beta_a^o$  is  $s$ -sparse (with  $s \ll p$ ). Then, both  $\mathcal{M}_a^r$  and  $\mathcal{M}_a^o$  (linear functions in  $\mathbb{R}^p$ ) have complexity  $c(\mathcal{M}_a^r) \approx c(\mathcal{M}_a^o) \approx p \log p$ , while  $\mathcal{D}_a^t$  (all  $s$ -sparse linear functions) has complexity  $c(\mathcal{D}_a^t) \approx s \log(p)$ . Using LASSO for estimation, the estimation errors are approximately:  $\Delta_2^2(\hat{\mu}_a^r, \mu_a^r) \approx \frac{p \log(p)}{n_a^r}$  and  $\Delta_2^2(\hat{\mu}_a^{o,t}, \mu_a^r) \approx \frac{p \log(p)}{n_a^o} + \frac{s \log(p)}{n_a^r}$ . When  $n_a^o \gg n_a^r$  and  $s \ll p$ , the two-stage R-OSCAR estimator achieves a substantially smaller estimation error than the direct RACER estimator, confirming the benefit of borrowing strength from OS data. Specifically, if  $\forall a$  :  $n_a^o \gtrsim (p/(p-s)) \cdot n_a^r$ , then R-OSCAR will have lower estimation error for the CMO than RACER.

### 2.3 Using the general framework under linear outcome means with sparse discrepancy

We now specialize our general framework to a parametric setting with linear potential outcome means and sparse linear mean discrepancies between populations. Specifically, we assume linearity of the RCT potential outcome means  $\mu_a^r(x) = \langle \gamma_a^r, x \rangle$  for both treatments  $a \in \{-1, +1\}$ , which implies linearity of the CATE ( $\tau^r(x)$ ), the CMO ( $\mu^r(x)$ ), and the function class  $\mathcal{F}$  in (3). Similarly, we assume the OS population's outcome means follow a linear form  $\mu^o(x, a) = \langle \gamma_a^o, x \rangle$ . The mean discrepancy function is defined as  $\delta_a(x) = \mu_a^r(x) - \mu^o(x, a) = \langle \delta_a, x \rangle$ , where  $\delta_a \equiv \gamma_a^r - \gamma_a^o$  is a sparse vector. For simplicity, we assume the RCT employs full randomization, where  $\pi_a^r(x^r) = \pi_a^r$ .

### 2.3.1 THE OSCAR FOR LINEAR OUTCOME MEAN WITH A SPARSE DISCREPANCY PARAMETERIZATION

The OSCAR estimation proceeds in two stages: first, estimating the OS outcome means  $\hat{\mu}^o(\cdot, a)$  using LASSO regression:

$$\hat{\gamma}_a^o = \underset{\gamma}{\operatorname{argmin}} \frac{1}{n_a^o} \sum_{i: A_i^o = a} [Y_i^o - \gamma^T X_i^o]^2 + \lambda_a^o \|\gamma\|_1, \quad (16)$$

where  $\gamma \in \mathbb{R}^p$  and the estimated outcome mean function is  $\hat{\mu}^o(x, a) \equiv \langle \hat{\gamma}_a^o, x \rangle$ . For the second step, we instantiate (11) to learn CMO and CATE jointly through the estimation of the sparse mean discrepancy coefficient vectors:

$$(\hat{\delta}_{-1}^j, \hat{\delta}_{+1}^j) = \underset{(d_{-1}, d_{+1})}{\operatorname{argmin}} \frac{1}{n^r} \sum_{i=1}^{n^r} \left[ \frac{A_i^r}{\pi_{A_i^r}^r} \left( Y_i^r - \sum_a \pi_{-a}^r (\hat{\gamma}_a^o + d_a)^T X_i^r \right) - \sum_a a (\hat{\gamma}_a^o + d_a)^T X_i^r \right]^2 + \sum_a \lambda_a^j \|d_a\|_1. \quad (17)$$

The objective in (17) can be simplified by recognizing that it is separable in the  $d_a$  terms:

$$\hat{\delta}_a^j = \underset{d_a}{\operatorname{argmin}} \frac{1}{n_a^r} \sum_{i: A_i^r = a} \left[ \left( \frac{a}{\pi_a^r} Y_i^r - a(1 + \alpha^a) \hat{\gamma}_a^{oT} X_i^r \right) - a(1 + \alpha^a) d_a^T X_i^r \right]^2 + \lambda_a^j \|d_a\|_1, \quad (18)$$

where  $\alpha \equiv \frac{\pi_{-1}^r}{\pi_{+1}^r}$ . The optimization in (18) represents two separate  $\ell_1$ -regularized regressions, each using data from the corresponding treatment arm  $a$  to estimate  $\delta_a$ . In total, our approach requires solving four LASSO problems: one for each treatment arm in the OS data (16) and one for each discrepancy term using RCT data in (18). All tuning parameters are chosen via cross validation.

### 2.3.2 THE ROBUST OSCAR FOR POSSIBLY MISSPECIFIED MODELS

To enhance the robustness of OSCAR against model misspecification, we introduce a two-stage calibration process. First, we estimate the per-treatment-arm linear coefficient vectors from the OS data, following (16). Then, we calibrate the OS outcome means to the RCT outcomes by estimating a sparse mean discrepancy vector for each treatment arm:

$$\hat{\delta}_a^t = \underset{d_a}{\operatorname{argmin}} \frac{1}{n_a^r} \sum_{i: A_i^r = a} [Y_i^r - (\hat{\gamma}_a^o + d_a)^T X_i^r]^2 + \lambda_a^t \|d_a\|_1. \quad (19)$$

This step produces a CMO and a preliminary CATE estimate tailored for the RCT. In the second stage, we refine the preliminary CATE by estimating a sparse linear discrepancy function to correct any residual misspecification using the calibrated CMO as the augmentation function:

$$\hat{\delta} = \underset{d}{\operatorname{argmin}} \frac{1}{n^r} \sum_{i=1}^{n^r} \left[ \frac{A_i^r}{\pi_{A_i^r}^r} \left( Y_i^r - \sum_a \pi_{-a}^r \hat{\gamma}_a^{o,tT} X_i^r \right) - \left( \hat{\gamma}_{+1}^{o,t} - \hat{\gamma}_{-1}^{o,t} + d \right)^T X_i^r \right]^2 + \lambda \|d\|_1, \quad (20)$$

where  $\hat{\gamma}_a^{o,t} \equiv \hat{\gamma}_a^o + \hat{\delta}_a^t$  represents the calibrated OS coefficients for arm  $a$ , and  $\hat{\mu}^{o,t}(x) = \sum_a \pi_{-a}^r (\hat{\gamma}_a^o + \hat{\delta}_a^t)^T x$  is the calibrated CMO. The  $\ell_1$ -norm regularization induces sparsity in the discrepancy estimates. The final CATE estimate is given by:  $\hat{\tau}_{\text{R-OSCAR}}(x) = \langle [(\hat{\gamma}_{+1}^{o,t} + \hat{\delta}_{+1}^t) - (\hat{\gamma}_{-1}^{o,t} + \hat{\delta}_{-1}^t)] + \hat{\delta}, x \rangle$ ,

Table 3: **CATE estimation with data simulated from the proposed model.** At three-decimal precision, OSCAR surpasses all alternatives, with R-OSCAR a close second. Notably, OSCAR and R-OSCAR with  $n_r = 250$  achieve comparable or better performance than the RCT-only methods (Naive and RACER) with  $n_r = 1000$ , demonstrating a 75% reduction in required RCT sample size when leveraging OS data.

	Number of Samples per Study		
	250	500	1000
$n_r$			
$n_o$	10,000	10,000	10,000
Method	RMSE of CATE Estimation		
Naive, (2.2.1)	$1.25 \pm 0.29$	$1.03 \pm 0.28$	$0.74 \pm 0.19$
RACER, (2.2.2)	$0.31 \pm 0.04$	$0.21 \pm 0.03$	$0.15 \pm 0.02$
OSCAR, (2.3.1)	<b><math>0.15 \pm 0.03</math></b>	<b><math>0.11 \pm 0.02</math></b>	$0.08 \pm 0.02$
R-OSCAR, (2.3.2)	$0.16 \pm 0.03$	<b><math>0.11 \pm 0.02</math></b>	<b><math>0.08 \pm 0.01</math></b>
R-OSCAR, Cross Fit	$0.18 \pm 0.03$	$0.12 \pm 0.02$	$0.09 \pm 0.01$

where  $\hat{\delta}_a^t$  is estimated during the outcome mean calibration step, and  $\hat{\delta}$  is estimated during the CATE calibration step. If the preliminary CATE estimate is already accurate, the calibration step shrinks all elements of  $\hat{\delta}$  to zero, avoiding unnecessary corrections.

### 3 Numerical Experiments

We conduct a comprehensive set of simulation studies to evaluate the performance of the proposed CATE estimators under diverse conditions, including covariate and outcome shifts, model misspecification, and unmeasured confounding. All simulations are based on a common generative setup described below, with specific variations introduced in each experiment.

#### 3.1 Simulation Setup

Each simulation generates paired OS and RCT datasets. Covariates  $x \in \mathbb{R}^{100}$  are drawn from a multivariate Gaussian distribution with zero mean and a dense covariance matrix (i.e., all off-diagonal entries are nonzero), inducing correlation among features. Potential outcomes are defined separately for each treatment arm, which allows us to compute the ground truth CATE before treatment assignment.

In the OS dataset, treatment is assigned via a logistic model using ten randomly selected covariates. The model includes an intercept term adjusted to ensure that at least one-third of samples receive the treatment. Treatment assignment in the RCT is randomized with probability  $\pi_1^r(x) = 0.5$ . For each arm  $a \in \{-1, 1\}$  in the OS, the potential outcome is generated linearly as  $y^o(a) = \langle \beta_a^o, x \rangle + \epsilon$ , where the coefficient vector  $\beta_a^o$  has a support size of 10, with non-zero entries drawn from  $\text{Uniform}(-2/3, -1/3) \cup \text{Uniform}(1/3, 2/3)$ . The noise term  $\epsilon \sim \mathcal{N}(0, 1/9)$ .

To introduce domain shift in the RCT, we simulate both covariate and outcome shifts:

- **Covariate shift:** Ten randomly selected covariates are shifted by adding a value sampled from  $\text{Uniform}(-1/2, -1/4) \cup \text{Uniform}(1/4, 1/2)$ .
- **Outcome shift:** Two non-zero coefficients in each  $\beta_a^r$  are perturbed by adding values from  $\text{Uniform}(-1, -1/2) \cup \text{Uniform}(1/2, 1)$ .

### 3.2 Estimators Compared

We compare five estimators for the CATE. All methods use LASSO regression with tuning parameter chosen via 10-fold cross-validation via `cv.glmnet` for model fitting unless otherwise specified:

- **Naive** ( $\hat{\tau}_{\text{Naive}}$ ): Fits arm-specific LASSO regressions using only RCT data, then contrasts predictions to estimate CATE.
- **RACER** ( $\hat{\tau}_{\text{RACER}}$ ): Estimates potential outcome means from RCT data and plugs them into the CMO expression in Equation (7).
- **OSCAR** ( $\hat{\tau}_{\text{OSCAR}}$ ): Uses OS data to estimate outcome models via Equation (16), then solves for the CMO and CATE jointly using Equation (18).
- **R-OSCAR** ( $\hat{\tau}_{\text{R-OSCAR}}$ ): Calibrates OS-based outcome models to the RCT population using Equation (19), then estimates CATE using Equation (20).
- **Cross-Fitted R-OSCAR**: Applies five-fold cross-fitting to R-OSCAR: one fold is used for calibration, and the remaining for estimation; final results average across folds.

### 3.3 Varying Sample Sizes Under Linear Outcome Models

We begin by evaluating performance under a correctly specified linear outcome model, focusing on how RCT and OS sample sizes affect CATE estimation accuracy.

**Varying RCT size.** In the first experiment, we fix the OS sample size at  $n^o = 10,000$  and vary the RCT sample size  $n^r$  from 250 to 1,000. Table 3 reports the RMSE of CATE estimation averaged over 100 replicates. As expected, estimation accuracy improves with larger RCT sample sizes. However, both OSCAR and R-OSCAR consistently outperform Naive and RACER, even when the latter have access to four times more RCT data (compare the last column of Naive and RACER rows with the first column of OSCAR and R-OSCAR in Table 3). The cross-fitted R-OSCAR performs comparably to its non-cross-fitted counterpart but is computationally more expensive, so we omit it in later experiments.

**Varying OS size and covariate dimension.** We next fix the RCT sample size at  $n^r = 300$  and systematically vary both the OS sample size  $n^o$  and the covariate dimension  $p$ . In each setting, we designate  $0.10p$  covariates as true effect modifiers in the OS population, and we allow  $s = 0.02p$  of modifiers to differ in the RCT model, which means, the number of effect modifiers in the RCT will be between  $\lceil 0.10p \rceil$  and  $\lfloor 0.12p \rfloor$  effect modifiers, so all of the OS modifiers are shared. For small  $p$ , this means only zero or one modifier differs (e.g. when  $p = 50$ , there are 5 OS modifiers and  $s = 1$ , so the RCT has 5 or 6 modifiers).

Figure 1A plots the RMSE difference between R-OSCAR and RACER as a function of  $n^o$  (averaged over 100 replicates), where negative values indicate superior performance by R-OSCAR. The results align closely with the theoretical predictions from Example 1, confirming two key behaviors:

1. **Threshold behavior around  $n^o = 300$ .** R-OSCAR begins to outperform RACER once  $n_a^o \approx n_a^r$ , matching the theoretical condition  $n_a^o \gtrsim \frac{p}{p-s} n_a^r$ . With  $s = 0.02p$ , this simplifies to  $n_a^o \approx n_a^r$ .
2. **Increasing advantage with higher dimension.** For fixed sample sizes and  $s = 0.02p$ , Example 1 shows that the MSE gap between RACER and R-OSCAR grows linearly with  $p$ , specifically

Table 4: **CATE estimation under model misspecification.** RMSE (with standard deviation) is reported for each method based on 100 replicates. “Without RF” columns correspond to settings where all methods—including baselines—use LASSO regression. “With RF” columns show results when Naive and RACER use random forest to model outcomes, while OSCAR and R-OSCAR remain fixed with LASSO. In both misspecification settings, R-OSCAR achieves the lowest RMSE. Notably, baseline methods perform better under LASSO-regularized regression than random forests.

Method	Quadratic		Sinusoidal	
	Without RF	With RF	Without RF	With RF
Naive	$1.99 \pm 0.25$	$5.06 \pm 0.33$	$0.90 \pm 0.24$	$2.23 \pm 0.67$
RACER	$1.54 \pm 0.16$	$1.87 \pm 0.12$	$0.32 \pm 0.03$	$1.24 \pm 0.15$
OSCAR	$1.73 \pm 0.19$	—	$0.32 \pm 0.03$	—
R-OSCAR	$1.50 \pm 0.19$	—	$0.28 \pm 0.03$	—

in proportion to  $p(0.98(1/n^r) - 1/n^o)$ . As a result, the relative RMSE benefit of R-OSCAR increases approximately with  $\sqrt{p}$ , which is reflected in the widening performance gap at higher dimensions in the figure.

### 3.4 Robustness to Model Misspecification

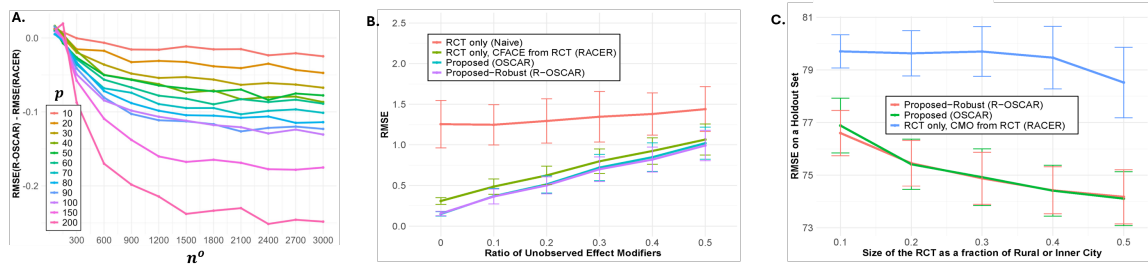
In this experiment, we test robustness to nonlinear effects by adding misspecified terms to the linear outcome model. For each effect modifier  $x_j \in \text{supp}(\beta_a^s)$ , we generate outcomes as:  $y_a^s = \sum_j \beta_a^s(j)x_j + m_a^s(j)g(x_j) + \epsilon$ , where  $g(x_j)$  is either a quadratic term  $x_j^2$  or sinusoidal term  $\sin(x_j)$ , and  $m_a^s(j) \sim \text{Uniform}(0.25, 0.5)$ .

We fix the number of RCT and OS samples at  $n_r = 1,000$  and  $n_o = 10,000$ , respectively, and repeat each simulation setting 100 times. In these misspecified scenarios, we allow Naive and RACER to use random forests for outcome modeling, while OSCAR and R-OSCAR deliberately retain LASSO, remaining misspecified by design. As shown in Table 4, R-OSCAR consistently outperforms all baseline variants—including both LASSO- and random forest-based implementations—demonstrating strong robustness to model misspecification. OSCAR performs slightly worse than RACER in the quadratic setting and comparably in the sinusoidal setting. This underscores the value of R-OSCAR’s two-stage design: by learning the mean discrepancy functions  $\delta_a(x)$  during calibration and then estimating the final CATE discrepancy  $\delta(x)$  separately, R-OSCAR maintains resilience to nonlinearity even when outcome models are misspecified.

Interestingly, in both the quadratic and sinusoidal settings, the LASSO versions of Naive and RACER outperform their random forest counterparts. This somewhat counterintuitive result likely stems from the challenge tree-based models face in capturing smooth nonlinear transformations—such as sine functions—in high-dimensional settings (Breiman, 2001). In contrast, LASSO continues to provide stable performance through effective regularization and sparsity induction.

### 3.5 Effect of Unmeasured Effect Modifiers

In this experiment, we investigate the impact of unmeasured effect modifiers on CATE estimation. We fix the RCT and OS sample sizes at 250 and 10,000, respectively. Starting from the base linear outcome model, we randomly remove 10% to 50% of the true effect modifiers from both datasets, corresponding to the removal of up to 8 covariates (out of approximately 15 total modifiers in each replication).



**Figure 1: CATE estimation accuracy.** (A) Difference in RMSE between R-OSCAR and RACER as the OS sample size  $n^o$  increases (with fixed  $n^r = 300$ ). Negative values indicate superior performance by R-OSCAR. Each curve corresponds to a different covariate dimension  $p$ , where 10% of covariates are true effect modifiers and 2% of these modifiers differ between the RCT and OS models (so  $s = 0.02p$ ). Two phenomena are noteworthy, both of which align with our theoretical results from Example 1: (i) a clear threshold around  $n^o \approx n^r$ , and (ii) an increasing advantage of R-OSCAR at higher  $p$ . Points are averages over 100 replicates; error bars are omitted for clarity. (B) Sensitivity to unmeasured effect modifiers. To simulate latent confounding, a fraction of true effect modifiers is removed from the observed covariates. RMSE is reported for each method across 100 replicates. All methods degrade with increasing unobserved confounding, but R-OSCAR consistently achieves the lowest error. (C) Results on the STAR dataset with varying RCT cohort sizes. Each point shows average RMSE across 30 replicates. Both OSCAR and R-OSCAR outperform the RCT-only baseline (RACER). The naive estimator is omitted due to substantially worse performance.

As shown in Figure 1B, estimation performance deteriorates across all methods as more modifiers are omitted. R-OSCAR consistently outperforms the baselines, though its margin of improvement narrows as the number of unmeasured effect modifiers increases. Notably, even the removal of a single covariate (roughly 10% of effect modifiers) leads to a sharp decline in performance for methods that rely on CMO estimation. In contrast, the Naive method exhibits relatively stable performance, likely because it does not depend on modeling outcome means across datasets. These findings highlight both the promise and limitations of leveraging OS data: while borrowing information can substantially improve accuracy, its effectiveness hinges on the completeness of key effect modifiers.

#### 4 Real-World Experiment: The STAR Study

We evaluate our estimators using the Tennessee Student/Teacher Achievement Ratio (STAR) dataset, following the setup in Kallus et al. (2018). Their goal was to use RCT data to reduce bias in OS-based CATE estimates; ours is to reduce variance in RCT-based CATE estimation using OS data. While the inference targets differ, the population construction (see below) is agnostic to this distinction, so we replicate their sampling strategy.

STAR was a randomized study of class size effects on performance. Students were randomly assigned to small classes, regular classes, or regular classes with a teacher’s aide, and remained in these assignments through third grade. Performance was assessed via standardized tests in kindergarten, first, and third grade. We focus on 4,218 students with complete first-grade data, estimating the CATE of small (treatment) vs. regular (control) class assignment on average first-grade test scores, conditional on gender, race, birth date, free lunch status, and teacher ID.

To simulate limited RCT data, we randomly sampled a fraction  $q \in \{0.1, 0.2, \dots, 0.5\}$  of students from rural or inner-city schools (30 replicates per  $q$ ). OS data was then constructed by including: (1) rural/inner-city students in regular classes not selected for the RCT; (2) rural/inner-city students in small classes not in the RCT whose scores were in the bottom 50%; (3) all urban/suburban students in regular classes; and (4) urban/suburban students in small classes with scores in the bottom 50%. This induces artificial confounding; see Kallus et al. (2018).

We compared four estimators of Table 2. Following Kallus et al. (2018), we defined ground-truth CATE values based on full STAR data, but with respect to the RCT population. Figure 1C shows the results.  $\hat{\tau}_{\text{Naive}}$  performed substantially worse and is omitted from the plot for readability. Both OSCAR and R-OSCAR clearly outperform RACER, with nearly identical performance.

## 5 Conclusion

This work presents a novel framework for improving conditional average treatment effect (CATE) estimation in RCTs by leveraging data from an observational study. Our approach addresses a fundamental challenge in precision medicine: RCTs provide unbiased causal estimates but often lack the sample size needed for precise heterogeneous treatment effect estimation, while observational studies offer abundant data but suffer from confounding bias. By reformulating CATE estimation as a supervised learning problem with pseudo-outcomes, we demonstrate that the counterfactual mean outcome (CMO) serves as the optimal augmentation function for variance reduction. Our key insight is that accurate CMO estimation—achieved by borrowing strength from observational data while calibrating for population differences—directly translates to improved CATE estimation through established risk bounds. The proposed R-OSCAR estimator employs a two-stage calibration strategy that relaxes the strong transportability assumptions typically required in data integration, instead modeling outcome shifts through sparse discrepancy functions that accommodate differences in the data generation process between RCT and observational study. Extensive simulations and real-world validation on the STAR dataset demonstrate that our methods achieve substantial efficiency gains, often matching the performance of RCT-only estimators with four times the sample size. This framework opens new possibilities for leveraging existing observational databases to enhance the precision of future clinical trials, ultimately supporting more reliable personalized treatment decisions while preserving the gold standard of randomized evidence.

## Acknowledgments and Disclosure of Funding

This work was supported in part by the Patient-Centered Outcomes Research Institute (PCORI) under award ME-2023C1-32148. A.A. acknowledges support from the National Human Genome Research Institute of the National Institutes of Health under award R00HG011367.



## Appendix A. Technical Proofs

### A.1 Proof of Proposition 1

**Proof** [Proof of Proposition 1] The first-order optimality condition requires implies that  $f^*(x) = \mathbb{E} [\tau_m(x, A, Y) | X = x]$ . Here, we demonstrate that this conditional expectation is equivalent to the true CATE:

$$\mathbb{E}_{A,Y} \left[ \frac{A(Y - m(X))}{\pi_A(X)} \middle| X \right] = \mathbb{E}_{A,Y(-1),Y(+1)} \left[ \frac{A(Y(A) - m(X))}{\pi_A(X)} \middle| X \right] \quad (21)$$

$$\begin{aligned} &= \mathbb{E}_A \mathbb{E}_{Y(-1),Y(+1)|A} \left[ \frac{A(Y(A) - m(X))}{\pi_A(X)} \middle| X \right] \\ &= \sum_{a \in \{-1, +1\}} \pi_a(X) \mathbb{E}_{Y(-1),Y(+1)|A=a} \left[ \frac{a(Y(a) - m(X))}{\pi_a(X)} \middle| X \right] \\ &= \sum_{a \in \{-1, +1\}} \mathbb{E}_{Y(-1),Y(+1)|A=a} \left[ a(Y(a) - m(X)) \middle| X \right] \\ &= \mathbb{E} [(Y(+1) | X)] - m(\mathbf{X}) - \mathbb{E} [(Y(-1) | X)] + m(X) \\ &= \tau(X), \end{aligned} \quad (22)$$

where we used SUTVA in (21) and conditional ignorability in (22). ■

### A.2 Proof of Theorem 2

The variance of  $\tau_m$  is:  $\mathbb{V}(\tau_m(x, A, Y) | X = x) = \mathbb{E}(\tau_m^2(x, A, Y) | X = x) - \mathbb{E}^2(\tau_m(x, A, Y) | X = x)$  where per Proposition 1, under the assumed conditions, the second term is equal to  $\tau^2(x)$ . So, only the first term,  $\mathbb{E}(\tau_m^2(x, A, Y) | X = x)$  depends on  $m(X)$ . The first order optimality condition implies that, for any  $x$ , the optimal  $m$  (calling it  $m^*$ ) should satisfy  $\mathbb{E}_{A,Y}((Y - m^*(x))/\pi_A^2(x)) = 0$ , which we expand below:

$$\begin{aligned} \mathbb{E}_{A,Y} \left[ \frac{Y - m^*(X)}{\pi_A^2(X)} \middle| X \right] &= \mathbb{E}_A \mathbb{E}_{Y(-1),Y(+1)|A} \left[ \frac{Y(A) - m^*(X)}{\pi_A^2(X)} \middle| X \right] \\ &= \sum_{a \in \{-1, +1\}} \pi_a(X) \mathbb{E}_{Y(-1),Y(+1)|A=a} \left[ \frac{Y(a) - m^*(X)}{\pi_a^2(X)} \middle| X \right] \\ &= \sum_{a \in \{-1, +1\}} \mathbb{E}_{Y(-1),Y(+1)|A=a} \left[ \frac{Y(a) - m^*(X)}{\pi_a(X)} \middle| X \right] \\ &= \mathbb{E}_{Y(+1)} \left[ \frac{Y(+1) - m^*(X)}{\pi_{+1}(X)} \middle| X \right] + \mathbb{E}_{Y(-1)} \left[ \frac{Y(-1) - m^*(X)}{\pi_{-1}(X)} \middle| X \right] \\ &= 0 \end{aligned} \quad (23)$$

Then for any  $X = x$ , the optimal function  $m^*(x)$  should satisfy:

$$\begin{aligned}
 & \mathbb{E}_{A,Y} \left[ \frac{Y - m^*(x)}{\pi_A^2(x)} \mid X = x \right] \\
 &= \frac{\mathbb{E}[Y(+1) \mid X = x]}{\pi_{+1}(x)} - \frac{m^*(x)}{\pi_{+1}(x)} + \frac{\mathbb{E}[Y(-1) \mid X = x]}{\pi_{-1}(x)} - \frac{m^*(x)}{\pi_{-1}(x)} \\
 &= \frac{\mathbb{E}[Y(+1) \mid X = x]}{\pi_{+1}(x)} + \frac{\mathbb{E}[Y(-1) \mid X = x]}{\pi_{-1}(x)} - m^*(x) \left( \frac{1}{\pi_{+1}(x)} + \frac{1}{\pi_{-1}(x)} \right) \\
 &= \frac{\mathbb{E}[Y(+1) \mid X = x]}{\pi_{+1}(x)} + \frac{\mathbb{E}[Y(-1) \mid X = x]}{\pi_{-1}(x)} - m^*(x) \left( \frac{1}{\pi_{+1}(x)\pi_{-1}(x)} \right) = 0
 \end{aligned}$$

Solving for  $m^*$  completes the proof:  $m^*(x) = \pi_{-1}(x)\mathbb{E}[Y(+1) \mid X = x] + \pi_{+1}(x)\mathbb{E}[Y(-1) \mid X = x] = \mu(x)$ .

### A.3 More Context for Augmentation in CATE Estimation

Here we provide a few more results that demonstrate the role of augmentation in estimation of the CATE using our pseudo-outcomes.

**Proposition 11** *The potential outcome mean can be decomposed into the counterfactual mean outcome (CMO) and the conditional average treatment effect (CATE) as follows:*

$$\mu_A(X) = \mathbb{E}(Y(A) \mid X) = \mu(X) + w(X, A)\tau(X),$$

where the weight function  $w(X, A) \equiv \frac{(1+A)\pi_{+1}(X) - (1-A)\pi_{-1}(X)}{2}$  is the only term that depends on  $A$ .

**Proof** [Proof of Proposition 11] Let's expand  $\mu(X) + w(X, A)\tau(X)$ :

$$\begin{aligned}
 \mu(X) + w(X, A)\tau(X) &= [\pi_{+1}(X)\mu_{-1}(X) + \pi_{-1}(X)\mu_{+1}(X)] \\
 &\quad + \frac{(1+A)\pi_{+1}(X) - (1-A)\pi_{-1}(X)}{2} [\mu_{+1}(X) - \mu_{-1}(X)]
 \end{aligned}$$

Collecting terms for  $\mu_{+1}(X)$  and  $\mu_{-1}(X)$ , the coefficients are:

$$\begin{aligned}
 \mu_{+1}(X) : \pi_{-1}(X) + \frac{(1+A)\pi_{+1}(X) - (1-A)\pi_{-1}(X)}{2} &= \begin{cases} \pi_{-1}(X) + \pi_{+1}(X) = 1 & A = +1 \\ \pi_{-1}(X) - \pi_{+1}(X) = 0 & A = -1 \end{cases} \\
 \mu_{-1}(X) : \pi_{+1}(X) - \frac{(1+A)\pi_{+1}(X) - (1-A)\pi_{-1}(X)}{2} &= \begin{cases} \pi_{+1}(X) - \pi_{+1}(X) = 0 & A = +1 \\ \pi_{+1}(X) + \pi_{-1}(X) = 1 & A = -1 \end{cases}
 \end{aligned}$$

This verifies that  $\mu(X) + w(X, A)\tau(X) = \mu_A(X)$  for both values of  $A$ . ■

As a known special case, consider an RCT with equal treatment probabilities. Here, the decomposition simplifies to:  $\mu^r(X, A) = \mu_A^r(X) = \mu^r(X) + \frac{A}{2}\tau^r(X)$ , where the terms  $\mu^r(X)$  and  $\frac{A}{2}\tau^r(X)$  have previously been coined in the literature as the main effect of covariates and the treatment effect, respectively (Chen et al., 2017).

The following corollary combines Proposition 1 (on bias) and Theorem 2 (on variance):

**Corollary 12** *Assume (A1)-(A3) hold and the true CMO  $\mu(X)$  is known. For a given dataset at point  $x$ ,  $\{(x, A_i, Y_i)\}_{i=1}^n$ , the estimator  $\hat{\tau}(x) = \frac{1}{n} \sum_{i=1}^n \tau_\mu(x, A_i, Y_i)$  is unbiased for the CATE at point  $x$  and minimizes the population variance.*

#### A.4 Proof of Theorem 4

Per Proposition 1, under the assumed conditions,  $\tau(X) = \mathbb{E}(\tau_m(X, A, Y)|X)$  and therefore we can write the variance of CATE as  $\mathbb{V}(\tau(X)) = \mathbb{E}\left(\frac{(Y - m(X))^2}{\pi_A^2(X)}|X\right) - \tau^2(X)$ . Only the first term of variance depends on  $m(X)$ . Our goal is to determine how much variance increases if  $m(X) := \hat{\mu}(X) = \mu(X) + d(X)$  instead of  $m(X) := \mu(X)$  which is the theoretical minimum.

$$\begin{aligned}
 & \mathbb{V}(\tau_{\hat{\mu}}(X, A, Y) | X = x) \\
 &= \mathbb{E}_{A,Y} \left( \frac{(Y - \hat{\mu}(X))^2}{\pi_A^2(X)} \middle| X \right) - \tau^2(X) \\
 &= \mathbb{E}_{A,Y} \left( \frac{(Y - \mu(X) - d(X))^2}{\pi_A^2(X)} \middle| X \right) - \tau^2(X) \\
 &= \mathbb{E}_{A,Y} \left( \frac{(Y - \mu(X))^2 - 2(Y - \mu(X))d(X) + d^2(X)}{\pi_A^2(X)} \middle| X \right) - \tau^2(X) \\
 &= \mathbb{V}(\tau_\mu(X, A, Y) | X = x) + \mathbb{E}_{A,Y} \left( \frac{-2(Y - \mu(X))d(X) + d^2(X)}{\pi_A^2(X)} \middle| X \right) \\
 &= \mathbb{V}(\tau_\mu(X, A, Y) | X = x) - 2d(x)\mathbb{E}_{A,Y} \left( \frac{(Y - \mu(X))}{\pi_A^2(X)} \middle| X \right) + d^2(x)\mathbb{E}_A \left( \frac{1}{\pi_A^2(X)} \middle| X \right)
 \end{aligned}$$

The middle term is zero because  $\mu(X)$  is the minimizer of the variance and  $\mathbb{E}_{A,Y} \left( \frac{(Y - \mu(X))}{\pi_A^2(X)} \middle| X \right)$  is the exact optimality condition used in (23). Moreover,  $\mathbb{E}_A \left( \left( \frac{1}{\pi_A^2(X)} \right) \middle| X = x \right) = \frac{1}{\pi_{+1}(x)\pi_{-1}(x)}$ . Together, they result in:

$$\begin{aligned}
 & \mathbb{V}(\tau_{\hat{\mu}}(X, A, Y) | X = x) = \mathbb{V}(\tau_\mu(X, A, Y) | X = x) + \frac{d^2(x)}{\pi_{+1}(x)\pi_{-1}(x)} \\
 & \Rightarrow \mathbb{V}(\tau_\mu | X = x) + \frac{d^2(x)}{(1-\rho)^2} \leq \mathbb{V}(\tau_{\hat{\mu}} | X = x) \leq \mathbb{V}(\tau_\mu | X = x) + \frac{d^2(x)}{\rho^2}.
 \end{aligned}$$

The last inequality holds because  $\rho \leq \pi_{+1}(x) \leq 1 - \rho$  and  $\pi_{-1}(x) = 1 - \pi_{+1}(x)$  and therefore  $\frac{1}{(1-\rho)^2} \leq \frac{1}{\pi_{+1}(x)\pi_{-1}(x)} \leq \frac{1}{\rho^2}$ . Taking expectation with respect to  $X$  completes the proof for the population case.

#### A.5 Proof of Theorem 6

We begin by analyzing the risk of an arbitrary function:

**Lemma 13 (Decomposition of Risk)** *Let  $\tau_m(X, A, Y)$  be an unbiased estimator of  $\tau(X)$ , i.e.,  $\mathbb{E}_{Y,A|X}[\tau_m(X, A, Y)] = \tau(X)$ . For any function  $f : \mathcal{X} \rightarrow \mathbb{R}$ , the risk*

$$R_m(f) \equiv \mathbb{E}_X \mathbb{E}_{Y,A|X} \left[ (\tau_m(X, A, Y) - f(X))^2 \right]$$

can be decomposed into irreducible error and estimation error:

$$R_m(f) = \mathbb{E}_X[\mathbb{V}(\tau_m | X)] + \mathbb{E}_X[(f(X) - \tau(X))^2].$$

**Proof** We decompose the risk by expanding the squared difference:

$$\begin{aligned} R_m(f) &= \mathbb{E}_X \mathbb{E}_{Y,A|X} [(\tau_m(X, A, Y) - f(X))^2] \\ &= \mathbb{E}_X \mathbb{E}_{Y,A|X} [(\tau_m(X, A, Y) - \tau(X) + \tau(X) - f(X))^2] \\ &= \mathbb{E}_X \mathbb{E}_{Y,A|X} [(\tau_m(X, A, Y) - \tau(X))^2 + (f(X) - \tau(X))^2 \\ &\quad + 2(\tau_m(X, A, Y) - \tau(X))(\tau(X) - f(X))]. \end{aligned}$$

The cross term vanishes because

$$\begin{aligned} &\mathbb{E}_X \mathbb{E}_{Y,A|X} [(\tau_m(X, A, Y) - \tau(X))(\tau(X) - f(X))] \\ &= \mathbb{E}_X [(\tau(X) - f(X)) \mathbb{E}_{Y,A|X} [\tau_m(X, A, Y) - \tau(X)]] = 0, \end{aligned}$$

where the last equality follows from the unbiasedness assumption.

Therefore,

$$\begin{aligned} R_m(f) &= \mathbb{E}_X \mathbb{E}_{Y,A|X} [(\tau_m(X, A, Y) - \tau(X))^2] + \mathbb{E}_X [(f(X) - \tau(X))^2] \\ &= \mathbb{E}_X[\mathbb{V}(\tau_m | X)] + \mathbb{E}_X [(f(X) - \tau(X))^2], \end{aligned}$$

where we use the definition of conditional variance in the last step. ■

We are interested in three key functions:

- True CATE:  $\tau(X) = \mathbb{E}_{Y,A}[\tau_m(X, A, Y) | X]$
- Best CATE estimator: The function in class  $\mathcal{F}$  that minimizes the mean squared error  $\tau^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathbb{E}_X [(f(X) - \tau(X))^2]$
- Empirical Risk Minimizer (ERM):  $\hat{\tau} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n [\tau_m(X_i, A_i, Y_i) - f(X_i)]^2$

Our primary objective is to bound the mean squared error of the ERM:  $\mathbb{E}_X [(\hat{\tau}(X) - \tau(X))^2]$ . In statistical learning theory, this is achieved by bounding the excess risk (also known as approximation error or generalization gap):  $R_m(\hat{\tau}) - R_m(\tau^*)$ , which quantifies the additional risk incurred by using the ERM instead of the best function in the function class. We leverage classic generalization bounds to establish an upper bound for the MSE of the CATE plug-in estimator.

**Lemma 14 (MSE Upper Bound of ERM)** *Let  $\tau(X)$  be the true CATE,  $\hat{\tau}$  be the ERM estimator in function class  $\mathcal{F}$ , and  $\tau^*$  be the best CATE estimator in  $\mathcal{F}$ . Define the estimation error for any estimator  $f$  as  $\Delta_\tau^2(f) = \mathbb{E}_X [(f(X) - \tau(X))^2]$ , and the minimum estimation error in class  $\mathcal{F}$  as  $\Delta_2^2(\mathcal{F}, \tau) = \inf_{f \in \mathcal{F}} \Delta_\tau^2(f)$ . Then,*

$$\Delta_2^2(\hat{\tau}, \tau) \leq 2G_m(\mathcal{F}, \mathcal{D}) + \Delta_2^2(\mathcal{F}, \tau),$$

where  $G_m(\mathcal{F}, \mathcal{D}) \equiv \sup_{f \in \mathcal{F}} |R_m(f) - \hat{R}_m(f)|$  is the supremum of the empirical risk gap over  $\mathcal{F}$  for training dataset  $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n$ .

**Proof** First, observe that the irreducible error cancels out in the excess risk:

$$\begin{aligned} R_m(\hat{\tau}) - R_m(\tau^*) &= \mathbb{E}_X[\mathbb{V}(\tau_m | X)] + \mathbb{E}_X[(\hat{\tau}(X) - \tau(X))^2] - \mathbb{E}_X[\mathbb{V}(\tau_m | X)] - \mathbb{E}_X[(\tau^*(X) - \tau(X))^2] \\ &= \Delta_2^2(\hat{\tau}, \tau) - \Delta_2^2(\mathcal{F}, \tau). \end{aligned}$$

Therefore, the estimation error can be decomposed as:

$$\Delta_2^2(\hat{\tau}, \tau) = R_m(\hat{\tau}) - R_m(\tau^*) + \Delta_2^2(\mathcal{F}, \tau).$$

To bound the excess risk  $R_m(\hat{\tau}) - R_m(\tau^*)$ , we decompose it as:

$$\begin{aligned} R_m(\hat{\tau}) - R_m(\tau^*) &= R_m(\hat{\tau}) - \hat{R}_m(\hat{\tau}) + \hat{R}_m(\hat{\tau}) - \hat{R}_m(\tau^*) + \hat{R}_m(\tau^*) - R_m(\tau^*) \\ &\leq R_m(\hat{\tau}) - \hat{R}_m(\hat{\tau}) + \hat{R}_m(\tau^*) - R_m(\tau^*) \\ &\leq |R_m(\hat{\tau}) - \hat{R}_m(\hat{\tau})| + |\hat{R}_m(\tau^*) - R_m(\tau^*)| \\ &\leq 2 \sup_{f \in \mathcal{F}} |R_m(f) - \hat{R}_m(f)| = 2G_m(\mathcal{F}, \mathcal{D}), \end{aligned}$$

where the first inequality follows from  $\hat{R}_m(\hat{\tau}) - \hat{R}_m(\tau^*) \leq 0$  by definition of ERM.

Combining these results yields the stated bound:  $\Delta_2^2(\hat{\tau}, \tau) \leq 2G_m(\mathcal{F}, \mathcal{D}) + \Delta_2^2(\mathcal{F}, \tau)$ .  $\blacksquare$

Next, as a warm-up, we present a preliminary Theorem that presents our results under restrictive settings.

**Theorem 15 (Generalization Bound via Rademacher Complexity)** *Let  $\mathcal{F}$  be a function class, and  $\hat{\tau}$  be the ERM estimator in  $\mathcal{F}$ . Assume the following:*

1. *The loss function  $L(\cdot, \cdot)$  is bounded in  $[0, l_\infty]$*
2. *For any sample  $(X_i, A_i, Y_i) \in \mathcal{D}$  and functions  $f \in \mathcal{F}$  and  $m$ :  $|\tau_m(X_i, A_i, Y_i) - f(X_i)| \leq C$*

*Then, with probability at least  $1 - \varepsilon_1$ :*

$$\Delta_2^2(\hat{\tau}, \tau) \leq \Delta_2^2(\mathcal{F}, \tau) + 4C\mathcal{R}_n(\mathcal{F}) + \frac{l_\infty}{\sqrt{2n}} \sqrt{\log \frac{2}{\varepsilon_1}},$$

*where  $\mathcal{R}_n(\mathcal{F})$  is the Rademacher complexity of function class  $\mathcal{F}$ :  $\mathcal{R}_n(\mathcal{F}) \equiv \mathbb{E}_{\epsilon, \mathcal{D}} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right]$ , with  $\epsilon_i \in \{-1, +1\}$  being i.i.d. Rademacher random variables.*

**Proof** The proof proceeds in three steps:

1. First, we bound the expected empirical risk gap using Rademacher complexity:

$$\mathbb{E}_{\mathcal{D}}[G_m(\mathcal{F}, \mathcal{D})] \leq \mathcal{R}_n(\mathcal{H}_m) \leq 2C\mathcal{R}_n(\mathcal{F}), \quad (24)$$

where  $\mathcal{H}_m = \{h : (x, a, y) \mapsto L(\tau_m(x, a, y), f(x)), f \in \mathcal{F}\}$ . The second inequality follows from the fact that for any loss function  $L(\cdot, \cdot)$  that is  $G$ -Lipschitz in its second argument (i.e.,  $|L(y_0, x') - L(y_0, x)| \leq G|x' - x|$ ), we have  $\mathcal{R}_n(\mathcal{H}_m) \leq G\mathcal{R}_n(\mathcal{F})$ . For the  $\ell_2$  loss  $L(y, x) =$

$(y - x)^2$ , under the assumption  $|\tau_m(X_i, A_i, Y_i) - f(X_i)| \leq C$ , we can show it is  $2C$ -Lipschitz in its second argument:

$$\begin{aligned}
 |L(y, x') - L(y, x)| &= |(y - x')^2 - (y - x)^2| \\
 &= |(x - x')(2y - x - x')| \\
 &= |x - x'| |2y - x - x'| \\
 &= |x - x'| (|y - x| + |y - x'|) \\
 &\leq 2C |x' - x|
 \end{aligned} \tag{25}$$

2. By McDiarmid's inequality (Bach, 2024), with probability at least  $1 - \varepsilon_1$ :

$$G_m(\mathcal{F}, \mathcal{D}) \leq 2\mathcal{R}_n(\mathcal{H}_m) + \frac{l_\infty}{\sqrt{2n}} \sqrt{\log \frac{2}{\varepsilon_1}}. \tag{26}$$

3. Combining bounds (24) and (26) with Lemma 14 yields the result. ■

**Remark 16 (On the Boundedness Assumption)** *The assumption  $|\tau_m(X_i, A_i, Y_i) - f(X_i)| \leq C$  is crucial but potentially restrictive. Unlike typical learning theory settings where boundedness can be achieved by assuming bounded outcomes (i.e.,  $Y$  in  $Y = f(X) + \epsilon$  setting) and function values ( $f(X)$ ), here  $\tau_m$  is an arbitrary transformation that depends on function  $m$ .*

*In what follows, we take a more practical approach by relaxing this assumption. Instead of requiring boundedness of arbitrary transformations  $\tau_m$ , we focus on an assumption about the optimal transformation  $\tau_\mu$ , which is more reasonable and limited. The final result explicitly connects the constants in our bound to the distance of  $m(X)$  to the true CMO  $\mu(X)$ .*

To relax the boundedness assumption on the transformed outcome  $\tau_m$ , we first show in Lemma 17 that, with high probability, the optimally transformed outcome  $\tau_\mu$  remains uniformly close to the true CATE  $\tau(X)$ . As a direct consequence, in Lemma 19 we relax the loss-boundedness assumption used in Theorem 15. Finally, in Lemma 20, we derive a new Lipschitz constant for the  $\ell_2$  loss to replace the constant  $C$  in Theorem 15. This new constant depends on how well the augmentation function  $m(X)$  approximates the true CMO  $\mu(X)$  under two reasonable assumptions: (1) the true treatment effect is bounded, i.e.  $|\tau(X)| \leq B$ , and (2) all functions in the class  $\mathcal{F}$  have bounded outputs, i.e.  $|f(X)| \leq B$ .

**Lemma 17** *Let  $X$  represent the covariates,  $A \in \{-1, +1\}$  denote the treatment assignment,  $Y$  be the observed outcome, and  $\pi_A(X)$  represent the propensity score for treatment  $A$ . For observed variables  $(X, A, Y)$ , define the optimal transformed outcome  $\tau_\mu(X, A, Y)$  as:  $\tau_\mu(X, A, Y) = \frac{A(Y - \mu(X))}{\pi_A(X)}$ , where  $\mu(X) \equiv \pi_{-1}(X)\mu_{+1}(X) + \pi_{+1}(X)\mu_{-1}(X)$  is the true Counterfactual Mean Outcome (CMO), and  $\tau(X) \equiv \mu_{+1}(X) - \mu_{-1}(X)$  is the true CATE.*

*Then, for any  $\varepsilon_2 \in (0, 1)$ , with probability at least  $1 - \varepsilon_2$ :*

$$|\tau_\mu(X, A, Y) - \tau(X)| \leq \frac{\sigma}{\sqrt{\varepsilon_2}},$$

*where  $\sigma^2 = \sup_{x \in \mathcal{X}} \mathbb{V}(\tau_\mu \mid X = x)$  is the supremum of the conditional variance of the optimal transformed outcome over the domain of  $X$ .*

**Proof** First, note that  $\tau(X) = \mathbb{E}_{A,Y}[\tau_\mu(X, A, Y) \mid X]$  by construction of  $\tau_\mu$ .

For any fixed  $x \in \mathcal{X}$ , applying Chebyshev's inequality to the random variable  $\tau_\mu(x, A, Y)$  conditional on  $X = x$ :

$$\mathbb{P}\left(|\tau_\mu(x, A, Y) - \tau(x)| \geq \frac{\sigma}{\sqrt{\varepsilon_2}} \mid X = x\right) \leq \frac{\mathbb{V}(\tau_\mu \mid X = x)}{(\sigma/\sqrt{\varepsilon_2})^2} \leq \varepsilon_2,$$

where the last inequality follows from the definition of  $\sigma^2$  as the supremum of the conditional variance. The result follows by the law of total probability over  $X$ .  $\blacksquare$

**Remark 18** *The main assumption of Lemma 17 is that the variance of the most efficient estimator remains bounded over the entire domain of covariates. This requirement is not restrictive. Indeed, from the definitions of  $\tau_\mu(X, A, Y)$  and  $\tau(X)$ , one can see that it is equivalent to assuming*

$$\max(\mathbb{V}(Y(+1) - \mu_{+1}(X) \mid X = x), \mathbb{V}(Y(-1) - \mu_{-1}(X) \mid X = x)) < \infty,$$

*i.e., the per-arm conditional outcome variance is uniformly bounded. In the additive noise model  $Y(a) = f_a(X) + \epsilon_a$ , this condition simply means that the noise  $\epsilon_a$  has bounded variance.*

One of the consequences of Lemma 17 is that now we can bound the loss function and relax the boundedness assumption of Theorem 15:

**Lemma 19 (High-Probability Loss Bound)** *Let  $|\tau(X)| \leq B$  and  $|f(X)| \leq B$  for all  $f \in \mathcal{F}$ . Define  $\sigma^2 = \sup_{x \in \mathcal{X}} \mathbb{V}(\tau_\mu \mid X = x)$  as the supremum of the conditional variance of the optimal transformed outcome over the domain of  $X$ . Then, with probability at least  $1 - \varepsilon_1$ , the squared loss is bounded as:*

$$\ell_2(\tau_m(X, A, Y), f(X)) = (\tau_m(X, A, Y) - f(X))^2 \leq l_\infty,$$

where  $l_\infty = \left(2B + \frac{\sigma}{\sqrt{\varepsilon_1}} + \frac{1}{\rho} \sup_{x \in \mathcal{X}} |m(X) - \mu(X)|\right)^2$  and  $\rho$  is a lower bound on the propensity scores.

**Proof** For any given sample  $(X, A, Y)$  we have:

$$\begin{aligned} |\tau_m(X, A, Y) - f(X)| &= |\tau_m(X, A, Y) - \tau(X) + \tau(X) - f(X)| \\ &\leq |\tau_m(X, A, Y) - \tau(X)| + |\tau(X) - f(X)| \\ &\leq |\tau_m(X, A, Y) - \tau_\mu(X, A, Y) + \tau_\mu(X, A, Y) - \tau(X)| + 2B \\ (\text{Lemma 17}) &\leq |\tau_m(X, A, Y) - \tau_\mu(X, A, Y)| + \frac{\sigma}{\sqrt{\varepsilon_1}} + 2B \\ &\leq \frac{1}{\rho} |m(X) - \mu(X)| + \frac{\sigma}{\sqrt{\varepsilon_1}} + 2B \\ &\leq \frac{1}{\rho} \sup_{x \in \mathcal{X}} |m(X) - \mu(X)| + \frac{\sigma}{\sqrt{\varepsilon_1}} + 2B \end{aligned}$$

where the third inequality onward hold with probability at least  $1 - \varepsilon_1$ .

Therefore, with the same probability:

$$\begin{aligned} \forall X, A, Y : \quad \ell_2(\tau_m(X, A, Y), f(X)) &= (\tau_m(X, A, Y) - f(X))^2 \\ &\leq \left(2B + \frac{\sigma}{\sqrt{\varepsilon_1}} + \frac{1}{\rho} \sup_{x \in \mathcal{X}} |m(X) - \mu(X)|\right)^2 = l_\infty \end{aligned}$$

■

Finally, under relaxed assumptions, we derive a new Lipschitz constant to be used for Rademacher complexity in below Lemma.

**Lemma 20 (Rademacher Complexity Bound for Loss Class)** *Let  $\mathcal{F}$  be a function class where  $|f(X)| \leq B$  for all  $f \in \mathcal{F}$ , and assume  $|\tau(X)| \leq B$ . Let  $\mathcal{H}_m = \{h : z \mapsto (\tau_m(z) - f(x))^2, z = (x, a, y), f \in \mathcal{F}\}$  be the squared loss function class. Then, with probability at least  $1 - \varepsilon_2$ :*

$$\mathcal{R}_n(\mathcal{H}_m) \leq l_m \mathcal{R}_n(\mathcal{F}),$$

where  $l_m = \left(4B + \frac{2\sigma}{\sqrt{\varepsilon_2}} + \frac{2}{\rho} \sup_{x \in \mathcal{X}} |\mu(x) - m(x)|\right)$  and  $\rho$  is a lower bound on the propensity scores.

**Proof** For brevity, denote  $Z_i = (X_i, A_i, Y_i)$  and let  $\tau_m(Z_i) \equiv \tau_m(X_i, A_i, Y_i)$ . We first examine the Lipschitz continuity condition of the  $\ell_2$  loss from Equation (25):

$$\begin{aligned} & |(\tau_m(Z_i) - f(X_i))^2 - (\tau_m(Z_i) - f(X_j))^2| \\ & \leq |f(X_i) - f(X_j)| (|\tau_m(Z_i) - f(X_i)| + |\tau_m(Z_i) - f(X_j)|) \end{aligned}$$

Adding and subtracting the conditional expectation  $\mathbb{E}_{A,Y}[\tau_m \mid X_i] = \tau(X_i)$  and applying the triangle inequality:

$$\begin{aligned} & |(\tau_m(Z_i) - f(X_i))^2 - (\tau_m(Z_i) - f(X_j))^2| \\ & \leq |f(X_i) - f(X_j)| \left( |\tau_m(Z_i) - \mathbb{E}_{A,Y}[\tau_m \mid X_i]| + |\tau(X_i) - f(X_j)| + \right. \\ & \quad \left. |\tau_m(Z_i) - \mathbb{E}_{A,Y}[\tau_m \mid X_i]| + |\tau(X_i) - f(X_i)| \right) \\ & \leq |f(X_i) - f(X_j)| (4B + 2 |\tau_m(Z_i) - \mathbb{E}_{A,Y}[\tau_m \mid X_i]|) \end{aligned}$$

where we used the boundedness assumptions.

Decomposing via the optimal transformation  $\tau_\mu$ :

$$\begin{aligned} & |(\tau_m(Z_i) - f(X_i))^2 - (\tau_m(Z_i) - f(X_j))^2| \\ & \leq |f(X_i) - f(X_j)| \left( 4B + 2 |\tau_m(Z_i) - \tau_\mu(Z_i) + \tau_\mu(Z_i) - \tau(X_i)| \right) \\ & \leq |f(X_i) - f(X_j)| \left( 4B + \frac{2\sigma}{\sqrt{\varepsilon_2}} + 2 |\tau_m(Z_i) - \tau_\mu(Z_i)| \right) \end{aligned}$$

where the last inequality holds with probability  $1 - \varepsilon_2$  from Lemma 17.

Expanding the definition of transformation we have

$$\begin{aligned} |\tau_m(Z_i) - \tau_\mu(Z_i)| &= \left| \frac{A_i}{\pi_{A_i}(X_i)} [Y_i - m(X_i) - Y_i + \mu(X_i)] \right| \\ &= \left| \frac{A_i}{\pi_{A_i}(X_i)} \right| |\mu(X_i) - m(X_i)| \\ &\leq \frac{1}{\rho} |\mu(X_i) - m(X_i)| \end{aligned}$$



where we used the propensity score bound  $\pi_{A_i}(X_i) \geq \rho$  in the last inequality.

The final form of the Lipschitz condition will be:

$$|(\tau_m(Z_i) - f(X_i))^2 - (\tau_m(Z_i) - f(X_j))^2| \leq |f(X_i) - f(X_j)| \left( 4B + \frac{2\sigma}{\sqrt{\varepsilon_2}} + \frac{2}{\rho} \sup_{x \in \mathcal{X}} |\mu(x) - m(x)| \right)$$

Therefore, with probability at least  $1 - \varepsilon_2$ , the squared loss class  $\mathcal{H}_m$  is  $l_m$ -Lipschitz, and:

$$\mathcal{R}_n(\mathcal{H}_m) = \mathbb{E}_{\epsilon, \mathcal{D}} \left[ \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i) \right] \leq l_m \mathbb{E}_{\epsilon, \mathcal{D}} \left[ \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right] = l_m \mathcal{R}_n(\mathcal{F})$$

■

**Remark 21** While  $m(X)$  is estimated from data (to approximate  $\mu(X)$ ), we assume it is computed using an independent dataset through sample splitting. For instance, with  $n$  samples, one can use half to learn  $m(X)$  and the other half to learn  $\tau(X)$ . Cross-fitting, which swaps the roles of the samples and averages the learned parameters, can utilize the full dataset. This ensures independence of  $L_m$  from dataset  $\mathcal{D}$ , and therefore, we can take it out of the expectation.

**Theorem 22 (CATE Estimation Error Bound)** Let  $\mathcal{F}$  be a function class and  $\hat{\tau}$  be the ERM estimator in  $\mathcal{F}$ . Assume:

1. The true CATE and functions in  $\mathcal{F}$  are bounded:  $|\tau(X)|, |f(X)| \leq B$  for all  $f \in \mathcal{F}$
2. The propensity scores are bounded away from zero:  $\pi_A(X) \geq \rho > 0$
3. The conditional variance of the optimal transformed outcome is bounded:  $\sigma^2 \equiv \sup_{x \in \mathcal{X}} \mathbb{V}(\tau_\mu | X = x) < \infty$

Then, with probability at least  $1 - 2\varepsilon$ :

$$\Delta_2^2(\hat{\tau}, \tau) \leq \Delta_2^2(\mathcal{F}, \tau) + 2C(m, \varepsilon) \mathcal{R}_n(\mathcal{F}) + C^2(m, \varepsilon) \sqrt{\frac{\log(2/\varepsilon)}{n}},$$

where  $C(m, \varepsilon) = \left( 2B + \frac{\sigma}{\sqrt{\varepsilon}} + \frac{1}{\rho} \sup_{x \in \mathcal{X}} |\mu(x) - m(x)| \right)$  is the Lipschitz constant of the loss function.

**Proof** The proof combines four key results:

1. From Lemma 17, with probability at least  $1 - \varepsilon_1$ , the optimal transformed outcome concentrates around the true CATE with radius  $\sigma/\sqrt{\varepsilon_1}$ .
2. Consequently, from Lemma 19, with probability at least  $1 - \varepsilon_1$ , the loss is bounded with  $C^2(m, \varepsilon)$ .
3. From Lemma 20, with probability at least  $1 - \varepsilon_2$ , the squared loss class  $\mathcal{H}_m$  is  $C(m, \varepsilon)$ -Lipschitz with respect to  $\mathcal{F}$ , implying:  $\mathcal{R}_n(\mathcal{H}_m) \leq C(m, \varepsilon) \mathcal{R}_n(\mathcal{F})$
4. Applying McDiarmid's inequality with the bounded loss and the new Rademacher complexity while setting  $\varepsilon_1 = \varepsilon_2 = \varepsilon$  completes the proof.

■

Table 5: Rademacher Complexity Bounds for Linear and Convex Models

Function Class	Rademacher Complexity Bound	Key Assumptions
Linear, $\ell_2$ -constraint: $\mathcal{F} = \{x \mapsto w^\top x : \ w\ _2 \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B \cdot \sqrt{\text{tr}(\Sigma)/n}$	Data bounded in $\ell_2$ or covariance $\Sigma = \mathbf{E}[xx^\top]$ bounded.
Linear, $\ell_1$ -constraint: $\mathcal{F} = \{x \mapsto w^\top x : \ w\ _1 \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq \frac{B}{\max_i \ x_i\ _\infty} \sqrt{\ln d/n}$	Samples $x_i$ have bounded $\ \cdot\ _\infty$ norm.
Sparse linear ( $s$ -sparse): $\mathcal{F} = \{x \mapsto w^\top x : \ w\ _0 \leq s, \ w\ _2 \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B\sqrt{s \ln(d/s)/n}$	Bounded features; $d = \dim(x)$ .
Convex Lipschitz: $\mathcal{F} = \{f : \mathbf{X} \rightarrow \mathbf{R}, f \text{ convex}, \ f\ _{\text{Lip}} \leq L\}$	$\mathcal{R}_n(\mathcal{F}) \leq L \cdot \text{diam}(\mathbf{X})/\sqrt{n}$	Domain $\mathbf{X}$ has diameter $\text{diam}(\mathbf{X})$ .

Table 6: Neural Networks

Function Class	Rademacher Complexity Bound	Key Assumptions
One-hidden-layer, $H$ units: $\mathcal{F} = \left\{ x \mapsto \sum_{j=1}^H v_j \sigma(w_j^\top x) : \ w_j\ _2 \leq B_w, \ v\ _1 \leq B_v \right\}$	$\mathcal{R}_n(\mathcal{F}) \leq \frac{B_v B_w \ \mathbf{X}\ _F}{\sqrt{n}}$	$\sigma$ is 1-Lipschitz; $\ \mathbf{X}\ _F$ is the Frobenius norm.
Deep nets, depth $L$ : $\mathcal{F} = \{f : \ W_l\ _2 \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B^L \ \mathbf{X}\ _F \prod_{l=1}^{L-1} \ W_l\ _2 / \sqrt{n}$	Spectral-norm constraints on weight matrices.
ReLU nets, depth $L$ , width $H$ : $\mathcal{F} = \{f : \ W_l\ _F \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B^L \sqrt{H} \ \mathbf{X}\ _F / \sqrt{n}$	Frobenius norm bound on each weight matrix. Scales with $\sqrt{H}$ and $B^L$ .
Conv nets, bounded filters: $\mathcal{F} = \{f : \ W\ _F \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B\sqrt{d \ln d/n}$	$d$ = total number of parameters.

## A.6 Rademacher Complexity Bounds for Common Function Classes

### A.7 Proof of Proposition 7

**Proof** We prove parts (i) and (ii) separately.

**Part (i): MSE Convergence.** By definition, for any  $x \in \mathcal{X}$ , we have

$$\hat{\mu}(x) - \mu(x) = \pi_{-1}(x) [\hat{\mu}_{+1}(x) - \mu_{+1}(x)] + \pi_{+1}(x) [\hat{\mu}_{-1}(x) - \mu_{-1}(x)].$$

Table 7: Other Function Classes

Function Class	Rademacher Complexity Bound	Key Assumptions
$\ell_p$ -norm ball: $\mathcal{F} = \{x \mapsto w^\top x : \ w\ _p \leq r\}$	$\mathcal{R}_n(\mathcal{F}) \leq r\ \mathbf{X}\ _{p^*}/\sqrt{n}$	$p^*$ is the dual of $p$ ; $\ X\ _{p^*}$ is the dual norm of data matrix.
Depth- $D$ decision trees:	$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{D \ln(2n)/n}$	Assumes roughly balanced splits.
$k$ -nearest neighbors:	$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{k \ln n/n}$	Assumes bounded outputs.
Smooth functions: $\mathcal{F} = \{f : \ f\ _{\text{Lip}} \leq L, \ f\ _\infty \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq L \cdot \text{diam}(\mathbf{X})/\sqrt{n}$	Domain $\mathbf{X}$ of diameter $\text{diam}(\mathbf{X})$ .

Table 8: Non-parametric Methods

Function Class	Rademacher Complexity Bound	Key Assumptions
Hölder smooth functions: $\mathcal{F} = \{f \in C^\alpha(\mathbf{X}) : \ f\ _{C^\alpha} \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B \cdot (1/n)^{\alpha/(2\alpha+d)}$	Functions with $\alpha$ derivatives, input dimension $d$ . Optimal rate for non-parametric regression.
Sparse additive models (SpAM): $\mathcal{F} = \left\{ f(x) = \sum_{j \in S} f_j(x_j) :  S  \leq s, \ f_j\ _\infty \leq B \right\}$	$\mathcal{R}_n(\mathcal{F}) \leq B \sqrt{s \ln(d)/n}$	At most $s$ active features out of $d$ total features. Each component function bounded.
Besov spaces $B_{p,q}^\alpha$ : $\mathcal{F} = \{f \in B_{p,q}^\alpha : \ f\ _{B_{p,q}^\alpha} \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq \frac{B}{(1/n)^{\alpha/(2\alpha+d(1/p-1/2)_+)}}$	Wavelet-based smoothness spaces. Generalize Hölder and Sobolev spaces.
Sobolev spaces: $\mathcal{F} = \{f \in W^{k,p}(\mathbf{X}) : \ f\ _{W^{k,p}} \leq B\}$	$\mathcal{R}_n(\mathcal{F}) \leq B \cdot (1/n)^{k/(2k+d)}$	Functions with $k$ weak derivatives in $L^p$ . Domain $\mathbf{X} \subset \mathbb{R}^d$ .

Squaring both sides yields

$$\begin{aligned}
 \left( \hat{\mu}(x) - \mu(x) \right)^2 &= \pi_{-1}(x)^2 \left[ \hat{\mu}_{+1}(x) - \mu_{+1}(x) \right]^2 + \pi_{+1}(x)^2 \left[ \hat{\mu}_{-1}(x) - \mu_{-1}(x) \right]^2 \\
 &\quad + 2 \pi_{-1}(x) \pi_{+1}(x) \left| \hat{\mu}_{+1}(x) - \mu_{+1}(x) \right| \left| \hat{\mu}_{-1}(x) - \mu_{-1}(x) \right|.
 \end{aligned}$$

Since the propensity weights satisfy  $0 \leq \pi_a(x) \leq 1$ , it follows that

$$\left( \hat{\mu}(x) - \mu(x) \right)^2 \leq \left[ \hat{\mu}_{+1}(x) - \mu_{+1}(x) \right]^2 + \left[ \hat{\mu}_{-1}(x) - \mu_{-1}(x) \right]^2 \tag{27}$$

$$+ 2 \left| \hat{\mu}_{+1}(x) - \mu_{+1}(x) \right| \left| \hat{\mu}_{-1}(x) - \mu_{-1}(x) \right|. \tag{28}$$

Taking expectation with respect to  $X$  and applying the Cauchy–Schwarz inequality to the cross term, we obtain

$$\begin{aligned} \mathbb{E}\left[\left(\hat{\mu}(X) - \mu(X)\right)^2\right] &\leq \mathbb{E}\left[\left(\hat{\mu}_{+1}(X) - \mu_{+1}(X)\right)^2\right] + \mathbb{E}\left[\left(\hat{\mu}_{-1}(X) - \mu_{-1}(X)\right)^2\right] \\ &\quad + 2\sqrt{\mathbb{E}\left[\left(\hat{\mu}_{+1}(X) - \mu_{+1}(X)\right)^2\right] \cdot \mathbb{E}\left[\left(\hat{\mu}_{-1}(X) - \mu_{-1}(X)\right)^2\right]}. \end{aligned}$$

By the per-arm assumption,

$$\mathbb{E}\left[\left(\hat{\mu}_a(X) - \mu_a(X)\right)^2\right] = \mathcal{O}_p\left(r^2(n'_a)\right) \quad \text{for } a \in \{+1, -1\}.$$

Since  $n'_a \geq \eta n'$  for both arms, we have  $r(n'_a) = \mathcal{O}(r(n'))$ . Consequently, there exists a constant  $C > 0$  (depending on  $\eta$ ) such that

$$\mathbb{E}\left[\left(\hat{\mu}(X) - \mu(X)\right)^2\right] \leq C\left(r^2(n') + r(n')r(n')\right) = \mathcal{O}_p\left(r^2(n')\right).$$

This proves the MSE convergence claim.

**Part (ii): Uniform Convergence.** We need the following lemma—a standard result in functional analysis—to connect the two modes of convergence.

**Lemma 23 (MSE Implies Uniform Convergence)** *Let  $\mathcal{X} \subset \mathbb{R}^d$  be compact, and suppose  $\mu, \hat{\mu}_{n'} : \mathcal{X} \rightarrow \mathbb{R}$  are both  $L$ -Lipschitz functions. Then*

$$\begin{aligned} \Delta_2^2(\hat{\mu}_{n'}, \mu) = \mathbb{E}\left[(\hat{\mu}_{n'}(X) - \mu(X))^2\right] &= \mathcal{O}_p\left(r^2(n')\right) \\ \iff \Delta_\infty(\hat{\mu}_{n'}, \mu) = \sup_{x \in \mathcal{X}} |\hat{\mu}_{n'}(x) - \mu(x)| &= \mathcal{O}_p\left(r(n')\right) \end{aligned}$$

**Proof** We show both directions:

**1) ( $\implies$ ) MSE  $\implies$  Uniform Convergence.** Assume  $\mathbb{E}\left[(\mu(X) - m(X))^2\right] \leq \xi$  for some small  $\xi > 0$ . Our goal is to prove  $\sup_{x \in \mathcal{X}} |\mu(x) - m(x)| \leq C\sqrt{\xi}$ , where  $C$  depends on  $L$  and the diameter of  $\mathcal{X}$ .

1. By Lipschitz continuity of both  $\mu$  and  $m$ , for any  $x, y \in \mathcal{X}$ :

$$|(m(x) - \mu(x)) - (m(y) - \mu(y))| \leq |m(x) - m(y)| + |\mu(x) - \mu(y)| \leq 2L\|x - y\|.$$

Hence,  $m(x) - \mu(x)$  is at most  $2L$ -Lipschitz.

2. By compactness of  $\mathcal{X}$ , for any  $\epsilon > 0$  there exists a finite  $\epsilon$ -cover  $x_1, \dots, x_N \in \mathcal{X}$  such that for every  $x \in \mathcal{X}$  there is an  $x_{i(x)}$  with  $\|x - x_{i(x)}\| \leq \epsilon$ . The integer  $n$  depends on  $\epsilon$  and the dimension  $d$ .

3. For any  $x \in \mathcal{X}$ , pick  $i(x)$  such that  $\|x - x_{i(x)}\| \leq \epsilon$ . Then

$$|\mu(x) - m(x)| \leq |\mu(x_{i(x)}) - m(x_{i(x)})| + 2L\|x - x_{i(x)}\| \leq |\mu(x_{i(x)}) - m(x_{i(x)})| + 2L\epsilon.$$

Taking the supremum over  $x \in \mathcal{X}$  gives

$$\sup_{x \in \mathcal{X}} |\mu(x) - m(x)| \leq \max_{1 \leq i \leq N} |\mu(x_i) - m(x_i)| + 2L\epsilon.$$

4. By Markov's inequality, for any  $t > 0$ ,  $P(|\mu(X) - m(X)|) \leq \frac{\xi}{t^2}$ .

Applying a union bound across the  $n$  cover points,  $P(\max_{1 \leq i \leq N} |\mu(x_i) - m(x_i)|) \leq \frac{N\xi}{t^2}$ .

Choose  $t = \sqrt{\frac{N\xi}{\varepsilon}}$  for  $\varepsilon \in (0, 1)$ . Then with probability at least  $1 - \varepsilon$ ,

$$\max_{1 \leq i \leq N} |\mu(x_i) - m(x_i)| \leq \sqrt{\frac{N\xi}{\varepsilon}}.$$

Combining with the bound in Step 3,  $\sup_{x \in \mathcal{X}} |\mu(x) - m(x)| \leq \sqrt{\frac{N\xi}{\varepsilon}} + 2L\varepsilon$ .

5. Setting  $\varepsilon = \frac{\sqrt{\xi}}{2L}$  yields:  $\sup_{x \in \mathcal{X}} |\mu(x) - m(x)| \leq \sqrt{\frac{N\xi}{\varepsilon}} + 2L\frac{\sqrt{\xi}}{2L} = \sqrt{\frac{N}{\varepsilon}}\sqrt{\xi} + \sqrt{\xi} = C\sqrt{\xi}$ , where  $C = \sqrt{\frac{N}{\varepsilon}} + 1$  depends on the cover size  $n$  (hence on  $\dim(\mathcal{X})$  and the diameter of  $\mathcal{X}$ ).

Thus, small MSE at rate  $\xi$  implies  $\sup_{x \in \mathcal{X}} |\mu(x) - m(x)|$  is  $\mathcal{O}(\sqrt{\xi})$ .

**2) ( $\Leftarrow$ ) Uniform Convergence  $\implies$  MSE.** Now, for the other direction, assume

$$\sup_{x \in X} |\mu(x) - m(x)| \leq r(n') \quad \text{with high probability.}$$

Then pointwise we have  $|\mu(X) - m(X)|^2 \leq r(n')^2$ . Hence,  $\mathbb{E}[(\mu(X) - m(X))^2] \leq r(n')^2$ , so in probability we get  $\Delta_2^2(m, \mu) = \mathcal{O}_p(r^2(n'))$ .

Putting both directions together, we see that under the proposed assumptions,

$$\Delta_2^2(m, \mu) = \mathcal{O}_p(r^2(n')) \iff \Delta_\infty(m, \mu) = \mathcal{O}_p(r(n')).$$

This completes the proof. ■

Now, let's move back to the proof part (ii) of Proposition 7.

Assume in addition that for each  $a \in \{+1, -1\}$  the functions  $\mu_a$  and their estimators  $\hat{\mu}_a$  are  $L$ -Lipschitz on the compact set  $\mathcal{X}$ . Then, by the standard result stated in Lemma 23, the per-arm mean squared error control,  $\mathbb{E}[(\hat{\mu}_a(X) - \mu_a(X))^2] = \mathcal{O}_p(r^2(n'_a))$ , implies that  $\sup_{x \in \mathcal{X}} |\hat{\mu}_a(x) - \mu_a(x)| = \mathcal{O}_p(r(n'_a)) = \mathcal{O}_p(r(n'))$ .

Recall that the CMO estimator is constructed as  $\hat{\mu}(x) = \pi_{-1}(x) \hat{\mu}_{+1}(x) + \pi_{+1}(x) \hat{\mu}_{-1}(x)$ , and similarly for  $\mu(x)$ . Thus, for any  $x \in \mathcal{X}$ ,

$$\begin{aligned} |\hat{\mu}(x) - \mu(x)| &= |\pi_{-1}(x) [\hat{\mu}_{+1}(x) - \mu_{+1}(x)] + \pi_{+1}(x) [\hat{\mu}_{-1}(x) - \mu_{-1}(x)]| \\ &\leq |\hat{\mu}_{+1}(x) - \mu_{+1}(x)| + |\hat{\mu}_{-1}(x) - \mu_{-1}(x)|. \end{aligned}$$

Taking the supremum over  $\mathcal{X}$  yields

$$\Delta_\infty(\hat{\mu}, \mu) = \sup_{x \in \mathcal{X}} |\hat{\mu}(x) - \mu(x)| \leq \sup_{x \in \mathcal{X}} |\hat{\mu}_{+1}(x) - \mu_{+1}(x)| + \sup_{x \in \mathcal{X}} |\hat{\mu}_{-1}(x) - \mu_{-1}(x)| = \mathcal{O}_p(r(n')).$$

This completes the proof of uniform convergence.

Combining the two parts, we conclude that the CMO estimator  $\hat{\mu}(x)$  converges to  $\mu(x)$  in MSE at the rate  $\mathcal{O}_p(r^2(n'))$  (without extra smoothness assumptions), and if the per-arm functions are additionally  $L$ -Lipschitz on a compact domain, then the uniform convergence rate is  $\mathcal{O}_p(r(n'))$ . ■

### A.8 Proof of Proposition 9

**Proof** By Theorem 6, with probability at least  $1 - 2\varepsilon$ :  $\Delta_2^2(\hat{\tau}_{n^r}, \tau) \leq \Delta_2^2(\mathcal{F}, \tau) + 2C(m, \varepsilon)\mathcal{R}_{n^r}(\mathcal{F}) + C^2(m, \varepsilon)\sqrt{\frac{\log(2/\varepsilon)}{n^r}}$ . Under the assumption (A5),  $\Delta_2^2(\mathcal{F}, \tau) = 0$ . For any well-behaved function class  $\mathcal{F}$ , the Rademacher complexity term  $\mathcal{R}_n(\mathcal{F})$  converges to zero as  $n^r \rightarrow \infty$ . The final term, containing the deviation penalty  $\sqrt{\log(2/\varepsilon)}$ , also vanishes at rate  $1/\sqrt{n^r}$ . Therefore,  $\Delta_2^2(\hat{\tau}_{n^r}) \rightarrow 0$  as  $n^r \rightarrow \infty$ , establishing consistency. ■

### References

- Amir Asiaee, Samet Oymak, Kevin R Coombes, and Arindam Banerjee. High dimensional data enrichment: Interpretable, fast, and data-efficient. *arXiv preprint arXiv:1806.04047*, 2018.
- Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.
- Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.
- Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. ISSN 0090-5364, 2168-8966. doi: 10.1214/18-AOS1709.
- Francis Bach. *Learning Theory from First Principles*. The MIT Press, 2024.
- Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.
- Peter L Bartlett and Shahar Mendelson. Local rademacher complexities and empirical minimization. *Annals of Statistics*, 34, 2006.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.
- Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- Rui Chen, Jared D Huling, Guanhua Chen, and Menggang Yu. Robust sample weighting to facilitate individualized treatment rule learning for a target population. *Biometrika*, 111(1):309–329, 2024.
- Shuai Chen, Lu Tian, Tianxi Cai, and Menggang Yu. A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*, 73(4):1199–1209, 2017.
- David Cheng and Tianxi Cai. Adaptive combination of randomized and observational data. *arXiv preprint arXiv:2111.15012*, 2021.

- Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.
- Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.
- Issa J Dahabreh, Sarah E Robertson, Jon A Steingrimsso, Elizabeth A Stuart, and Miguel A Hernan. Extending inferences from a randomized trial to a new target population. *Statistics in medicine*, 39(14):1999–2014, 2020.
- Xiaowu Dai and Peter Chien. Another look at statistical calibration: A non-asymptotic theory and prediction-oriented optimality. *arXiv preprint arXiv:1802.00021*, 2018.
- Irina Degtiar and Sherri Rose. A review of generalizability and transportability. *Annual Review of Statistics and Its Application*, 10:501–524, 2023.
- Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1097–1104, 2011.
- Qingliang Fan, Yu-Chin Hsu, Robert P Lieli, and Yichong Zhang. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1): 313–327, 2022.
- Samuel M Gross and Robert Tibshirani. Data shared lasso: A novel tool to discover uplift. *Computational statistics & data analysis*, 101:226–235, 2016.
- P Richard Hahn, Jared S Murray, and Carlos M Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.
- Kai Helli, David Schnurr, Noah Hollmann, Samuel Müller, and Frank Hutter. Drift-resilient tabpfn: In-context learning regression temporal distribution shifts on tabular data. In *NeurIPS*, 2024.
- Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC., 2020.
- JG Ibrahim and MH Chen. Power prior distributions for regression models. *Statistical Science*, 15: 46–60, 2000.
- Daniel Jacob. Cate meets ml: Conditional average treatment effect and machine learning. *Digital Finance*, 3(2):99–148, 2021.
- A Kaizer, JS Koopmeiners, and BP Hobbs. Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics*, 19(2):169–184, 2018.
- Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

- Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.
- Marc C Kennedy and Anthony O’Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
- Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D’Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.
- Michael R Kosorok and Eric B Laber. Precision medicine. *Annual Review of Statistics and its Application*, 6:263–286, 2019.
- A Kotalik, DM Vock, EC Donny, DK Hatsukami, and JS Koopmeiners. Dynamic borrowing in the presence of treatment effect heterogeneity. *Biostatistics*, 22(4):789–804, 2021.
- Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018.
- Sören R Künzle, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.
- Christoph F Kurz. Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, 42(2):156–167, 2022.
- X Li and Y Song. Target population statistical inference with data integration across multiple sources—an approach to mitigate information shortage in rare disease clinical trials. *Statistics in Biopharmaceutical Research*, 12(3):322–333, 2020.
- Zhe Li, Eivind Kristoffersen, and Jingyue Li. Deep transfer learning for failure prediction across failure types. *Reliability Engineering & System Safety*, 223:108508, 2022.
- Y Liu, B Lu, R Foster, Y Zhang, ZJ Zhong, MH Chen, and P Sun. Matching design for augmenting the control arm of a randomized controlled trial using real-world data. *Biopharm Stat*, 32(1):124–140, 2022.
- B Neuenschwander, G Capkun-Niggli, M Branson, and DJ Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010.
- Anh Tuan Nguyen, Hyewon Jeong, Eunho Yang, and Sungju Hwang. Temporal probabilistic asymmetric multi-task learning. *International Conference on Learning Representations (ICLR)*, 2020.
- Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.
- Michael Oberst, Alexander D’Amour, Minmin Chen, Yuyan Wang, David Sontag, and Steve Yadowlowsky. Understanding the risks and rewards of combining unbiased and possibly biased estimators, with applications to causal inference. *arXiv preprint arXiv:2205.10467*, 2022.



- SJ Pocock. The combination of randomized and historical controls in clinical trials. *Journal of Chronic Diseases*, 29(3):175–188, 1976. ISSN 0021–9681. doi: [https://doi.org/10.1016/0021-9681\(76\)90044-8](https://doi.org/10.1016/0021-9681(76)90044-8).
- Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.
- Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180–1210, 2011.
- Jon A. Steingrimsso, Constantine Gatsonis, Bing Li, and Issa J. Dahabreh. Transporting a prediction model for use in a new target population. *Epidemiology*, 34(2):216–224, 2023.
- Jared Strauch and Amir Asiaee. Improving drug sensitivity prediction and inference by multi-task learning. *bioRxiv*, pages 2024–05, 2024.
- Harini Suresh, Jen J. Gong, and John V. Guttag. Learning tasks for multitask learning: Heterogeneous patient populations in the icu. *arXiv preprint arXiv:1806.02878*, 2019.
- Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.
- Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, 1st edition. 2nd printing. 2008 edition, November 2008. ISBN 978-0-387-79051-0.
- Stefan Wager. Causal inference: A statistical learning approach, 2024.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.
- Zhi Wang et al. Molecular pathways enhance drug response prediction using transfer learning from cell lines to tumors and patient-derived xenografts. *Scientific Reports*, 12(1):1–14, 2022.
- Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.
- Lili Wu and Shu Yang. Integrative  $\text{SR\$}$ -learner of heterogeneous treatment effects combining experimental and observational studies. In *Proceedings of the First Conference on Causal Learning and Reasoning*, pages 904–926. PMLR, June 2022. ISSN: 2640-3498.
- Shu Yang, Siyi Liu, Donglin Zeng, and Xiaofei Wang. Data fusion methods for the heterogeneity of the treatment effect and confounding function. *Bernoulli*, 2025. doi: 10.48550/arXiv.2007.12922. URL <http://arxiv.org/abs/2007.12922>. Accepted for publication.
- Zhenghao Zeng, Edward H Kennedy, Lisa M Bodnar, and Ashley I Naimi. Efficient generalization and transportation. *arXiv preprint arXiv:2302.00092*, 2023.
- Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Zhao Zhili, Qin Jian, Gou Zhuoyue, Zhang Yanan, and Yang Yi. Multi-task learning models for predicting active compounds. *Journal of Biomedical Informatics*, 107:103443, 2020.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76, 2020.