Novel data-adaptive statistical methods for observational and experimental studies with applications to HIV/AIDS research

Ernesto J Ulloa

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Marco Carone, Chair

Alex Luedtke, Chair

Jennifer Nelson

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Novel data-adaptive statistical methods for observational and experimental studies with applications to HIV/AIDS research

Ernesto J Ulloa

Co-Chairs of the Supervisory Committee:
Marco Carone
Department of Biostatistics

Alex Luedtke
Department of Statistics

Data-adaptive statistical methods can often be used to improve different attributes of estimators, such as their asymptotic variance, calibration, and predictive accuracy. In this dissertation, we (i) describe how careful estimation of the propensity score can be used to improve the efficiency of a matching estimator; (ii) propose and examine the asymptotic properties of isotonic calibration in the context of conditional average treatment effect estimation; and (iii) implement and assess a SuperLearner for HIV risk prediction that can be used to improve the precision of treatment effect estimators in HIV prevention trials. Matching adjusts for confounding by comparing individuals with similar propensity score; however, it often results in an inefficient estimator. In Chapter 2, we propose mitigating this inefficiency by including an optimal covariate —estimated via data adaptive methods— in the propensity score model. Chapter 3 describes how isotonic calibration, paired with flexible estimation of the pseudo-outcomes, can be used to calibrate a given CATE estimator. We provide rates of convergence of our method's calibration and predictive performance. Finally, in Chapter 4, we propose using flexible data adaptive methods to implement a Super Learner for HIV prediction and assess its performance when predicting HIV risk based on different combinations of baseline covariates and varying risk time windows.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This has been a long journey. As such, I have had the joy and luck to meet wonderful people along the way.

First, I want to thank my advisors Marco and Alex for their invaluable training and support. I've learnt so much working under their guidance. I want to thank Jennifer Nelson who not only was a fantastic RA supervisor, but also a great mentor and role-model. I am immensely grateful to have worked with Jen during most of my PhD. I also want to thank Ruth Etzioni, Lurdes Inoue, Scott Emerson, and Coley Yates for providing key guidance and training during my years at UW. I am also very thankful for all the care and support Gitana offered me during this long journey. Finally, I want to thank Alex Mackenzie for helping me arrive to this program.

I don't have enough words to thank my wife Sofi who has always been extremely supportive, and has brought so much love, and happiness to my life. I've learned so much from you, and I hope we keep growing as we share our journey together. I look forward to the next chapter in our lifes. Te amo, $<3^{\infty}$.

I thank my parents Angela and Alfredo for always being there for me. You two are fantastic role models, I learn so much from you every day. I am also very grateful for all the love and support I received from my dear family. Regina, your talent and creativity is a source of inspiration. Alfredo, your dedication and determination are attributes I will always aspire to. Rodrigo, your brotherly company, insights, and approach to life have made me grow so much as a person. Comadre Mónica, thank you for strengthening the roots of our family and for always being so considerate of others, including me. Mateo, thank you for all your smiles and hugs, you will always make my day. I hope to one day be one of your role

models. Dani, thank you for being a fantastic cuñada and for opening the doors to my new family. Dear Roberto, thank you for all your kind advice and support, and all the fun times we've spent during our chess evenings. Your friendship means so much to me. Finally, I thank my grandparents mamá Julie and papá Chuy for all their kindness and love they gave me since I was a tiny ϵ.

I am thankful for all the fantastic people I met at UW, and who made my experience outside of school much more enjoyable: Nathan, Tyler, Scott, Tracy, David, Austin, Arash, Lucy, Jean, Travis, Mareldi, Zack, Kelsey, Phuong, Arjun, Jon, Brian W, Adam, Natalie, Susana, Diego, Sunny, Kevin, Antonio, Brian K, and Itzue. I also want to thank my dear friends from home, Gonzalo (and his marzipans) and Rodrigo.

DEDICATION

To my wife Sofía, and my parents Angela & Alfredo.

"We are because we *are* seen; we *are* because we are loved. The world *is* because it is beheld and loved into being." R.M. Rilke

Chapter 1
INTRODUCTION

This dissertation consists of three projects. In the first project, we describe how augmentation of the propensity score model can be used to improve the efficiency of a matching estimator. In the next project, we propose and examine the asymptotic properties of isotonic calibration in the context of conditional average treatment effect estimation. In the final part of this dissertation, we assess the predictive accuracy of a SuperLearner implementation for predicting HIV risk based on different combinations of baseline covariates and varying risk time windows. In this chapter, we provide an overview of the contributions for each project.

1.0.1 Enhancing a matching estimator via propensity score augmentation

In causal inference, matching is a popular method used to estimate the causal effect of a binary exposure in the presence of confounding. To adjust for confounding, matching compares individuals with similar probability of treatment —known as the propensity score (PS). In practice, the PS is often estimated via logistic regression. Even when the parametric model is correctly specified, matching results in an inefficient estimator Abadie and Imbens (2006). However, estimating the propensity score yields an efficiency gain in the asymptotic variance of the estimator Abadie and Imbens (2016).

In Chapter 2, we propose to mitigate the inefficiency of the estimator by maximizing the efficiency gain through an augmentation of the PS model. We derive an optimal covariate that maximizes the gain in efficiency. Since the optimal augmentation covariate depends on unknown functions of the predictors, to perform this augmentation, we propose estimating it using flexible data-adaptive methods. Additionally, we derive the asymptotic distribution of our proposed method which accounts for estimation of the unknown optimal covariate. Then,

we assess the performance of the proposed method through simulation studies. Finally, we apply the augmented PS matching estimator to estimate the causal effect of circumcision on HIV-1 infection incidence.

1.0.2 Calibration of conditional average treatment effect estimators

In the health sciences, conditional average treatment effects (CATEs) are of high importance to understanding how certain treatments or interventions affect the health of sub-populations Dahabreh et al. (2016), Lee et al. (2020). However, CATE estimation can be challenging for many reasons, including that the nuisance parameters involved in its estimation can be highly non-smooth, high-dimensional, or otherwise difficult to model correctly. The implications of a biased method can be profound if the estimated CATE is used as a support tool to guide therapy (Van Calster et al. 2019). Therefore, it is desirable that a method produces unbiased estimates across all stratums defined by its predicted values. In prediction settings, this property is commonly known as *calibration*.

In Chapter 3, we propose a method that improves a given CATE estimator guaranteeing it is well calibrated while preserving its predictive power. In particular, we provide rates of convergence of our method's calibration performance, and show that its mean squared error is no larger than that of the initial CATE estimator plus an additional term that converges to 0. Additionally, we examine the performance of our method via a suite of simulation studies that consider a wide scope of scenarios, ranging from relatively simple CATE structures to more complex scenarios where the CATE is challenging to estimate.

1.0.3 Implementation and assessment of the performance of a SuperLearner in HIV risk prediction

Risk assessment tools are often used in HIV preventive trials to: improve recruitment strategies, understand site heterogeneity, reduce loss to follow-up bias, and improve precision. Current HIV-1 risk scores reported in the literature have been fitted via step-wise Cox Proportional Hazards (Cox-PH) models Wand et al. (2018), Balkus et al. (2016). Moreover, the

impact of adding or removing covariates when implementing the risk score models has not been studied. The latter is particularly relevant to inform variable collection in future HIV-1 prevention trials.

In Chapter 4, we implement an ensemble learning algorithm known as the SuperLearner to predict HIV-1 risk. Using the AUC as a performance metric, we assess the predictive accuracy of the SuperLearner when estimating HIV-1 risk across different time points. Finally, to provide insights on covariate collection in future HIV-1 prevention trials, we assess the impact of removing specific baseline variables when fitting the SuperLearner.

Chapter 2

PROPENSITY SCORE AUGMENTATION IN MATCHING-BASED ESTIMATION OF CAUSAL EFFECTS

2.1 Introduction

In many studies, the goal is to estimate the causal effect of a binary exposure, say treatment versus control, on an outcome of interest. In randomized trials, since by design treatment level is independent of patient characteristics, treatment effects can often be inferred through simple, unadjusted comparisons. For example, the average treatment effect can be estimated by the difference in mean observed outcomes across treatment levels. In contrast, in observational studies, patients receiving different treatment levels may also differ in baseline characteristics, thereby giving rise to potential confounding of the treatment-outcome relationship. In order to infer about treatment effects, a lack of balance in patient characteristics across treatment levels must be addressed using appropriate methods.

Matching methods are commonly used to infer causal effects using data from observational studies. Such methods outline how to turn the original dataset into a matched dataset in which the distribution of confounders is appropriately balanced across treatment groups. Often, this is accomplished by pairing each patient to one (or many) patients with similar confounder values in the opposite treatment arm. Once a matched dataset has been constructed, treatment effects can be estimated similarly as they would had the data been generated from a randomized trial. Since the conditional treatment probability given baseline covariates — referred to as the propensity score — is a balancing score, in that treatment level and baseline covariates are independent conditionally upon the propensity score value, matching on the propensity score suffices to balance the distribution of confounders between treatment groups (Rosenbaum and Rubin 1983). Propensity score matching has been popular in

practice because it provides a principled means of dimension reduction, allowing pairing of patients with dissimilar covariate values that may nevertheless have similar propensity score values. There are many ways to construct a matched set based on propensity score values — Austin (2011) provides a comprehensive review. Below, we focus on one to several (1:M) propensity score matching with replacement aimed at targeting the average treatment effect; throughout this chapter, this scheme is what we refer to as 1:M matching.

Matching estimators are appealing because of their simplicity, the ease with which their output can be visualized and communicated, and the wide availability of software (Stuart 2010). However, such estimators are generally inefficient, failing to achieve the nonparametric efficiency bound for the average treatment effect; this fact was formally shown in Abadie and Imbens (2016), who derived the large-sample properties of the 1:M matching estimator. Since there exist several alternative methods for estimating the average treatment effect that in fact do attain this bound (Robins et al. 1994, van der Laan and Rubin 2006), this limitation may discourage some practitioners from using the 1:M matching estimator. In this work, we address this limitation by showing that the inefficiency of this estimator can typically be reduced — sometimes substantially — by augmenting the propensity score model using a carefully-constructed synthetic covariate. Our proposal is a natural extrapolation of the fact that, even when the propensity score is known, the asymptotic variance of the matching estimator can be reduced by instead estimating the propensity score, as shown by Abadie and Imbens (2016). Here, we characterize the maximal efficiency gain possible through propensity score model augmentation, and show that this gain can be achieved with a one-dimensional augmentation. The augmented 1:M estimator we propose specifically leverages information from available prognostic variables, baseline covariates unrelated to treatment level but predictive of the outcome, which are often ignored by traditional matching methods since they do not feature in the propensity score model. Finally, we also show that our results suggest that certain implementations of the 1:M matching estimator may in fact be (nearly) nonparametric efficient.

This Chapter is structured as follows. In Section 2.2, we review existing results on the

large-sample behavior of the 1:M matching estimator. In Section 2.3, we study propensity score model augmentation, establish the maximal efficiency gain possible through augmentation, and use these results to propose a novel augmented 1:M matching estimator of the average treatment effect. We also establish the large-sample theory of our proposal. In Section 2.4, we show results from numerical studies conduct to illustrate potential efficiency gains as well as the behavior of our proposed estimator. Then, in Section 2.5, we illustrate the use of our method through an application in which we estimate the causal effect of circumcision on HIV infection using data from a HIV vaccine trial. We conclude with a discussion of our findings in Section 2.6. All proofs of technical results are provided in the Supplementary Material.

2.2 Review of results on the 1:M matching estimator

2.2.1 Statistical setup

Suppose that the available data consist of independent observations $O_1, O_2, \ldots, O_n$ from an unknown distribution P_0. Here, the data unit $O_i := (V_i, A_i, Y_i)$ represents the information gathered on patient i, with $V_i \in \mathcal{V} \subseteq \mathbb{R}^p$ a vector of baseline covariates, $A_i \in \{0, 1\}$ the treatment level received, and $Y_i \in \mathcal{Y} \subseteq \mathbb{R}$ the outcome of interest. For convenience, we also define $W_i := (1, V_i^\top)^\top \in \mathcal{W} \subseteq \mathbb{R}^{p+1}$. For each covariate level $w \in \mathcal{W}$, we denote the propensity score evaluated at w by $\pi_0(w) := P_0(A = 1 \,|\, W = w)$. Throughout this article, we assume nonparametric models for the conditional distribution of Y given (A, W) and of the marginal distribution of W. Furthermore, we assume that the true propensity score follows a logistic regression model, namely that $\pi_0 = \pi_{\theta_0}$ for some $\theta_0 \in \mathbb{R}^{p+1}$, where for each $\theta \in \mathbb{R}^{p+1}$ we define $\pi_\theta : w \mapsto \text{expit}(\theta^\top w)$. We will denote by $\mathcal{M}_0$ this propensity score model. Relaxation of this modeling assumption is discussed in Section 2.6.

The causal estimand we focus on in this chapter is the average treatment effect. To define it precisely, we make use of the potential outcomes framework, which allows us to define treatment effects (Rubin 1974). For $a \in \{0, 1\}$, we denote by $Y(a)$ the potential outcome

under treatment level a. Then, the average treatment effect is given by $\mathbb{E}[Y(1) - Y(0)]$, where $\mathbb{E}$ denotes expectation of the true joint distribution of counterfactual outcomes. We denote the outcome regression and propensity-reduced outcome regression, respectively, as $\mu_0(a, w) := E_0(Y \mid A = a, W = w)$ and $\bar{\mu}_0(a, p) := E_0\{Y \mid A = a, \pi_0(W) = p\}$. Here and below, we use the notation E_0, var_0 and cov_0 to denote, respectively, an expectation, variance and covariance computed under the true sampling distribution P_0. Under standard causal conditions, including:

1. (no interference) any patient's treatment level is independent of any other patient's outcome;

2. (consistency) the observed outcome is $Y = AY(1) + (1 - A)Y(0)$;

3. (positivity) $\pi_0(W) \in (0, 1)$ P_0-almost surely;

4. (exchangeability) $Y(a)$ and A are independent given W with P_0-almost surely for each $a \in \{0, 1\}$;

the average treatment effect can be identified as

$$\psi_0 := E_0\{\mu_0(1, W) - \mu_0(0, W)\} = E_0\{\bar{\mu}_0(1, \pi_0(W)) - \bar{\mu}_0(0, \pi_0(W))\},$$

a summary of the distribution P_0 of the observed data unit O.

2.2.2 Definition of the 1:M matching estimator

The identification formula above motivates matching on the propensity score instead of the entire covariate vector. If π is the propensity score used to match observations, the π-specific 1:M matching estimator of the average treatment effect can be written as

$$\psi_n(\pi) := \frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)\left[Y_i - \frac{1}{M}\sum_{j=1}^{n} I\{j \in J_M(i, \pi)\}\, Y_j\right],$$

where we denote by $J_M(i,\pi)$ the matched set for the i^{th} observation based on the propensity score π

$$\{j : A_j = 1 - A_i, \textstyle\sum_{k=1}^{n} I\{A_k = 1 - A_i\} I\left\{|\pi(W_j) - \pi(W_i)| \leq |\pi(W_k) - \pi(W_i)|\right\} \leq M\}.$$

Writing the number $K_{M,\pi}(i) := \sum_{j=1}^{n} I\{i \in J_M(j,\pi)\}$ of times observation i occurs as a match when propensity score π is used, the π-specific matching estimator can also be represented as

$$\psi_n(\pi) = \frac{1}{n}\sum_{i=1}^{n}(2A_i - 1)\left\{1 + \frac{K_{M,\pi}(i)}{M}\right\} Y_i \ .$$

Since the true index value θ_0 of the propensity score π_0 in the logistic regression model $\mathcal{M}_0$ is typically unknown, an estimator of θ_0 must be used instead. Suppose that $\theta_n := \text{argmax}_\theta \, \ell_n(\theta)$ is the maximum likelihood estimator of θ_0, where $\ell_n(\theta) := \sum_{i=1}^{n} [A_i \log \pi_\theta(W_i)) + (1 - A_i) \log$ is the log-likelihood of θ. Then, the resulting 1:M matching estimator available for use is $\psi_n := \psi_n(\pi_{\theta_n})$.

2.2.3 Large-sample behavior of the 1:M matching estimator

The large-sample properties of ψ_n and its oracle counterpart $\psi_n(\pi_0)$ were carefully established in Abadie and Imbens (2016) and Abadie and Imbens (2006), respectively. Here, we summarize two of their key findings, which our proposal heavily builds upon.

The first finding we focus on provides an asymptotic distributional result for the π_0-specific 1:M matching estimator. Specifically, under some regularity conditions, Abadie and Imbens (2006) showed that $n^{1/2}\{\psi_n(\pi_0) - \psi_0\}$ tends to a normal random variable with mean zero and variance $\sigma_M^2 := \sigma_1^2 + \sigma_{2,M}^2$, where we define

$$\begin{aligned}
\sigma_1^2 &:= E_0\left[\frac{\bar{\sigma}^2(1,\pi_0(W))}{\pi_0(W)} + \frac{\bar{\sigma}^2(0,\pi_0(W))}{1-\pi_0(W)} + \{\bar{\mu}_0(1,\pi_0(W)) - \bar{\mu}_0(0,\pi_0(W)) - \psi_0\}^2\right], \\
\sigma_{2,M}^2 &:= \frac{1}{2M} E_0\left[\bar{\sigma}^2(1,\pi_0(W))\left\{\frac{1}{\pi_0(W)} - \pi_0(W)\right\} + \bar{\sigma}^2(0,\pi_0(W))\left\{\frac{1}{1-\pi_0(W)} - 1 + \pi_0(W)\right\}\right],
\end{aligned}$$

and for each $a \in \{0,1\}$ and $w \in \mathcal{W}$, $\bar{\sigma}^2(a,\pi_0(w)) := var_0\{Y \mid A = a, \pi_0(W) = \pi_0(w)\}$ denotes the conditional variance of Y given $A = a$ and $\pi_0(W) = \pi_0(w)$ computed under P_0. In

particular, this result highlights that, as expected, larger values of M render $\psi_n(\pi_0)$ relatively less variable in large samples, but that, irrespective of M, $\psi_n(\pi_0)$ is not nonparametric efficient except in trivial cases.

The second finding central to the developments in this chapter describes the extension of the above distributional results to the 1:M matching estimator based on an estimated propensity score index. For technical reasons pertaining to a certain non-smoothness in the behavior of matching estimators (Andreou and Werker 2012), the estimator studied in Abadie and Imbens (2016) is based not on the maximum likelihood estimator θ_n but rather on a discretized version $\theta_{n,k}$ of θ_n, defined as $\theta_{n,k} := \lfloor kn^{1/2}\theta_n \rceil/(kn^{1/2})$, where k is a user-specified discretization constant and $\lfloor\cdot\rceil : \mathbb{R}^{p+1} \to \mathbb{Z}^{p+1}$ is the component-wise nearest-integer function. If either k or n is large, $\theta_{n,k}$ is close to θ_n in each sample. Under regularity conditions, they established that

$$\lim_{k\downarrow 0} \lim_{n\to\infty} P_0 \left[n^{1/2} \left\{ \psi_n(\pi_{\theta_{n,k}}) - \psi_0 \right\} / \sigma_{M,*} \leq z \right] = \Phi(z) \tag{2.1}$$

for each $z \in \mathbb{R}$, where we define the asymptotic variance $\sigma^2_{M,*} := \sigma^2_M - c(\theta_0)^\top \mathcal{I}(\theta_0)^{-1} c(\theta_0)$ and the $\mathbb{R}^{p+1}$-valued vector

$$\begin{aligned} c(\theta) \;:=\; & E_0 \left[\pi_0(W)\, cov_0 \left\{ W, \mu_0(0,W) \,|\, A=0, \pi_0(W) \right\} \right] \\ & + E_0 \left[\{1-\pi_0(W)\}\, cov_0 \left\{ W, \mu_0(1,W) \,|\, A=1, \pi_0(W) \right\} \right]. \end{aligned}$$

Here, $\mathcal{I}(\theta_0) := E_0 \left[\pi_0(W)\{1-\pi_0(W)\} W W^\top \right]$ is the Fisher information for θ at $\theta = \theta_0$ in the logistic regression model, and it is assumed to be invertible. Furthermore, Φ denotes the standard normal distribution function. This result suggests that the 1:M matching estimator $\psi_n(\pi_{\theta_{n,k}})$ based on a (discretization-regularized) estimator of the propensity score is expected to be more efficient than the θ_0-specific 1:M matching estimator $\psi_n(\pi_0)$, with a reduction in asymptotic variance given by $c(\theta_0)^\top \mathcal{I}(\theta_0)^{-1} c(\theta_0) \geq 0$. Despite this fact, as was the case for $\psi_n(\pi_0)$, $\psi_n(\pi_{\theta_{n,k}})$ also typically does not achieve the nonparametric efficiency bound.

2.3 Propensity score augmentation for the 1:M matching estimator

2.3.1 Using model augmentation to gain efficiency

Given that estimation of the propensity score increased the asymptotic efficiency of the 1:M matching estimator relative to the (oracle) 1:M matching estimator based on the true propensity score, it is natural to wonder whether further gains are possible by augmentation of the propensity score model. To investigate this question, we first make use of the second result of Abadie and Imbens (2016) stated above to characterize potential gains resulting from the use of a given augmentation of the propensity score model.

Suppose that $q := (q_1, q_2, \ldots, q_m)^\top$ is an arbitrary m-dimensional augmentation function, with $q_j : \mathcal{W} \to \mathbb{R}$ satisfying that $var_0\{q_j(W)\} < \infty$ for each $j = 1, 2, \ldots, m$. We consider the 1:M matching estimator based on a propensity score estimated using the augmented logistic regression model $\mathcal{M}(q) := \{\pi_{q,\theta,\gamma} : \theta \in \mathbb{R}^{p+1}, \gamma \in \mathbb{R}^m\}$, where we define

$$\pi_{q,\theta,\gamma} : w \mapsto \text{expit}\left\{\theta^\top w + \gamma^\top q(w)\right\} .$$

This corresponds to fitting a regression model not only with covariate vector W but also a synthetic covariate vector $q(W)$ created by transformation of W. For any choice of q, the propensity score model $\mathcal{M}(q)$ is correctly specified since π_0 is recovered by setting $\vartheta := (\theta, \gamma)$ equal to $\vartheta_0 := (\theta_0, 0)$, and indeed, $\pi_{q,\vartheta_0} = \pi_0$ irrespective of q. Among regularity conditions needed for identification in this augmented model, the Fisher information for ϑ in $\mathcal{M}(q)$ at $\vartheta = \vartheta_0$, given by

$$\mathcal{I}_q(\theta_0) := E_0\left[\pi_0(W)\{1 - \pi_0(W)\}\begin{bmatrix} WW^\top & Wq(W)^\top \\ q(W)W^\top & q(W)q(W)^\top \end{bmatrix}\right],$$

is assumed to be invertible. We denote by $\vartheta_{q,n} := \text{argmax}_\vartheta\, \ell_{q,n}(\vartheta)$ the maximum likelihood estimator of ϑ_0, where $\ell_{q,n}(\vartheta) := \sum_{i=1}^n [A_i \log \pi_{q,\vartheta}(W_i)) + (1 - A_i) \log\{1 - \pi_{q,\vartheta}(W_i)\}]$ is the log-likelihood of ϑ based on model $\mathcal{M}(q)$ for the propensity score. Using once more the results of Abadie and Imbens (2016), we find that the (sample size-scaled) asymptotic variance of the 1:M matching estimator $\psi_{n,q,k} := \psi_n(\pi_{q,\vartheta_{q,n,k}})$ using an estimated propensity score based

on model $\mathcal{M}(q)$ and a discretized regularization is given by $\sigma^2_{q,M} := \sigma^2_M - \text{gain}(q)$, where for convenience we define $\text{gain}(q) := c_q(\theta_0)^\top \mathcal{I}_q(\theta_0)^{-1} c_q(\theta_0)$ and write $c_q(\theta_0) = (c(\theta_0)^\top, \bar{c}_q(\theta_0)^\top)^\top$ with

$$\begin{aligned}\bar{c}_q(\theta_0) \;:=\; & E_0\left[\pi_0(W)\, cov_0\left\{q(W), \mu_0(0,W) \,|\, A=0, \pi_0(W)\right\}\right] \\ & + E_0\left[\{1-\pi_0(W)\}\, cov_0\left\{q(W), \mu_0(1,W) \,|\, A=1, \pi_0(W)\right\}\right].\end{aligned}$$

This implies, in particular, that there is no loss but only potential gains in terms of asymptotic variance from augmenting the propensity score model with some covariate vector $q(W)$.

2.3.2 Maximizing the efficiency gain

Given the potential efficiency gains resulting from augmentation by any covariate vector $q(W)$, we consider constructing a 1:M matching estimator with a propensity score estimated using an augmented model $\mathcal{M}(q)$. However, as the expression for $\text{gain}(q)$ implies, the efficiency gain possible depends on the choice of q. Thus, many candidates for q could be considered to increase or even perhaps maximize the efficiency gain. In the following theorem, we display an augmented propensity score model through which the efficiency gain is maximized. Before stating our formal result, we define the transformation $h_0 : \mathcal{W} \to \mathbb{R}$ pointwise as

$$h_0(w) := \frac{\mu_0(1,w) - \bar{\mu}_0(1, \pi_0(w))}{\pi_0(w)} + \frac{\mu_0(0,w) - \bar{\mu}_0(0, \pi_0(w))}{1-\pi_0(w)}\ .$$

We also denote by $\mathcal{Q}_0$ the set obtaining by collecting, for each $m \in \{1, 2, \ldots\}$, each augmentation function $q : \mathcal{W} \to \mathbb{R}^m$ such that $E_0\{q_j(W)^2\} < \infty$ for each $j = 1, 2, \ldots, m$ and $\mathcal{I}_q(\theta_0)$ is invertible. The following result describes the efficiency gain resulting from using

$$\mathcal{M}(h_0) = \left\{w \mapsto \text{expit}\left\{\theta^\top w + \gamma h_0(w)\right\} : \theta \in \mathbb{R}^{p+1}, \gamma \in \mathbb{R}\right\},$$

the propensity score model augmented with the synthetic covariate $h_0(W)$.

Theorem 2.1. *The following statements hold true:*

(a) $\text{gain}(h_0) = \sup_{q \in \mathcal{Q}_0} \text{gain}(q)$;

(b) $\text{gain}(h_0) = E_0\left[\pi_0(W)\{1-\pi_0(W)\}\{h_0(W)\}^2\right]$.

The implications of this result are significant since it indicates that a one-dimensional augmentation covariate suffices to achieve maximal efficiency gain among all possible augmented propensity score models. Moreover, it indicates that if $\mu_0(a,w) = \bar{\mu}_0(a,\pi_0(w))$ for P_0-almost every (a,w) — this occurs, for example, if W is one-dimensional — then there is no efficiency to be gained, and in fact, the 1:M matching estimator has the same asymptotic variance irrespective of whether the true propensity score is used or instead estimated. The optimal gain depends on the unknown distribution P_0 — we provide some insights on when larger gains can be expected in Section 2.4.2.

The following corollary provides a comparison between the asymptotic variance of the augmentation-based 1:M matching estimator and the nonparametric efficiency bound

$$\sigma^2_{NP} := E_0\left[\frac{\sigma^2(1,W)}{\pi_0(W)} + \frac{\sigma^2(0,W)}{1-\pi_0(W)} + \{\mu_0(1,W) + \mu_0(0,W) - \psi_0\}^2\right],$$

where $\sigma^2(a,w) := var_0\left(Y \,|\, A=a, W=w\right)$ is the conditional variance of Y given $(A,W) = (a,w)$.

Corollary 2.1. *The difference* $\delta_M := \sigma^2_M - \text{gain}(h_0) - \sigma^2_{NP}$ *between the asymptotic variance of the optimally-augmented 1:M matching estimator and the nonparametric efficiency bound is given by*

$$\delta_M \;=\; \frac{1}{2M}\,E_0\left[\left\{\frac{1}{\pi_0(W)} - \pi_0(W)\right\}\left\{\sigma^2(1,W) + \zeta(1,W)\right\} \right.$$
$$\left. + \left\{\frac{1}{1-\pi_0(W)} - 1 + \pi_0(W)\right\}\left\{\sigma^2(0,W) + \zeta(0,W)\right\}\right] \;\geq\; 0\;,$$

where we define $\zeta(a,w) := var_0\left\{\mu_0(a,W) \,|\, \pi_0(W) = \pi_0(w)\right\}$.

In particular, this result indicates that the asymptotic inefficiency of the optimally-augmented 1:M matching estimator — the amount by which its (sample size-scaled) asymptotic variance exceeds the nonparametric efficiency bound — can be made arbitrarily small by choosing a large value of M.

2.3.3 Estimating the optimal augmentation covariate

In practice, since the optimal augmentation function h_0 is unknown, it must be estimated. In this section, we establish that, under certain regularity conditions, the augmented 1:M matching estimators using an estimator of h_0 is asymptotically equivalent to the corresponding estimator using h_0 itself. We show this in two steps. First, in Proposition 2.1, we consider the behavior of the 1:M matching estimator resulting from use of the augmented propensity score model $\mathcal{M}(h_{0n})$, where $h_{01}, h_{02}, \ldots$ is a sequence of deterministic functions such that $\int \{h_{0n}(w) - h_0(w)\}^2 \, dP_{W,0}(w)$ tends to zero. Here, $P_{W,0}$ denotes the marginal distribution of W implied by P_0. This result allows accounting for the fact that, in applications, a proxy of h_0 would be used instead of h_0. It does not however account for the randomness of this proxy. Thus, in Corollary 2.2, we show that the results of Proposition 2.1 still hold when the proxy used for h_0 is a random function h_n constructed on an independent sample such that $\int \{h_n(w) - h_0(w)\}^2 \, dP_{W,0}(w)$ P_0-almost surely.

Our theoretical results establishing the asymptotic normality of our proposed matching estimator require several regularity conditions, stated in the Supplementary Material. An interpretation of these technical conditions is also provided in the Supplementary Material. We employ the strategy of Abadie and Imbens (2016) to derive large-sample distributional results for the 1:M matching estimator based on a discretization-regularized estimate using an augmented propensity score model. As before, given a discretization constant $k > 0$, we define a partition of cubes with sides of length $(kn^{1/2})^{-1}$ in $\mathbb{R}^{p+2}$. For a given augmentation function q, the discretized estimator $\vartheta_{q,n,k}$ is set to equal the midpoint of the cube to which $\vartheta_{q,n}$ belongs. In our notation, the discretized estimator is thus defined as $\vartheta_{q,n,k} := \lfloor kn^{1/2}\vartheta_{q,n} \rceil / (kn^{1/2})$, where, similarly as before, $s : \lfloor \cdot \rceil : \mathbb{R}^{p+2} \to \mathbb{Z}^{p+2}$ is the componentwise nearest-integer function. The 1:M matching estimator based on a discretization-regularized estimate in the augmented propensity score model with augmentation functions h_{0n} and h_n, respectively, are then given by $\psi_{n,h_{0n},k} := \psi_n(\pi_{h_{0n},\vartheta_{h_{0n},n,k}})$ and $\psi_{n,h_n,k} := \psi_n(\pi_{h_n,\vartheta_{h_n,n,k}})$.

Given the sequence $h_{01}, h_{02}, \ldots$ of deterministic functions approximating the optimal

augmentation function h_0, the following proposition states that $n^{1/2}\left(\psi_{h_{0n},n,k} - \psi\right)$ tends to a mean-zero normal distribution with oracle variance $\sigma^2_{M,\text{opt}} := \sigma^2_M - \text{gain}(h_0)$.

Proposition 2.1. *Suppose that Conditions 1–9 hold. Then, for each $z \in \mathbb{R}$, it holds that*

$$\lim_{k\downarrow 0} \lim_{n\to\infty} P_0\left[n^{1/2}\left(\psi_{n,h_{0n},k} - \psi_0\right)/\sigma_{M,\text{opt}} \leq z\right] = \Phi(z) \ .$$

The above result gives a large-sample distributional approximation relevant to the matching estimator based on the discretization-regularized propensity score estimator involving augmentation by h_{0n}. While the involved probability sequence is based on sampling from the true distribution P_0 at each sample size n, a stronger result wherein the probability sequence is based on sampling from certain local perturbations of P_0 can also be proven, and is established in the Supplementary Material. This strengthened result ascribes a certain level of regularity to the matching estimator studied, which is particularly relevant given the irregularity that may perhaps be expected given how highly non-smooth the matching operation is. A key distinction between this result and the main result of Abadie and Imbens (2016) is that the propensity score model used here incorporates a covariate — in our case, an augmentation covariate — that is changing with n but eventually settles down. Nevertheless, in practice, it would be the case both that h_0 is unknown and that a deterministic approximating sequence h_{0n} would not be available, limiting the applicability of the result. The next result overcomes this limitation by allowing use of a random approximating sequence h_n in the propensity score augmentation step.

To allow consideration of a random estimator h_n of the optimal augmentation function h_0, we require h_n to be constructed on a different dataset than the resulting (discretized) maximum likelihood estimator $\vartheta_{h_n,n,k}$ of $\vartheta_{h_n,0}$ in model $\mathcal{M}(h_n)$ and matching estimator $\psi_{n,h_n,k}$. In practice, this implies that the original dataset must be partitioned into two subsets: subset A_n used for estimation of h_0, say of size $1 < m_n < n$, and subset B_n used for the remainder of the matching estimation procedure, of size $n_{\text{eff}} := n - m_n$, with m_n tending to infinity with n. We are then able to prove that the same large-sample distributional result holds for the 1:M matching estimator based on a discretization-based estimate of the

propensity score model augmented with an estimator of the optimal augmentation function as with the (oracle) optimal augmentation function itself, except for the loss of effective sample size resulting for the need to estimate h_0 on a separate sample.

Corollary 2.2. *Suppose that Conditions 1–9 hold, and that* $\int \{h_n(w) - h_0(w)\}^2 dP_{W,0}(w) \to 0$ P_0*-almost surely. Then, for each* $z \in \mathbb{R}$*, it holds that*

$$\lim_{k \downarrow 0} \lim_{n \to \infty} P_0 \left[n_{\text{eff}}^{1/2} \left(\psi_{n,h_n,k} - \psi_0 \right) / \sigma_{M,\text{opt}} \leq z \,\middle|\, A_n \right] = \Phi(z)$$

P_0*-almost surely.*

To estimate the optimal augmentation function h_0, we must first estimate the outcome regression μ_0 and the propensity-reduced outcome regression $\bar{\mu}_0$. Flexible regression techniques can be used for this purpose. The nuisance functions μ_0 and $\bar{\mu}_0$ can naturally be estimated in sequence. First, an estimate of $\mu_0(a, w)$ can be obtained by using any regression technique for learning the mean of outcome Y given $W = w$ within the subset of observations with $A = a$. Regression tools used can include parametric, semiparametric, or nonparametric approaches, or a combination thereof pooled via an ensembling strategy. Then, based on the fact that

$$\bar{\mu}_0(a, p) = E_0 \left\{ Y \,|\, A = a, \pi_0(W) = \pi_0(w) \right\} = E_0 \left\{ \mu_0(a, W) \,|\, A = a, \pi_0(W) = \pi_0(w) \right\},$$

an estimate of $\bar{\mu}_0(a, p)$ can be obtained using a mean regression of outcome $\mu_n(A, W)$ on $\pi_n(W)$ within the subset of observations with $A = a$, where π_n is a model-based estimator of the propensity score. Here again, a variety of regression tools could be considered. This nested regression strategy is more likely to result in estimates of $\mu_0(a, w)$ and $\bar{\mu}_0(a, \pi_0(w))$ that are compatible with each other, and hence may be preferable in finite samples compared to the use of separate regressions to estimate these nuisance functions. We summarize the steps involved in constructing the final matching estimator in Algorithm 1.

In practice, the size m_n of A_n can be chosen to be relatively small compared to n_{eff} since, while h_n must be consistent in a suitable sense, there is no minimal convergence rate

1: **partition data into two subsets A and B of size n_A and n_B, respectively;**

2: **using subset A_n of data, obtain h_n as follows:**

3: compute MLE θ_n of θ_0 in $\mathcal{M}_0$ by performing logistic regression of A on W, and set $\pi_n := \pi_{\theta_n}$;

4: for $a \in \{0, 1\}$, obtain $\mu_n(a, \cdot)$ by regressing Y on W using observations with $A = a$;

5: for $a \in \{0, 1\}$, obtain $\bar{\mu}_n(a, \cdot)$ by regressing $\mu_n(A, W)$ on $\pi_n(W)$ using observations with $A = a$.

6: define

$h_n : w \mapsto \{\mu_n(1, w) - \bar{\mu}_n(1, \pi_n(w)\}/\pi_n(w) + \{\mu_n(0, w) - \bar{\mu}_n(0, \pi_n(w)\}/\{1 - \pi_n(w)\}$;

7: **using subset B_n of data, compute $\psi_{n,h_n,k}$ as follows:**

8: compute MLE ϑ_n of ϑ_0 in $\mathcal{M}(h_n)$ by performing logistic regression of A on $(W, h_n(W))$;

9: compute 1:M matching estimator $\psi_{n,h_n,k}$ based on matching using $\pi_{h_n,\vartheta_{h_n,n,k}}$.

Algorithm 1: Optimally-augmented propensity score 1:M matching with sample-splitting.

required. In our simulation studies, we selected $m_n = 0.05 \times n$ to minimize the variance inflation due to sample-splitting. We note that relatively poor estimation of h_0, due, for example, to the use of insufficiently flexible regression techniques or an overly small value of m_n, is not expected to be catastrophic since any augmentation of the propensity score model is expected to possibly help but never hurt the performance of the resulting matching estimator.

2.4 Numerical studies

2.4.1 Motivating questions

The theory provided in this chapter suggests that efficiency gains are possible via propensity score model augmentation, and that a regularization of the resulting 1:M matching estimator has desirable large-sample properties. Nevertheless, there are critical questions that are best addressed through numerical studies, including:

1. how substantial can the theoretical efficiency gains derived from augmentation be, and when can meaningful gains be expected?

2. does the proposed matching estimator achieve predicted efficiency gains in practice, and is the normal approximation derived useful to describe the finite-sample behavior of the estimator?

3. does the use of sample-splitting to estimate the optimal augmentation function and calculate the resulting estimator appear critical to the validity of the derived normal approximation?

We begin by investigating the first question in the context of a simple data-generating mechanism for which certain closed-form results can be derived based on our analytic expressions. We then perform simulation studies using a collection of scenarios in order to provide additional insight on the first question and to address the second and third questions as well.

2.4.2 Illustration of predicted gains

The efficiency gain resulting from propensity score model augmentation depends on various aspects of the data-generating mechanism P_0. Based on the form of the optimal augmentation function, we expect that the gain will tend to be higher when the contrast between the outcome regression and propensity-reduced outcome regression is large. This occurs, for

example, when W includes variables that are strongly associated with the outcome but less so with treatment.

We considered a simple data-generating mechanism based upon which to perform some analytic and numeric efficiency gain computations. Specifically, we considered a setting in which $W := (W_1, W_2)$ consists of two independent mean-zero normal random variables with unit variance, A follows a Bernoulli distribution with success probability $\pi_0(w_1, w_2) := \text{expit}\left(\theta_0 + \theta_1 w_1 + \theta_2 w_2\right)$ conditionally upon $W = (w_1, w_2)$, and Y is normal random variable with mean

$$\mu_0(a, w) := a\left(\beta_0 + \beta_1 w_1 + \beta_2 w_2\right) + (1 - a)\left(\gamma_0 + \gamma_1 w_1 + \gamma_2 w_2\right)$$

and variance σ^2 conditionally upon $W = (w_1, w_2)$ and $A = a$. The average treatment effect equals $\beta_0 - \gamma_0$. Furthermore, the propensity-reduced outcome regression $\bar{\mu}_0(a, \pi_0(w))$ can be shown to equal

$$a\left\{\beta_0 + (\beta_1\theta_1 + \beta_2\theta_2)\left(\frac{\theta_1 w_1 + \theta_2 w_2}{\theta_1^2 + \theta_2^2}\right)\right\} + (1 - a)\left\{\gamma_0 + (\gamma_1\theta_1 + \gamma_2\theta_2)\left(\frac{\theta_1 w_1 + \theta_2 w_2}{\theta_1^2 + \theta_2^2}\right)\right\},$$

which implies, for example, that

$$\mu_0(1, w) - \bar{\mu}_0(1, \pi_0(w)) = \frac{\theta_2(\beta_1\theta_2 - \beta_2\theta_1)w_1 + \theta_1(\beta_2\theta_1 - \beta_1\theta_2)w_2}{\theta_1^2 + \theta_2^2}.$$

A similar expression holds for $\mu_0(0, w) - \bar{\mu}_0(0, \pi_0(w))$, replacing β_0, β_1, β_2 by γ_0, γ_1, γ_2, respectively. Suppose that W_2 is a precision variable, in the sense that it has an effect on Y but not on A; in the context of this data-generating mechanism, this is achieved by setting $\theta_2 = 0$ but ensuring that either $\beta_2 \neq 0$ or $\gamma_2 \neq 0$. In this case, the difference in outcome regressions simplify to $\mu_0(1, w) - \bar{\mu}_0(1, \pi_0(w)) = \beta_2 w_2$ and $\mu_0(0, w) - \bar{\mu}_0(0, \pi_0(w)) = \gamma_2 w_2$. For simplicity, we may consider the case in which $\gamma_2 = \beta_2$ so that there is no effect modification by W_2. In this case, we can explicitly calculate the nonparametric efficiency bound to be

$$\sigma^2_{NP} = 2\sigma^2\left\{1 + \exp(\theta_1^2/2)\right\} + (\gamma_1 - \beta_1)^2 \ .$$

The asymptotic variances of the 1:M matching estimator based on unaugmented and optimally augmented propensity score models are, respectively, $\sigma^2_{M,*} = \sigma^2_M - \text{gain}(\varnothing)$ and

$\sigma^2_{M,\text{opt}} = \sigma^2_M - \text{gain}(h_0)$, where we can explicitly compute that

$$\sigma^2_M = \sigma^2_{NP} + 2\beta_2^2 \left\{1 + \exp(\theta_1^2/2)\right\} + \delta_M$$

with $\delta_M := (\sigma^2 + \beta_2^2)\left\{1 + 2\exp(\theta_1^2/2)\right\}/(2M)$, and furthermore, as expected, that

$$\text{gain}(\varnothing) = \frac{\beta_2^2}{E_0\left[\pi_0(W)\{1-\pi_0(W)\}\right]} \;\leq\; \beta_2^2\, E_0\left[\frac{1}{\pi_0(W)\{1-\pi_0(W)\}}\right] = \text{gain}(h_0)\ .$$

Additionally, further simplification yields that $\text{gain}(h_0) = 2\beta_2^2\left\{1 + \exp(\theta_1^2/2)\right\}$. We thus find that the relative efficiency of the unaugmented 1:M matching estimator to its augmented counterpart equals

$$\text{relative efficiency} \;=\; 1 + \bar{\beta}_2^2\left[\frac{2\{1+\exp(\theta_1^2/2)\} - 1/E_0\left[\pi_0(W)\{1-\pi_0(W)\}\right]}{2\{1+\exp(\theta_1^2/2)\} + (\bar{\gamma}_1 - \bar{\beta}_1)^2 + \frac{1}{2M}(1+\bar{\beta}_2^2)\{1+2\exp(\theta_1^2/2)\}}\right],$$

where we have defined the standardized coefficients $\bar{\beta}_1 := \beta_1/\sigma$, $\bar{\beta}_2 := \beta_2/\sigma$ and $\bar{\gamma}_1 := \gamma_1/\sigma$. In Figure 2.4.2, for different choices of M, we display the relative efficiency as a function of θ_1 for $\bar{\beta}_2 = 1$ and as a function of $\bar{\beta}_2$ for $\theta_1 = 1$. In these displays, for simplicity, we have set $\bar{\beta}_1 = \bar{\gamma}_1$, corresponding to the absence of effect modification by W_1. We observe that the relative gains in efficiency resulting from the use of augmentation compared to standard propensity score-based 1:M matching grow with the strength of the associations between Y and W_2 and between A and W_1, and that these gains can be substantial even under moderate associations. If any of these two associations is null, then no gains in efficiency result from augmentation. We also note that, as expected, the relative efficiency of augmentation grows with the number M of matches.

2.4.3 Simulation studies with sample-splitting

In order to verify that the proposed estimation procedure behaves as suggested by our theoretical results, we ran simulations under each of six different scenarios. In each simulation scenario, data were generated as follows. First, a covariate vector $W := (W_1, W_2)$ was generated, where W_1 is a mean-zero normal random variable with unit variance, and W_2 is a standard binary random variable. Then, given $W = (w_1, w_2)$, a Bernoulli random variable

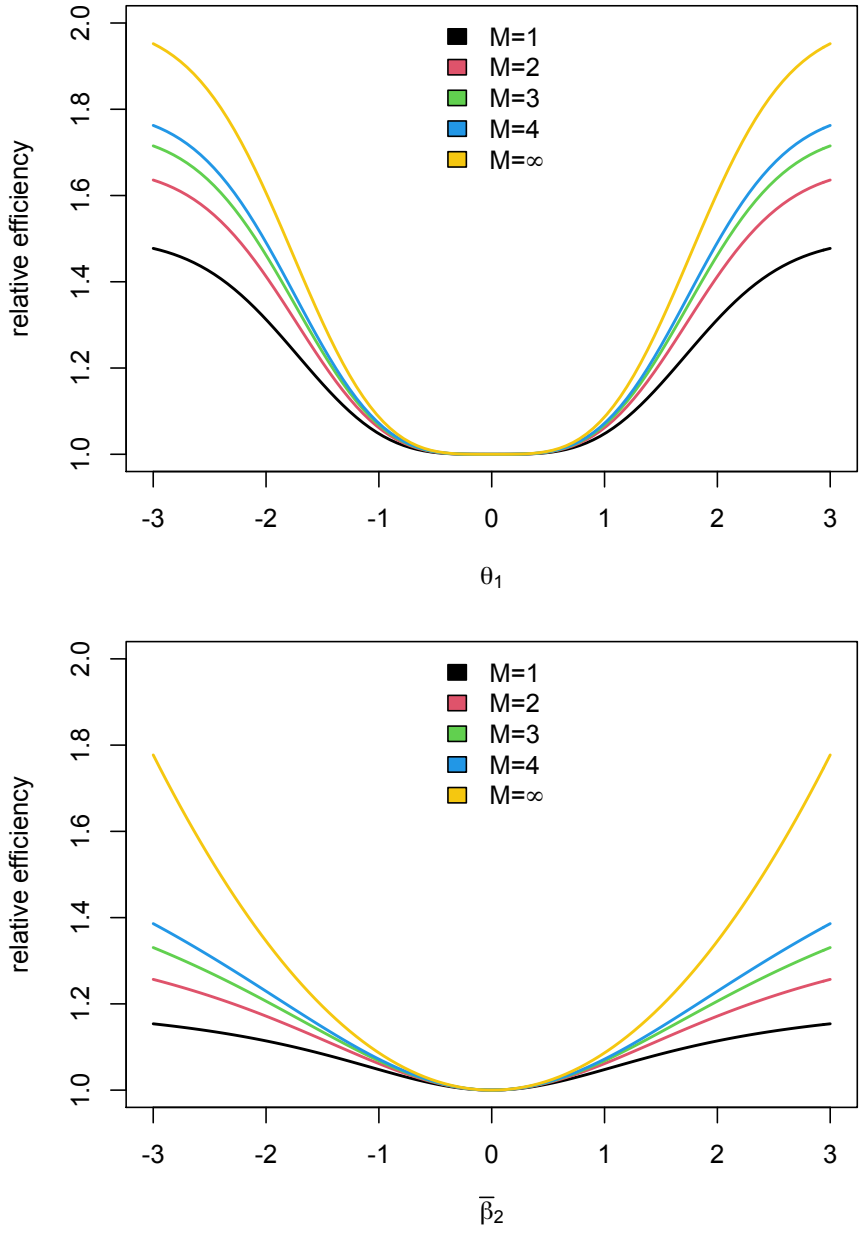

Figure 2.1: Relative efficiency of 1:M matching estimator with optimal propensity augmentation versus unaugmented propensity in analytic example described in Section 4.1. The left plot displays the relative efficiency as a function of θ_1 for $\bar{\beta}_2 = 1$. The right plot displays the relative efficiency as a function of β_2 for $\theta_1 = 1$. In both plots, we have set $\bar{\beta}_1 = \bar{\gamma}_1 = 1$, and show the relative efficiency value for different values of M.

scenario	$\text{logit}\,\pi_0(w)$	$\mu_0(1,w)$	$\mu_0(0,w)$
1	$0.2+1.5w_1+0.1w_2$	$2.2+4.5w_1+5.1w_2$	$2+3w_1+5w_2$
2	$0.2+1.5w_1-0.5w_2$	$1+3w_2$	$5w_2$
3	$0.2+1.5w_1$	$4+w_1+3w_2$	$2-w_1+4w_2$
4	$-1+w_1+2w_2$	$5+w_1+3w_2$	$2w_1+w_2$
5	$0.4+1.5w_1-1.5w_2$	$2+w_1+w_2$	w_1+w_2
6	$1-w_1-w_2$	$4+3w_1+2w_2$	$2+w_1+w_2$

Table 2.1: Description of data-generating mechanism for various simulation scenarios considered.

A with success probability $\pi_0(w_1, w_2)$ was drawn. Finally, given $W = (w_1, w_2)$ and $A = a$, a normal random variable Y with mean $\mu_0(a, w)$ and unit variance was generated. Table 2.1 provides the specific form of $\pi_0(w_1, w_2)$ and $\mu_0(a, w)$ defining each simulation scenario considered. We note that W_1 is an instrument in scenario 2, and W_2 is a precision variable in scenario 3 and is nearly a precision variable in scenario 1. In all other settings, W_1 and W_2 are both confounders. In all settings except for scenario 5, W_1 and W_2 also serve as effect modifiers.

To validate numerically our asymptotic theory, we generated 1,000 random datasets of size 5,000 in each scenario outlined. We implemented a 95/5 sample-splitting scheme. Specifically, we split each dataset randomly into two parts, one with $n_{\text{eff}} = 4,750$ observations and the other with $m_n = 250$. The smaller sample was used to obtain an estimate h_n of the optimal augmentation function h_0, whereas the larger sample was used to compute the 1:1 matching estimator based on propensity score augmentation by h_n. All comparisons between empirical and theoretical variances were based on sample size n_{eff}. To construct h_n, we fitted (correctly-specified) parametric regression models to estimate μ_0 and π_0, and estimated $\bar{\mu}_0$ using the highly adaptive lasso estimator (**?**) with 10-fold cross validation to select the

scenario	efficiency bound	theoretical var. (augm.)	empirical var. (augm.)	empirical var. (unaugm.)
1	11.86	39.52	35.5	41.6
2	9.88	27.96	26.9	27.0
3	14.43	32.76	28.5	35.4
4	9.96	15.03	13.8	13.0
5	11.07	19.27	18.5	14.8
6	12.19	15.91	14.7	14.6

Table 2.2: Theoretical versus empirical variance of the 1:1 matching with augmented propensity score estimator and 95/5 sample-splitting across simulation scenarios at sample size 5,000. The second and third columns provide the variance suggested by theory for a nonparametric efficient estimator and the optimally augmented matching estimator based on sample-splitting. The fourth and fifth columns display the empirical Monte Carlo variance of the augmented (sample-split) matching estimator and the unaugmented matching estimator implemented over 1,000 simulated datasets.

bound on the variation norm.

Results of this simulation are summarized in Table 2.2. Overall, the empirical variance of the augmented 1:1 matching estimator (fourth column) was close to the theoretical variance predicted by our theoretical results and adjusted for sample size depletion due to sample-splitting (third column), though in general the empirical variance was slightly below the theoretical variance. In these scenarios, none of the estimation procedure considered achieved nonparametric efficiency. In scenarios 1 and 3, in which W_2 was a strong precision variable (either strictly or practically), augmentation led to a significant improvement in efficiency, even despite the reduction in effective sample size resulting from sample-splitting. In scenarios 2, 4 and 6, the augmented and unaugmented estimators had a similar variance.

In scenario 5, augmentation resulted in a small inflation in variance.

2.4.4 Simulation studies without sample-splitting

The requirement for sample-splitting in our proposed augmentation method can be problematic for at least two reasons. First, it results in a loss in effective sample size. Second, it requires specification of the proportion of the data to use for estimation of the optimal augmentation function, which serves as a tuning parameter for the entire procedure. This observation makes it natural to consider foregoing sample-splitting altogether.

Through a simulation study, we examined the extent to which the behavior of the augmented propensity score 1:1 matching estimator differs when sample-splitting is used versus not. While our theoretical results for the augmented 1:M matching estimator have only been shown to be valid when sample-splitting is used, as highlighted above, it is of great practical interest to explore whether this is necessary or merely sufficient for the sake of our technical arguments. In this simulation study, we considered the same six data-generating mechanisms previously described. For each scenario, we evaluated the performance of the augmented matching estimator without sample-splitting to the unaugmented matching estimator. The implementation of the augmented estimator was the same as described in the previous subsection except for the use of the full sample for estimation of both the optimal augmentation function and the average treatment effect.

Results are summarized in Table 2.3. Once more, the Monte Carlo variance of the augmented matching estimator was relatively close to the variance predicted by our theoretical results, with some under-dispersion seen in certain settings. In contrast to results shown in Table 2.2, and likely in view of the larger effective sample size, the (sample size-scaled) variance of the augmented matching estimator was never larger and in several cases substantially lower than its unaugmented counterpart. While bias is not explicitly reported on in this table, we found all estimators to have negligible bias, and the percent reduction in empirical variance reported in Table 2.3 was essentially equal to the percent reduction in empirical mean squared error in these simulations. In Figure 2.2, the empirical percent change in (sam-

scenario	theoretical var. (augm.)	empirical var. (augm.)	empirical var. (unaugm.)	empirical change in var.
1	37.55	30.7	41.6	–26%
2	26.57	22.7	27.0	–16%
3	31.12	27.8	35.4	–21%
4	14.28	12.8	13.0	–2%
5	18.31	14.8	14.8	+0%
6	15.12	14.8	14.6	+1%

Table 2.3: Theoretical versus empirical (sample size-scaled) variance of the 1:1 matching with augmented propensity score estimator and no sample-splitting across simulation scenarios at sample size 5,000. The second column provide the variance suggested by theory for an optimally augmented matching estimator based on the known augmentation function. The third and fourth columns display the empirical Monte Carlo variance of the augmented matching estimator (without sample-splitting) and the unaugmented matching estimator implemented over 1,000 simulated datasets. The fifth column shows the percent change in variance between the augmented and unaugmented estimators.

ple size-scaled) variance and mean squared error resulting from augmentation is displayed not only for sample size $n = 5000$ but also for smaller sample sizes $n \in \{500, 1000, 2500\}$. In most scenarios considered, Wald-type confidence intervals for the average treatment effect based on the augmented 1:1 matching estimator (without sample-splitting) and on estimates of the theoretical variance suffered from moderate undercoverage in smaller sample sizes considered, and slight overcoverage at sample size $n = 5000$. The overcoverage seen in large samples mirrors the fact that the empirical variance of the estimator was somewhat smaller than predicted by theory.

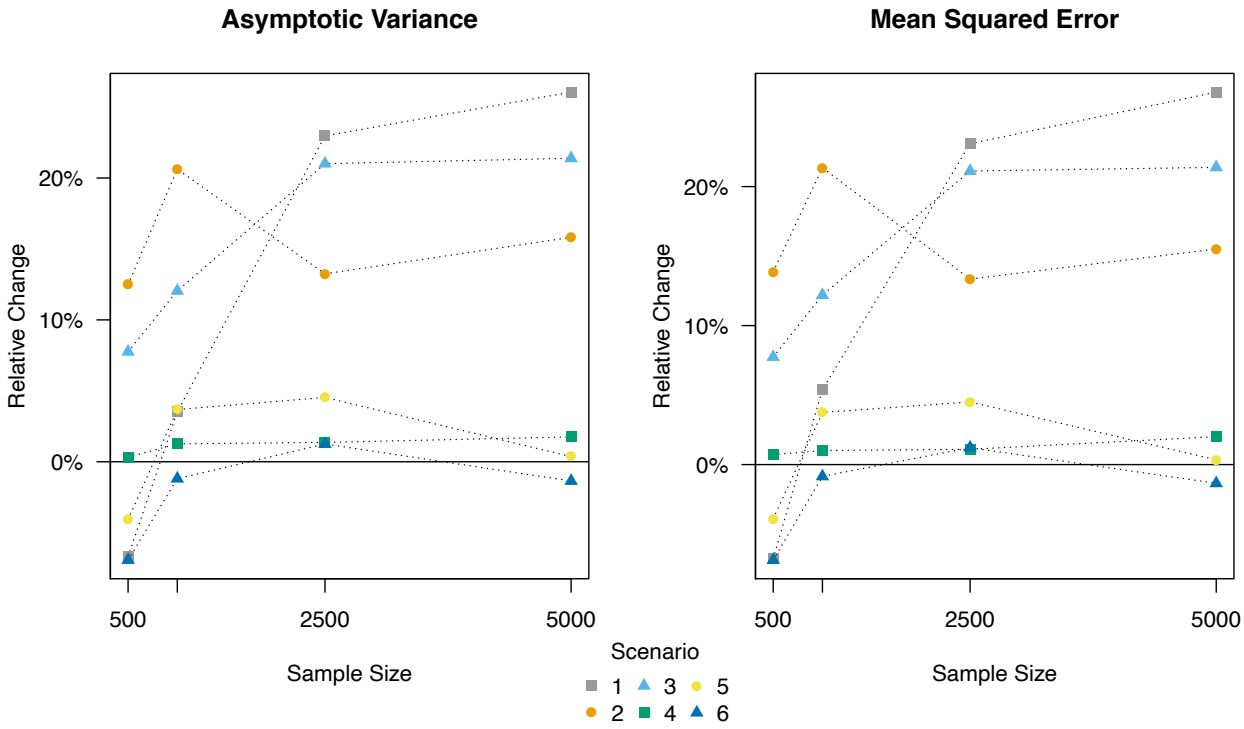

Figure 2.2: Percent reduction in (sample size-scaled) variance and mean squared error comparing the augmented 1:1 matching estimator (without sample-splitting) and its unaugmented counterpart across scenarios and sample sizes. A positive relative reduction indicates improvements resulting from augmentation.

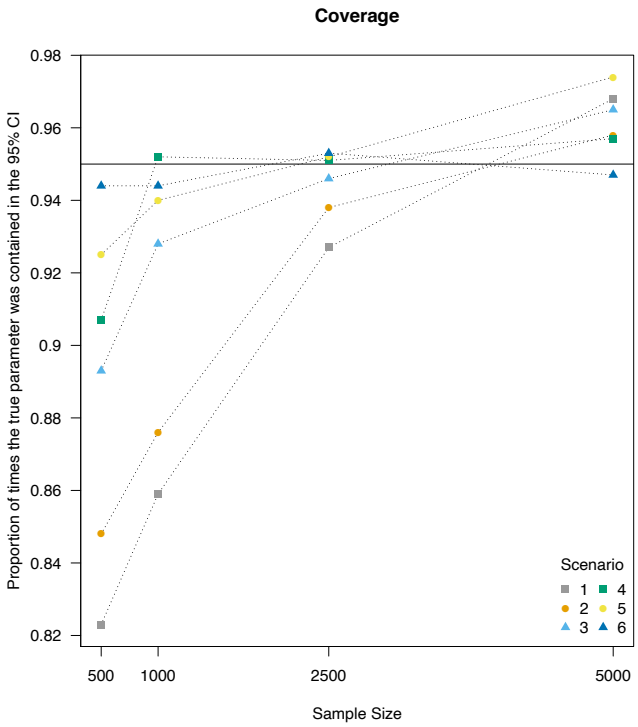

Figure 2.3: Empirical coverage of Wald-type confidence intervals based on the augmented 1:1 matching estimator (without sample-splitting) across scenarios and sample sizes.

2.5 Analyzing the effect of circumcision on risk of HIV infection

We have employed our proposed propensity score augmentation matching method to estimate the causal effect of circumcision on risk of HIV-1 infection. To do so, we analyzed data from the Step Study, a randomized trial conducted by the HIV Vaccine Trials Network to assess the efficacy of a recombinant adenovirus type-5 (Ad5) HIV vaccine. The trial enrolled seronegative men between 18 and 46 years of age at 27 different sites in North America, the Caribbean, South America, and Australia. More details on the study design can be found in Duerr et al. (2012).

Exposure, outcome of interest, and measured potential confounders were defined as follows. Circumcision, the exposure of interest, was self-reported initially but reassessed by clinicians through physical exam in the subsequent follow-up study. The outcome considered is HIV-1 infection occurring within the first two years since enrollment. Potential confounders recorded at baseline were participant age (< 30 or ≥ 30), region (North America + Australia, or Other), ethnicity (Black, Hispanic, Mestizo/Mestiza, White, or Other), attenuated Ad5 titer, evidence of sexually transmitted disease, and sexual risk factors, including number of male partners (0, < 4, or ≥ 4), unprotected insertive anal sex (yes or no), and unprotected receptive anal sex (yes or no) in the last three months before enrollment.

To estimate the causal effect of circumcision on risk of HIV-1 infection, we conducted 1:1 propensity score matching, and first estimated the propensity score through logistic regression with all measured potential confounders as main terms. We compared the results obtained to matching with the augmented propensity score, which included the same main terms and the estimated optimal augmentation covariate. We estimated μ_0 using the Super Learner algorithm (van der Laan et al. 2007), and $\bar{\mu}_0$ using the nested procedure and the univariable highly adaptive lasso algorithm (Benkeser and van der Laan 2016). In our analysis, sample-splitting was not implemented. Point estimates were obtained using the `Matching` package in `R` (Sekhon 2008). The variance of resulting matching estimators were estimated by subtracting from the estimator of σ^2_M provided by the `Matching` package an empirical

plug-in estimator of gain(h_0) based on μ_n, $\bar{\mu}_n$ and π_n.

A total of 1,488 study participants were followed for at least two years and provided complete covariate information. Of these participants, 828 were circumcised, and the other 660 were not. In all, 109 participants developed an HIV-1 infection in the first two years of follow-up, 63 of which were circumcised and the other 46 were not. The average treatment effect was estimated to be -0.01 using the standard 1:1 propensity score-based matching estimator, with estimated standard error 0.028, and -0.04 using the augmented 1:1 propensity score-based matching estimator, with estimated standard error 0.025.

2.6 Discussion

Our findings suggest that 1:M matching using a propensity score estimated through the augmented model $\mathcal{M}(h_0)$ may yield substantial efficiency gains. Brookhart et al. (2006) found that including precision variables in propensity score regression and sub-classification methods increased their precision. This is also true for matching estimators, and the resulting gain in efficiency from including precision variables is even larger when the optimal augmentation covariate is included. Thus, we recommend including any precision variable available into the augmentation covariate. Interestingly, a slight modification of Theorem 2.1 shows that there is no need to include any precision variable in the propensity score model, since doing so or not leads to the same asymptotic variance after optimal augmentation. Thus, if researchers have at their disposal several precision variables, they can extract all the relevant information they contain by including them all in the augmentation covariate, without any need to increase the number of parameters in the propensity score model beyond the single augmentation covariate. This result can be practically advantageous in settings in which the number of precision variables is large and sample size is limited. (Brookhart et al. 2006) also found that including instruments in propensity score regression and sub-classification methods can be detrimental to their performance. This result aligns with our findings in Section 2.4.2, where we showed that the variance of the matching estimator increased as the strength of the instrument increased. The use of the optimal augmentation covariate did

attenuate this loss in precision. Nevertheless, we recommend excluding instruments in both the propensity score model and the augmentation covariate.

Nearest-neighbor matching estimators were recently shown to be nonparametric efficient under certain regularity conditions, provided M suitably tends to infinity with n (Lin et al. 2021). This is not true of the unaugmented propensity score-based matching estimator, for which inefficiency typically persists even for large M. However, our theoretical findings suggest that the (sample size-scaled) asymptotic variance of the 1:M matching estimator with an optimally-augmented propensity score estimator can be arbitrarily close to the efficiency bound if M is large enough. As such, optimal augmentation appears to recover the 'large M' nonparametric efficiency enjoyed by nearest-neighbor but not propensity score-based matching. Currently, we consider M to be fixed and chosen a priori by a practitioner. However, it would be interesting to explore the use of empirical performance criteria to select M in a data-driven fashion. This would require extending our theory to deal with a sample-dependent choice $M = M_n$. It is worth noting that, while other methods (Robins et al. (1994), van der Laan and Rubin (2006)) have been shown to achieve the efficiency bound without complicated tuning, we expect the optimally-augmented 1:M matching estimator to be appealing to practitioners since it may be able to recover statistical efficiency while retaining many of the desirable properties of matching estimation.

While our proposed approach provides a means of performing propensity score matching with greater statistical efficiency, it does suffer from three potential limitations. First, our approach builds upon correct specification of the logistic regression model for the propensity score. Nevertheless, because any propensity score can be approximated by a logistic regression model based on covariate transformations, our methodology could be implemented with W replaced by the evaluation of a rich collection of basis functions on W. Alternatively, as in Abadie and Imbens (2016), it appears possible to extend our theoretical results to arbitrary parametric regression models for the propensity score. Second, the need for sample-splitting in order to estimate h_0 in our procedure results in a reduced effective sample size. Nevertheless, our simulation results suggest that using the complete sample to build an estimator

of h_0 and then to estimate ψ_0 does not invalidate the asymptotic distributional approximation provided in our theory. Whether sample-splitting is fundamentally necessary or only an artifact of our technical proofs remains to be formally studied. Third, similarly as in the work of Abadie and Imbens (2016), our theoretical results only pertain to the augmented $1{:}M$ matching estimator incorporating a discretization regularization, but not to its unregularized counterpart, even though our simulation studies suggest that this is perhaps not a significant issue for practical purposes. Here again, it is of interest to determine whether our results can also be established for the unregularized augmentated $1{:}M$ matching estimator.

While we have focused on the average treatment effect in this work, the same ideas used here could be explored in the context of the average treatment effect among the treated. Abadie and Imbens (2016) showed that when estimating the latter causal contrast, estimation of the propensity score may or may not result in efficiency improvements. Identifying the optimal augmentation covariate and elucidating its impact on efficiency remains an open question.

APPENDIX

2.A Preliminaries

2.A.1 Corresponding notation in relation to main text

In this Appendix, we will use different notation to the one presented in the main text of Chapter 1. We describe the most important modifications in this section. In subsequent sections, all relevant definitions will be explicitly stated, making the notation in this Appendix self-contained.

First, the true parameter of the propensity score model is denoted by θ^*, where θ^* is a $p+1$ dimensional real valued vector. In contrast, it was defined as θ_0 in the main text. Additionally, we let π_{θ^*} be the propensity score evaluated at θ^*, whereas previously it was denoted as π_0. Augmentation of the propensity score model will be denoted in a different way too. Previously, for an augmentation covariate $q : \mathcal{W} \to \mathbb{R}^m$, we let the augmented model be equal to $\mathcal{M}(q) = \{\pi_{q,\theta,\gamma}; \theta \in \mathbb{R}^{p+1}, \gamma \in \mathbb{R}^m\}$, where $\pi_{q,\theta,\gamma} : w \mapsto \text{expit}\{\theta^\top w + \gamma^\top q(w)\}$. In this section, we will use $\mathcal{M}^q := \{\text{logit}[\pi^q_{\bar\theta}(w)] = \bar\theta^\top q(w); \bar\theta \in \mathbb{R}^{p+m+1}\}$ to denote the same augmentation model, where $\pi^q_{\bar\theta}$ is the augmented propensity score, and $\bar\theta$ is a real valued vector of appropriate dimensions. In the main text we defined $\vartheta = (\theta, \gamma)$ to be the vector that conformed the parameters of the augmented model. In this section we do not append the parameters, but rather include a bar on top of the parameter θ to denote augmentation.

Additionally, we will denote the vector of true parameters by $\bar\theta^*_m$, whereas previously it was denoted by ϑ_0. Next, we will denote the MLE of $\bar\theta^*_m$ based on a dataset $\mathcal{D}_n$ by $\tilde\theta_{n,q}(\mathcal{D}_n)$. Previously, this estimator was denoted by $\vartheta_{q,n}$. For a discretization constant k, the discretized versions of the aforementioned MLEs are denoted by $\bar{\tilde\theta}_{k,n,q}(\mathcal{D}_n)$ and $\vartheta_{q,n,k}$, respectively. The matching estimator based on augmentation covariate q and discretized MLE $\bar{\tilde\theta}_{k,n,q}(\mathcal{D}_n)$, will be denoted by $\hat\tau(\bar{\tilde\theta}_{k,n,q}(\mathcal{D}_n), q)$. In the main text, this estimator was defined as $\psi_n(\pi_{q,\vartheta_{q,n,k}})$.

Additionally, the sequence of deterministic functions $\{h_{0n}\}_{n=1}^{\infty}$ will be denoted by $\{\hbar_n\}_{n=1}^{\infty}$. Finally, in this section, the true ATE is defined as τ, whereas in the main text is was denoted by ψ_0.

2.A.2 Notation

Fix a sequence of natural numbers $\{m_n\}_{n=1}^{\infty}$ for which $\lim_{n\to\infty} \frac{m_n}{n} = 0$. In our upcoming arguments, we will generally condition on the iid sequence $S := \{(W_{-i}, A_{-i}, Y_{-i})\}_{i=1}^{\infty}$ from P. The first m_n observations of this sequence are used to obtain the estimate h_n of $\hbar^*$. Some of our arguments will also rely on deterministic sequences in $\mathcal{H}$ that satisfy certain properties. As we will show later on, when the random sequence $(h_n)_{n=1}^{\infty}$ satisfies these properties with probability one, the conclusions of these arguments will also hold with probability one. We always denote deterministic sequences via $(\hbar_n)_{n=1}^{\infty}$, whereas $(h_n)_{n=1}^{\infty}$ denotes the (random) sequence of estimators of $\hbar^*$.

Let P denote the distribution of each of the $n + m_n$ iid copies of (W, A, Y) that were observed. Let $\mathcal{W}, \mathcal{Y}$ be the support of W and Y, respectively. We follow the convention that all vectors are column vectors and we let v_k denote the k-th entry of a vector v. By assumption there exists some $\theta^* \in \mathbb{R}^{p+1}$ such that

$$\operatorname{logit} P(A = 1|W) = (\theta^*)^\top W.$$

Hereafter we let $\bar{\theta}^* := (\theta^*, 0)$, $\pi(w) := P(A = 1|W = w)$, and $\rho(x) := \frac{d}{dx}\operatorname{expit}(x) = \frac{e^x}{(1+e^x)^2}$. For any $\theta \in \mathbb{R}^{p+2}$ and $\hbar : \mathbb{R}^{p+1} \to \mathbb{R}$, $r_\hbar(w) := (w, \hbar(w))$, $\pi_{\theta,\hbar}(w) := \operatorname{expit}[\theta^\top r_\hbar(w)]$, and let $P_{\theta,\hbar}$ denote the distribution for which

$$\frac{dP_{\theta,\hbar}}{dP}(w, a, y) = \frac{\pi_{\theta,\hbar}(w)^a[1 - \pi_{\theta,\hbar}(w)]^{1-a}}{\pi(w)^a[1 - \pi(w)]^{1-a}}. \tag{2.2}$$

We write $E_{\theta,\hbar}$ to denote the expectation operator under $P_{\theta,\hbar}$. Unless otherwise stated, $E[\cdot]$ will be used to denote $E_{\bar{\theta}^*,\hbar^*}[\cdot]$. Note that $P = P_{\bar{\theta}^*,\hbar}$, regardless of the chosen function $\hbar$. Moreover, under sampling (W, A, Y) from $P_{\theta,\hbar}$, W has the same marginal distribution as it would under sampling from P, and $Y \mid A, W$ has the same conditional distribution

as it would under sampling from P. The two distributions only differ in the conditional distribution of $A \mid W$. In particular, $P_{\theta,\hbar}(A = 1 \mid W = w)$ is equal to $\pi_{\theta,\hbar}(w)$, rather than to $\pi(w)$.

Fix $g \in \mathbb{R}^{p+2}$ and let $\bar{P}_{\theta,\hbar}$ denote the marginal distribution of (W, A) under sampling from $P_{\theta,\hbar}$. Let $\theta_n := \bar{\theta}^* + g/\sqrt{n}$. In the following definitions, we let θ and θ' be generic elements of $\mathbb{R}^{p+2}$, $\hbar$ be a $\mathbb{R}^{p+1} \to \mathbb{R}$ function, $z := (w, a)$ denote a realization of a covariate-treatment tuple, $\mathcal{Z}$ its support, and $d_n := \{z_i\}_{i=1}^n$ denote a dataset consisting of n such tuples. We let $\ell_\hbar(\theta; z)$ denote the log-likelihood for $\theta \in \mathbb{R}^{p+2}$ that would arise under sampling from $\bar{P}_{\theta,\hbar}$, $U_\hbar(\theta; z) := \frac{\partial}{\partial\theta}\ell_\hbar(\theta; z)$ denote the score function for θ, and $I_{\theta,\hbar} := E_{\theta,\hbar}[U_\hbar(\theta; Z)U_\hbar(\theta; Z)^\top]$ be the information matrix for θ. In our arguments, we will use that the aforementioned score function takes the following form:

$$U_\hbar(\theta; z) = a\nabla_\theta \log \pi_{\theta,\hbar}(w) + (1 - a)\nabla_\theta \log\{1 - \pi_{\theta,\hbar}(w)\}. \tag{2.3}$$

Additionally, define the following function to denote the Jacobian of the score function:

$$K_h(\theta; z) = \frac{\partial}{\partial\theta} U_\hbar(\theta; z). \tag{2.4}$$

In a slight abuse of notation, we also define $\ell_\hbar(\theta; d_n) := \sum_{i=1}^n \ell_\hbar(\theta; z_i)$ and $U_\hbar(\theta; d_n) := \sum_{i=1}^n U_\hbar(\theta; z_i)$, $K_h(\theta; d_n) = \sum_{i=1}^n K_h(\theta; z_i)$. Finally, we let $\Lambda_\hbar(\theta, \theta'; d_n) := \ell_\hbar(\theta; d_n) - \ell_\hbar(\theta'; d_n)$.

Finally, let $\bar{p}_{\theta,\hbar} := d\bar{P}_{\theta,\hbar}/d\bar{P}$, where $\bar{P}$ is the marginal distribution of (W, A) under sampling from P. Additionally, let P_W denote the marginal distribution of W under P. Note that, by (2.2), $\bar{p}_{\theta,\hbar}(w, a) = \pi_{\theta,\hbar}(w)^a[1 - \pi_{\theta,\hbar}(w)]^{1-a}/\{\pi(w)^a[1 - \pi(w)]^{1-a}\}$.

The following section outlines the sufficient conditions for the proofs of the results in our work to hold.

2.A.3 Conditions

Condition 1. *Identifiability: (i) consistency: for each observation* $Y = AY(A) + (1 - A)Y(1 - A)$*; (ii) for* $a = 0, 1$ $Y(a) \perp\!\!\!\perp A|W$*; (iii) no interference: the random variables* $Y(1), Y(0)$ *and* A *are such that for every* $i \neq j$*, we have that* $Y_i(0) \perp\!\!\!\perp A_j$*, and* $Y_i(1) \perp\!\!\!\perp A_j$*.*

Condition 2. *Positivity: for some constants u and l such that $0 < l < u < 1$, we have that $l < \pi_{\bar{\theta}^*,\hbar^*}(W) < u$ almost surely.*

Condition 3. *Boundedness: $\mathcal{W}$ is a uniformly bounded subset of $\mathbb{R}^p$.*

Condition 4. *Conditions on $\mathcal{H}$: (i) $\mathcal{H}$ is a collection of bounded $\mathcal{W} \to \mathbb{R}$ functions, so that $\sup_{\hbar\in\mathcal{H}} \sup_{w\in\mathcal{W}} |\hbar(w)| < \infty$. (ii) The class of functions $\mathcal{H}$ is totally bounded in $L^2(P_W)$. (iii) The class of functions $\mathcal{H}$ is such that for every $\hbar \in \mathcal{H}$, $E(r_\hbar(W)r_\hbar(W)^\top)$ is positive definite.*

Condition 5. *Consistency: $\bar{\theta}^*$ is identifiable, in the sense that for two observations $\{a_i, w_i\}$ and $\{a_j, w_j\}$, we have that if $\theta \neq \bar{\theta}^*$, then $P_{\theta,\hbar}(a_i, w_i) \neq P_{\bar{\theta}^*,\hbar}(a_i, w_i)$.*

Condition 6. *The evaluation of the propensity score $\pi_{\bar{\theta}^*,\hbar^*}$ at W is continuously distributed, with continuous density.*

Condition 7. *There exists an $\tilde{\epsilon} > 0$ such that, for all $a \in \{0, 1\}$ and $\tilde{\theta} \in \{\theta \in \mathbb{R}^{p+2} : ||\theta - \bar{\theta}^*|| \leq \tilde{\epsilon}\}$, both of the following hold: $E_{\tilde{\theta},\hbar}[Y|A = a, \pi_{\tilde{\theta},\hbar}(W) = p]$ is L-Lipschitz continuous in p and uniformly bounded over $\tilde{\theta}$, $\hbar$, and p; and there exists a $\tilde{\delta} > 0$ such that $E_{\tilde{\theta},\hbar}[|Y|^{2+\tilde{\delta}}|A = a, \pi_{\tilde{\theta},\hbar}(W) = p]$ is uniformly bounded over $\tilde{\theta}$, $\hbar$, and p.*

Condition 8. *For every sequence of vectors $(\theta_n)_{n=1}^\infty$ such that $\theta_n = \bar{\theta}^* + g/\sqrt{n}$ for some $g \in \mathbb{R}^{p+1}$ and every sequence of functions $(\hbar_n)_{n=1}^\infty$ such that $\|\hbar_n - \hbar^*\|_{L_2(P)} \to 0$, we have that, for every $\epsilon > 0$,*

$$P_{\theta_n,\hbar_n}\left(\left|\frac{\lambda_n}{\lambda^*} - 1\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{2.5}$$

Above λ_n and λ^* are as defined in Equations 2.50 and 2.51, which can be found in the proof of Lemma 2.11.

Condition 9. *Given d_n, θ, and θ', let $\Lambda_\hbar(\theta, \theta'; d_n) = \ell_\hbar(\theta; d_n) - \ell_\hbar(\theta'; d_n)$ be the log-likelihood ratio. Then, for any θ_n, θ'_n, sequence of vectors in $\mathbb{R}^{p+2}$ contiguous to $\bar{\theta}^*$ such that $\delta_n = \sqrt{n}(\theta_n - \bar{\theta}^*)$ and $\delta'_n = \sqrt{n}(\theta' - \bar{\theta}^*)$ are bounded sequences in $\mathbb{R}^{p+2}$, there exists a sequence of*

processes $\{\Delta_{\hbar_n}(\theta) : \theta \in \mathbb{R}^{p+2}\}_{n=1}^{\infty}$ *such that*

$$\begin{aligned}\Lambda_{\hbar_n}(\theta_n'|\theta_n) &= (\delta_n' - \delta_n)^\top \Delta_{\hbar_n}(\bar{\theta}^* + \delta_n/\sqrt{n}) - \frac{1}{2}(\delta_n' - \delta_n)^\top I_{\bar{\theta}^*,\hbar^*}(\delta_n' - \delta_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1) \\ &= (\delta_n' - \delta_n)^\top \Delta_{\hbar_n}(\bar{\theta}^* + \delta_n'/\sqrt{n}) + \frac{1}{2}(\delta_n' - \delta_n)^\top I_{\bar{\theta}^*,\hbar^*}(\delta_n' - \delta_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1).\end{aligned}$$

Moreover, the sequence $\Delta_{\hbar_n}(\theta_n)$ *is asymptotically normally distributed with zero mean and variance* $I_{\bar{\theta}^*,\hbar^*}$ *under* $P_{\theta_n,\hbar_n}$ *as* $n \to \infty$.

2.A.4 Interpretation of conditions

In the first part of our proof, we show a uniform version of Theorem 7.2 in van der Vaart (2000) over probability models indexed by θ and $\hbar \in \mathcal{H}$, where $\mathcal{H}$ is a function class. Specifically, the uniformity is over the function class $\mathcal{H}$. In order to show uniform quadratic mean differentiability we require Conditions 3 and 4, which impose restrictions on $\mathcal{W}$ and $\mathcal{H}$, respectively. Next, in Condition 5, we require that the probability model is identifiable for the maximum likelihood estimator to be consistent. Conditions 6 to 8 are used to show joint asymptotic normality of a martingale whose definition is based on the score function at θ_n, and $\hat{\tau}(\theta_n, h_{0n})$, where $\theta_n = \bar{\theta}^* + g/\sqrt{n}$, for some $g \in \mathbb{R}^{p+2}$. Specifically, Condition 6 pertains to the distribution of the propensity score evaluated at W and is required to leverage the results shown in Abadie and Imbens (2016). Condition 7, which is similar to Assumption 4 in Abadie and Imbens (2016), imposes restrictions on the conditional mean of Y given the propensity score and treatment value a. Condition 8 helps us control the asymptotic variance of the aforementioned martingale. Based on the asymptotic normality of the martingale, we then show joint asymptotic normality of the matching estimator, the maximum likelihood estimator, and the log likelihood ratio between $\bar{\theta}^*$ and its shift θ_n. Finally, in order to obtain the marginal distribution of the matching estimator, we use the discretization technique from Andreou and Werker (2012), which requires Condition 9.

2.B Supporting lemmas

As a first step, we want to show that model $\{\bar{P}_{\theta,\hbar}; \theta \in \mathbb{R}^{p+2}, \hbar \in \mathcal{H}\}$ is uniformly differentiable in quadratic mean (UDQM) at every $\theta \in \Theta$, where Θ is an open subset of $\mathbb{R}^{p+2}$. By UDQM, we mean that

$$\sup_{\hbar \in \mathcal{H}} \int \left[\sqrt{\bar{p}_{\theta+g,\hbar}(z)} - \sqrt{\bar{p}_{\theta,\hbar}(z)} - \frac{1}{2} g^t U_\hbar(\theta; z) \sqrt{\bar{p}_{\theta,\hbar}(z)} \right]^2 d\bar{P}(z) = o(||g||^2), \; g \to 0$$

Before we do this, we establish quadratic mean differentiability for each fixed $\hbar$.

Lemma 2.1. *Assume Conditions 2, 3, and 4 hold, then for each $\hbar \in \mathcal{H}$, $\{\bar{P}_{\theta,\hbar}; \theta \in \mathbb{R}^{p+2}\}$ is quadratic mean differentiable.*

Proof of Lemma 2.1. Fix $\hbar \in \mathcal{H}$. In what follows, we will show that the conditions of Lemma 7.6 in van der Vaart (2000) are satisfied, which will give the result.

Let $s_{\theta,\hbar} := \sqrt{\bar{p}_{\theta,\hbar}}$. We first show that, for each z, $\theta \mapsto s_{\theta,\hbar}(z)$ is continuously differentiable. Fix z, noting that $\pi_{\theta,\hbar}(w) = \text{expit}[\theta^\top r_\hbar(w)]$, we see that $\theta \mapsto \pi_{\theta,\hbar}(w)$ is the composition of two continuously differentiable functions, namely $\text{expit}(\cdot)$ and $\theta \mapsto \theta^\top r_\hbar(w)$, and so is itself continuously differentiable. Next, let $\dot{p}_{\theta,\hbar}(z) := \nabla_\theta \bar{p}_{\theta,\hbar}(z)$. It can be shown that

$$\dot{p}_{\theta,\hbar}(z) = \frac{(-1)^a \nabla_\theta \pi_{\theta,\hbar}(w)}{\pi(w)^a [1-\pi(w)]^{1-a}} = \frac{(-1)^a \rho(\theta^\top r_\hbar(w)) r_\hbar(w)}{\pi(w)^a [1-\pi(w)]^{1-a}}.$$

Hence, $\theta \mapsto \dot{p}_{\theta,\hbar}(z)$ is continuous in θ. Finally, let $\dot{s}_{\theta,\hbar}(z) := \nabla_\theta s_{\theta,\hbar}(z)$ be the gradient of $\theta \mapsto s_{\theta,\hbar}(z)$. Noting that $p_{\theta,\hbar}(z) > 0$ for all θ, we see that

$$\dot{s}_{\theta,\hbar}(z) := \nabla_\theta s_{h,\theta}(z) = \nabla_\theta \sqrt{\bar{p}_{\theta,\hbar}(z)} = \dot{p}_{\theta,\hbar}(z) / \left[2 s_{\theta,\hbar}(z)\right].$$

Noting that $\theta \mapsto \bar{p}_{\theta,\hbar}(z)$ is a strictly positive, continuous function and that $x \mapsto \sqrt{x}$ and $x \mapsto 1/x$ are continuous on $(0, \infty)$, we see that $\theta \mapsto \bar{p}_{\theta,\hbar}(z)^{-1/2}$ is continuous. Hence, $z \mapsto \dot{s}_{\theta,\hbar}(z)$ is continuous, and so the map $\theta \mapsto s_{\theta,\hbar}$ is continuously differentiable.

We now show that the elements of the matrix $I_{\theta,\hbar} = \int (\dot{p}_{\theta,\hbar}/\bar{p}_{\theta,\hbar})(\dot{p}_{\theta,\hbar}^\top/\bar{p}_{\theta,\hbar}) \bar{p}_{\theta,\hbar} d\mu$ are well

defined and continuous in θ. Recalling that $\dot{p}_{\theta,\hbar}(z) = \frac{(-1)^a \rho(\theta^\top r_\hbar(w)) r_\hbar(w)}{\pi(w)^a [1-\pi(w)]^{1-a}}$, we see that

$$\begin{aligned}(\dot{p}_{\theta,\hbar}/\bar{p}_{\theta,\hbar})(\dot{p}_{\theta,\hbar}^\top/\bar{p}_{\theta,\hbar}) &= \frac{1}{[\pi_{\theta,\hbar}(w)^a(1-\pi_{\theta,\hbar}(w))^{1-a}]^2}\rho(\theta^\top r_\hbar(w))r_\hbar(w)r_\hbar(w)^\top\rho(\theta^\top r_\hbar(w)) \\ &= \frac{\rho(\theta^\top r_\hbar(w))^2}{[\pi_{\theta,\hbar}(w)^a(1-\pi_{\theta,\hbar}(w))^{1-a}]^2} r_\hbar(w)r_\hbar(w)^\top\end{aligned}$$

Below we use $E_{\bar{P}_W}$ to denote the expectation operator under $\bar{P}_W$. We have that

$$\begin{aligned}I_{\theta,\hbar} &= \int (\dot{p}_{\theta,\hbar}/\bar{p}_{\theta,\hbar})(\dot{p}_{\theta,\hbar}^\top/\bar{p}_{\theta,\hbar}) d\bar{P}_{\theta,\hbar} \\ &= \int_{\mathcal{W}} \left(\sum_{a\in\{1,0\}} \frac{\rho(\theta^\top r_\hbar(w))^2 r_\hbar(w)r_\hbar(w)^\top}{[\pi_{\theta,\hbar}(w)^a(1-\pi_{\theta,\hbar}(w))^{1-a}]^2} [\pi_{\theta,\hbar}(w)^a(1-\pi_{\theta,\hbar}(w))^{1-a}] \right) d\bar{P}_W \\ &= \int_{\mathcal{W}} \left(\sum_{a\in\{1,0\}} \frac{\rho(\theta^\top r_\hbar(w))^2 r_\hbar(w)r_\hbar(w)^\top}{\pi_{\theta,\hbar}(w)^a(1-\pi_{\theta,\hbar}(w))^{1-a}} \right) d\bar{P}_W \\ &= \int_{\mathcal{W}} \frac{\rho(\theta^\top r_\hbar(w))^2 r_\hbar(w)r_\hbar(w)^\top}{\pi_{\theta,\hbar}(w)(1-\pi_{\theta,\hbar}(w))} d\bar{P}_W \\ &= E_{\bar{P}_W}\left[\frac{\rho(\theta^\top r_\hbar(W))^2 r_\hbar(W)r_\hbar(W)^\top}{\pi_{\theta,\hbar}(W)(1-\pi_{\theta,\hbar}(W))} \right],\end{aligned}$$

the third equality uses that $1/t + 1/(1-t) = 1/[t(1-t)]$. By assumption, $\hbar(\cdot)$ is bounded. Hence, there exists a finite constant C_u such that

$$0 \leq \frac{\rho(\theta^\top r_\hbar(w))^2}{\pi_{\theta,\hbar}(w)(1-\pi_{\theta,\hbar}(w))} \leq C_u \text{ for all } w \in \mathcal{W}, \text{ and all } \theta \in \mathbb{R}^{p+2}.$$

Combining the bound on $\hbar(\cdot)$ with the fact that $\mathcal{W}$ is bounded shows that the entries of the matrix $r_\hbar(w)r_\hbar(w)^\top$ are also bounded in w. Therefore, $I_{\theta,\hbar}$ is well-defined and finite. Next, note that the entries of the matrix

$$r_\hbar(w)r_\hbar(w)^\top \frac{\rho_{\theta,\hbar}(w)}{\pi_{\theta,\hbar}(w)^2(1-\pi_{\theta,\hbar}(w))}$$

correspond to the composition of continuous functions in θ and, as such, are also continuous in θ for each w. Therefore, for every w and every $\mathbb{R}^{p+2}$-valued sequence $\{\theta_n\}_{n=1}^\infty$ that converges to θ, we have that

$$r_\hbar(w)r_\hbar(w)^\top \frac{\rho(\theta_n^\top r_\hbar(w))^2}{\pi_{\theta_n,\hbar}(w)(1-\pi_{\theta_n,\hbar}(w))} \longrightarrow r_\hbar(w)r_\hbar(w)^\top \frac{\rho(\theta^\top r_\hbar(w))^2}{\pi_{\theta,\hbar}(w)(1-\pi_{\theta,\hbar}(w))}.$$

Because each entry of $r_\hbar(w)r_\hbar(w)^\top \frac{\rho(\theta_n^\top r_\hbar(w))^2}{\pi_{\theta_n,\hbar}(w)(1-\pi_{\theta_n,\hbar}(w))}$ is bounded by a constant, the Dominated Convergence Theorem implies that

$$E_{\bar{P}_W}\left[\frac{\rho(\theta_n^\top r_\hbar(W))^2 r_\hbar(W)r_\hbar(W)^\top}{\pi_{\theta_n,\hbar}(W)(1-\pi_{\theta_n,\hbar}(W))}\right] \longrightarrow E_{\bar{P}_W}\left[\frac{\rho(\theta^\top r_\hbar(W))^2 r_\hbar(W)r_\hbar(W)^\top}{\pi_{\theta,\hbar}(W)(1-\pi_{\theta,\hbar}(W))}\right]$$

Therefore. $I_{\theta,\hbar}$ is continuous in θ.

By Lemma 7.6 in van der Vaart (2000), the model $\{\bar{P}_{\theta,\hbar}; \theta \in \mathbb{R}^{p+2}\}$ is quadratic mean differentiable. □

Lemma 2.2. *If the conditions of Lemma 2.1 hold, then $\{\bar{P}_{\theta,\hbar}; \theta \in \mathbb{R}^{p+2}, \hbar \in \mathcal{H}\}$ is uniformly differentiable in quadratic mean.*

Before proving the above result, we establish the following useful lemma. Concisely stating this lemma requires a slight abuse of notation. We first fix an open, bounded, convex set H that contains the image of $\hbar$. For each $\eta \in H$, we then define $\pi_{\theta,\eta}(w) := \text{expit}(\theta^\top(w,\eta))$, $\bar{p}_{\theta,\eta}(z) := \pi_{\theta,\eta}(w)^a(1-\pi_{\theta,\eta}(w))^{1-a}/\{\pi(w)^a(1-\pi(w))^{1-a}\}$, and $U_\eta(\theta;z) = a\frac{\partial}{\partial\theta}\log\pi_{\theta,\eta}(w) + (1-a)\frac{\partial}{\partial\theta}\log\{1-\pi_{\theta,\eta}(w)\}$. For each $\theta, g \in \mathbb{R}^{p+2}$ and $z \in \mathcal{Z}$, define the following $H \to \mathbb{R}$ function:

$$f_{\theta,g,z}(\eta) := \left[\bar{p}_{\theta+g,\eta}(z)^{1/2} - \bar{p}_{\theta,\eta}(z)^{1/2} - \frac{1}{2}g^t U_\eta(\theta;z)\bar{p}_{\theta,\eta}(z)^{1/2}\right]^2. \tag{2.6}$$

Lemma 2.3. *Fix $\theta \in \mathbb{R}^{p+2}$ and a bounded neighborhood $\mathcal{N} \subset \mathbb{R}^{p+2}$ of zero. If the conditions of Lemma 2.1 hold, there exists an $L > 0$ such that, for all $g \in \mathcal{N}$ and $z \in \mathcal{Z}$, $f_{\theta,g,z}(\cdot)$ is L-Lipschitz.*

Proof of Lemma 2.3. Fix $\theta \in \mathbb{R}^{p+2}$ and a bounded neighborhood $\mathcal{N}$ of zero. Though not indicated in the notation, all constants displayed in this proof may depend on the choice of θ and $\mathcal{N}$. We show that $\left|\frac{\partial}{\partial\eta}f_{\theta,g,z}(\eta)\right|$ is uniformly bounded over $g \in \mathcal{N}$, $\eta \in H$, and $z \in \mathcal{Z}$.

For each $v \in \mathbb{R}^{p+2}$, $\eta \in H$, and $z = (w,a) \in \mathcal{Z}$, let $q_{v,\eta}(z) := [\frac{\partial}{\partial\eta}\bar{p}_{v,\eta}(z)]/[\bar{p}_{v,\eta}(z)^{1/2}]$. It can be shown that

$$q_{v,\eta}(z) = \frac{1}{\{\pi(w)^a[1-\pi(w)]^{1-a}\}^{1/2}}\left\{\frac{(-1)^{1-a}\pi_{v,\eta}(w)[1-\pi_{v,\eta}(w)]}{\{\pi_{v,\eta}(w)^a[1-\pi_{v,\eta}(w)]^{1-a}\}^{1/2}}\right\}v_{p+2}. \tag{2.7}$$

Further,

$$\begin{aligned}\frac{\partial}{\partial\eta}U_\eta(\theta;z) &= \frac{\partial}{\partial\eta}\left\{a\nabla_\theta\log\{\pi_{\theta,\eta}(w)\}+(1-a)\nabla_\theta\log\{1-\pi_{\theta,\eta}(w,\eta)\}\right\}\\ &= \frac{\partial}{\partial\eta}\left\{(w,\eta)(a-\pi_{\theta,\eta}(w))\right\}\\ &= -\rho(\theta^\top(w,\eta))\theta_{p+2}(w,\eta-[a-\pi_{\theta,\eta}(w)]).\end{aligned}$$

For each $v\in\mathbb{R}^{p+2}$, $\eta\in H$, $z\in\mathcal{Z}$, and $g\in\mathcal{N}$, let $\beta_{\theta,g,\eta,z} := \bar{p}_{\theta+g,\eta}(z)^{1/2}-\bar{p}_{\theta,\eta}(z)^{1/2}-\frac{1}{2}g^\top U_\eta(\theta;z)\bar{p}_{\theta,\eta}(z)^{1/2}$. We have that

$$\begin{aligned}&\frac{\partial}{\partial\eta}f_{\theta,g,z}(\eta) = \qquad (2.8)\\ &= 2\beta_{\theta,g,\eta,z}\left\{\frac{d}{d\eta}\bar{p}_{\theta+g,\eta}(z)^{1/2}-\frac{d}{d\eta}\bar{p}_{\theta,\eta}(z)^{1/2}-\frac{d}{d\eta}[\frac{1}{2}g^tU_\eta(\theta;z)\bar{p}_{\theta,\eta}(z)^{1/2}]\right\}\\ &= 2\beta_{\theta,g,\eta,z}\left\{\frac{\frac{\partial}{\partial\eta}\bar{p}_{\theta+g,\eta}(z)}{2\bar{p}_{\theta+g,\eta}(z)^{1/2}}-\frac{\frac{\partial}{\partial\eta}\bar{p}_{\theta,\eta}(z)}{2\bar{p}_{\theta,\eta}(z)^{1/2}}\right.\\ &\qquad\left.-\frac{1}{2}g^\top\left(\frac{\partial}{\partial\eta}U_\eta(\theta;z)\right)\bar{p}_{\theta,\eta}(z)^{1/2}-\frac{1}{2}g^\top U_\eta(\theta;z)\left(\frac{\frac{\partial}{\partial\eta}\bar{p}_{\theta,\eta}(z)}{2\bar{p}_{\theta,\eta}(z)^{1/2}}\right)\right\}\\ &= \beta_{\theta,g,\eta,z}\left\{\frac{\frac{\partial}{\partial\eta}\bar{p}_{\theta+g,\eta}(z)}{\bar{p}_{\theta+g,\eta}(z)^{1/2}}-\frac{\frac{\partial}{\partial\eta}\bar{p}_{\theta,\eta}(z)}{\bar{p}_{\theta,\eta}(z)^{1/2}}-g^\top\left(\frac{\partial}{\partial\eta}U_\eta(\theta;z)\right)\bar{p}_{\theta,\eta}(z)^{1/2}-\frac{1}{2}g^\top U_\eta(\theta;z)\left(\frac{\frac{\partial}{\partial\eta}\bar{p}_{\theta,\eta}(z)}{\bar{p}_{\theta,\eta}(z)^{1/2}}\right)\right\}\\ &= \beta_{\theta,g,\eta,z}\left\{q_{\theta+g,\eta}(z)-q_{\theta,\eta}(z)-g^\top\left(\frac{\partial}{\partial\eta}U_\eta(\theta;z)\right)\bar{p}_{\theta,\eta}(z)^{1/2}-\frac{1}{2}g^\top U_\eta(\theta;z)q_{\theta,\eta}(z)\right\}\\ &= \beta_{\theta,g,\eta,z}\left\{q_{\theta+g,\eta}(z)-q_{\theta,\eta}(z)-g^\top\left(\frac{\partial}{\partial\eta}U_\eta(\theta;z)\right)[\bar{p}_{\theta,\eta}(z)^{1/2}+q_{\theta,\eta}(z)]\right\}\\ &= \beta_{\theta,g,\eta,z}\left\{q_{\theta+g,\eta}(z)-q_{\theta,\eta}(z)+\rho(\theta^\top(w,\eta))\theta_{p+2}g^\top(w,\eta-[a-\pi_{\theta,\eta}(w)])[\bar{p}_{\theta,\eta}(z)^{1/2}+q_{\theta,\eta}(z)]\right\}.\end{aligned}$$

Because all $\hbar\in\mathcal{H}$ are bounded functions, the domain of $f_{\theta,g,z}$, namely H, is bounded. Moreover, because $\mathcal{W}$ is bounded, we see that $\sup_{w\in\mathcal{W},g\in\mathcal{N}}|(\theta+g)^\top(w,\eta)|<\infty$, and so there exists a $\delta>0$ so that

$$\delta\le\inf_{w\in\mathcal{W},g\in\mathcal{N}}\pi_{\theta+g,\eta}(w)\le\sup_{w\in\mathcal{W},g\in\mathcal{N}}\pi_{\theta+g,\eta}(w)\le 1-\delta. \qquad (2.9)$$

Plugging this observation and a similar observation about the propensity π into (2.7) shows that there exist finite constants C_1, C_2 such that, for all $g\in\mathcal{N}$, $\eta\in H$, and $z\in\mathcal{Z}$,

$|q_{\theta+g,\eta}(z)| \leq C_1(|\theta_{p+2}| + |g_{p+2}|) \leq C_1|\theta_{p+2}| + C_2$, where the latter inequality used that $\mathcal{N}$ is bounded. Let $C := C_1|\theta_{p+2}| + C_2$.

Next, note that there exists an $M < \infty$ such that $\sup_{g\in\mathcal{N},\eta\in H,z\in\mathcal{Z}} |g^\top(w, \eta-[a-\pi_{\theta,\eta}(w)])| < M$. Moreover, because $\rho(x) \leq 1$ for all $x \in \mathbb{R}$ such that

$$\sup_{g\in\mathcal{N},\eta\in H,z\in\mathcal{Z}} |\rho(\theta^\top(w,\eta))\theta_{p+2}g^\top(w, \eta - [a - \pi_{\theta,\eta}(w)])| < M|\theta_{p+2}|.$$

Lastly, by the aforementioned arguments, there exists a $D < \infty$ such that $\sup_{g\in\mathcal{N},\eta\in H,z\in\mathcal{Z}} \bar{p}_{\theta+g,\eta}(z) \leq D$. Therefore, for all $g \in \mathcal{N}$, $\eta \in H$, and $z \in \mathcal{Z}$,

$$\begin{aligned}
&q_{\theta+g,\eta}(z) - q_{\theta,\eta}(z) + \rho(\theta^\top(w,\eta))\theta_{p+2}g^\top(w, \eta - [a - \pi_{\theta,\eta}(w)])[\bar{p}_{\theta,\eta}(z)^{1/2} + q_{\theta,\eta}(z)] \\
&\leq |q_{\theta+g,\eta}(z)| + |q_{\theta,\eta}(z)| + |\rho(\theta^\top(w,\eta))\theta_{p+2}g^\top(w, \eta - [a - \pi_{\theta,\eta}(w)])||\bar{p}_{\theta,\eta}(z)^{1/2} + q_{\theta,\eta}(z)| \\
&\leq 2C + |\theta_{p+2}|M(D^{1/2} + C),
\end{aligned}$$

and also

$$\beta_{\theta,g,\eta,z} = |\bar{p}_{\theta+g,\eta}(z)^{1/2} - \bar{p}_{\theta,\eta}(z)^{1/2} - \frac{1}{2}g^\top U_\eta(\theta; z)\bar{p}_{\theta,\eta}(z)^{1/2}| \leq 2D^{1/2} + |\theta_{p+2}|MD^{1/2}.$$

Combining the above bounds with the display in equation 2.8, we have that

$$\sup_{g\in\mathcal{N},\eta\in H,z\in\mathcal{Z}} \left|\frac{\partial}{\partial\eta} f_{\theta,g,z}(\eta)\right| < (2D^{1/2} + |\theta_{p+2}|MD^{1/2})[2C + |\theta_{p+2}|M(D^{1/2} + C)] =: L.$$

Hence, $f_{\theta,g,z}(\cdot)$ is L-Lipschitz for all $g \in \mathcal{N}$ and $z \in \mathcal{Z}$. □

Proof of Lemma 2.2. Fix $\theta \in \mathbb{R}^{p+2}$. We show that

$$\sup_{\hbar\in\mathcal{H}} \int \left(\sqrt{\bar{p}_{\theta+g,\hbar}(z)} - \sqrt{\bar{p}_{\theta,\hbar}(z)} - \frac{1}{2}g^t U_\hbar(\theta; z)\sqrt{\bar{p}_{\theta,\hbar}(z)}\right)^2 d\bar{P}_Z(z) \to 0 \text{ as } g \to 0.$$

For $g \in \mathcal{N}$ and $\hbar_1, \hbar_2 \in \mathcal{H}$, define

$$\Psi_g(\hbar_1, \hbar_2) := \left|\int [f_{\theta,g,z} \circ \hbar_1(w) - f_{\theta,g,z} \circ \hbar_2(w)] d\bar{P}_Z(z)\right|$$

By two applications of Jensen's inequality and Lemma 2.3, the following holds for all $g \in \mathcal{N}$ and $\hbar_1, \hbar_2 \in \mathcal{H}$:

$$\begin{aligned}
\Psi_g(\hbar_1, \hbar_2) &\leq \int |f_{\theta,g,z} \circ \hbar_1(w) - f_{\theta,g,z} \circ \hbar_2(w)| d\bar{P}_Z(z) \\
&\leq L_\theta \int |\hbar_1(w) - \hbar_2(w)| dP_W(w) \leq L_\theta \|\hbar_1 - \hbar_2\|_{L^2(P_W)}, \qquad (2.10)
\end{aligned}$$

where $\|\cdot\|_{L^2(P_W)}$ denotes the $L^2(P_W)$ norm, defined as $\|f\|_{L^2(P_W)} := [\int f(w)^2 dP_W(w)]^{1/2}$. Because $(\mathcal{H}, L^2(P_W))$ is totally bounded, for every $\delta > 0$, there exists a finite δ-cover of $\mathcal{H}$. Now, fix $\epsilon > 0$. By (2.10), $\|\hbar_1 - \hbar_2\|_{L^2(P_W)} \leq \epsilon/L_\theta$ implies $\Psi_g(\hbar_1, \hbar_2) \leq \epsilon$. Let $\mathcal{H}_\epsilon$ be a minimal ϵ/L_θ-cover of $\mathcal{H}$ with finite cardinality. We then define a $\mathcal{H} \to \mathcal{H}_\epsilon$ map H_ϵ so that, for all $\hbar \in \mathcal{H}$, $H_\epsilon(\hbar)$ is such that $\|\hbar - H_\epsilon(\hbar)\|_{L^2(P_W)} \leq \epsilon/L_\theta$, and note that, by the above $\Psi_g(\hbar, H_\epsilon(\hbar)) \leq \epsilon$ for all $\hbar \in \mathcal{H}$. Combining this with the triangle inequality and the fact that $f_{\theta,g,z}$ is a nonnegative function shows that

$$\begin{aligned}
0 \leq \sup_{\hbar\in\mathcal{H}} \int f_{\theta,g,z} \circ \hbar(w) d\bar{P}_Z(z) &\leq \sup_{\hbar\in\mathcal{H}} \int f_{\theta,g,z} \circ H_\epsilon(\hbar)(w) d\bar{P}_Z(z) \\
&\quad + \sup_{\hbar\in\mathcal{H}} \left| \int \left[f_{\theta,g,z} \circ \hbar(w) - f_{\theta,g,z} \circ H_\epsilon(\hbar)(w) \right] d\bar{P}_Z(z) \right| \\
&= \sup_{\hbar\in\mathcal{H}} \int f_{\theta,g,z} \circ H_\epsilon(\hbar)(w) d\bar{P}_Z(z) + \sup_{\hbar} \Psi_g(\hbar, H_\epsilon(\hbar)) \\
&\leq \sup_{\hbar\in\mathcal{H}} \int f_{\theta,g,z} \circ H_\epsilon(\hbar)(w) d\bar{P}_Z(z) + \epsilon \\
&= \sup_{\hbar\in\mathcal{H}_\epsilon} \int f_{\theta,g,z} \circ \hbar(w) d\bar{P}_Z(z) + \epsilon.
\end{aligned}$$

By Lemma 2.1, for each $\hbar \in \mathcal{H}_\epsilon$, $\int f_{\theta,g,z} \circ \hbar(w) d\bar{P}_Z(z) \to 0$ as $g \to 0$. Thus, the supremum on the right-hand side above, which is a supremum over a finite set, converges to zero. Hence,

$$\sup_{\hbar\in\mathcal{H}} \int f_{\theta,g,z} \circ \hbar(w) d\bar{P}_Z(z) \leq \epsilon.$$

As $\epsilon > 0$ was arbitrary, the LHS of the above equation converges to 0. Recalling the definition of $f_{\theta,g,z}$ from (2.6) gives the result. □

Lemma 2.4. *Assume that the conditions of Lemma 2.1 hold. Let $g \in \mathbb{R}^{p+2}$ and $\theta_n = \bar{\theta}^* + g/\sqrt{n}$. Then, for all $\hbar \in \mathcal{H}$, $E_{\bar{\theta}^*,\hbar}[U_\hbar(\bar{\theta}^*; z)] = 0$, and $I_{\bar{\theta}^*,\hbar}$ exists. Furthermore, for every $\epsilon > 0$ as $n \to \infty$*

$$\sup_{\hbar\in\mathcal{H}} P_{\bar{\theta}^*,\hbar} \left(\left| \Lambda_\hbar(\theta_n, \bar{\theta}^*; D_n) - g^\top \frac{1}{\sqrt{n}} U_\hbar(\bar{\theta}^*; D_n) + \frac{1}{2} g^\top I_{\bar{\theta}^*,\hbar} g \right| > \epsilon \right) \to 0 \tag{2.11}$$

Proof. Let $\hbar \in \mathcal{H}$. By assumption, $P_{\bar{\theta}^*,\hbar}$ is quadratic mean differentiable. Then, by Theorem 7.2 in van der Vaart (2000) $E_{\bar{\theta}^*,\hbar}[U(\bar{\theta}^*; z)] = 0$ and $I_{\bar{\theta}^*,\hbar}$ exists.

Now we show equation 2.11. First, we introduce some notation: let $\bar{p}_{n,\hbar}(z) := \bar{p}_{\bar{\theta}^*+g/\sqrt{n},\hbar}(z)$ and $s_\hbar(z) := g^\top U_\hbar(\bar{\theta}^*, z)$. Note that, since we use $\bar{P}_Z$ as a dominating measure, $\bar{p}_{\bar{\theta}^*,\hbar}(z) = 1$. Now, define the random variable $T_{ni,\hbar} := 2[\sqrt{\bar{p}_{n,\hbar}(Z_i)} - 1]$ and note that

$$\begin{aligned}
&\sup_{\hbar\in\mathcal{H}} E\left[\left\{\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right\}^2\right] \\
&\leq \sup_{\hbar\in\mathcal{H}} \operatorname{var}\left(\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right) \\
&\quad + \sup_{\hbar\in\mathcal{H}} E\left[\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right]^2
\end{aligned}$$

We show these two quantities converge to 0. First, by properties of variances and Lemma 2.2:

$$\begin{aligned}
0 &\leq \sup_{\hbar\in\mathcal{H}} \operatorname{var}\left(\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right) \\
&= \sup_{\hbar\in\mathcal{H}} \operatorname{var}\left(\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i)\right) = \sup_{\hbar\in\mathcal{H}} \sum_{i=1}^n \operatorname{var}\left(T_{ni,\hbar} - n^{-1/2}s_\hbar(Z_i)\right) \\
&= \sup_{\hbar\in\mathcal{H}} \operatorname{var}\left(n^{1/2}T_{n1,\hbar} - s_\hbar(Z_1)\right) \leq \sup_{\hbar\in\mathcal{H}} E\left[\left(n^{1/2}T_{n1,\hbar} - s_\hbar(Z_1)\right)^2\right] \\
&= 2\sup_{\hbar\in\mathcal{H}} \int n\left(\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)} - \frac{1}{2}\frac{g^\top}{\sqrt{n}}U_\hbar(\bar{\theta}^*, z)\sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)}\right)^2 d\bar{P}_Z(z) \to 0. \qquad (2.12)
\end{aligned}$$

Next, note that

$$\begin{aligned}
\sup_{\hbar\in\mathcal{H}} E\left[\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right]^2 &= \sup_{\hbar\in\mathcal{H}} E\left[\sum_{i=1}^n T_{ni,\hbar} + \frac{1}{4}Ps_\hbar^2\right]^2 \\
&= \sup_{\hbar\in\mathcal{H}} \left(2n\int[\sqrt{\bar{p}_{n,\hbar}(z)} - 1]d\bar{P}_Z(z) + \frac{1}{4}Ps_\hbar^2\right)^2 \\
&= \sup_{\hbar\in\mathcal{H}} \left(-n\int\left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)}\right]^2 d\bar{P}_Z(z) + \frac{1}{4}Ps_\hbar^2\right)^2,
\end{aligned}$$

where for a fixed $\hbar$ we have that

$$
\begin{aligned}
&\int \left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)}\right]^2 d\bar{P}_Z(z) \\
&= \int \left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)}\right]^2 d\bar{P}_Z(z) - \int \left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)} - n^{-1/2}\frac{1}{2}s_\hbar(z)\right]^2 d\bar{P}_Z(z) + o(n^{-1}) \\
&= -n^{-1}\frac{1}{4}Ps_\hbar^2 + 2\int \left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)}\right]\left[n^{-1/2}\frac{1}{2}s_\hbar(z)\right] d\bar{P}_Z(z) + o(n^{-1}) \\
&= -n^{-1}\frac{1}{4}Ps_\hbar^2 + n^{-1}\frac{1}{2}\int s_\hbar(z)^2 d\bar{P}_Z(z) \\
&\quad + 2\int \left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)} - n^{-1/2}\frac{1}{2}s_\hbar(z)\right]\left[n^{-1/2}\frac{1}{2}s_\hbar(z)\right] d\bar{P}_Z(z) + o(n^{-1}) \qquad (2.13) \\
&= n^{-1}\frac{1}{4}Ps_\hbar^2 + o(n^{-1}),
\end{aligned}
$$

where the integral in (2.13) is $o(n^{-1})$ by Cauchy-Schwarz and Lemma 2.2. By the UQMD property of $\{P_{\theta,\hbar} : \theta \in \mathbb{R}^{p+2}, \hbar \in \mathcal{H}\}$, the $o(n^{-1})$ terms above are uniform over the choice of $\hbar \in \mathcal{H}$. Therefore,

$$
\sup_{\hbar\in\mathcal{H}} E\left[\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right]^2 \to 0. \qquad (2.14)
$$

By (2.12), (2.14), and Markov's inequality, we have that, for every $\epsilon > 0$,

$$
\begin{aligned}
&\sup_{\hbar\in\mathcal{H}} P\left[\left|\sum_{i=1}^N T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right| > \epsilon\right] \\
&\le \epsilon^{-2}\sup_{\hbar\in\mathcal{H}} E\left[\left\{\sum_{i=1}^n T_{ni,\hbar} - n^{-1/2}\sum_{i=1}^n s_\hbar(Z_i) + \frac{1}{4}Ps_\hbar^2\right\}^2\right] \overset{n\to\infty}{\longrightarrow} 0. \qquad (2.15)
\end{aligned}
$$

Next, we express $\Lambda_\hbar(\theta_n, \bar{\theta}^*; d_n)$ through a Taylor expansion of the logarithm. Define the $(-2,\infty) \to \infty$ function

$$
R(x) = \begin{cases} \frac{\log\left(1+\frac{x}{2}\right) - x/2 + x^2/8}{x^2/4}, & \text{if } x \neq 0, \\ 0, & \text{otherwise.} \end{cases}
$$

Note that $\log(1+x) = x - \frac{1}{2}x^2 + x^2 R(2x)$. Moreover, by a Taylor expansion of $x \mapsto \log(1+x)$

about 0, it can be shown that $R(x) \to 0$ as $x \to 0$. Then, for a given $\hbar \in \mathcal{H}$ we have that

$$\Lambda_\hbar(\theta_n, \bar{\theta}^*; d_n) = 2\sum_{i=1}^{n} \log\left(1 + \frac{1}{2}T_{ni,\hbar}\right) = \sum_{i=1}^{n} T_{ni,\hbar} - \frac{1}{4}\sum_{i=1}^{n} T_{ni,\hbar}^2 + \frac{1}{2}\sum_{i=1}^{n} T_{ni,\hbar}^2 R(T_{ni,\hbar}). \quad (2.16)$$

We have already shown uniform convergence of the first term of the right hand side of the above, in the sense described in (2.15). We now show that $\sum_{i=1}^{n} T_{ni,\hbar}^2 \to Ps_\hbar^2$ in an analogous uniform sense. To do this, we will show that both terms of the right hand side in the equation below converge to 0,

$$\sup_{\hbar\in\mathcal{H}} E\left[\left\{\sum_{i=1}^{n} T_{ni,\hbar}^2 - Ps_\hbar^2\right\}^2\right] \leq \sup_{\hbar\in\mathcal{H}} \operatorname{var}\left\{\sum_{i=1}^{n} T_{ni,\hbar}^2 - Ps_\hbar^2\right\} + \sup_{\hbar\in\mathcal{H}} E\left[\sum_{i=1}^{n} T_{ni,\hbar}^2 - Ps_\hbar^2\right]^2. \quad (2.17)$$

Let $T_{n,\hbar}(z) := 2[\sqrt{\bar{p}_{n,\hbar}(z)} - 1]$, and note that $T_{ni,\hbar} = T_{n,\hbar}(Z_i)$. We begin by showing that the variance term converges to 0 uniformly. To do this, we first upper bound it as follows:

$$\begin{aligned}\sup_{\hbar\in\mathcal{H}} \operatorname{var}\left\{\sum_{i=1}^{n} T_{ni,\hbar}^2 - Ps_\hbar^2\right\} = \sup_{\hbar\in\mathcal{H}} \operatorname{var}\left\{\sum_{i=1}^{n} T_{ni,\hbar}^2\right\} &= \sup_{\hbar\in\mathcal{H}} \sum_{i=1}^{n} \operatorname{var}\left\{T_{ni,\hbar}^2\right\} \leq n \sup_{\hbar\in\mathcal{H}} \int T_{n,\hbar}(z)^4 dP_Z(z) \\ &\leq n\left(\sup_{\hbar\in\mathcal{H}, z\in\mathcal{Z}} T_{n,\hbar}(z)^2\right) \sup_{\hbar\in\mathcal{H}} \int T_{n,\hbar}(z)^2 dP_Z(z).\end{aligned}$$

Now, by Lemma 2.5, $\sup_{\hbar\in\mathcal{H}, z\in\mathcal{Z}} T_{n,\hbar}(z)^2 \to 0$ as $n \to \infty$, and because of the arguments in (2.13) from Lemma 2.4, we have that $\sup_{\hbar\in\mathcal{H}} \int T_{n,\hbar}(z)^2 dP_Z(z) = O(n^{-1})$. Hence, $\sup_{\hbar\in\mathcal{H}} \operatorname{var}\left\{\sum_{i=1}^{n} T_{ni,\hbar}^2 - Ps_\hbar^2\right\} \to 0$. For the second term in (2.17), note that

$$0 \leq \sup_{\hbar\in\mathcal{H}} E\left[\sum_{i=1}^{n} T_{ni,\hbar}^2 - Ps_\hbar^2\right]^2 = \sup_{\hbar\in\mathcal{H}} \left(nE[T_{n1,\hbar}^2] - Ps_\hbar^2\right)^2.$$

Note that by the arguments from (2.13) in Lemma 2.5,

$$\begin{aligned}|nE[T_{ni,\hbar}^2] - Ps_\hbar^2| &= \left|n4\int\left[\sqrt{\bar{p}_{n,\hbar}(z)} - \sqrt{\bar{p}_{\bar{\theta}^*,\hbar}(z)}\right]^2 d\bar{P}_Z(z) - Ps_\hbar^2\right| = \\ &= \left|n4[n^{-1}\frac{1}{4}Ps_\hbar^2 + o(n^{-1})] - Ps_\hbar^2\right| = o(n^{-1}) \to 0,\end{aligned}$$

uniformly over the choice of $\hbar$. Hence, $\sup_{\hbar\in\mathcal{H}}\left(nE[T_{n1,\hbar}^2]-Ps_{\hbar}^2\right)^2\to 0$. Combining the two preceding arguments we have that

$$\sup_{\hbar\in\mathcal{H}}\operatorname{var}\left\{\sum_{i=1}^{n}T_{ni,\hbar}^2-Ps_{\hbar}^2\right\}\to 0.$$

Hence, by Markov's inequality and the aforementioned arguments, we have that for every $\epsilon>0$

$$\sup_{\hbar\in\mathcal{H}}P\left(\left|\sum_{i=1}^{n}T_{ni,\hbar}^2-Ps_{\hbar}^2\right|>\epsilon\right)\overset{n\to\infty}{\longrightarrow}0. \tag{2.18}$$

Next, we show that for any $\epsilon>0$

$$\sup_{\hbar\in\mathcal{H}}P_{\bar{\theta}^*,\hbar}\left\{\left|\sum_{i=1}^{n}T_{ni,\hbar}^2R(T_{ni,\hbar})\right|>\epsilon\right\}\overset{n\to\infty}{\longrightarrow}0.$$

First, note that,

$$E\left[\left|\sum_{i=1}^{n}T_{ni,\hbar}^2R(T_{ni,\hbar})\right|\right]\leq\sum_{i=1}^{n}E\left[\left|T_{ni,\hbar}^2R(T_{ni,\hbar})\right|\right]\leq n\sup_{z\in\mathcal{Z},\hbar\in\mathcal{H}}R(T_{n,\hbar}(z))E\left[T_{ni,\hbar}^2\right]. \tag{2.19}$$

We now argue that $\sup_{z\in\mathcal{Z},\hbar\in\mathcal{H}}R(T_{n,\hbar}(z))\to 0$ as $n\to\infty$. We first note that R is continuous and nondecreasing. Also, because $\mathcal{W}$ and $\mathcal{H}$ are bounded, $\sup_{z\in\mathcal{Z},\hbar\in\mathcal{H}}R(T_{n,\hbar}(z))<\infty$. Hence,

$$\sup_{z\in\mathcal{Z},\hbar\in\mathcal{H}}R(T_{n,\hbar}(z))=R\left(\sup_{z\in\mathcal{Z},\hbar\in\mathcal{H}}T_{n,\hbar}(z)\right).$$

By Lemma 2.5, $\sup_{\hbar\in\mathcal{H},z\in\mathcal{Z}}T_{n,\hbar}(z)^2\to 0$, which implies that $\sup_{\hbar\in\mathcal{H},z\in\mathcal{Z}}T_{n,\hbar}(z)\to 0$. By the above, this implies that $\sup_{z\in\mathcal{Z},\hbar\in\mathcal{H}}R(T_{n,\hbar}(z))\to 0$ as $n\to\infty$. Next, because the arguments outlined in Lemma 2.4 and (2.13), we have that $\sup_{\hbar\in\mathcal{H}}E[T_{ni,\hbar}^2]=O(n^{-1})$. Combining the preceding arguments with (2.19) and applying Markov's inequality, we have that, for any $\epsilon>0$,

$$\sup_{\hbar\in\mathcal{H}}P_{\bar{\theta}^*,\hbar}\left\{\left|\sum_{i=1}^{n}T_{ni,\hbar}^2R(T_{ni,\hbar})\right|>\epsilon\right\}\leq\epsilon^{-1}\sup_{\hbar\in\mathcal{H}}E\left|\sum_{i=1}^{n}T_{ni,\hbar}^2R(T_{ni,\hbar})\right|\overset{n\to\infty}{\longrightarrow}0. \tag{2.20}$$

Combining the identity in (2.16) with the results shown in (2.15), (2.18), and (2.20), and by the triangle inequality, we have that for every $\epsilon>0$, so

$$\sup_{\hbar\in\mathcal{H}}P_{\bar{\theta}^*,\hbar}\left(\left|\Lambda_{\hbar}(\theta_n,\bar{\theta}^*;D_n)-\frac{g^{\top}}{\sqrt{n}}U_{\hbar}(\bar{\theta}^*;D_n)+\frac{1}{2}g^{\top}I_{\bar{\theta}^*,\hbar}g\right|>\epsilon\right)\overset{n\to\infty}{\longrightarrow}0.$$

□

Lemma 2.5. *Fix $\theta, g \in \mathbb{R}^{p+2}$. If the conditions of Lemma 2.4 hold, then*

$$\sup_{\hbar \in \mathcal{H}, z \in \mathcal{Z}} T_{n,\hbar}(z)^2 \to 0.$$

Proof. Fix $z \in \mathcal{Z}$ and $\hbar \in \mathcal{H}$ and define the $\mathbb{R}^{p+2} \to \mathbb{R}$ function $v_{z,\hbar}(\theta) := \bar{p}_{\theta,\hbar}(z)^{1/2}$. We show that v is Lipschitz continuous in θ for all $\hbar \in \mathcal{H}$ and $z \in \mathcal{Z}$. We do so by showing that its derivative is uniformly upper bounded entrywise. First, note that

$$\left[\pi(w)^a(1-\pi(w))^{1-a}\right] \frac{\partial}{\partial \theta} \bar{p}_{\theta,\hbar}(z) = \frac{\partial}{\partial \theta} \left\{\pi_{\theta,\hbar}(w)^a(1-\pi_{\theta,\hbar}(w))^{1-a}\right\} = \begin{cases} \frac{\partial}{\partial \theta}\pi_{\theta,\hbar}(w) \text{ if } a = 1 \\ -\frac{\partial}{\partial \theta}\pi_{\theta,\hbar}(w) \text{ if } a = 0, \end{cases}$$

and because $\frac{\partial}{\partial \theta}\pi_{\theta,\hbar}(w) = r_\hbar(w)\rho(\theta^\top w)$, we have that,

$$\begin{aligned} 2\left|\left[\pi(w)^a(1-\pi(w))^{1-a}\right]^{1/2} \frac{\partial}{\partial \theta} \bar{p}_{\theta,\hbar}(z)^{1/2}\right| &= \left|\left[\pi(w)^a(1-\pi(w))^{1-a}\right]^{1/2} \frac{\frac{\partial}{\partial \theta}\bar{p}_{\theta,\hbar}(z)}{\bar{p}_{\theta,\hbar}(z)^{1/2}}\right| \\ = \left|\frac{r_\hbar(w)\rho(\theta^\top r_\hbar(w))}{\left\{\pi_{\theta,\hbar}(w)^a[1-\pi_{\theta,\hbar}(w)]^{1-a}\right\}^{1/2}}\right| &= \frac{\left|r_\hbar(w)\rho(\theta^\top r_\hbar(w))\right|}{\left\{\pi_{\theta,\hbar}(w)^a[1-\pi_{\theta,\hbar}(w)]^{1-a}\right\}^{1/2}}. \end{aligned}$$

By the definition of ρ, we know that $\rho(\theta^\top r_\hbar(w)) < 1$ regardless of the value of w and $\hbar$. Because of the boundedness of $\mathcal{W}$ and $\mathcal{H}$, there exists a finite vector C_1 and a finite constant C_2 such that $\sup_{\hbar,w} |r_\hbar(w)\rho(\theta^\top r_\hbar(w))| < C_1$ entrywise and $\sup_{z,\hbar} \left\{\pi_{\theta,\hbar}(w)^a[1-\pi_{\theta,\hbar}(w)]^{1-a}\right\}^{-1/2} < C_2$. Finally, by the positivity assumption, there exists a $\delta > 0$ such that $\pi(w) > \delta$ for all $w \in \mathcal{W}$. Hence, $\sup_{z,\hbar} \left|\frac{\partial}{\partial \theta}\bar{p}_{\theta,\hbar}(z)^{1/2}\right| \leq \delta^{-1/2}C_1C_2^{1/2}$, which implies that $v_{z,\hbar}$ is Lipschitz for all $z \in \mathcal{Z}$ and $\hbar \in \mathcal{H}$ with Lipschitz constant $\delta^{-1/2}C_1C_2^{1/2}$. Therefore,

$$\begin{aligned} 0 \leq \sup_{\hbar,z} T_{n,\hbar}(z)^2 = \sup_{\hbar,z} [\bar{p}_{n,\hbar}(z)^{1/2} - \bar{p}_{n,\hbar}(z)^{1/2}]^2 &= \sup_{\hbar,z} [v_{z,\hbar}(\theta + g/\sqrt{n}) - v_{z,\hbar}(\theta)]^2 \\ &\leq \delta^{-1/2}C_1C_2^{1/2}\|(\theta + g/\sqrt{n}) - \theta\|^2 = \delta^{-1/2}C_1C_2^{1/2}\|g\|^2/\sqrt{n} \overset{n\to\infty}{\longrightarrow} 0, \end{aligned}$$

which is the desired result. □

Before stating the next Lemma, we state the definition of contiguity of the measures $P_{\theta_n,\hbar_n}$ and $P_{\bar{\theta}^*,\hbar^*}$. We say that the sequence of measures $P_{\theta_n,\hbar_n}$ is contiguous to $P_{\bar{\theta}^*,\hbar^*}$ if $P_{\bar{\theta}^*,\hbar^*}(A_n) \to 0$ implies that $P_{\theta_n,\hbar_n}(A_n) \to 0$ for every sequence of measurable sets A_n. Recall

that we treat $\hbar_n$ as an element of the deterministic sequence of functions $\{\hbar_n\}_{n=1}^{\infty}$. If $P_{\theta_n,\hbar_n}$ is contiguous to $P_{\bar{\theta}^*,\hbar^*}$ and $P_{\theta_n,\hbar_n}(A_n) \to 0$ implies that $P_{\bar{\theta}^*,\hbar^*}(A_n) \to 0$ for every sequence of measurable sets A_n, then we say that the sequence of measures $P_{\theta_n,\hbar_n}$ is mutually contiguous to $P_{\bar{\theta}^*,\hbar^*}$.

We also introduce additional notation. First, let $\|\cdot\|_{L^1(P)}$ denote the $L^1(P)$ norm and $\|\cdot\|$ denote the $L^2(P)$ norm. Moreover, for a given measurable function f, let $\|f\|_{L^1(P_{\theta^*,\hbar^*})} := E_{P_{\theta^*,\hbar^*}}|f(W,A,Y)|$. Finally, for a given matrix M, we let $(M)_{l,k}$ denote the value in its k-th column and l-th row.

Lemma 2.6. *Fix $g \in \mathbb{R}^{p+2}$ and let $\theta_n = \bar{\theta}^* + g/\sqrt{n}$. Assume that the conditions of Lemma 2.1 hold and $\|\hbar_n - \hbar^*\|_{L^1(P_{\bar{\theta}^*,\hbar^*})} \to 0$. Then, for every $\epsilon > 0$,*

$$P_{\theta_n,\hbar_n}\left(\left|\Lambda_{\hbar_n}(\bar{\theta}^*,\theta_n;D_n) + g^\top \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n;D_n) + \frac{1}{2} g^\top I_{\bar{\theta}^*,\hbar^*} g\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{2.21}$$

Proof. First, note that $P_{\bar{\theta}^*,\hbar^*} = P_{\bar{\theta}^*,\hbar_n}$ and that, by Lemma 2.1, $\{P_{\theta,\hbar} : \theta \in \mathbb{R}^{p+2}, \hbar \in \mathcal{H}\}$ is

uniformly differentiable in quadratic mean. By the triangle inequality,

$$
\begin{aligned}
&P_{\bar{\theta}^*,\hbar^*}\left(\left|\Lambda_{\hbar_n}(\bar{\theta}^*,\theta_n;D_n)+g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\theta_n;D_n)+\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\epsilon\right)\\
=&P_{\bar{\theta}^*,\hbar_n}\left(\left|\Lambda_{\hbar_n}(\theta_n,\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\theta_n;D_n)-\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\epsilon\right)\\
\leq&P_{\bar{\theta}^*,\hbar_n}\left(\left|\Lambda_{\hbar_n}(\theta_n,\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*;D_n)+\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar_n}g\right|>\epsilon\right)\\
&+P_{\bar{\theta}^*,\hbar^*}\left(\left|g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\theta_n;D_n)-\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar_n}g-\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\epsilon\right)\\
\leq&P_{\bar{\theta}^*,\hbar_n}\left(\left|\Lambda_{\hbar_n}(\theta_n,\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*;D_n)+\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar_n}g\right|>\epsilon\right)\\
&+P_{\bar{\theta}^*,\hbar^*}\left(\left|g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar_n}(\theta_n;D_n)-g^\top I_{\bar{\theta}^*,\hbar_n}g\right|>\epsilon\right)\\
&+I\left\{\frac{1}{2}\left|g^\top I_{\bar{\theta}^*,\hbar_n}g-g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\epsilon\right\}\\
\leq&\sup_{\hbar\in\mathcal{H}}P_{\bar{\theta}^*,\hbar}\left(\left|\Lambda_{\hbar}(\theta_n,\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar}(\bar{\theta}^*;D_n)+\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar}g\right|>\epsilon\right)\\
&+\sup_{\hbar\in\mathcal{H}}P_{\bar{\theta}^*,\hbar}\left(\left|g^\top\frac{1}{\sqrt{n}}U_{\hbar}(\bar{\theta}^*;D_n)-g^\top\frac{1}{\sqrt{n}}U_{\hbar}(\theta_n;D_n)-g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\epsilon\right)\\
&+I\left\{\frac{1}{2}\left|g^\top I_{\bar{\theta}^*,\hbar_n}g-g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\epsilon\right\}. \qquad (2.22)
\end{aligned}
$$

By Lemmas 2.7 and 2.4, the first two terms in the last inequality converge to 0 as $n\to\infty$. For the third term, one can use that $\bar{\theta}^*_{p+2}=0$, and therefore $\pi_{\bar{\theta}^*,\hbar_n}=\pi_{\bar{\theta}^*,\hbar^*}$, to show that

$$
\left[I_{\bar{\theta}^*,\hbar_n}-I_{\bar{\theta}^*,\hbar^*}\right]_{l,k}=\begin{cases}0 & l\leq p,\ k\leq p\\ E\left\{W_k\left[\hbar_n(W)-\hbar^*(W)\right]\pi_{\bar{\theta}^*,\hbar^*}(W)(1-\pi_{\bar{\theta}^*,\hbar^*}(W))\right\} & l=p+2,\ k\leq p+1\\ E\left\{W_l\left[\hbar_n(W)-\hbar^*(W)\right]\pi_{\bar{\theta}^*,\hbar^*}(W)(1-\pi_{\bar{\theta}^*,\hbar^*}(W))\right\} & k=p+2,\ l\leq p+1\\ E\left\{[\hbar_n^2(W)-\hbar^{*2}(W)]\pi_{\bar{\theta}^*,\hbar^*}(W)(1-\pi_{\bar{\theta}^*,\hbar^*}(W))\right\} & k=p+2,\ l=p+2,\end{cases}
$$

where W_k is the $k-th$ covariate. By similar arguments to those in Lemma 2.7, and because $\pi_{\bar{\theta}^*,\hbar^*}(w)$ is uniformly bounded, there exists a finite constant M that does not depend on n and is such that

$$
\frac{1}{2}\left|g^\top I_{\bar{\theta}^*,\hbar_n}g-g^\top I_{\bar{\theta}^*,\hbar^*}g\right|\leq M\left|E\left[\hbar_n(W)-\hbar^*(W)\right]\right|\leq M\|\hbar_n-\hbar^*\|_{L^1(P_{\theta^*,\hbar^*})}.
$$

By the above display and because $\|\hbar_n - \hbar^*\|_{L^1(P_{\theta^*,\hbar^*})} \to 0$, we have that

$$I\left\{\frac{1}{2}\left|g^\top I_{\bar{\theta}^*,\hbar_n} g - g^\top I_{\bar{\theta}^*,\hbar^*} g\right| > \epsilon\right\} \overset{n\to\infty}{\longrightarrow} 0.$$

Therefore, for every $\epsilon > 0$,

$$P_{\bar{\theta}^*,\hbar^*}\left(\left|\Lambda_{\hbar_n}(\bar{\theta}^*, \theta_n; D_n) + g^\top \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) + \frac{1}{2} g^\top I_{\bar{\theta}^*,\hbar^*} g\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Because of Lemma 2.8 the sequence of measures $(P^n_{\theta_n,\hbar_n})_{n=1}^\infty$ and $(P^n_{\bar{\theta}^*,\hbar^*})_{n=1}^\infty$ are mutually contiguous. The latter implies that

$$P_{\theta_n,\hbar_n}\left(\left|\Lambda_{\hbar_n}(\bar{\theta}^*, \theta_n; D_n) + g^\top \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) + \frac{1}{2} g^\top I_{\bar{\theta}^*,\hbar^*} g\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0,$$

which is the desired result. □

Lemma 2.7. *Under the conditions of Lemma 2.6, for every $\epsilon > 0$,*

$$\sup_{\hbar\in\mathcal{H}} P_{\bar{\theta}^*,\hbar^*}\left(\left| - g^\top \frac{1}{\sqrt{n}} U_\hbar(\theta_n; D_n) + g^\top \frac{1}{\sqrt{n}} U_\hbar(\bar{\theta}^*; D_n) - g^\top I_{\bar{\theta}^*,\hbar} g\right| > \epsilon\right) \to 0$$

Proof. Fix $\hbar \in \mathcal{H}$ and $g \in \mathbb{R}^{p+2}$. A Taylor expansion of the function $\theta \mapsto \frac{g^\top}{\sqrt{n}} U_\hbar(\theta; d_n)$ around $\bar{\theta}^*$ yields

$$\begin{aligned}
\frac{g^\top}{\sqrt{n}} U_\hbar(\theta_n, d_n) &= \frac{g^\top}{\sqrt{n}} U_\hbar(\bar{\theta}^*; d_n) + \frac{g^\top}{\sqrt{n}} \left(K_\hbar(\theta; d_n)\Big|_{\theta=\bar{\theta}^*}\right)(\theta_n - \bar{\theta}^*) + R_\hbar(\theta_n, \bar{\theta}^*, d_n) \\
&= \frac{g^\top}{\sqrt{n}} U_\hbar(\bar{\theta}^*; d_n) + \frac{g^\top}{n} \left(K_\hbar(\theta; d_n)\Big|_{\theta=\bar{\theta}^*}\right) g + R_\hbar(\theta_n, \bar{\theta}^*, d_n),
\end{aligned}$$

where $R_\hbar(\theta_n, \bar{\theta}^*, d_n)$ is the remainder term of the Taylor expansion, and $K_\hbar(\theta; d_n)$ is the Jacobian of $U_\hbar(\theta; d_n)$. First, we will show that

$$\sup_{\hbar\in\mathcal{H}} P_{\bar{\theta}^*}\left(\left|g^\top \left[\frac{1}{n} K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0.$$

Note that, for a given d_n,

$$K_\hbar(\theta; d_n)\Big|_{\theta=\bar{\theta}^*} = \sum_{i=1}^n \left\{r_\hbar(w_i) r_\hbar(w_i)^\top [\pi_{\bar{\theta}^*,\hbar}(w_i)(1 - \pi_{\bar{\theta}^*,\hbar}(w_i))]\right\}.$$

Now, for a fixed $w \in \mathcal{W}$, define the $\mathcal{H} \to \mathbb{R}$ function

$$f_{w,g}(\hbar) = g^\top [r_\hbar(w) r_\hbar(w)^\top [\pi_{\bar{\theta}^*,\hbar}(w)(1 - \pi_{\bar{\theta}^*,\hbar}(w))] - I_{\bar{\theta}^*,\hbar}] g.$$

We will show that there exists a finite constant Q such that for $\hbar_1, \hbar_2 \in \mathcal{H}$

$$E\left[|f_{w,g}(\hbar_1) - f_{w,g}(\hbar_2)|\right] \leq 2Q \int |\hbar_1(w) - \hbar_2(w)| d\bar{P}_W(w).$$

Given a matrix M, let $(M)_{k,l}$ its value in row k and column l, and note that

$$\left[r_{\hbar_1}(w) r_{\hbar_1}(w)^\top - r_{\hbar_2}(w) r_{\hbar_2}(w)^\top\right]_{l,k} = \begin{cases} 0 & l \leq p,\ k \leq p \\ w_k \left[\hbar_1(w) - \hbar_2(w)\right] & l = p+2,\ 1 \leq k \leq p+1 \\ w_l \left[\hbar_1(w) - \hbar_2(w)\right] & k = p+2,\ 1 \leq l \leq p+1 \\ \hbar_1(w)^2 - \hbar_2(w)^2 & k = p+2,\ l = p+2, \end{cases}$$

where w_j is the j-th covariate. Given the above, and because $g \in \mathbb{R}^{p+2}$, W has bounded support, and $\mathcal{H}$ is bounded, we have that there exists a constant $K < \infty$ such that, for $\hbar_1, \hbar_2 \in \mathcal{H}$

$$g^\top |r_{\hbar_1}(w) r_{\hbar_1}(w)^\top - r_{\hbar_2}(w) r_{\hbar_2}(w)^\top| g \leq K |\hbar_1(w) - \hbar_2(w)|.$$

Additionally, using that $\mathcal{W}$ is bounded and the positivity assumption, we have that for all $w \in \mathcal{W}$ $\pi_{\bar{\theta}^*,\hbar_1}(w)(1 - \pi_{\bar{\theta}^*,\hbar_1}(w)) < C$ for some $C < \infty$. Note also that $\pi_{\bar{\theta}^*,\hbar_1}(w) = \pi_{\bar{\theta}^*,\hbar_2}(w)$ since $\bar{\theta}^*_{p+2} = 0$. Combining the preceding arguments, we have that

$$\begin{aligned} &g^\top [r_{\hbar_1}(w) r_{\hbar_1}(w)^\top [\pi_{\bar{\theta}^*,\hbar_1}(w)(1 - \pi_{\bar{\theta}^*,\hbar_1}(w))] - [r_{\hbar_2}(w) r_{\hbar_2}(w)^\top [\pi_{\bar{\theta}^*,\hbar_2}(w)(1 - \pi_{\bar{\theta}^*,\hbar_2}(w))] g \\ &= [\pi_{\bar{\theta}^*,\hbar_1}(w)(1 - \pi_{\bar{\theta}^*,\hbar_1}(w))] g^\top [r_{\hbar_1}(w) r_{\hbar_1}(w)^\top - r_{\hbar_2}(w) r_{\hbar_2}(w)^\top] g \\ &\leq (K \times C) |\hbar_1(w) - \hbar_2(w)|. \end{aligned}$$

Because $I_{\bar{\theta}^*,\hbar} = E\left[r_\hbar(W) r_\hbar(W)^\top \pi_{\bar{\theta}^*,\hbar}(W)(1 - \pi_{\bar{\theta}^*,\hbar}(W))\right]$ and by the aforementioned arguments, we have that, for $\hbar_1, \hbar_2 \in \mathcal{H}$

$$|g^\top [I_{\bar{\theta}^*,\hbar_1} - I_{\bar{\theta}^*,\hbar_2}] g| \leq K \times C \|\hbar_1(w) - \hbar_2(w)\|_{L^1(P)}.$$

Therefore, letting $Q = K \times C$ we have that,

$$E[|f_{W,g}(\hbar_1) - f_{W,g}(\hbar_2)|] \leq 2Q \int |\hbar_1(w) - \hbar_2(w)| d\bar{P}_W(w).$$

Next, because $\mathcal{W}$ and $\mathcal{H}$ are bounded, $E\left|g^\top [r_\hbar(W) r_\hbar(W)^\top [\pi_{\bar{\theta}^*,\hbar}(W)(1 - \pi_{\bar{\theta}^*,\hbar}(W))] g\right| < \infty$. Hence, by the Weak Law of Large Numbers, for all $\hbar \in \mathcal{H}$,

$$P_{\bar{\theta}^*}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left\{g^\top\left[r_\hbar(W_i) r_\hbar(W_i)^\top [\pi_{\bar{\theta}^*,\hbar}(W_i)(1 - \pi_{\bar{\theta}^*,\hbar}(W_i))] - I_{\bar{\theta}^*,\hbar}\right] g\right\}\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \quad (2.23)$$

Fix $\delta > 0$ and let $\mathcal{H}_\delta$ be a minimal δ/Q-cover of $\mathcal{H}$ with respect to the $L^1(P)$ pseudometric — this cover has finite cardinality since $\mathcal{H}_\delta$ is totally bounded in $L^2(P)$ and the $L^2(P)$ norm is stronger than the $L^1(P)$ norm. Define a $\mathcal{H} \to \mathcal{H}_\delta$ map H_δ so that, for all $\hbar \in \mathcal{H}$, $H_\delta(\hbar)$ is such that $\|\hbar - H_\delta(\hbar)\|_{L^1(\bar{P}_W)} \leq \delta/Q$. Hence, by the triangle inequality and Markov's inequality, we have that

$$\begin{aligned}
0 &\leq \sup_{\hbar\in\mathcal{H}} P_{\bar{\theta}^*}\left(\left|g^\top\left[\frac{1}{n}K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right) \\
&\leq \sup_{\hbar\in\mathcal{H}} P\left(|f_{w,g}(\hbar) - f_{w,g} \circ H_\epsilon(\hbar)| > \epsilon\right) + \sup_{\hbar\in\mathcal{H}_\delta} P_{\bar{\theta}^*,\hbar}\left(\left|g^\top\left[\frac{1}{n}K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right) \\
&\leq \epsilon^{-1} \sup_{\hbar\in\mathcal{H}} \int |f_{w,g}(\hbar) - f_{w,g} \circ H_\epsilon(\hbar)| \, d\bar{P}_W(w) + \sup_{\hbar\in\mathcal{H}_\delta} P_{\bar{\theta}^*,\hbar}\left(\left|g^\top\left[\frac{1}{n}K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right) \\
&\leq \epsilon^{-1} \sup_{\hbar\in\mathcal{H}} Q \int |\hbar(w) - H_\epsilon(\hbar)(w)| d\bar{P}_W(w) + \sup_{\hbar\in\mathcal{H}_\delta} P_{\bar{\theta}^*,\hbar}\left(\left|g^\top\left[\frac{1}{n}K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right) \\
&\leq \delta/\epsilon + \sup_{\hbar\in\mathcal{H}_\delta} P_{\bar{\theta}^*,\hbar}\left(\left|g^\top\left[\frac{1}{n}K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right)
\end{aligned}$$

Taking a limit superior as $n \to \infty$, Equation 2.23 and the fact that $\mathcal{H}_\delta$ has finite cardinality show that

$$0 \leq \limsup_{n\to\infty} \sup_{\hbar\in\mathcal{H}} P_{\bar{\theta}^*}\left(\left|g^\top\left[\frac{1}{n}K_\hbar(\theta; D_n)\Big|_{\theta=\bar{\theta}^*} - I_{\bar{\theta}^*,\hbar}\right] g\right| > \epsilon\right) \leq \delta/\epsilon.$$

Since $\delta > 0$ was arbitrary, the limit superior above is zero. Next, we show that

$$\sup_{\hbar\in\mathcal{H}} P_{\bar{\theta}^*,\hbar}\left(\left|R_\hbar(\theta_n, \bar{\theta}^*, D_n)\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \quad (2.24)$$

Fix $\hbar \in \mathcal{H}$ and $d_n \in \mathcal{Z}^n$. Let $\mathbb{B}_r(x)$ denote a p+2-dimensional open ball of radius r centered at x. Writing $U(\theta; d_n)_j$ to denote the j-th entry of $U(\theta; d_n)$, we note that, for each j, the function $\theta \mapsto U(\theta; d_n)_j$ is twice differentiable with continuous Hessian on $\mathbb{B}_{R_n}(\bar{\theta^*})$, where $R_n = 2g/\sqrt{n}$. Moreover, for arbitrary $j, k \in \{1, , ...p+2\}$ and letting $w_{i,j}$ denote the value of the j-th covariate of the i-th observation, it can be shown that

$$\frac{\partial^2}{\partial\theta_k \partial\theta_j}\left\{\frac{g^\top}{\sqrt{n}} U_\hbar(\theta; d_n)\right\} = \frac{g^\top}{\sqrt{n}} \sum_{i=1}^{n} \left\{ w_{i,j} w_{i,k} r_\hbar(w_i) \rho(\theta^\top r_\hbar(w_i)) \left[\frac{1 - \exp(\theta^\top r_\hbar(w_i)}{1 + \exp(\theta^\top r_\hbar(w_i)}\right]\right\}.$$

Because $g \in \mathbb{R}$, $\mathcal{W}$ and $\mathcal{H}$ are bounded, and $[1 - \exp(x)]/[(1 + \exp(x)] < 1$ for all $x \in \mathbb{R}$, there exists a finite constant M such that

$$\frac{\partial^2}{\partial\theta_k \partial\theta_j}\left\{\frac{g^\top}{\sqrt{n}} U_\hbar(\theta; d_n)\right\} \leq \frac{M}{\sqrt{n}} n = M\sqrt{n}.$$

Hence, the second partial derivatives of the function $\theta \mapsto g^\top/\sqrt{n} U_\hbar(\theta; d_n)$ can be bounded (up to $\sqrt{n}$) by a finite constant M which does not depend on $\hbar$ or d_n. By Taylor's Theorem we have that

$$\left|R_\hbar(\theta_n, \bar{\theta}^*, d_n)\right| \leq \frac{M\sqrt{n}}{2!} \left\| \frac{g}{\sqrt{n}} \right\|_1^2 = \frac{M}{2!} n^{-1/2} \|g\|_1^2 \leq \sqrt{p+2} \frac{M}{2!} n^{-1/2} \|g\|^2,$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm and, as throughout, $\|\cdot\|$ denotes the ℓ_2 norm. Taking a supremum on both sides yields

$$\sup_{\hbar \in \mathcal{H}} \left|R_\hbar(\theta_n, \bar{\theta}^*, d_n)\right| \leq \sqrt{p+2} \frac{M}{2!} n^{-1/2} \|g\|^2.$$

Therefore, for any $\hbar \in \mathcal{H}$,

$$P_{\bar{\theta}^*, \hbar}\left(\sup_{\hbar' \in \mathcal{H}} \left|R_{\hbar'}(\theta_n, \bar{\theta}^*, D_n)\right| > \epsilon\right) \leq I\left\{\sqrt{p+2} \frac{M}{2!} n^{-1/2} \|g\|^2 > \epsilon\right\}.$$

Taking a supremum over $\hbar \in \mathcal{H}$ and a limit as $n \to \infty$ then shows that

$$\sup_{\hbar \in \mathcal{H}} P_{\bar{\theta}^*, \hbar}\left(\sup_{\hbar' \in \mathcal{H}} \left|R_{\hbar'}(\theta_n, \bar{\theta}^*, D_n)\right| > \epsilon\right) \leq I\left\{\sqrt{p+2} \frac{M}{2!} n^{-1/2} \|g\|^2 > \epsilon\right\} \to 0.$$

Combining this with the fact that

$$\sup_{\hbar \in \mathcal{H}} P_{\bar{\theta}^*, \hbar}\left(\left|r_\hbar(\theta_n, \bar{\theta}^*, D_n)\right| > \epsilon\right) \leq \sup_{\hbar \in \mathcal{H}} P_{\bar{\theta}^*, \hbar}\left(\sup_{\hbar' \in \mathcal{H}} \left|R_{\hbar'}(\theta_n, \bar{\theta}^*, D_n)\right| > \epsilon\right)$$

establishes (2.24). □

Lemma 2.8. *Assume that the conditions of Lemma 2.1 hold and that the $\mathcal{H}$-valued sequence $(\hbar)_{n=1}^{\infty}$ is such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$ as $n \to \infty$. Let $g \in \mathbb{R}^{p+2}$, and $\theta_n = \bar{\theta}^* + g/\sqrt{n}$. Then, the sequences of probability measures $(P^n_{\theta_n,\hbar_n})_{n=1}^{\infty}$ and $(P^n_{\bar{\theta}^*,\hbar^*})_{n=1}^{\infty}$ are mutually contiguous, where, for a distribution Q, Q^n denotes the n-fold product measure of Q.*

Proof. Fix a sequence $(\hbar)_{n=1}^{\infty}$ that is such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\theta^*,\hbar^*})} \to 0$. We show that the ratio of the log-likelihoods, namely $\log\left\{\prod_{i=1}^n \frac{p_{\theta_n,\hbar_n}(Z_i)}{p_{\bar{\theta}^*,\hbar^*}(Z_i)}\right\}$, is asymptotically normal under $P_{\bar{\theta}^*,\hbar^*}$. Recall that, because $\bar{\theta}^*_{p+2} = 0$, it holds that $P_{\bar{\theta}^*,\hbar^*} = P_{\bar{\theta}^*,\hbar_n}$. A Taylor expansion of the log-likelihood ratio thus gives that, for some intermediate value $\theta'(D_n)$ between $\bar{\theta}^*$ and θ_n,

$$\begin{aligned}
\log\left\{\prod_{i=1}^n \frac{p_{\theta_n,\hbar_n}(Z_i)}{p_{\bar{\theta}^*,\hbar^*}(Z_i)}\right\} &= \sum_{i=1}^n \left\{\log\left(p_{\theta_n,\hbar_n}(Z_i)\right) - \log\left(P_{\bar{\theta}^*,\hbar}(Z_i)\right)\right\} \\
&= \sum_{i=1}^n \left\{\log\left(p_{\theta_n,\hbar_n}(Z_i)\right) - \log\left(p_{\bar{\theta}^*,\hbar_n}(Z_i)\right)\right\} \\
&= (\theta_n - \bar{\theta}^*)^\top \left\{\sum_{i=1}^n U_{\hbar_n}(\bar{\theta}^*; Z_i)\right\} + \frac{1}{2}(\theta_n - \bar{\theta}^*)^\top \left\{\sum_{i=1}^n K_{\hbar_n}(\theta'(D_n); Z_i)\right\}(\theta_n - \bar{\theta}^*) \\
&= n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1/2}\sum_{i=1}^n U_{\hbar_n}(\bar{\theta}^*; Z_i)\right\} \qquad (2.25) \\
&\quad + \frac{1}{2}n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1}\sum_{i=1}^n K_{\hbar_n}(\theta'(D_n); Z_i)\right\} n^{1/2}(\theta_n - \bar{\theta}^*)
\end{aligned}$$

where $K(\theta, z)$ is the Jacobian of the score function evaluated at (θ, z) as defined in (2.4). We analyze the two sums in Equation 2.25 as follows. For the first sum, we have that

$$\begin{aligned}
n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1/2}\sum_{i=1}^n U_{\hbar_n}(\bar{\theta}^*; Z_i)\right\} &= n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1/2}\sum_{i=1}^n U_{\hbar^*}(\bar{\theta}^*; Z_i)\right\} \\
+ n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1/2}\sum_{i=1}^n \left\{U_{\hbar_n}(\bar{\theta}^*; Z_i) - U_{\hbar^*}(\bar{\theta}^*; Z_i)\right\}\right\}&.
\end{aligned}$$

Next, note that for all $z = (w, a)$,

$$\begin{aligned}
U_{\hbar_n}(\bar{\theta}^*; z) &= r_{\hbar_n}(w)\frac{a - \pi_{\bar{\theta}^*,\hbar_n}(w)}{\pi_{\bar{\theta}^*,\hbar_n}(w)\left(1 - \pi_{\bar{\theta}^*,\hbar_n}(w)\right)}\rho(\bar{\theta}^{*\top} r_{\hbar_n}(w)) \\
&= r_{\hbar_n}(w)\frac{-\pi_{\bar{\theta}^*,\hbar^*}(w)}{\pi_{\bar{\theta}^*,\hbar^*}(w)\left(1 - \pi_{\bar{\theta}^*,\hbar^*}(w)\right)}\rho(\bar{\theta}^{*\top} r_{\hbar^*}(w)).
\end{aligned}$$

The latter implies

$$U_{\hbar_n}(\bar{\theta}^*; z) - U_{\hbar^*}(\bar{\theta}^*; z) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (\hbar_n(w) - \hbar^*(w)) \frac{a - \pi_{\bar{\theta}^*, \hbar^*}(w)}{\pi_{\bar{\theta}^*, \hbar^*}(w)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(w)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(w)) \end{pmatrix}.$$

Then, because $\sqrt{n}(\theta_n - \bar{\theta}^*) = g$, we have that

$$n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left(n^{-1/2} \sum_{i=1}^{n} \left\{ U_{\hbar_n}(\bar{\theta}^*; Z_i) - U_{\hbar^*}(\bar{\theta}^*; Z_i) \right\} \right) = \tag{2.26}$$
$$= g_{p+2} n^{-1/2} \sum_{i=1}^{n} \left\{ [\hbar_n(W_i) - \hbar^*(W_i)] \frac{A_i - \pi_{\bar{\theta}^*, \hbar^*}(W_i)}{\pi_{\bar{\theta}^*, \hbar^*}(W_i)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(W_i)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(W_i)) \right\},$$

The expectation of each element in the above sum is

$$E\left[[\hbar_n(W) - \hbar^*(W)] \frac{A - \pi_{\bar{\theta}^*, \hbar^*}(W)}{\pi_{\bar{\theta}^*, \hbar^*}(W)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(W)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(W)) \right]$$
$$= E\left\{ E\left[[\hbar_n(W) - \hbar^*(W)] \frac{A - \pi_{\bar{\theta}^*, \hbar^*}(W)}{\pi_{\bar{\theta}^*, \hbar^*}(W)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(W)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(W)) | W \right] \right\}$$
$$= E\left\{ \frac{[\hbar_n(W) - \hbar^*(W)] \rho(\bar{\theta}^{*\top} r_{\hbar^*}(W))}{\pi_{\bar{\theta}^*, \hbar^*}(W)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(W)\right)} E\left[A - \pi_{\bar{\theta}^*, \hbar^*}(W) | W \right] \right\} = 0.$$

Therefore, the expectation of (2.26) is equal to 0. Next, the variance of the right-hand side in Equation 2.26 is equal to

$$\text{var}\left(g_{p+2} n^{-1/2} \sum_{i=1}^{n} \left\{ [\hbar_n(W_i) - h(W_i)] \frac{A_i - \pi_{\bar{\theta}^*, \hbar^*}(W_i)}{\pi_{\bar{\theta}^*, \hbar^*}(W_i)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(W_i)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(W_i)) \right\} \right) \tag{2.27}$$
$$= g_{p+2}^2 E\left\{ [\hbar_n(W) - h(W)]^2 \left(\frac{A - \pi_{\bar{\theta}^*, \hbar^*}(W)}{\pi_{\bar{\theta}^*, \hbar^*}(W)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(W)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(W)) \right)^2 \right\}.$$

We show that the above variance converges to 0 as $n \to \infty$. Because $\mathcal{W}$ is bounded and functions in $\mathcal{H}$ are uniformly bounded, and since $0 < \pi_{\bar{\theta}^*, \hbar}(w) < 1$ for all $w \in \mathcal{W}$, the quantity $\left(\frac{a - \pi_{\bar{\theta}^*, \hbar^*}(w)}{\pi_{\bar{\theta}^*, \hbar^*}(w)\left(1 - \pi_{\bar{\theta}^*, \hbar^*}(w)\right)} \rho(\bar{\theta}^{*\top} r_{\hbar^*}(w)) \right)^2$ can be uniformly bounded by a constant $M < \infty$. Next,

by assumption, $\|\hbar_n - \hbar^*\|_{L^2(P_{\theta^*,\hbar^*})} \to 0$. Hence, the variance in (2.27) can be upper bounded by a term that converges to 0, therefore it also converges to 0. Because the mean of (2.26) is equal to 0 and its variance converges to 0, Chebyshev's inequality shows that, for any $\epsilon > 0$,

$$\begin{aligned}
&P_{\bar{\theta}^*,\hbar^*}\left(\left|n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1/2}\sum_{i=1}^n \left[U_{\hbar_n}(\bar{\theta}^*; Z_i) - U_{\hbar^*}(\bar{\theta}^*; Z_i)\right]\right\}\right| > \varepsilon\right) \\
&\leq \frac{\operatorname{var}\left(g_{p+2} n^{-1/2}\sum_{i=1}^n \left\{[\hbar_n(W_i) - h(W_i)]\frac{A_i - \pi_{\bar{\theta}^*,\hbar^*}(W_i)}{\pi_{\bar{\theta}^*,\hbar^*}(W_i)\left(1-\pi_{\bar{\theta}^*,\hbar^*}(W_i)\right)}\rho(\bar{\theta}^{*\top} r_{\hbar^*}(W_i))\right\}\right)}{\epsilon^2} \\
&\overset{n\to\infty}{\longrightarrow} 0. \qquad (2.28)
\end{aligned}$$

Next, by the central limit theorem, under $P_{\bar{\theta}^*,\hbar^*}$,

$$n^{-1/2}\sum_{i=1}^n U_{\hbar^*}(\bar{\theta}^*; Z_i) \rightsquigarrow N\left(\vec{0}, I_{\bar{\theta}^*}\right). \qquad (2.29)$$

By continuous mapping theorem,

$$n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1/2}\sum_{i=1}^n U_{\hbar^*}(\bar{\theta}^*; Z_i)\right\} \rightsquigarrow N\left(0, g^\top I_{\bar{\theta}^*,\hbar^*} g\right).$$

Combining the arguments shown in (2.28) and (2.29), we have that, under $P_{\bar{\theta}^*,\hbar^*}$,

$$(\theta_n - \bar{\theta}^*)^\top \left\{\sum_{i=1}^n U_{\hbar_n}(\bar{\theta}^*; Z_i)\right\} \rightsquigarrow N\left(0, g^\top I_{\bar{\theta}^*,\hbar^*} g\right). \qquad (2.30)$$

This concludes the study of the leading term in (2.25). Now, the second term in (2.25) can be reexpressed as

$$\begin{aligned}
&\frac{1}{2} n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1}\sum_{i=1}^n K_{\hbar_n}(\theta'(D_n); Z_i)\right\} n^{1/2}(\theta_n - \bar{\theta}^*) \\
&= \frac{1}{2} n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left\{n^{-1}\sum_{i=1}^n K_{\hbar^*}(\bar{\theta}^*; Z_i)\right\} n^{1/2}(\theta_n - \bar{\theta}^*) \qquad (2.31) \\
&+ \frac{1}{2} n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left[n^{-1}\sum_{i=1}^n \left\{K_{\hbar_n}(\bar{\theta}^*; Z_i) - K_{\hbar^*}(\bar{\theta}^*; Z_i)\right\}\right] n^{1/2}(\theta_n - \bar{\theta}^*) \\
&+ \frac{1}{2} n^{1/2}(\theta_n - \bar{\theta}^*)^\top \left[n^{-1}\sum_{i=1}^n \left\{K_{\hbar_n}(\theta'(D_n); Z_i) - K_{\hbar_n}(\bar{\theta}^*; Z_i)\right\}\right] n^{1/2}(\theta_n - \bar{\theta}^*).
\end{aligned}$$

By the weak law of large numbers, $n^{-1}\sum_{i=1}^{n} K_{\hbar^*}(\bar{\theta}^*; Z_i) \longrightarrow -I_{\bar{\theta}^*\hbar^*}$ in probability under $P_{\bar{\theta}^*,\hbar^*}$. Therefore, by the continuous mapping theorem,

$$P_{\bar{\theta}^*,\hbar^*}\left(\left|\frac{1}{2}n^{1/2}(\theta_n-\bar{\theta}^*)^\top\left\{n^{-1}\sum_{i=1}^{n}K_{\hbar^*}(\bar{\theta}^*;Z_i)\right\}n^{1/2}(\theta_n-\bar{\theta}^*)+\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar^*}g\right|>\varepsilon\right)\overset{n\to\infty}{\longrightarrow}0 \tag{2.32}$$

Next, note that for all z,

$$\begin{aligned}K_{\hbar_n}(\bar{\theta}^*;z) &= [\pi_{\bar{\theta}^*,\hbar_n}(w)][1-\pi_{\bar{\theta}^*,\hbar_n}(w)]r_{\hbar_n}(w)r_{\hbar_n}(w)^\top \\ &= [\pi_{\bar{\theta}^*,\hbar^*}(w)][1-\pi_{\bar{\theta}^*,\hbar^*}(w)]r_{\hbar_n}(w)r_{\hbar_n}(w)^\top,\end{aligned}$$

where we have used that $\bar{\theta}^*_{p+2}=0$ implies that $\pi_{\bar{\theta}^*,\hbar}=\pi_{\bar{\theta}^*,\hbar^*}$ for any $\hbar\in\mathcal{H}$. Hence,

$$\begin{aligned}&\frac{1}{2}n^{1/2}(\theta_n-\bar{\theta}^*)^\top\left[n^{-1}\sum_{i=1}^{n}\left\{K_{\hbar_n}(\bar{\theta}^*;Z_i)-K_{\hbar^*}(\bar{\theta}^*;Z_i)\right\}\right]n^{1/2}(\theta_n-\bar{\theta}^*)= \\ &=\frac{g^\top}{2}\left[n^{-1}\sum_{i=1}^{n}\left(r_{\hbar_n}(W_i)r_{\hbar_n}(W_i)^\top-r_{\hbar^*}(W_i)r_{\hbar^*}(W_i)^\top\right)\left[\pi_{\bar{\theta}^*,\hbar^*}(W_i)(1-\pi_{\bar{\theta}^*,\hbar^*}(W_i))\right]\right]g.\end{aligned}$$

We now show that the above term converges to 0 in probability by showing that the expected value of each entry of the matrix

$$\left|n^{-1}\sum_{i=1}^{n}\left(r_{\hbar_n}(w_i)r_{\hbar_n}(W_i)^\top-r_{\hbar^*}(w_i)r_{\hbar^*}(w_i)^\top\right)\left[\pi_{\bar{\theta}^*,\hbar^*}(W_i)(1-\pi_{\bar{\theta}^*,\hbar^*}(W_i))\right]\right|$$

converges to 0, where $|\cdot|$ denotes the entrywise absolute value. First, note that for a given w, the matrix $r_{\hbar_n}(w)r_{\hbar_n}(w)^\top - r_{\hbar^*}(w)r_{\hbar^*}(w)^\top$ consists of the following entries:

$$\left(r_{\hbar_n}(w)r_{\hbar_n}(w)^\top-r_{\hbar^*}(w)r_{\hbar^*}(w)^\top\right)_{l,k}=\begin{cases}0 & l\le p,\ k\le p\\ w_k\left(\hbar_n(w)-\hbar^*(w)\right) & l=p+2,\ 1\le k\le p+1\\ w_l\left(\hbar_n(w)-\hbar^*(w)\right) & k=p+2,\ 1\le l\le p+1\\ \hbar_n(w)^2-\hbar^*(w)^2 & k=p+2,\ l=p+2\end{cases}$$

where w_j is the $j-th$ observed covariate. For those entries that are different from 0, we

apply Cauchy-Schwartz inequality:

$$E\left|\left(r_{\hbar_n}(W)r_{\hbar_n}(W)^\top - r_{\hbar^*}(W)r_{\hbar^*}(W)^\top\right)_{l,k}\left[\pi_{\bar\theta^*,\hbar^*}(W)(1-\pi_{\bar\theta^*,\hbar^*}(W))\right]\right|$$
$$\leq E\left\{\left(r_{\hbar_n}(W)r_{\hbar_n}(W)^\top - r_{\hbar^*}(W)r_{\hbar^*}(W)^\top\right)^2_{l,k}\right\}^{1/2} E\left\{\left[\pi_{\bar\theta^*,\hbar^*}(W)(1-\pi_{\bar\theta^*,\hbar^*}(W))\right]^2\right\}^{1/2}$$

Because $\|\hbar_n - \hbar^*\|_{L^2(P_{\theta^*,\hbar^*})} \to 0$ and functions in $\mathcal{H}$ are uniformly bounded, we have that $E\{[\hbar_n(W)^2 - \hbar^*(W)^2]^2\} = E\{[\hbar_n(W)+\hbar^*(W)]^2[\hbar_n(W)-\hbar^*(W)]^2\} \overset{n\to\infty}{\longrightarrow} 0$. Additionally, because $\mathcal{W}$ is bounded, $E\{[W_l(\hbar_n(W)-\hbar^*(W))]^2\} \overset{n\to\infty}{\longrightarrow} 0$ for all $l \in \{1,...,p+1\}$. The latter statements combined with the fact that that $E\left\{\left[\pi_{\bar\theta^*,\hbar^*}(W)(1-\pi_{\bar\theta^*,\hbar^*}(W))\right]^2\right\} < 1$ imply that

$$E\left[\left(r_{\hbar_n}(W)r_{\hbar_n}(W)^\top - r_{\hbar^*}(W)r_{\hbar^*}(W)^\top\right)^2_{l,k}\right]^{1/2} E\left\{\left[\pi_{\bar\theta^*,\hbar^*}(W)(1-\pi_{\bar\theta^*,\hbar^*}(W))\right]^2\right\}^{1/2} \overset{n\to\infty}{\longrightarrow} 0.$$

Let $0_{(p+2)\times(p+2)}$ denote a $(p+2)\times(p+2)$ matrix of 0's. Then, the above limit and Markov's inequality imply that

$$n^{-1}\sum_{i=1}^n \left(r_{\hbar_n}(W_i)r_{\hbar_n}(W_i)^\top - r_{\hbar^*}(W_i)r_{\hbar^*}(W_i)^\top\right)\left[\pi_{\bar\theta^*,\hbar^*}(W_i)(1-\pi_{\bar\theta^*,\hbar^*}(W_i))\right] \overset{n\to\infty}{\longrightarrow} 0_{(p+2)\times(p+2)}$$

in probability under $P_{\bar\theta^*,\hbar^*}$. By the continuous mapping theorem,

$$\frac{1}{2}n^{1/2}(\theta_n-\bar\theta^*)^\top\left[n^{-1}\sum_{i=1}^n\left\{K_{\hbar_n}(\bar\theta^*;Z_i) - K_{\hbar^*}(\bar\theta^*;Z_i)\right\}\right]n^{1/2}(\theta_n-\bar\theta^*) = o_{P_{\bar\theta^*,\hbar^*}}(1). \quad (2.33)$$

Finally, we show that the last term in Equation 2.31 converges to 0 in probability. First note that this term is equal to

$$\frac{1}{2}n^{1/2}(\theta_n-\bar\theta^*)^\top\left[n^{-1}\sum_{i=1}^n\left\{K_{\hbar_n}(\theta'(D_n);Z_i) - K_{\hbar_n}(\bar\theta^*;Z_i)\right\}\right]n^{1/2}(\theta_n-\bar\theta^*)$$
$$= \frac{g^\top}{2}\left[n^{-1}\sum_{i=1}^n r_{\hbar_n}(W_i)r_{\hbar_n}(W_i)^\top\left(\left[\pi_{\theta'(D_n),\hbar_n}(W_i)(1-\pi_{\theta'(D_n),\hbar_n}(W_i))\right] - \left[\pi_{\bar\theta^*,\hbar_n}(W_i)(1-\pi_{\bar\theta^*,\hbar_n}(W_i))\right]\right)\right]g.$$

Fix $w \in \mathcal{W}$ and $\hbar \in \mathcal{H}$ and define the $\mathbb{R}^{p+2} \to \mathbb{R}$ function $v_{w,\hbar}(\theta) := \pi_{\theta,\hbar}(w_i)(1-\pi_{\theta,\hbar}(w_i))$. We show that v is Lipschitz continuous in θ uniformly over $\hbar \in \mathcal{H}$ and $w \in \mathcal{W}$. We do so

by showing that its derivative is uniformly upper bounded entry-wise. The absolute value of the derivative is

$$\left|\frac{\partial}{\partial\theta}v_{w,\hbar}(\theta)\right| = \left|\frac{\partial}{\partial\theta}\pi_{\theta,\hbar}(w) - \frac{\partial}{\partial\theta}\pi_{\theta,\hbar}(w)^2\right| = \left|\rho(\theta^\top r_\hbar(w))r_\hbar(w) - 2\pi_{\theta,\hbar}(w)\rho(\theta^\top r_\hbar(w))r_\hbar(w)\right|$$
$$= \left|\rho(\theta^\top r_\hbar(w))r_\hbar(w)[1-2\pi_{\theta,\hbar}(w)]\right|.$$

By definition $|\rho(\theta^\top r_\hbar(w))| < 1$ for all $w \in \mathcal{W}$ and $\hbar \in \mathcal{H}$. Next, because $w \in \mathcal{W}$ and $\hbar \in \mathcal{H}$ are bounded, we can upper bound each entry of $r_\hbar(w)$ by a finite constant L that depends neither on w nor $\hbar$. Finally, because $0 < \pi_{\theta,\hbar}(w) < 1$ regardless of w or $\hbar$, we have that $|1 - 2\pi_{\theta,\hbar}(w)| \leq 1$. Hence, for all θ,

$$\sup_{w\in\mathcal{W},\hbar\in\mathcal{H}}\left|\frac{\partial}{\partial\theta}v_{w,\hbar}(\theta)\right| \leq L,$$

which implies that $v_{w,\hbar}(\cdot)$ is L-Lipschitz uniformly over $w \in \mathcal{W}$ and $\hbar \in \mathcal{H}$. Combining the preceding arguments,

$$E\left|\frac{g^\top}{2}\left[n^{-1}\sum_{i=1}^n r_{\hbar_n}(W_i)r_{\hbar_n}(W_i)^\top\left(\left[\pi_{\theta'(D_n),\hbar_n}(W_i)(1-\pi_{\theta'(D_n),\hbar_n}(W_i))\right] - \left[\pi_{\bar\theta^*,\hbar_n}(W_i)(1-\pi_{\bar\theta^*,\hbar_n}(W_i))\right]\right)\right]g\right|$$
$$\leq L^2||g||^2 E\left[||\theta'(D_n) - \bar\theta^*||\right].$$

By Markov's inequality and the above display,

$$P_{\bar\theta^*,\hbar^*}\left(\left|\frac{1}{2}n^{1/2}(\theta_n-\bar\theta^*)^\top\left[n^{-1}\sum_{i=1}^n\left\{K_{\hbar_n}(\theta'(D_n);Z_i) - K_{\hbar_n}(\bar\theta^*;Z_i)\right\}\right]n^{1/2}(\theta_n-\bar\theta^*)\right| > \varepsilon\right) \tag{2.34}$$
$$\leq \varepsilon^{-1}L^2\|g\|^2 E\left[\|\theta'(D_n)-\bar\theta^*\|\right] \leq \varepsilon^{-1}L^2\|g\|^2\|\theta_n-\bar\theta^*\| = \varepsilon^{-1}L^2\frac{\|g\|^3}{\sqrt{n}} \overset{n\to\infty}{\longrightarrow} 0.$$

Therefore, by Equations 2.31, 2.32, 2.33, and 2.34, we have that

$$P_{\bar\theta^*,\hbar^*}\left(\left|\frac{1}{2}n^{1/2}(\theta_n-\bar\theta^*)^\top\left\{n^{-1}\sum_{i=1}^n K_{\hbar_n}(\theta'(D_n);Z_i)\right\}n^{1/2}(\theta_n-\bar\theta^*) + \frac{1}{2}g^\top I_{\bar\theta^*,\hbar^*}g\right| > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0 \tag{2.35}$$

Combining the equality in 2.25 along with the results stated in Equations 2.30 and 2.35, implies that

$$\log\left\{\prod_{i=1}^{n}\frac{p_{\theta_n,\hbar_n}(Z_i)}{p_{\bar{\theta}^*,\hbar^*}(Z_i)}\right\} \rightsquigarrow N\left(-\frac{1}{2}g^\top I_{\bar{\theta}^*,\hbar^*}g, g^\top I_{\bar{\theta}^*,\hbar^*}g\right)$$

under $P_{\bar{\theta}^*,\hbar^*}$. Hence, by Le Cam's first lemma — see Example 6.5 in van der Vaart (2000) for an appropriate version — $(P^n_{\theta_n,\hbar_n})_{n=1}^\infty$ and $(P^n_{\bar{\theta}^*,\hbar^*})_{n=1}^\infty$ are mutually contiguous. □

The following lemmas describe the asymptotic behavior of the maximum likelihood estimator of $\bar{\theta}^*$ along the sequence of models $\{P_{\theta,\hbar_n} : \theta \in \mathbb{R}^{p+2}\}_{n=1}^\infty$. Recall that $\ell_{\hbar_n}(\theta; d_n)$ is the log-likelihood function given d_n and $\hbar_n \in (\hbar)_{n=1}^\infty$,

$$\ell_{\hbar_n}(\theta; d_n) := \frac{1}{n}\sum_{i=1}^{n}\left[a_i \log \pi_{\theta,\hbar_n}(w_i) + (1-a_i)\log(1-\pi_{\theta,\hbar_n}(w_i))\right]. \tag{2.36}$$

Then, the maximum likelihood estimator of θ based on d_n and $\hbar_n$ is defined as

$$\tilde{\theta}_{n,\hbar_n}(d_n) := \underset{\theta}{\operatorname{argmax}}\, \ell_{\hbar_n}(\theta; d_n),$$

where this maximizer exists and is unique with probability one under our conditions. Finally, let $L(\theta, \hbar)$ denote the expectation of $\ell_n(\theta, \hbar, D_n)$,

$$L(\theta, \hbar) := E\left[A \log \pi_{\theta,\hbar}(W) + (1-A)\log(1-\pi_{\theta,\hbar}(W))\right].$$

The following lemma shows that $\tilde{\theta}_{n,\hbar_n}(D_n)$ is consistent for $\bar{\theta}^*$.

Lemma 2.9. *Assume that the $\mathcal{H}$-valued sequence $(\hbar_n)_{n=1}^\infty$ is such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$ as $n \to \infty$ and that the conditions of Proposition 2.2 hold. Let $\tilde{\theta}_{n,\hbar_n}(D_n) = \operatorname{argmax}_\theta \ell_{\hbar_n}(\theta; D_n)$ be the maximum likelihood estimator of $\bar{\theta}^*$. Then, for every $\epsilon > 0$,*

$$P_{\bar{\theta}^*,\hbar^*}\left(||\tilde{\theta}_{n,\hbar_n}(D_n) - \bar{\theta}^*|| \geq \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{2.37}$$

Proof. Let D_n be a dataset arising as n iid draws from $P_{\bar{\theta}^*,\hbar^*}$. It can be shown that, under the conditions of this lemma, it is true that (i) $L(\cdot, \hbar^*)$ is uniquely maximized at $\bar{\theta}^*$, (ii) $\ell_{\hbar_n}(\cdot, D_n)$ is concave almost surely, (iii) $\ell_{\hbar_n}(\theta, D_n) \to L(\theta, \hbar^*)$ in probability for all $\theta \in \mathbb{R}^{p+2}$ — details are omitted here for brevity. Hence, Theorem 2.7 in Newey and McFadden (1994) gives the desired result. □

One can also show that the maximum likelihood estimator based on $\hbar^*$, $\tilde{\theta}_{n,\hbar^*}(D_n)$, is consistent for $\bar{\theta}^*$. We omit the proof for brevity, and because it is very similar to the one shown Lemma 2.9. Next, the following lemma shows that the maximum likelihood estimator $\tilde{\theta}_{n,\hbar_n}(D_n)$ is asymptotically linear. For brevity, in the remaining of this supplement we omit the dependency on d_n by writing $\tilde{\theta}_{n,\hbar}$ rather than $\tilde{\theta}_{n,\hbar}(d_n)$.

Lemma 2.10. *Assume that the $\mathcal{H}$-valued sequence $(\hbar_n)_{n=1}^{\infty}$ is such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$ as $n \to \infty$ and that the conditions of Proposition 2.2 hold. Let $\tilde{\theta}_{n,\hbar_n} = \operatorname{argmax}_\theta \ell_{\hbar_n}(\theta; d_n)$ be the maximum likelihood estimator of $\bar{\theta}^*$, $g \in \mathbb{R}^{p+2}$, and $\theta_n = \bar{\theta}^* + g/\sqrt{n}$. Then, under sampling from $P^n_{\theta_n,\hbar_n}$,*

$$\sqrt{n}[\tilde{\theta}_{n,\hbar_n} - \theta_n] = I^{-1}_{\bar{\theta}^*,\hbar^*} \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) + o_p(1).$$

Proof. We first show that, under $P_{\bar{\theta}^*,\hbar^*}$, the following holds:

$$\sqrt{n}[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*] = I^{-1}_{\bar{\theta}^*,\hbar^*} \frac{1}{\sqrt{n}} U_{\hbar^*}(\bar{\theta}^*; D_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1).$$

Let $Z_0(\theta, \hbar) := \frac{\partial}{\partial\theta} L(\theta, \hbar)$, $\tilde{\theta}_{n,\hbar^*} := \operatorname{argmax}_\theta \ell_{\hbar^*}(\theta; d_n)$ be the maximum likelihood estimator when $\hbar^*$ is known, and $Z_n(\theta, \hbar) := U_\hbar(\theta, d_n)$. Applying Triangle Inequality and by definition of the MLE's we have that

$$\begin{aligned}
\|Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_n(\tilde{\theta}_{n,\hbar^*}, \hbar^*)\| &= \|Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar_n)\| \\
&\leq \|Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar^*) + Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar_n) - Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar_n)\| \\
&\quad + \|Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar_n)\|. \qquad (2.38)
\end{aligned}$$

We begin by showing that, for every $\epsilon > 0$,

$$P_{\theta^*,\hbar^*}\left(\|Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar^*) + Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar_n) - Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar_n)\| > n^{-1/2}\epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \quad (2.39)$$

By consistency of $\tilde{\theta}_{n,\hbar_n}$ and the assumption on $(\hbar_n)_{n=1}^{\infty}$ we know that there exists a sequence δ_n that converges to zero sufficiently slowly so that $P_{\bar{\theta}^*,\hbar^*}\left\{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| > \delta_n\right\} \to 0$ and $\|\hbar_n - \hbar^*\| \leq \delta_n$. Thus, it suffices to show that

$$P_{\bar{\theta}^*,\hbar^*}\left(\sup_{\theta \in D_{\delta_n}} \|Z_n(\theta, \hbar^*) - Z_0(\theta, \hbar^*) + Z_0(\theta, \hbar_n) - Z_n(\theta, \hbar_n)\| > n^{-1/2}\epsilon\right) \overset{n\to\infty}{\longrightarrow} 0.$$

where $D_\delta = \{\theta \in \mathbb{R}^{p+2} : \|\theta - \bar{\theta}^*\| \leq \delta\}$. Now, because

$$\begin{aligned}
&\sup_{\theta \in D_{\delta_n}} \|Z_n(\theta, \hbar^*) - Z_0(\theta, \hbar^*) + Z_0(\theta, \hbar_n) - Z_n(\theta, \hbar_n)\| \\
&= \sup_{\theta \in D_{\delta_n}} \|Z_n(\theta, \hbar^*) - Z_0(\theta, \hbar^*) + [Z_n(\bar{\theta}^*, \hbar^*) - Z_n(\bar{\theta}^*, \hbar^*)] + Z_0(\theta, \hbar_n) - Z_n(\theta, \hbar_n)| \\
&\leq \sup_{\theta \in D_{\delta_n}} \|Z_n(\theta, \hbar^*) - Z_n(\bar{\theta}^*, \hbar^*) - Z_0(\theta, \hbar^*))\| + \sup_{\theta \in D_{\delta_n}} \|Z_n(\bar{\theta}^*, \hbar^*) - Z_n(\theta, \hbar_n) + Z_0(\theta, \hbar_n)\|
\end{aligned}$$

and since $\hbar^* \in \mathcal{H}_{\delta_n}$, we can instead show that

$$\sup_{\hbar \in \mathcal{H}_{\delta_n}} P_{\bar{\theta}^*, \hbar^*} \left(\sup_{\theta \in D_{\delta_n}} \|Z_n(\theta, \hbar) - Z_n(\bar{\theta}^*, \hbar^*) - Z_0(\theta, \hbar)\| > n^{-1/2}\epsilon \right) \overset{n \to \infty}{\longrightarrow} 0. \tag{2.40}$$

Fix $\hbar \in \mathcal{H}_{\delta_n}$ and let

$$f_{\theta,\hbar}(z) := a\frac{\rho(\theta^\top r_\hbar(w))}{\pi_{\theta,\hbar}(w)} r_\hbar(w) + (a-1)\frac{\rho(\theta^\top r_\hbar(w))}{1 - \pi_{\theta,\hbar}(w)} r_\hbar(w) - a\frac{\rho(\theta^{*\top} r_{\hbar^*}(w))}{\pi_{\theta^*,\hbar^*}(w)} r_\hbar(w) - (a-1)\frac{\rho(\theta^{*\top} r_{\hbar^*}(w))}{1 - \pi_{\theta^*,\hbar^*}(w)} r_{\hbar^*}(w).$$

Additionally, let $(f_{\theta,\hbar}(z))_k$ denote the k-th entry of the function $f_{\theta,\hbar}(\cdot)$ and $(M)_k$ denote the k-th column of a matrix M. Further, we also define the class of functions $\mathcal{F}^k_{\hbar,\delta} = \{z \mapsto f^k_{\theta,\hbar}(z) : \theta \in \mathbb{R}^{p+2}, \theta \in D_\delta\}$, and let $N_{[]}(\epsilon, \mathcal{F}^k_{\hbar,\delta}, L_2(P))$ denote its bracketing number of radius ϵ with respect to the $L_2(P)$ norm. We will show that $E[|\sup_{f \in \mathcal{F}^k_{\delta_n,\hbar}} \mathbb{G}_n f|] \longrightarrow 0$ as $n \to \infty$. Fix $z \in \mathcal{Z}$ and $k \in \{1, ..., p+2\}$ and let $\nu^k_{z,\hbar} : \theta \mapsto (f_{\theta,\hbar}(z))_k$ be a function that maps θ to the k-th entry of $f_{\theta,\hbar}(z)$. Note that each entry of the derivative

$$\left| \frac{\partial}{\partial \theta} \nu^k_{\theta,\hbar} \right| = \left| (\rho(\theta^\top r_\hbar(w)) r_\hbar(w) r_\hbar(w)^\top)_k \right|$$

can be uniformly bounded by a finite constant $C > 0$ because both $\mathcal{W}$ and $\mathcal{H}$ are bounded. Therefore, the function $\nu^k_{z,\hbar}(\theta)$ is Lipschitz in θ. Hence, for $\theta_1, \theta_1 \in D_{\delta_n}$ we have that

$$|f^k_{\theta_1,\hbar}(z) - f^k_{\theta_2,\hbar}(z)| \leq C||\theta_1 - \theta_2|| \leq 2C\delta_n \tag{2.41}$$

Then, by Example 19.7 in van der Vaart (2000), we have that there exists a constant K that depends on p and such that

$$N_{[]}(\epsilon C, \mathcal{F}^k_{\hbar,\delta_n}, L_2(P)) \leq K \left(\frac{2\delta_n}{\epsilon} \right)^{p+2}, \text{ for every } 0 < \epsilon < 2\delta_n.$$

Next, note that if $\theta = \bar{\theta}^*$, then $f^k_{\theta,\hbar} = 0$. Applying the bound in Equation 2.41 with $\theta_2 = \bar{\theta}^*$, we have that $|f^k_{\theta_1,\hbar}(z)| \leq C||\theta_1 - \bar{\theta}^*|| \leq C\delta_n$. Because the previous bound is irrespective of θ_1 and $\hbar$, we have that for every $z \in \mathcal{Z}$, $\sup_{\hbar\in\mathcal{H}} \sup_{f_1\in\mathcal{F}^k_{\hbar,\delta_n}} \sup_z |f_1(z)| \leq C\delta_n$. Then, let $F^k_n(z) = C\delta_n$ be the envelope function of the function class $\mathcal{F}^k_{\hbar,\delta_n}$. Therefore, for every $0 < \epsilon < 2$, the bracketing integral is finite because it can be upper bounded as follows,

$$\int_0^1 \sqrt{1 + \log N_{[]}(\epsilon C\delta_n, \mathcal{F}^k_{\hbar,\delta_n}, L_2(P))}d\epsilon \leq \int_0^1 \sqrt{1 + \log K + (p+2)\log\left(\frac{2}{\epsilon}\right)}d\epsilon =: K' < \infty.$$

By Theorem 2.14.2 in van der Vaart and Wellner (1996), $E|\sup_{f\in\mathcal{F}^k_{\hbar,\delta_n}} \mathbb{G}_n f| \leq K'C\delta_n \overset{n\to\infty}{\longrightarrow} 0$. Because the index k was arbitrary and the upper bound $K'C\delta_n$ does not depend on $\hbar$, we can establish that (2.40) holds via Markov's inequality. Next, we show that

$$\|Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_0(\tilde{\theta}_{n,\hbar_n}, \hbar_n)\| = o_p(\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|). \tag{2.42}$$

Using similar notation for multivariate functions as before, we will let $Z^k_0(\theta, \hbar)$ denote the k-th value of one of its outputs. First note that,

$$Z^k_0(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z^k_0(\tilde{\theta}_{n,\hbar_n}, \hbar_n) = Z^k_0(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z^k_0(\bar{\theta}^*, \hbar^*) + Z^k_0(\bar{\theta}^*, \hbar_n) - Z^k_0(\tilde{\theta}_{n,\hbar_n}, \hbar_n)$$

since $Z^k_0(\bar{\theta}^*, \hbar^*) = Z^k_0(\bar{\theta}^*, \hbar_n) = 0$. Next, by the fundamental theorem of calculus, we have that, for any $\hbar$, θ,

$$Z^k_0(\theta, \hbar) - Z^k_0(\bar{\theta}^*, \hbar) = \int_0^{\|\theta-\bar{\theta}^*\|} \frac{\partial}{\partial\delta} Z^k_0\left(\bar{\theta}^* + \delta\frac{\theta - \bar{\theta}^*}{\|\theta - \bar{\theta}^*\|}, \hbar\right)\Big|_{\delta=\epsilon} d\epsilon.$$

Hence,

$$
\begin{aligned}
&Z_0^k(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_0^k(\bar{\theta}^*, \hbar^*) + Z_0^k(\bar{\theta}^*, \hbar_n) - Z_0^k(\tilde{\theta}_{n,\hbar_n}, \hbar_n) \\
&= \int_0^{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|} \frac{\partial}{\partial \delta} \left[Z_0^k \left(\bar{\theta}^* + \delta \frac{\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*}{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|}, \hbar^* \right) - Z_0^k \left(\bar{\theta}^* + \delta \frac{\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*}{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|}, \hbar_n \right) \right] \Bigg|_{\delta=\epsilon} d\epsilon \\
&\leq \int_0^{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|} \left| \frac{\partial}{\partial \delta} \left[Z_0^k \left(\bar{\theta}^* + \delta \frac{\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*}{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|}, \hbar^* \right) - Z_0^k \left(\bar{\theta}^* + \delta \frac{\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*}{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|}, \hbar_n \right) \right] \Bigg|_{\delta=\epsilon} \right| d\epsilon \\
&\leq \|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| \sup_{\epsilon \in [0,1]} \left| \frac{\partial}{\partial \delta} \left[Z_0^k \left(\bar{\theta}^* + \delta \frac{[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*]}{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|}, \hbar^* \right) - Z_0^k \left(\bar{\theta}^* + \delta [\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar_n \right) \right] \Bigg|_{\delta=\epsilon} \right| \\
&= \|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| \sup_{\epsilon \in [0,1]} \left| \frac{[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*]}{\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|} \left[(I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar_n} - I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar^*})_k \right] \right| \\
&\leq \sup_{\epsilon \in [0,1]} \|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| \|(I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar_n} - I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar^*})_k\| \\
&= \|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| \sup_{\epsilon \in [0,1]} \|(I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar_n} - I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar^*})_k\| \\
&\leq \|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|^2 \sup_{\epsilon \in [0,1]} \|(I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar_n} - I_{\bar{\theta}^* + \epsilon[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*], \hbar^*})_k\|
\end{aligned}
$$

By similar arguments outlined in Lemma 2.6, one can show that there exists a neighborhood of $\bar{\theta}^*$ in which $\hbar \to I_{\theta,\hbar}$ is Lipschitz with constant M, where M does not depend on the value of θ. Hence,

$$
Z_0^k(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_0^k(\bar{\theta}^*, \hbar^*) + Z_0^k(\bar{\theta}^*, \hbar_n) - Z_0^k(\tilde{\theta}_{n,\hbar_n}, \hbar_n) \leq M \|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\|^2 \|\hbar_n - \hbar^*\|_{L_1(P)} \quad (2.43)
$$

implies Equation 2.42 by Markov's inequality. The inequality in Equation 2.38, along with the results from Equations 2.39 and 2.42 imply that

$$
\|Z_n(\tilde{\theta}_{n,\hbar_n}, \hbar^*) - Z_n(\tilde{\theta}_{n,\hbar^*}, \hbar^*)\| = o_{P_{\bar{\theta}^*,\hbar^*}}(\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| + n^{-1/2}). \quad (2.44)
$$

Next, note that by the boundedness of $\mathcal{Z}$ and $\mathcal{H}$, $E[\|Z_0(\theta, \hbar_n)\|^2] < \infty$. Moreover, by Lemma 2.4 along with regularity conditions of the model $\{P_{\theta,\hbar} : \theta \in \mathbb{R}^{p+2}, \hbar \in \mathcal{H}\}$ the mapping $\theta \mapsto E[U_{\hbar^*}(\theta, z)]$ is differentiable at $\bar{\theta}^*$ with derivative matrix $-I_{\bar{\theta}^*,\hbar^*}$, which is invertible by assumption. These properties, along with the Lipschitz result shown in equation (2.40) and

the consistency of $\tilde{\theta}_{n,\hbar_n}$ imply that the conditions in Theorem 5.21 in van der Vaart (2000) are satisfied, with the exception that our rate in Equation 2.44 is not necessarily $o_{P_{\bar{\theta}^*,\hbar^*}}(n^{-1/2})$, but instead $o_{P_{\bar{\theta}^*,\hbar^*}}(\|\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*\| + n^{-1/2})$. However, by inspecting the proof of Theorem 5.21 in that reference, it is clear that — in the notation of that work — this result holds under our condition that $\mathbb{P}_n\psi_{\hat{\theta}_n} = o_P(\|\hat{\theta}_n - \theta_0\| + n^{-1/2})$, which holds in our setting. Therefore, we have that

$$\sqrt{n}[\tilde{\theta}_{n,\hbar_n} - \bar{\theta}^*] = -I^{-1}_{\bar{\theta}^*,\hbar^*}\frac{1}{\sqrt{n}}U_{\hbar^*}(\bar{\theta}^*; D_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1). \tag{2.45}$$

By the results shown in Lemma 2.8 (particularly the identity in Equation 2.26 and the convergence result in Equation 2.28), it is true that

$$\frac{1}{\sqrt{n}}U_{\hbar^*}(\bar{\theta}^*; D_n) = \frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*; D_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1),$$

and so, by the continuous mapping theorem,

$$I^{-1}_{\bar{\theta}^*,\hbar^*}\frac{1}{\sqrt{n}}U_{\hbar^*}(\bar{\theta}^*; D_n) = I^{-1}_{\bar{\theta}^*,\hbar^*}\frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*; D_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1). \tag{2.46}$$

Combining Equations 2.45 and 2.46, and because $\theta_n = \bar{\theta}^* + g/\sqrt{n}$, it follows that

$$\sqrt{n}[\tilde{\theta}_{n,\hbar_n} - \theta_n] = I^{-1}_{\bar{\theta}^*,\hbar^*}\frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*; D_n) - g + o_{P_{\bar{\theta}^*,\hbar^*}}(1).$$

Using very similar techniques as those that show Lemma 2.7 and the fact that $I_{\bar{\theta}^*,\hbar_n} \longrightarrow I_{\bar{\theta}^*,\hbar^*}$, it can be shown that

$$\frac{1}{\sqrt{n}}U_{\hbar_n}(\theta_n; D_n) - \frac{1}{\sqrt{n}}U_{\hbar_n}(\bar{\theta}^*; D_n) + I_{\bar{\theta}^*,\hbar^*}g = o_{P_{\bar{\theta}^*,\hbar^*}}(1),$$

which implies that

$$\frac{1}{\sqrt{n}}I^{-1}_{\bar{\theta}^*,\hbar^*}U_{\hbar_n}(\bar{\theta}^*; D_n) - g = \frac{1}{\sqrt{n}}I^{-1}_{\bar{\theta}^*,\hbar^*}U_{\hbar_n}(\theta_n; D_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1).$$

Hence,

$$\sqrt{n}[\tilde{\theta}_{n,\hbar_n} - \theta_n] = I^{-1}_{\bar{\theta}^*,\hbar^*}\frac{1}{\sqrt{n}}U_{\hbar_n}(\theta_n; D_n) + o_{P_{\bar{\theta}^*,\hbar^*}}(1).$$

Lemma 2.8 shows that $(P^n_{\theta_n,\hbar_n})_{n=1}^{\infty}$ is contiguous with respect to $(P^n_{\bar{\theta}^*,\hbar^*})_{n=1}^{\infty}$, and applying this to the above yields the desired result. □

The following lemmas pertain to the asymptotic distribution of the matching estimator. Given $\theta \in \mathbb{R}^{p+2}$, $\hbar \in \mathcal{H}$, and a sample D_n drawn from $P_{\theta,\hbar}$ and consisting of n iid observations, the matching estimator $\hat{\tau}(\theta, \hbar)$ is defined as

$$\hat{\tau}(\theta, \hbar) = \frac{1}{n} \sum_{i=1}^{n} (2a_i - 1) \left(y_i - \frac{1}{M} \sum_{j=1}^{n} I(j \in \mathcal{J}_M(i, \theta, \hbar)) y_j \right), \tag{2.47}$$

where $\mathcal{J}_M(i, \theta, \hbar)$ is the matched set of observation i based on the propensity score $\pi_{\theta,\hbar}(\cdot)$:

$$\mathcal{J}_M(i, \theta, \hbar) = \left\{ j : a_i = 1 - a_j, \sum_{k=1}^{n} I(a_i = 1 - a_k) I\left(|\pi_{\theta,\hbar}(w_j) - \pi_{\theta,\hbar}(w_i)| \leq |\pi_{\theta,\hbar}(w_k) - \pi_{\theta,\hbar}(w_i)| \right) \leq M \right\}.$$

If we let $K_{M,\theta,\hbar}(i)$ denote the number of times observation i is used as a match,

$$K_{M,\theta,\hbar}(i) = \sum_{j=1}^{n} I(i \in \mathcal{J}(j, \theta, \hbar)),$$

then the matching estimator can be represented as

$$\hat{\tau}(\theta, \hbar) = \frac{1}{n} \sum_{i=1}^{n} (2A_i - 1) \left(1 + \frac{K_{M,\theta,\hbar}(i)}{M} \right) y_i.$$

Let $\bar{\mu}_{\theta,\hbar}(a, p) := E_{\theta,\hbar}[Y|A = a, \pi_{\theta,\hbar}(W) = p]$ denote the conditional expectation — under $P_{\theta,\hbar}$ — of Y given a and propensity score p, and let $\mu_{\theta,\hbar}(a, w) := E_{\theta,\hbar}[Y|A = a, W = w]$ denote the conditional mean of Y under treatment a and covariates w. Similarly, we let $\sigma^2(a, w) := \text{var}(Y|W = w, A = a)$ be the conditional mean and variance of Y given treatment level a and covariates w, and $\bar{\sigma}^2(a, p) := \text{var}(Y|A = a, \pi_{\theta,\hbar}(W) = p)$ be the conditional mean and variance of Y given treatment level a and propensity score value p. Finally, let $\tau := E[E(Y|A = 1, W) - E(Y|A = 0, W)]$, which under our assumptions also equals $E[E(Y|A = 1, \pi_{\theta,\hbar}(W) = p) - E(Y|A = 0, \pi_{\theta,\hbar}(W) = p)]$, where expectations are under $P_{\theta,\hbar}$.

Following Abadie and Imbens (2016), the root-n scaled matching estimator can be expressed as the sum of two terms, namely B_n and R_n, as follows:

$$\sqrt{n}[\hat{\tau}(\theta, \hbar) - \tau] = B_n(\theta, \hbar) + R_n(\theta, \hbar),$$

where

$$B_n(\theta, \hbar) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \bar{\mu}_{\theta,\hbar,} \left(1, \pi_{\theta,\hbar}(w_i)\right) - \bar{\mu}_{\theta,\hbar} \left(0, \pi_{\theta,\hbar}(w_i)\right) - \tau \right.$$
$$\left. + (2a_i - 1) \left(1 + \frac{K_{M,\theta,\hbar}(i)}{M}\right) \left(Y_i - \mu_{\theta,\hbar} \left(a_i, \pi_{\theta,\hbar}(w_i)\right)\right) \right\},$$

$$R_n(\theta, \hbar) := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ (2a_i - 1) \left(\bar{\mu}_{\theta,\hbar} \left(1 - a_i, \pi_{\theta,\hbar}(w_i)\right) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i,\theta,\hbar)} \bar{\mu}_{\theta,\hbar} \left(1 - a_i, \pi_{\theta,\hbar}\left(w_j\right)\right) \right) \right\}.$$

The following lemma shows that R_n converges to 0 in probability under sampling from $P_{\theta_n,\hbar_n}$, and that B_n and $n^{-1/2}U_{\hbar_n}(\theta_n, D_n)$ are jointly asymptotically normal where the covariance of their asymptotic distribution will depend on the following quantities:

$$c_{\bar{\theta}^*,\hbar^*} := E_{\bar{\theta}^*,\hbar^*} \left\{ \mathrm{Cov}[r_{\hbar^*}(W), \mu_{\bar{\theta}^*,\hbar^*}(A, W) | \bar{\pi}_1(W)(W), A] \rho_{\bar{\theta}^*,\hbar^*}(W) \right.$$
$$\left. \left(\frac{A}{\pi_{\bar{\theta}^*,\hbar^*}(W)^2} + \frac{1 - A}{(1 - \pi_{\bar{\theta}^*,\hbar^*}(W))^2} \right) \right\}$$

$$\sigma_M^2 := E_{\bar{\theta}^*,\hbar^*} \left[\left(\bar{\mu}_{\bar{\theta}^*,\hbar}(1, \pi_{\bar{\theta}^*\hbar^*}(W))) - \bar{\mu}(0, \pi_{\bar{\theta}^*\hbar^*}(W)) - \tau \right)^2 \right]$$
$$+ E_{\bar{\theta}^*,\hbar^*} \left[\bar{\sigma}^2(1, \pi_{\bar{\theta}^*,\hbar^*}(W)) \left(\frac{1}{\pi_{\bar{\theta}^*,\hbar^*}(W)} + \frac{1}{2M} \left(\frac{1}{\pi_{\bar{\theta}^*,\hbar^*}(r_{\hbar^*}(W))} - \pi_{\bar{\theta}^*,\hbar^*}(W) \right) \right) \right],$$
$$+ E_{\bar{\theta}^*,\hbar^*} \left[\bar{\sigma}^2(0, \pi_{\bar{\theta}^*,\hbar^*}(W)) \left(\frac{1}{1 - \pi_{\bar{\theta}^*,\hbar^*}(W)} \right.\right.$$
$$\left.\left. + \frac{1}{2M} \left(\frac{1}{1 - \pi_{\bar{\theta}^*,\hbar^*}(r_{\hbar^*}(W))} - (1 - \pi_{\bar{\theta}^*,\hbar^*}(W)) \right) \right) \right].$$

Lemma 2.11. *Assume that the $\mathcal{H}$-valued sequence $(\hbar_n)_{n=1}^{\infty}$ is such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$ as $n \to \infty$ and that the conditions of Proposition 2.2 hold. Let $g \in \mathbb{R}^{p+2}$, and $\theta_n = \bar{\theta}^* + g/\sqrt{n}$. Then, under sampling from $P_{\theta_n,\hbar_n}$,*

$$\begin{pmatrix} B_n(\theta_n, \hbar_n) \\ \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & c_{\bar{\theta}^*,\hbar^*} \\ c_{\bar{\theta}^*,\hbar^*} & I_{\bar{\theta}^*,\hbar^*} \end{pmatrix} \right) \tag{2.48}$$

and $R_n(\theta_n, \hbar_n) \to 0$ in probability.

Proof. Fix $g \in \mathbb{R}^{p+2}$. Let $z_1 \in \mathbb{R}$, $z_2 \in \mathbb{R}^{p+2}$, and define the linear combination $C_n := z_1 B_n(\theta_n, \hbar_n) + z_2^\top \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n)$. Note that

$$\begin{aligned} C_n(\theta_n, \hbar_n) = z_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ & \bar{\mu}_{\theta_n,\hbar_n}(1, \pi_{\theta_n,\hbar_n}(W_i)) - \bar{\mu}_{\theta_n,\hbar_n}(0, \pi_{\theta_n,\hbar_n}(W_i)) - \tau \} \\ & + z_1 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ (2A_i - 1) \left(1 + \frac{K_{M,\theta_n,\hbar_n}(i)}{M} \right) \times [Y_i - \bar{\mu}_{\theta_n,\hbar_n}(A_i, \pi_{\theta_n,\hbar_n}(W_i))] \right\} \\ & + z_2^\top \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ r_{\hbar_n}(W_i) \frac{A_i - \pi_{\theta_n,\hbar_n}(W_i)}{\pi_{\theta_n,\hbar_n}(W_i)(1 - \pi_{\theta_n,\hbar_n}(W_i))} \rho_{\theta_n,\hbar_n}(W_i) \right\}. \end{aligned}$$

Furthermore, $C_n(\theta_n, \hbar_n)$ can be re-expressed as $C_n(\theta_n, \hbar_n) = \sum_{k=1}^{3n} \xi_{n,k}(\theta_n, \hbar_n)$ where for $1 \leq k \leq n$,

$$\begin{aligned} \xi_{n,k}(\theta_n, \hbar_n) = z_1 \frac{1}{\sqrt{n}} & \{ \bar{\mu}_{\theta_n,\hbar_n}(1, \pi_{\theta_n,\hbar_n}(W_k)) - \bar{\mu}_{\theta_n,\hbar_n}(0, \pi_{\theta_n,\hbar_n}(W_k)) - \tau \} \\ & + \frac{1}{\sqrt{n}} z_2^\top E[r_{\hbar_n}(W_k) \mid \pi_{\theta_n,\hbar_n}(W_k)] \left[\frac{A_k - \pi_{\theta_n,\hbar_n}(W_k)}{\pi_{\theta_n,\hbar_n}(W_k)(1 - \pi_{\theta_n,\hbar_n}(W_k))} \rho_{\theta_n,\hbar_n}(W_k) \right], \end{aligned}$$

for $n + 1 \leq k \leq 2n$,

$$\begin{aligned} \xi_{n,k}(\theta_n, \hbar_n) = z_2^\top \frac{1}{\sqrt{n}} & (r_{\hbar_n}(W_{k-n}) - E[r_{\hbar_n}(W_{k-n}) \mid \pi_{\theta_n,\hbar_n}(W_{k-n})]) \\ & \left[\frac{A_{k-n} - \pi_{\theta_n,\hbar_n}(W_{k-n})}{\pi_{\theta_n,\hbar_n}(W_{k-n})(1 - \pi_{\theta_n,\hbar_n}(W_{k-n}))} \rho_{\theta_n,\hbar_n}(W_{k-n}) \right] \\ & + z_1 \frac{1}{\sqrt{n}} (2A_{k-n} - 1) \left(1 + \frac{K_{M,\theta_n,\hbar_n}(k-n)}{M} \right) \\ & \cdot [\mu_{\theta_n,\hbar_n}(A_{k-n}, W_{k-n}) - \bar{\mu}_{\theta_n,\hbar_n}(A_{k-n}, \pi_{\theta_n,\hbar_n}(W_{k-n}))], \end{aligned}$$

and, for $2n + 1 \leq k \leq 3n$,

$$\xi_{n,k}(\theta_n, \hbar_n) = z_1 \frac{1}{\sqrt{n}} (2A_{k-2n} - 1) \left(1 + \frac{K_{M,\theta_n,\hbar_n}(k-2n)}{M} \right) (Y_{k-2n} - \mu_{\theta_n,\hbar_n}((A_{k-2n}, W_{k-2n})).$$

Now, it can be shown that

$$\left\{ \sum_{j=1}^{i} \xi_{n,j}(\theta_n, \hbar_n), \mathcal{F}_{n,i}, 1 \leq i \leq 3n \right\} \tag{2.49}$$

is a martingale difference sequence with respect to the filtrations:

$$\begin{cases} \mathcal{F}_{n,k} = \sigma\{A_1, ...A_k, \theta_n^\top r_{\hbar_n}(W_1), ..., \theta_n^\top r_{\hbar_n}(W_k), (\hbar_j)_{j=1}^n\}, \text{ for } 1 \leq k \leq n, \\ \mathcal{F}_{n,k} = \sigma\{r_{\hbar_n}(W_1), ..., r_{\hbar_n}(W_{k-n}), A_1, ...A_n, \theta_n^\top r_{\hbar_n}(W_1), ..., \theta_n^\top r_{\hbar_n}(W_n), (\hbar_j)_{j=1}^n\}, \text{ for } n+1 \leq k \leq 2n, \\ \mathcal{F}_{n,k} = \sigma\{r_{\hbar_n}(W_1), ..., r_{\hbar_n}(W_n), A_1, ...A_n, Y_1, ...Y_{k-2n}, (\hbar_j)_{j=1}^n\}, \text{ for } 2n+1 \leq k \leq 3n; \end{cases}$$

Next, let

$$\lambda_n := \sum_{k=1}^{3n} E_{\theta_n,\hbar_n}[\xi_{n,k}(\theta_n, \hbar_n)^2 | \mathcal{F}_{n,k-1}], \text{ and} \tag{2.50}$$

$$\lambda^* := z_1^2\sigma_M^2 + z_2^\top I_{\bar{\theta}^*,\hbar^*} z_2 + 2z_2^\top c_{\bar{\theta}^*,\hbar^*} z_1. \tag{2.51}$$

By assumption, for every $\epsilon > 0$,

$$P_{\theta_n,\hbar_n}\left(\left|\frac{\lambda_n}{\lambda^*} - 1\right| > \epsilon\right) \overset{n\to\infty}{\longrightarrow} 0. \tag{2.52}$$

Additionally, by Lemma 2.12, under $P_{\theta_n,\hbar_n}$, for each $\epsilon > 0$,

$$\sum_{k=1}^{\infty} E_{\theta_n,\hbar_n}\left[(\lambda^*)^{-1/2}\xi_{n,k}(\theta_n, \hbar_n)^2 I(|\xi_{n,k}(\theta_n, \hbar_n)| > \epsilon)\right] \longrightarrow 0. \tag{2.53}$$

Hence, for an $\epsilon > 0$ by Markov's inequality and the preceding result we have that

$$\begin{aligned} &\lim_{n\to\infty} P_{\theta_n,\hbar_n}\left(\sum_{k=1}^{n} E_{\theta_n,\hbar_n}\left[(\lambda^*)^{-1/2}\xi_{n,k}(\theta_n, \hbar_n)^2 I(|\xi_{n,k}(\theta_n, \hbar_n)| > \epsilon)|\mathcal{F}_{n,k-1}\right] > \epsilon\right) \\ &\leq \lim_{n\to\infty} \frac{E_{\theta_n,\hbar_n}\left[(\lambda^*)^{-1/2}\sum_{k=1}^{n} E_{\theta_n,\hbar_n}\left[\xi_{n,k}(\theta_n, \hbar_n)^2 I(|\xi_{n,k}(\theta_n, \hbar_n)| > \epsilon|\mathcal{F}_{n,k-1})\right]\right]}{\epsilon} \\ &= \lim_{n\to\infty} \frac{(\lambda^*)^{-1/2} E_{\theta_n,\hbar_n}\left[\xi_{n,k}(\theta_n, \hbar_n)^2 I(|\xi_{n,k}(\theta_n, \hbar_n)| > \epsilon)\right]}{\epsilon} \longrightarrow 0 \end{aligned}$$

Therefore, the conditions stated in Theorem 2 from Gaenssler and Haeusler (1986) hold. Hence, under $P_{\theta_n,\hbar_n}$,

$$\sum_{k=1}^{\infty} (\lambda^*)^{-1/2}\xi_{n,k}(\theta_n, \hbar_n) \rightsquigarrow N(0,1).$$

By the Cramer-Wold device,

$$\begin{pmatrix} B_n(\theta_n, \hbar_n) \\ n^{-1/2} U_{\hbar_n}(\theta_n; D_n) \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & c_{\bar{\theta}^*, \hbar^*} \\ c_{\bar{\theta}^*, \hbar^*} & I_{\bar{\theta}^*, \hbar^*} \end{pmatrix} \right).$$

which shows Equation 2.48. Next, we show that, for all $\epsilon > 0$

$$P_{\theta_n, \hbar_n} \left(|R_n(\theta_n, \hbar_n)| > \epsilon \right) \overset{n \to \infty}{\longrightarrow} 0. \tag{2.54}$$

Let

$$R_{n,a}(\theta_n, \hbar_n) := \frac{(2a-1)}{\sqrt{n}} \sum_{\{i: a_i = a\}} \left\{ \bar{\mu}_{\theta_n, \hbar_n}(1-a, \pi_{\theta_n, \hbar_n}(w_i)) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \theta_n, \hbar_n)} \bar{\mu}_{\theta_n, \hbar_n}(1-a, \pi_{\theta_n, \hbar_n}(w_j)) \right\}.$$

Then $R_n(\theta_n, \hbar_n) = R_{n,1}(\theta_n, \hbar_n) + R_{n,0}(\theta_n, \hbar_n)$. We show that $R_{n,1}(\theta_n, \hbar_n) = o_{P_{\theta_n, \hbar_n}}(1)$, the proof for $R_{n,0}(\theta_n, \hbar_n) = o_{P_{\theta_n, \hbar_n}}(1)$ is analogous. Fix $\hbar \in \mathcal{H}$, $\theta \in \mathbb{R}^{p+2}$, and let n_a be the number of observations in treatment arm a. Note that n_0 is lower and upper bounded by $M \leq n_0 \leq n - M$, and that $n_1 = n - n_0$. Without loss of generality, assume that the observations are ordered such that the observations with values $a_i = 1$ are first. Then, the expectation of $R_{n,1}(\theta, \hbar)$ conditional on the event $N_0 = n_0$ can be bounded as follows:

$$\begin{aligned}
&E_{\theta, \hbar} \left[|R_{n,1}(\theta, \hbar)| | N_0 = n_0 \right] \\
&\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n_1} E_{\theta, \hbar} \left[\left| \bar{\mu}_{\theta, \hbar}(0, \pi_{\theta, \hbar}(W_i)) - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \theta, \hbar)} \bar{\mu}_{\theta, \hbar}(0, \pi_{\theta, \hbar}(W_j)) \right| | N_0 = n_0 \right] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n_1} E_{\theta, \hbar} \left[\left| \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \theta, \hbar)} \{ \bar{\mu}_{\theta, \hbar}(0, \pi_{\theta, \hbar}(W_i)) - \bar{\mu}_{\theta, \hbar}(0, \pi_{\theta, \hbar}(W_j)) \} \right| | N_0 = n_0 \right] \\
&\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n_1} \left\{ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \theta, \hbar)} E_{\theta, \hbar} \left[|\bar{\mu}_{\theta, \hbar}(0, \pi_{\theta, \hbar}(W_i)) - \bar{\mu}_{\theta, \hbar}(0, \pi_{\theta, \hbar}(W_j))| \, | N_0 = n_0 \right] \right\} \\
&\leq \frac{L}{\sqrt{n}} \sum_{i=1}^{n_1} \left\{ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i, \theta, \hbar)} E_{\theta, \hbar} \left[|\pi_{\theta, \hbar}(W_i) - \pi_{\theta, \hbar}(W_j)| \, | N_0 = n_0 \right] \right\},
\end{aligned}$$

where the final inequality follows by the assumption that $\bar{\mu}_{\theta, \hbar}(a, p)$ is L-Lipschitz in p. Let $U_{n_0, n_1, i}(m)$ be the mth order statistic of $\{|\pi_{\theta, \hbar}(W_i) - \pi_{\theta, \hbar}(W_j)| : A_j = 0\}$. Then, the last

expression, up to the constant L, is equal to

$$\frac{1}{\sqrt{n}}\sum_{i:a_i=1}^{n}\frac{1}{M}\sum_{j=1}^{M}E_{\theta,\hbar}\left[U_{n_1,n_0,i}(j)\right]=E_{\theta,\hbar}\left[\frac{1}{\sqrt{n}}\sum_{i:a_i=1}^{n}\frac{1}{M}\sum_{j=1}^{M}U_{n_1,n_0,i}(j)\right].$$

By our assumptions and Lemmas S.1 and S.2 in Abadie and Imbens (2016), there exists a finite constant r that depends exclusively on the support of $\mathcal{W}$ and the boundedness of $\mathcal{H}$ such that

$$\begin{aligned}E_{\theta,\hbar}\left[\frac{1}{\sqrt{n}}\sum_{i:A_i=1}^{n}\frac{1}{M}\sum_{j=1}^{M}U_{n_1,n_0,i}(j)\right]&\leq r\frac{n_1}{\sqrt{n}\lfloor n_0^{3/4}\rfloor}+M\frac{n_1}{\sqrt{n}}n_0^{M-1/4}\exp(-n_0^{1/4})\\&=\frac{rn_1\lfloor n_0^{-3/4}\rfloor n^{3/4}+Mn_1n_0^{M+1/2}\exp(-n_0^{1/4})}{\sqrt{n}n_0^{3/4}}.\end{aligned}$$

Moreover, $\lfloor n_0^{-3/4}\rfloor n^{3/4}$, $n_0^{M+1/2}\exp(-n_0^{1/4})$, r, and M are bounded. Therefore, there is a constant K such that

$$E_{\theta,\hbar}\left[\frac{1}{\sqrt{n}}\sum_{i:A_i=1}^{n}\frac{1}{M}\sum_{j=1}^{M}U_{n_1,n_0,i}(j)\right]\leq\frac{K}{n^{1/4}}\left(\frac{n_1}{n}\frac{n^{3/4}}{n_0^{3/4}}\right).$$

Note that because r does not depend on θ nor $\hbar$, neither does K. Combining the previous results, we have that

$$E_{\theta,\hbar}\left[|R_{n,1}(\theta,\hbar)||N_0=n_0\right]\leq\frac{LK}{n^{1/4}}\left(\frac{n_1}{n}\frac{n^{3/4}}{n_0^{3/4}}\right).$$

By the tower property and and because $n_1/n<1$, we have that

$$E_{\theta,\hbar}\left[|R_{n,1}(\theta,\hbar)|\right]\leq\frac{LK}{n^{1/4}}E_{\theta,\hbar}\left[\left(\frac{n}{n_0}\right)^{3/4}\right]$$

By Lemma S.3 in Abadie and Imbens (2016), there exists a constant C that does not depend on θ nor $\hbar$ such that

$$E_{\theta,\hbar}\left[\left(\frac{n}{n_0}\right)^{3/4}\right]\leq C.$$

Therefore,

$$E_{\theta,\hbar}\left[|R_{n,1}(\theta,\hbar)|\right]\leq\frac{LK}{n^{1/4}}C.$$

Because the above uppper bound is uniform over θ and $\hbar$, we have that $\sup_{\theta,\hbar} E_{\theta,\hbar}\left[|R_{n,1}(\theta,\hbar)|\right] \overset{n\to\infty}{\longrightarrow} 0$. Similarly, $\sup_{\theta,\hbar} E_{\theta,\hbar}\left[|R_{n,0}(\theta,\hbar)|\right] \overset{n\to\infty}{\longrightarrow} 0$. Applying Markov's inequality followed by the triangle inequality to $P_{\theta_n,\hbar_n}\left(|R_n(\theta_n,\hbar_n)| > \epsilon\right)$ along with the latter convergence result shows Equation 2.54. □

Lemma 2.12. *Let*

$$\left\{\sum_{j=1}^{i} \xi_{n,j}(\theta_n,\hbar_n), \mathcal{F}_{n,i}, 1 \le i \le 3n\right\}$$

be the martingale difference sequence defined in Lemma 2.11, Equation 2.49. Assume that the conditions in Lemma 2.11 hold. Then, for every $\epsilon > 0$

$$\sum_{k=1}^{\infty} E_{\theta_n,\hbar_n}\left[\xi_{n,k}(\theta_n,\hbar_n)^2 I(|\xi_{n,k}(\theta_n,\hbar_n) > \epsilon|)\right] \longrightarrow 0 \tag{2.55}$$

Proof. Instead of showing Equation 2.55 directly, we show Lyapunov's condition instead, which implies the desired result. That is, we show that

$$\lim_{n\to\infty} \sum_{k=1}^{3n} E_{\theta_n,\hbar_n}\left[|\xi_{n,k}(\theta_n,\hbar_n)|^{2+\delta}\right] = 0, \quad \text{for some } \delta > 0.$$

First, let $\tilde{\epsilon}$ be as defined in Condition 7. Define the following upper bounds: let $\bar{w}$ denote the p+2-dimensional such that $\bar{w} = \sup_{w\in\mathcal{W},\hbar\in\mathcal{H}} r_\hbar(w)$ entry-wise, $[\underline{p},\overline{p}]$ be the upper and lower bounds of $\pi_{\theta,\hbar}(w)$ for all $w \in \mathcal{W}$ that are uniform over $\hbar \in \mathcal{H}$ and θ belonging to the ball $D_{\tilde{\epsilon}} = \{\theta \in \mathbb{R}^{p+2} : \|\theta - \bar{\theta}^*\| < \tilde{\epsilon}\}$, and let $C_{\bar{\mu}} = \sup_{a\in\{0,1\},\theta\in D_{\tilde{\epsilon}},\hbar\in\mathcal{H}} \bar{\mu}_{\theta,\hbar}(a,\pi_{\theta,\hbar}(w))$. Note that $C_{\bar{\mu}} < \infty$ by Condition 7. Further, recall that $0 < \rho_\theta(r_\hbar(w)) < 1$ for all $\theta \in \mathbb{R}^{p+2}, \hbar \in \mathcal{H}$, and $w \in \mathcal{W}$.

Fix $\theta \in D_{\tilde{\epsilon}}$, $\hbar \in \mathcal{H}$. Let δ be some positive real number such that there exists a $C_\delta < \infty$ such that $E_{\theta,\hbar}[|Y|^{2+\delta}|A = a, \pi_{\theta,\hbar}(W) = p]$ is uniformly bounded by C_δ — such a δ is guaranteed to exist by Condition 7. In the following equations all expectations are taken with respect to $P_{\theta,\hbar}$.

We divide $\sum_{k=1}^{3n} E\left[|\xi_{n,k}(\theta,\hbar)|^{2+\delta}\right]$ into three separate sums and show that each converge

to 0. For the first sum, we have that

$$\begin{aligned}
\sum_{k=1}^{n} E\left[|\xi_{n,k}(\theta,\hbar)|^{2+\delta}\right] &\leq \sum_{k=1}^{n} E\left[\left|z_1 \frac{1}{\sqrt{n}}(\bar{\mu}_{\theta,\hbar}(1,\pi_{\theta,\hbar}(W_k)) - \bar{\mu}_{\theta,\hbar}(0,\pi_{\theta,\hbar}(W_k)) - \tau)\right|^{2+\delta}\right] \\
&\quad + \sum_{k=1}^{n} E\left[\left| \left(\frac{1}{\sqrt{n}} z_2^\top E[r_\hbar(W_k)|\pi_{\theta,\hbar}(W_k)]\right)\right.\right. \\
&\qquad \left.\left. \left(\frac{A_k - \pi_{\theta,\hbar}((W_k))}{\pi_{\theta,\hbar}(W_k)(1-\pi_{\theta,\hbar}(W_k))}\rho_\theta(r_\hbar(W_k))\right)\right|^{2+\delta}\right] \\
&\leq \frac{|2z_1 C_{\bar{\mu}} + z_1|\tau||^{2+\delta}}{n^{\delta/2}} + \sum_{k=1}^{n} E\left[\left| \left(\frac{1}{\sqrt{n}} z_2^\top E_{\theta,\hbar}[r_\hbar(W_k)|\pi_{\theta,\hbar}(W_k)]\right)\right.\right. \\
&\qquad \left.\left. \left(\frac{A_{n,k} - \pi_{\theta,\hbar}(W_k)}{\pi_{\theta,\hbar}(W_k)(1-\pi_{\theta,\hbar}(W_k))}\rho_\theta(r_\hbar(W_k))\right)\right|^{2+\delta}\right] \\
&\leq \frac{|2z_1 C_{\bar{\mu}} + z_1|\tau||^{2+\delta}}{n^{\delta/2}} + \frac{|z_2^\top \bar{w}|^{2+\delta}}{n^{\delta/2}} \left(\frac{1}{\underline{p}(1-\overline{p})}\right)^{2+\delta}, \qquad (2.56)
\end{aligned}$$

where the first inequality holds due to the triangle inequality. Note that the latter upper bound converges to 0 as $n \to \infty$. To bound the second sum, first let $\Delta_{\theta,\hbar}(a,w,\pi) := E[Y|A=a, W=w] - E[Y|A=a, \pi_{\theta,\hbar}(W) = \pi_{\theta,\hbar}(w)]$. Then, $E\left[|\ \Delta_{\theta,\hbar}(A,W,\pi)\ |^{2+\delta}\ |\pi_{\theta,\hbar}(W), A\right]$ can be bounded as follows:

$$\begin{aligned}
E\left[|\Delta_{\theta,\hbar}(A,W,\pi)|^{2+\delta}|\pi_{\theta,\hbar}(W),A\right] &\leq E\left[|E(Y|A,W)|^{2+\delta}|\pi_{\theta,\hbar},A\right] + E\left[|E(Y|A,\pi_{\theta,\hbar}(W))|^{2+\delta}|\pi_{\theta,\hbar}(W),A\right] \\
&\leq E\left[E(|Y|^{2+\delta}|A,W)|\pi_{\theta,\hbar}(W),A\right] + E[|Y|^{2+\delta}|\pi_{\theta,\hbar}(W),A] \\
&= 2E[|Y|^{2+\delta}|\pi_{\theta,\hbar}(W)=p,A]] \leq 2C_\delta
\end{aligned}$$

Let $C_\Delta := 2C_\delta$ where, by assumption, C_Δ does not depend on θ, nor $\hbar$, nor p. Then, we

have that

$$\sum_{k=n+1}^{2n} E[|\xi_{n,k}(\theta,\hbar)|^{2+\delta}] \leq \sum_{k=n+1}^{2n} E\Bigg\{\Bigg|\left[z_2^\top \frac{1}{\sqrt{n}}(r_\hbar(W_{k-n}) - E[r_\hbar(W_{k-n})|\pi_{\theta,\hbar}(W_{k-n})])\right] \left[\frac{A_{k-n}-\pi_{\theta,\hbar}(W_{k-n})}{\pi_{\theta,\hbar}(W_{k-n})(1-\pi_{\theta,\hbar}(W_{k-n}))}\rho_\theta(r_\hbar(W_{k-n}))\right]\Bigg|^{2+\delta}\Bigg\}$$
$$+ \sum_{k=n+1}^{2n} E\left\{\left|\frac{z_1}{\sqrt{n}}(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\Delta_{\theta,\hbar}(A_{k-n},W_{k-n},\pi_{\theta,\hbar})\right|^{2+\delta}\right\}$$
$$\leq \sum_{k=n+1}^{2n} E\left\{\left|z_2^\top \frac{1}{\sqrt{n}}(r_\hbar(W_{k-n}))\left(\frac{A_{k-n}-\pi_{\theta,\hbar}(W_{k-n})}{\pi_{\theta,\hbar}(W_{k-n})(1-\pi_{\theta,\hbar}(W_{k-n}))}\right)\rho_\theta(r_\hbar(W_{k-n}))\right|^{2+\delta}\right\}$$
$$+ \sum_{k=n+1}^{2n} E\left\{\left|z_2^\top \frac{1}{\sqrt{n}}(E[r_\hbar(W_{k-n})|\pi_{\theta,\hbar}(W_{k-n})])\left(\frac{A_{k-n}-\pi_{\theta,\hbar}(W_{k-n})}{\pi_{\theta,\hbar}(W_{k-n})(1-\pi_{\theta,\hbar}(W_{k-n}))}\right)\rho_\theta(r_\hbar(W_{k-n}))\right|^{2+\delta}\right\}$$
$$+ \sum_{k=n+1}^{2n} E\left\{\left|\frac{z_1}{\sqrt{n}}(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\Delta_{\theta,\hbar}(A_{k-n},W_{k-n},\pi_{\theta,\hbar})\right|^{2+\delta}\right\}$$
$$\leq 2\frac{|z_2^\top \bar{w}|^{2+\delta}}{n^{\delta/2}}\left|\frac{1}{\underline{p}(1-\overline{p})}\right|^{2+\delta}$$
$$+ \sum_{k=n+1}^{2n} E\left\{\left|\frac{z_1}{\sqrt{n}}(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\Delta_{\theta,\hbar}(A_{k-n},W_{k-n},\pi_{\theta,\hbar})\right|^{2+\delta}\right\},$$

where the first and second inequalities are due to the triangle inequality and the third is due to the boundedness of $\mathcal{W}$ and $\mathcal{H}$, and the fact that $\theta \in D_{\tilde{\epsilon}}$. Next, let A^n denote the treatment values for all n observations. Similarly, let $\pi^n_{\theta,\hbar}$ denote all the propensity score values for all n observations. Note that, when conditioning on A^n and $\pi^n_{\theta,\hbar}$, each $K_{M,\theta_n,\hbar}(i)$ term is constant for all $i \in \{n+1,...,2n\}$. Hence,

$$\sum_{k=n+1}^{2n} E\left\{\left|\frac{z_1}{\sqrt{n}}(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\Delta_{\theta,\hbar}(A_{k-n},W_{k-n},\pi_{\theta,\hbar})\right|^{2+\delta}\right\}$$
$$= \sum_{k=n+1}^{2n} E\left[E\left\{\left|\frac{z_1}{\sqrt{n}}(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\Delta_{\theta,\hbar}(A_{k-n},W_{k-n},\pi_{\theta,\hbar})\right|^{2+\delta}\Bigg|A^n,\pi^n_{\theta,\hbar}\right\}\right]$$
$$= \frac{z_1^2}{n^{1+\delta/2}}\sum_{k=n+1}^{2n} E\left[\left|(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\right|^{2+\delta} E\left\{\Delta_{\theta,\hbar}(A_{k-n},W_{k-n},\pi_{\theta,\hbar})^{2+\delta}|A^n,\pi^n_{\theta,\hbar}\right\}\right]$$
$$\leq \frac{C_\Delta z_1^2}{n^{1+\delta/2}}\sum_{k=n+1}^{2n} E\left[\left|(2A_{k-n}-1)\left(1+\frac{K_{M,\theta,\hbar}(k-n)}{M}\right)\right|^{2+\delta}\right].$$

Nearly identical arguments to those used to prove Lemma S.8 in Abadie and Imbens (2016) show that $E[K_{M,\theta,\hbar}^{\lceil 2+\delta \rceil}(i)|A_i = a]$ is uniformly bounded in n by a finite constant that neither depends on θ nor $\hbar$ — we denote this constant by L. Moreover, because $K_{M,\theta,\hbar}(i)$ is a positive integer, we have that $E[K_{M,\theta,\hbar}^{2+\delta}(i)|A_i = a] \leq E[K_{M,\theta,\hbar}^{\lceil 2+\delta \rceil}(i)|A_i = a]$, where $\lceil \cdot \rceil$ denotes the ceiling function. Therefore,

$$\sum_{k=n+1}^{2n} E[|\xi_{n,k}(\theta, \hbar)|^{2+\delta}] \leq \frac{LC_\Delta z_1^2}{n^{\delta/2}}, \tag{2.57}$$

and recall that neither L nor C_Δ depend on θ or $\hbar$. The last sum can be bounded as follows:

$$\begin{aligned}
&\sum_{k=2n+1}^{3n} E[|\xi_{n,k}(\theta, \hbar)|^{2+\delta}] \\
&= \sum_{k=2n+1}^{3n} E\left[\left| z_1 \frac{1}{\sqrt{n}}(2A_{k-2n} - 1)\left(1 + \frac{K_{M,\theta_n,\hbar_n}(k-2n)}{M}\right) \times (Y_{k-2n} - \mu(A_{k-2n}, W_{k-2n}))\right|^{2+\delta}\right] \\
&\leq \sum_{k=2n+1}^{3n} E\left[\left| z_1 \frac{1}{\sqrt{n}}(2A_{k-2n} - 1)\left(1 + \frac{K_{M,\theta_n,\hbar_n}(k-2n)}{M}\right) Y_{k-2n}\right|^{2+\delta}\right] \\
&\quad + \sum_{k=2n+1}^{3n} E\left[\left| z_1 \frac{1}{\sqrt{n}}(2A_{k-2n} - 1)\left(1 + \frac{K_{M,\theta_n,\hbar_n}(k-2n)}{M}\right) \mu(A_{k-2n}, W_{k-2n})\right|^{2+\delta}\right].
\end{aligned}$$

Similar arguments as those employed to show Equation 2.56 (broadly, conditioning on A^n and $\pi_{\theta,\hbar}^n$ followed by bounding the terms in the sum) can be used to show that

$$E\left[\left| \sum_{k=2n+1}^{3n} |\xi_{n,k}(\theta, \hbar)|^{2+\delta}\right] \leq \frac{z_1^2 M_\delta}{n^{\delta/2}}, \tag{2.58}$$

where M_δ depends on neither θ nor $\hbar$.

By the triangle inequality and since the bounds on Equations 2.55, 2.56, 2.57, and 2.58 are uniform over θ and $\hbar$, we have that $\sup_{\theta \in D_{\tilde{\epsilon}}, \hbar \in \mathcal{H}} E\left[\sum_{i=1}^{3n} |\xi_{n,k}(\theta_n, \hbar_n)|^{2+\delta}\right] \to 0$. Finally, because $P_{\theta_n,\hbar_n}(\theta_n \notin D_{\tilde{\epsilon}}) \overset{n\to\infty}{\longrightarrow} 0$,

$$\sum_{k=1}^{3n} E_{\theta_n,\hbar_n}\left[|\xi_{n,k}(\theta_n, \hbar_n)|^{2+\delta}\right] \overset{n\to\infty}{\longrightarrow} 0.$$

□

2.C Proof of Proposition 1

Similar to the results in Abadie and Imbens (2016), our main asymptotic normality theorem provides guarantees for the discretized version of the maximum likelihood estimator. That is, we employ the same result from Andreou and Werker (2012) that allows us to show the asymptotic distribution of the matching estimator when using the following transformation of $\tilde{\theta}_{n,\hbar}(d_n)$ to match. Let $k > 0$ be the discretization constant that we use to define a grid of cubes in $\mathbb{R}^{p+2}$ of sides of length $k/\sqrt{n}$. Informally, the discretized estimator $\bar{\theta}_{n,\hbar}(d_n)$ is obtained by taking the midpoint of the cube $\tilde{\theta}_{n,\hbar}(d_n)$ belongs to. Formally, if $r : \mathbb{R}^{p+2} \to \mathbb{Z}^{p+2}$ is a function that rounds each entry of a vector to its nearest integer, then $\bar{\tilde{\theta}}_{k,n,\hbar}(d_n) := kr(\sqrt{n}\tilde{\theta}_{n,\hbar}(d_n)/k)/\sqrt{n}$. The following proposition shows that asymptotic distribution of the matching estimator when carrying matching via the discretized estimator $\bar{\tilde{\theta}}_{k,n,\hbar_n}(d_n)$ is asymptotically normal under a set of conditions. It is worth noting that our asymptotic result is based on taking the limit with respect to the probability measure based on the shifted discretized version of the truth defined as $\bar{\theta}^*_{n,k} := kr(\sqrt{n}\bar{\theta}^*/k)/\sqrt{n} + g/\sqrt{n}$, where $g \in \mathbb{R}^{p+2}$.

Proposition 2.2. *Suppose that Conditions 1-9 hold. Let $(\hbar_n)_{n=1}^{\infty}$ be an $\mathcal{H}-$ valued sequence such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$, and let $g \in \mathbb{R}^{p+2}$. Additionally, let: $\tilde{\theta}_{n,\hbar_n}(D_n) = \operatorname{argmax}_\theta \ell_{\hbar_n}(\theta; D_n)$ be the maximum likelihood estimator based on observations $D_n = \{A_i, W_i, \hbar_n(W_i)\}_{i=1}^n$, and $\bar{\tilde{\theta}}_{k,n,\hbar_n}(D_n)$ denote its discretized version with discretization constant k. Let $\bar{\theta}^*_{k,n} = kr(\sqrt{n}\bar{\theta}^*/k)/\sqrt{n} + g/\sqrt{n}$ be the shifted discretized version of $\bar{\theta}^*$. Then,*

$$\lim_{k\downarrow 0}\lim_{n\to\infty} P_{\bar{\theta}^*_{k,n},\hbar_n}\left(\sqrt{n}(\sigma^2_M - c^\top_{\bar{\theta}^*,\hbar^*} I_{\bar{\theta}^*,\hbar^*} c_{\bar{\theta}^*,\hbar^*})^{-1/2}(\hat{\tau}(\bar{\tilde{\theta}}_{k,n,\hbar_n}(D_n), \hbar_n) - \tau) \le z\right) = \Phi(z).$$

where Φ is the cumulative distribution function of a standard normal random variable.

Proof. Let $\theta_n = \bar{\theta}^* + g/\sqrt{n}$. By Lemma 2.11, we have that, under $P_{\theta_n,\hbar_n}$,

$$\begin{pmatrix} B_n(\theta_n, \hbar_n) \\ \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_M & c_{\hbar^*} \\ c_{\hbar^*} & I_{\bar{\theta}^*,\hbar^*} \end{pmatrix}\right).$$

Moreover, also because of Lemma 2.11, we have that $R_n(\theta_n, \hbar_n) \to 0$ in probability. By Slutsky's Theorem, under $P_{\theta_n,\hbar_n}$,

$$\begin{pmatrix} \sqrt{n}(\hat{\tau}(\theta_n, \hbar_n) - \tau) \\ \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & c_{\hbar^*} \\ c_{\hbar^*} & I_{\bar{\theta}^*,\hbar^*} \end{pmatrix} \right). \tag{2.59}$$

Applying the continuous mapping theorem to (3.26) with the bivariate function $f(x, y) = (x, I_{\bar{\theta}^*}^{-1} y, -g^\top y - g^\top I_{\bar{\theta}^*} g/2)$, we have that under $P_{\theta_n,\hbar_n}$,

$$\begin{pmatrix} \sqrt{n}(\hat{\tau}(\theta_n, \hbar_n) - \tau) \\ I_{\bar{\theta}^*}^{-1} \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) \\ -g^\top \frac{1}{\sqrt{n}} U_{\hbar_n}(\theta_n; D_n) - g^\top I_{\bar{\theta}^*} g/2 \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 0 \\ 0 \\ -g^\top I_{\bar{\theta}^*,\hbar^*} g/2 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & c_{\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} & -c_{\hbar^*}^\top g \\ I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\hbar^*} & I_{\bar{\theta}^*,\hbar^*}^{-1} & -g \\ -g^\top c_{\hbar^*} & -g^\top & g^\top I_{\bar{\theta}^*,\hbar^*} g \end{pmatrix} \right).$$

By Slutsky's theorem, Lemma 2.6, and Lemma 2.10, we have that under $P_{\theta_n,\hbar_n}$,

$$\begin{pmatrix} \sqrt{n}(\hat{\tau}(\theta_n, \hbar_n) - \tau) \\ \sqrt{n}\left(\tilde{\theta}_{n,\hbar_n}(D_n) - \theta_n \right) \\ \Lambda\left(\bar{\theta}^* | \theta_n \right) \end{pmatrix} \rightsquigarrow N\left(\begin{pmatrix} 0 \\ 0 \\ -g^\top I_{\bar{\theta}^*,\hbar^*} g/2 \end{pmatrix}, \begin{pmatrix} \sigma_M^2 & c_{\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} & -c_{\hbar^*}^\top g \\ I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\hbar^*} & I_{\bar{\theta}^*,\hbar^*}^{-1} & -g \\ -g^\top c_{\hbar^*} & -g^\top & g^\top I_{\bar{\theta}^*,\hbar^*} g \end{pmatrix} \right).$$

Therefore, Condition AN in Andreou and Werker (2012) is satisfied. Next, because we assumed that Condition 9 is true, the (ULAN) condition in Andreou and Werker (2012) applies in our setting.

Now, recalling the definition of $r : \mathbb{R}^{p+2} \to \mathbb{Z}^{p+2}$ as a function that rounds each entry of a vector to their nearest integer, let $\bar{\theta}_{k,n}^* = kr(\sqrt{n}\bar{\theta}^*/k)/\sqrt{n} + g/\sqrt{n}$ be the shifted discretized version of $\bar{\theta}^*$. Then, we have that,

$$\lim_{k \downarrow 0} \lim_{n \to \infty} P_{\bar{\theta}_{k,n}^*,\hbar_n}\left(\sqrt{n}(\sigma_M^2 - c_{\bar{\theta}^*,\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\bar{\theta}^*,\hbar^*})^{-1/2} (\hat{\tau}(\bar{\tilde{\theta}}_{k,n,\hbar_n}(D_n), \hbar_n) - \tau) \leq z \right) = \Phi(z).$$

□

2.D Proof of Corollary 2.3

In the following corollary we show the convergence statement from Proposition 2.2 when the sequence of $\mathcal{H}$- valued functions is random and obtained from a dataset $S = \{O_{-i}\}_{i=1}^{\infty}$

consisting of infinitely many iid draws from $P_{\bar{\theta}^*,\hbar^*}$. Let $\{H_n\}_{n=1}^{\infty}$ denote a sequence of operators that take as input a realization s of the independent dataset and output an estimate $H_n(s) \in \mathcal{H}$ of $\hbar^*$. We mostly have in mind the case where $H_n(S)$ is an estimate of $\hbar^*$ based on the first $m_n < \infty$ observations $\{O_{-i}\}_{i=-1}^{-m_n}$ in S, where $(m_n)_{n=1}^{\infty}$ is some increasing sequence, though our specification of $\{H_n\}_{n=1}^{\infty}$ enables consideration of more general specifications of the functions H_n. We will establish the convergence in probability of the sequence of matching estimators, assuming that $\|H_n(S) - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$ holds almost surely under sampling of S. Hereafter we will denote $H_n(S)$ as h_n. As in Proposition 2.2, we will establish a distributional result regarding the sequence of matching estimators under sampling from a discretized distribution, namely the distribution indexed by the unshifted discretized parameter $\bar{\theta}^*_{k,n} := kr(\sqrt{n}\bar{\theta}^*/k)/\sqrt{n}$. Since the final entry of $\bar{\theta}^*$ is zero by assumption, the final entry of $\bar{\theta}^*_{k,n}$ is zero as well. Hence, the value of $P_{\bar{\theta}^*_{k,n},h}$ does not depend on the value of $h \in \mathcal{H}$. Hereafter we shall write $P_{\bar{\theta}^*_{k,n}}$, rather than $P_{\bar{\theta}^*_{k,n},h}$, to make this clear.

Our distributional result will hold conditional on S. Before showing the Corollary, we clarify what we mean by this notion of convergence. For each n, let F_n denote a random $\mathcal{D}^n \to \mathbb{R}$ function that is measurable with respect to the σ-field generated by S, $\mathcal{D} := \mathbb{R}^{p+2} \times \{0,1\} \times \mathbb{R}$ contain the support of each $P_{\theta,h}$, and D_n denote a random dataset of n iid draws from $P_{\bar{\theta}^*_{k,n}}$, where this dataset is independent of S. We say that, under $P_{\bar{\theta}^*_{k,n}}$ and conditionally on S, $F_n(D_n)$ converges to a random variable Z if

$$P_{\bar{\theta}^*,\hbar^*}\left(\lim_{k\downarrow 0}\lim_{n\to\infty} \Pr(F_n(D_n) \leq z \mid S) = G_Z(z)\right) = 1,$$

where G_Z is the cumulative distribution function of Z and $\Pr(\cdot \mid S)$ denotes the conditional distribution of (D_n, S) given S. Note that the display above depends on the the probability measure $P_{\bar{\theta}^*_{k,n}}$ through D_n, which is an iid sample from $P_{\bar{\theta}^*_{k,n}}$.

Corollary 2.3. *Suppose that the conditions in Proposition 2.2 hold. Suppose that* $\|h_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$ *almost surely. Let* $\tilde{\theta}_{n,h_n}(D_n) = \operatorname{argmax}_\theta \ell_{h_n}(\theta; D_n)$ *be the maximum likelihood estimator based on observations* $D_n = \{A_i, W_i, h_n(W_i)\}_{i=1}^n$, *and* $\bar{\tilde{\theta}}_{k,n,h_n}(D_n) = kr(\sqrt{n}\tilde{\theta}_{n,h_n}(D_n)/k)/\sqrt{n}$

denote its discretized version with discretization constant k. Then, under $P_{\bar{\theta}^*_{k,n}}$ *and conditionally on S,*

$$\sqrt{n}(\sigma_M^2 - c_{\bar{\theta}^*,\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\bar{\theta}^*,\hbar^*})^{-1/2}(\hat{\tau}(\bar{\bar{\theta}}_{k,n,h_n}(D_n), h_n) - \tau)$$

converges to a standard normal random variable.

Proof. Let Φ denote the cumulative distribution function of a standard normal random variable. By Proposition 2.2, taking the shift g in that result to be equal to zero, we have that

$$\lim_{k\downarrow 0}\lim_{n\to\infty} P_{\bar{\theta}^*_{k,n}}\left(\sqrt{n}(\sigma_M^2 - c_{\bar{\theta}^*,\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\bar{\theta}^*,\hbar^*})^{-1/2}(\hat{\tau}(\bar{\bar{\theta}}_{k,n,\hbar_n}(D_n), \hbar_n) - \tau) \le z\right) = \Phi(z),$$

for every $\mathcal{H}$-valued deterministic sequence $\{\hbar_n\}_{n=1}^\infty$ such that $\|\hbar_n - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0$. Hence,

$$\left\{s \in \mathcal{S} : \lim_{n\to\infty} \|H_n(s) - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0\right\}$$
$$\subseteq \left\{s \in \mathcal{S} : \lim_{k\downarrow 0}\lim_{n\to\infty} P_{\bar{\theta}^*_{k,n}}\left(\sqrt{n}(\sigma_M^2 - c_{\bar{\theta}^*,\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\bar{\theta}^*,\hbar^*})^{-1/2}(\hat{\tau}(\bar{\bar{\theta}}_{k,n,H_n(s)}(D_n), H_n(s)) - \tau) \le z\right) = \Phi(z)\right\}.$$

By assumption, the event $\left\{s \in \mathcal{S} : \lim_{n\to\infty} \|H_n(s) - \hbar^*\|_{L^2(P_{\bar{\theta}^*,\hbar^*})} \to 0\right\}$ has probability one under sampling the variates in S independently from $P_{\bar{\theta}^*,\hbar^*}$. Hence, the event on the right-hand side also has probability one. Finally, Because S is independent of D_n, the probability in the event on the right-hand side is equal to

$$\Pr\left(\sqrt{n}(\sigma_M^2 - c_{\bar{\theta}^*,\hbar^*}^\top I_{\bar{\theta}^*,\hbar^*}^{-1} c_{\bar{\theta}^*,\hbar^*})^{-1/2}(\hat{\tau}(\bar{\bar{\theta}}_{k,n,h_n}(D_n), h_n) - \tau) \le z \,\middle|\, S\right).$$

□

2.E Proof of Theorem 2.2

The following theorem and its corollary pertain to the efficiency gain term in the asymptotic variance of the matching estimator. Before stating the theorem, we introduce additional notation and definitions as well as remind the reader of previous definitions.

Let $\bar{q}(w) := (q_1(w), \ldots, q_{p+k+1}(w))$ denote a $p+k+1$-dimensional vector that consists of the evaluations of $p+k+1$ real-valued that belong to the $L^2(P_W)$ space at w, and are such

that the first $p+1$ functions are a mapping to the covariate with the same index, that is, $q_j(w) = w_j$.

We let $\pi^{\bar{q}}_{\bar{\theta}}(w) := \text{expit}[\bar{\theta}^\top \bar{q}(w)]$, where $\bar{\theta} \in \mathbb{R}^{k+p+1}$. In the special case where $k = 1$ and $q_{p+2} = \hbar$, it holds that $\pi^{\bar{q}}_{\bar{\theta}} = \pi_{\bar{\theta},\hbar}$. Recall that $\rho(x) := \frac{d}{dx}\text{expit}(x) = \frac{e^x}{(1+e^x)^2}$, let $\rho^{\bar{q}}_{\bar{\theta}}(w) := \frac{e^{\bar{\theta}^\top \bar{q}(w)}}{(1+e^{\bar{\theta}^\top \bar{q}(w)})^2}$. Additionally, we will let $\bar{\theta}^*_k$ be a vector of dimension $k+p+1$, where its first $p+1$ entries of have the same values as those of θ^*, and the remaining k are 0. Let $P^{\bar{q}}_{\bar{\theta}}$ denote the distribution for which

$$\frac{dP^{\bar{q}}_{\bar{\theta}}}{dP}(w,a,y) = \frac{\pi^{\bar{q}}_{\bar{\theta}}(w)^a[1-\pi^{\bar{q}}_{\bar{\theta}}(w)]^{1-a}}{\pi(w)^a[1-\pi(w)]^{1-a}}.$$

Let $\bar{\mu}^{\bar{q}}_{\bar{\theta}}(a,p) := E^{\bar{q}}_{\bar{\theta}}[Y|A=a, \pi^{\bar{q}}_{\bar{\theta}}(W)=p]$ denote the conditional expectation under $P^{\bar{q}}_{\bar{\theta}}$ of Y given a and the propensity score evaluated at p, and let $\mu^{\bar{q}}_{\bar{\theta}}(a,w) = E^{\bar{q}}_{\bar{\theta}}[Y|A=a, W=w]$ denote the conditional mean of Y given treatment a and covariates w, also under $P^{\bar{q}}_{\bar{\theta}}$.

Given augmentation covariates $\bar{q}(w)$, we will denote the corresponding model of the propensity score by

$$\begin{aligned}\mathcal{M}^{\bar{q}} := \{&\text{logit}[\pi^{\bar{q}}_{\bar{\theta}}(w)] = \bar{\theta}^\top \bar{q}(w) : \bar{\theta} \in \mathbb{R}^{p+k+1},\ p,k \in \mathbb{N},\\ &\bar{q}(w) = (q_1(w), \ldots, q_{p+k+1}(w)),\ q_j(w) = w_j\ \forall j \in \{1,\ldots,p+1\},\\ &q_i \in L^2(P_W),\ I^{\bar{q}}_{\bar{\theta}^*_k} \text{ is invertible}\},\end{aligned}$$

where $I^{\bar{q}}_{\bar{\theta}} := E^{\bar{q}}_{\bar{\theta}}[\pi^{\bar{q}}_{\bar{\theta}}(W)(1-\pi^{\bar{q}}_{\bar{\theta}}(W))\bar{q}(W)\bar{q}(W)^\top]$ denotes the information matrix for $\bar{\theta}$. Let $(I^{\bar{q}}_{\bar{\theta}^*_k})^{-1}$ be equal to the inverse of the information matrix.

Then, by Abadie and Imbens (2016), $\text{Eff}(\mathcal{M}^{\bar{q}})$ is equal to $c^{\bar{q}\top}_{\bar{\theta}^*_k}(I^{\bar{q}}_{\bar{\theta}^*_k})^{-1}c^{\bar{q}}_{\bar{\theta}^*_k}$, where

$$c^{\bar{q}}_{\bar{\theta}^*_k} := E^{\bar{q}}_{\bar{\theta}^*_k}\left[\text{Cov}[\bar{q}(W), \mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W), A]\rho^{\bar{q}}_{\bar{\theta}^*_k}(W)\left(\frac{A}{\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)^2} + \frac{1-A}{(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))^2}\right)\right].$$

Finally, recall that $\hbar^*$ is defined as

$$\hbar^*(w) = \frac{1}{\pi_{\bar{\theta}^*,\hbar^*}(w)}\left[\mu_{\bar{\theta}^*,\hbar^*}(1,w) - \bar{\mu}(1,\pi_{\bar{\theta}^*,\hbar^*}(w))\right] + \frac{1}{1-\pi_{\bar{\theta}^*,\hbar^*}(w)}\left[\mu_{\bar{\theta}^*,\hbar^*}(0,w) - \bar{\mu}(0,\pi_{\bar{\theta}^*,\hbar^*}(w))\right]. \tag{2.60}$$

Note that Conditions 2, 3, and 7 imply that $\hbar^* \in L^2(P_W)$.

Theorem 2.2. *Let*

$$\mathcal{Q}_P := \{\bar{q}(w) = (q_1(w), \ldots, q_{p+k+1}(w)) : \; q_j(w) = w_j \; \forall w \in \mathcal{W} \text{ and } j \in \{1, \ldots, p+1\},$$
$$q_i \in L^2(P_W), \; I^{\bar{q}}_{\bar{\theta}^*_k} \textit{ is invertible}, \; p, k \in \mathbb{N}\},$$

Then, $\text{Eff}(\mathcal{M}_{\hbar^*}) = E\left[\left(\sqrt{\pi_{\bar{\theta}^*,\hbar^*}(W)(1-\pi_{\bar{\theta}^*,\hbar^*}(W))}\hbar^*(W)\right)^2\right]$, *and*

$$\sup_{\bar{q}\in\mathcal{Q}_P} \text{Eff}(\mathcal{M}^{\bar{q}}) = \text{Eff}(\mathcal{M}_{\hbar^*}).$$

Proof. Fix $\bar{q} \in \mathcal{Q}_P$. By Abadie and Imbens (2016), $\text{Eff}(\mathcal{M}^{\bar{q}})$ is given by $c^{\bar{q}\top}_{\bar{\theta}^*_k} I^{\bar{q}-1}_{\bar{\theta}^*_k} c^{\bar{q}}_{\bar{\theta}^*_k}$, where $I^{\bar{q}-1}_{\bar{\theta}^*} = E^{\bar{q}}_{\bar{\theta}^*}\left[\pi^{\bar{q}}_{\bar{\theta}^*}(W)(1-\pi^{\bar{q}}_{\bar{\theta}^*}(W))\bar{q}(W)\bar{q}(W)^\top\right]^{-1}$, and

$$c^{\bar{q}}_{\bar{\theta}^*_k} = \begin{pmatrix} E\left[\text{Cov}(q_1(W), \mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W), A)\rho^{\bar{q}}_{\bar{\theta}^*_k}(W)\left(\frac{A}{\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)^2} + \frac{1-A}{(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)^2}\right)\right] \\ \vdots \\ E\left[\text{Cov}(q_{k+p+1}(W), \mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W), A)\rho^{\bar{q}}_{\bar{\theta}^*_k}(W)\left(\frac{A}{\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)^2} + \frac{1-A}{(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)^2}\right)\right] \end{pmatrix}. \tag{2.61}$$

Let $g^{\bar{q}}_{\bar{\theta}^*}(a,w) := \frac{a(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(w))}{\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)} + \frac{(1-a)\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)}{1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)}$. Note that conditioning on a, and $\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)$, $g^{\bar{q}}_{\bar{\theta}^*}(a,w)$

is fixed. Then, for an arbitrary q_j we have that

$$
\begin{aligned}
&E\left\{\mathrm{Cov}(q_j(W),\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A)\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))\left(\frac{A}{\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)^2}+\frac{1-A}{(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))^2}\right)\right\}\\
=&E\left\{\mathrm{Cov}(q_j(W),\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A)\left(\frac{A(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))}{\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)}+\frac{(1-A)\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)}{(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))}\right)\right\}\\
=&E\left\{\mathrm{Cov}(q_j(W),\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A)g^{\bar{q}}_{\bar{\theta}^*}(A,W)\right\}\\
=&E\Big\{E\left[q_j(W)\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A)\right]g^{\bar{q}}_{\bar{\theta}^*}(A,W)\\
&\quad-E\left[q_j(W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A\right]E\left[\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A\right]g^{\bar{q}}_{\bar{\theta}^*}(A,W)\Big\}\\
=&E\Big\{E\left[q_j(W)\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A\right]g^{\bar{q}}_{\bar{\theta}^*}(A,W)\\
&\quad-E\left[q_j(W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A\right]\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(A,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))g^{\bar{q}}_{\bar{\theta}^*}(A,W)\Big\}\\
=&E\left\{E\left[q_j(W)\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)g^{\bar{q}}_{\bar{\theta}^*}(A,W)|\pi^{\bar{q}}_{\bar{\theta}^*_k}(W),A\right]-E\left[q_j(W)\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(A,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))g^{\bar{q}}_{\bar{\theta}^*}(A,W)|A,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)\right]\right\}\\
=&E\left\{q_j(W)g^{\bar{q}}_{\bar{\theta}^*}(A,W)\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)-q_j(W)\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(A,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W))g^{\bar{q}}_{\bar{\theta}^*}(A,W)\right\}\\
=&E\left\{q_j(W)g^{\bar{q}}_{\bar{\theta}^*}(A,W)(\mu^{\bar{q}}_{\bar{\theta}^*_k}(A,W)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(A,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)))\right\}\\
=&E\left\{q_j(W)g^{\bar{q}}_{\bar{\theta}^*}(1,W)(\mu^{\bar{q}}_{\bar{\theta}^*_k}(1,W)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(1,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)))\pi_{\bar{\theta}^*,\hbar^*}(W)\right\}\\
&+E\left\{q_j(W)\left[g^{\bar{q}}_{\bar{\theta}^*}(0,W)(\mu^{\bar{q}}_{\bar{\theta}^*_k}(0,W)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(0,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)))(1-\pi_{\bar{\theta}^*,\hbar^*}(W))\right]\right\}\\
=&E\left\{q_j(W)\left[(\mu^{\bar{q}}_{\bar{\theta}^*_k}(1,W)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(1,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)))(1-\pi_{\bar{\theta}^*,\hbar^*}(W))+(\mu^{\bar{q}}_{\bar{\theta}^*_k}(0,W)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(0,\pi^{\bar{q}}_{\bar{\theta}^*_k}(W)))\pi_{\bar{\theta}^*,\hbar^*}(W)\right]\right\}
\end{aligned}
\tag{2.62}
$$

Next, note that by definition of $\hbar^*$ in (2.60), and because $\pi^{\bar{q}}_{\bar{\theta}^*_k}(w))=\pi_{\bar{\theta}^*,\hbar^*}(w)$ for all $w\in\mathcal{W}$, we have that

$$
(\mu^{\bar{q}}_{\bar{\theta}^*_k}(1,w)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(1,\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)))(1-\pi_{\bar{\theta}^*,\hbar^*}(w))+(\mu^{\bar{q}}_{\bar{\theta}^*_k}(0,w)-\bar{\mu}^{\bar{q}}_{\bar{\theta}^*}(0,\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)))\pi_{\bar{\theta}^*,\hbar^*}(w)=\left[s^{\bar{q}}_{\bar{\theta}^*}(w)\right]^2\hbar^*(w),
$$

where $s^{\bar{q}}_{\bar{\theta}^*}(w):=\sqrt{\pi^{\bar{q}}_{\bar{\theta}^*_k}(w)(1-\pi^{\bar{q}}_{\bar{\theta}^*_k}(w))}$. Therefore, the left-hand side of (2.62) is equal to

$$
E\left\{[s^{\bar{q}}_{\bar{\theta}^*}(W)q_j(W)][s^{\bar{q}}_{\bar{\theta}^*}(W)\hbar^*(W)]\right\}.
$$

The above identity implies that the vector $c^{\bar{q}}_{\bar{\theta}^*_k}$ can be re-expressed as

$$c^{\tilde{q}}_{\bar{\theta}^*_k} := \begin{pmatrix} E\left[\tilde{q}_1(W)\tilde{\hbar}(W)\right] \\ \vdots \\ E\left[\tilde{q}_{p+k+1}(W)\tilde{\hbar}(W)\right] \end{pmatrix}, \tag{2.63}$$

where, for each $i \in \{1, \ldots, p+k+1\}$ and $w \in \mathcal{W}$, $\tilde{q}_i(w) := s^{\bar{q}}_{\bar{\theta}^*}(w) q_i(w)$, and $\tilde{\hbar}(w) := s^{\bar{q}}_{\bar{\theta}^*} \hbar^*(w)$. Next, let $\tilde{q}(w) := (\tilde{q}_1(w), ...\tilde{q}_{p+k+1}(w))$ denote the $p+k+1$ real valued-vector, where for each $i \in \{1, \ldots, p+k+1\}$, each entry is equal to the corresponding $\tilde{q}_i$ function evaluated at w. Finally, note that $I^{\bar{q}}_{\bar{\theta}^*_k}$ can be re-expressed as $E\left[\tilde{q}(W)\tilde{q}(W)^\top\right]$.

Let $I^{\tilde{q}}_{\bar{\theta}^*_k} := E\left[\tilde{q}(W)\tilde{q}(W)^\top\right]$. Because $I^{\tilde{q}}_{\bar{\theta}^*_k}$ is invertible, the covariates $\{\tilde{q}_1(w), \ldots, \tilde{q}_{p+k+1}(w)\}$ must be linearly independent. Thus, we can construct an orthonormal basis of the space spanned by the functions in $\{\tilde{q}_1, \ldots, \tilde{q}_{p+k+1}\}$ as follows. Let B be a matrix such that $B^2 = (I^{\tilde{q}}_{\bar{\theta}^*})^{-1}$. Then, for every $i \in \{1, \ldots, p+k+1\}$, let a_i be a $\mathcal{W} \to \mathbb{R}$ function defined as $a_i(w) := \sum_{j=1}^{p+k+1} B_{ij}\tilde{q}_j(w)$, and, for a given w, let $\bar{a}(w) := (a_1(w), \ldots a_{p+k+1}(w))$. By construction, the functions $\{a_1, ...a_{p+k+1}\}$ are an orthonormal basis of the space spanned by the functions $\{\tilde{q}_1, \ldots, \tilde{q}_{p+k+1}\}$ with Gram matrix $I^{\tilde{q}}_{\bar{\theta}^*}$. Additionally, note that

$$E[a_i(W)\tilde{\hbar}(W)] = \sum_{j=1}^{p+k+1} B_{ij} E[\tilde{q}_j(W)\tilde{\hbar}(W)]. \tag{2.64}$$

Before proceeding further, we introduce some notation. For $f_1, \ldots, f_j \in L^2(P_W)$, let $\text{span}\{f_1, \ldots, f_j\}$ denote the linear span of $f_1, \ldots, f_j$. For a subspace $\mathcal{S}$ of $L^2(P_W)$, let $\mathcal{S}^\perp$ denote the orthogonal complement of $\mathcal{S}$ in $L^2(P_W)$ and let $\Pi(f|\mathcal{S})$ denote the $L^2(P_W)$-projection of f onto $\mathcal{S}$. When $\mathcal{S}$ is a one-dimensional space spanned by the function a_i, we write $\Pi(f \mid a_i)$ to mean $\Pi(f \mid \text{span}\{a_i\})$. Given this notation, we can decompose $\tilde{\hbar}$ as $\tilde{\hbar} = \Pi(\tilde{\hbar}|\text{span}\{a_1, \ldots, a_{k+p+1}\}^\perp) + \sum_{i=1}^{k+p+1} \Pi(\tilde{\hbar}|a_i)$. Then, letting

$$c^{\bar{a}}_{\bar{\theta}^*_k} := \begin{pmatrix} E\left[a_1(W)\tilde{\hbar}(W)\right] \\ \vdots \\ E\left[a_{p+k+1}(W)\tilde{\hbar}(W)\right] \end{pmatrix}, \tag{2.65}$$

and by Equations 3.13 and 3.14, $\text{Eff}(\mathcal{M}^{\bar{\text{q}}})$ can be upper bounded as follows:

$$\begin{aligned}\text{Eff}(\mathcal{M}^{\bar{\text{q}}}) &= c_{\bar{\theta}^*_k}^{\bar{q}\top}(I_{\bar{\theta}^*}^{\bar{q}})^{-1}c_{\bar{\theta}^*_k}^{\bar{q}} = c_{\bar{\theta}^*_k}^{\tilde{q}\top}(I_{\bar{\theta}^*}^{\tilde{q}})^{-1}c_{\bar{\theta}^*_k}^{\tilde{q}} = c_{\bar{\theta}^*_k}^{\tilde{q}\top}BBc_{\bar{\theta}^*_k}^{\tilde{q}} = c_{\bar{\theta}^*_k}^{\bar{a}\top}c_{\bar{\theta}^*_k}^{\bar{a}} \\ &= \sum_{i=1}^{k+p+1} E^2[a_i(W)\tilde{\hbar}(W)] = \sum_{i=1}^{k+p+1} E^2\left[a_i(W)\Pi(\tilde{\hbar}|a_i)(W)\right] \\ &\leq \sum_{i=1}^{k+p+1} E[a_i^2(W)]E\left[\Pi(\tilde{\hbar}|a_i)^2(W)\right] = \sum_{i=1}^{k+p+1} E\left[\Pi(\tilde{\hbar}|a_i)^2(W)\right] \leq E\left[\tilde{\hbar}^2(W)\right].\end{aligned} \tag{2.66}$$

Now, consider the sub-model that contains $\hbar^*$ as an augmentation covariate:

$$\mathcal{M}_{\hbar^*} = \{\text{logit}(\pi_{\bar{\theta}}^{\hbar^*}(w)) = \theta_1^\top w + \theta_2\hbar^*(w) :\ \theta_1 \in \mathbb{R}^{p+1},\ p \in \mathbb{N}, \theta_2\ \in \mathbb{R}, I_{\bar{\theta}^*,\hbar^*} \text{ is invertible}\}.$$

For $\bar{q}(w) = (w_1, w_2, \ldots, w_{p+1}, \hbar^*(w))$ it holds that $\mathcal{M}_{\hbar^*} = \mathcal{M}^{\bar{q}}$. Applying (2.66) for this particular choice of $\bar{q}$ and then using Parseval's identity shows that

$$\text{Eff}(\mathcal{M}_{\hbar^*}) = \sum_{i=1}^{p+2} E^2\left[a_i(W)\Pi(\tilde{\hbar}|a_i)(W)\right] = E\left[\Pi(\tilde{\hbar} \mid \text{span}\{a_1, \ldots, a_{p+2}\})(W)^2\right].$$

Because $\bar{q}_{p+2} = \hbar^*$, it holds that $\tilde{\hbar} = \Pi(\tilde{\hbar} \mid \text{span}\{a_1, \ldots, a_{p+2}\})$. Hence, the above yields that $\text{Eff}(\mathcal{M}_{\hbar^*}) = \text{E}[\tilde{\hbar}^2(\text{W})]$, which shows that the inequality in (2.66) is saturated for this particular choice of $\bar{q}$.

□

2.F Proof of Corollary 2.4

In the statement and proof of the next corollary, we let $\bar{\pi}(a|w) := \pi_{\bar{\theta}^*,\hbar^*}(w)^a[1-\pi_{\bar{\theta}^*,\hbar^*}(w)]^{1-a}$. We also write $\bar{\pi}_1(W)$ to mean $\bar{\pi}(1|w)$. Additionally, for notational brevity, and because all expectations are taken with respect to $P_{\bar{\theta}^*,\hbar^*}$, we will drop the subscripts $\bar{\theta}^*, \hbar^*$ from our notation. For instance, we will use the notation $\mu(a, w)$ and $\bar{\mu}(a, \pi_1(w))$ to denote $\mu_{\bar{\theta}^*,\hbar^*}(a, w)$ and $\bar{\mu}_{\bar{\theta}^*,\hbar^*}(a, \pi_1(w))$, respectively.

Corollary 2.4. *Let $\hat{\tau}_e$ be the efficient asymptotic estimator of the average treatment effect, with influence function $(w, a, y) \mapsto \frac{2a-1}{\bar{\pi}(a|w)}(Y - \mu(a, w)) + (\mu(1, w) - \mu(0, w) - \tau)$, and let σ_e^2*

denote its asymptotic variance. It holds that

$$\sigma_M^2 - \mathrm{Eff}(\mathcal{M}_{\hat{\pi}^*}) = \sigma_e^2 + g_M, \tag{2.67}$$

where, for $\Sigma(a,w) = \mathrm{Var}[\mu(a,W)|\bar{\pi}_1(W) = \bar{\pi}_1(w)]$,

$$g_M = \frac{1}{2M}\sum_{a=0}^{1} E\left\{\left[\frac{1}{\bar{\pi}(a|W)} - \bar{\pi}(a|W)\right]\left[\sigma^2(a,W) + \Sigma(a,W)\right]\right\}. \tag{2.68}$$

Proof. Let $\gamma(w) = \mu(1,w) - \mu(0,w)$, and $\bar{\gamma}(w) = \bar{\mu}(1,w) - \bar{\mu}(0,w)$. Now, let s_M, be a $(0,1) \to \mathbb{R}$ function defined as

$$s_M(p) := p^{-1} + (2M)^{-1}\left(p^{-1} - p\right).$$

We write $s_M \circ \bar{\pi}(a|w)$ to mean $s_M(\bar{\pi}(a|w))$. Then, σ_M^2 can be expressed as

$$\sigma_M^2 = E\left[(\bar{\gamma}(W) - \tau)^2\right] + \sum_{a=0}^{1} E\left[\bar{\sigma}^2(a, \bar{\pi}_1(W)) s_M \circ \bar{\pi}(a|W)\right].$$

Next, note that the variance of an asymptotically efficient estimator is equal to

$$\sigma_e^2 = E\left[\left(\frac{2A-1}{\bar{\pi}(A|W)}(Y - \mu(A,W)) + (\gamma(W) - \tau)\right)^2\right] = E\left[\sum_{a=0}^{1}\frac{\sigma^2(a,W)}{\bar{\pi}(a|W)}\right] + E\left[(\gamma(W) - \tau)^2\right].$$

Hence,

$$\begin{aligned}
\sigma_M^2 - \sigma_e^2 &= E\left[(\bar{\gamma}(W) - \tau)^2\right] + \sum_{a=0}^{1} E\left[\bar{\sigma}^2(a, \bar{\pi}_1(W)) s_M \circ \bar{\pi}(a|W)\right] \\
&\quad - \left\{E\left[\sum_{a=0}^{1}\frac{\sigma^2(a,W)}{\bar{\pi}(a|W)}\right] + E\left[(\gamma(W) - \tau)^2\right]\right\} \\
&= E\left[(\bar{\gamma}(W) - \tau)^2\right] - E\left[(\gamma(W) - \tau)^2\right] + \sum_{a=0}^{1} E\left[\bar{\sigma}^2(a, \bar{\pi}_1(W)) s_M \circ \bar{\pi}(a|W) - \frac{\sigma^2(a,W)}{\bar{\pi}(a|W)}\right].
\end{aligned}$$

Let $\Gamma(w) := \mathrm{Var}(\gamma(W)|\bar{\pi}_1(W) = \bar{\pi}_1(w))$. Noting that $\tau = E[\bar{\gamma}(W)] = E[\gamma(W)]$ and $E[\gamma(W)|\bar{\pi}_1(W)] = \bar{\gamma}(W)$ almost surely, it is not difficult to show that $E\left[(\bar{\gamma}(W) - \tau)^2\right] - E\left[(\gamma(W) - \tau)^2\right] = -E[\Gamma(W)]$. Hence, the above rewrites as follows:

$$\sigma_M^2 - \sigma_e^2 = -E[\Gamma(W)] + \sum_{a=0}^{1} E\left[\bar{\sigma}^2(a, \bar{\pi}_1(W)) s_M \circ \bar{\pi}(a|W) - \frac{\sigma^2(a,W)}{\bar{\pi}(a|W)}\right]. \tag{2.69}$$

We now begin to study the latter term above. By the law of total variance, the following holds for P-almost all w:

$$\sum_{a=0}^{1}\left[\bar{\sigma}^2(a,\bar{\pi}_1(w))s_M\circ\bar{\pi}(a|w)\right]=\sum_{a=0}^{1}\left[E\left\{\sigma^2(a,W)|\bar{\pi}_1(W)=\bar{\pi}_1(w)\right\}s_M\circ\bar{\pi}(a|w)+\Sigma(a,w)s_M\circ\bar{\pi}(a|w)\right].$$

Integrating w on both sides against the marginal distribution under P and applying the law of total expectation then shows that

$$\sum_{a=0}^{1}E\left[\bar{\sigma}^2(a,\bar{\pi}_1(W))s_M\circ\bar{\pi}(a|W)\right]=\sum_{a=0}^{1}E\left[\sigma^2(a,W)s_M\circ\bar{\pi}(a|W)+\Sigma(a,W)s_M\circ\bar{\pi}(a|W)\right].$$

Plugging this into the latter term in (2.69) and simplifying, we find that

$$\sum_{a=0}^{1}E\left[\bar{\sigma}^2(a,\bar{\pi}_1(W))s_M\circ\bar{\pi}(a|W)-\frac{\sigma^2(a,W)}{\bar{\pi}(a|W)}\right]=\sum_{a=0}^{1}E\left[\frac{\Sigma(a,W)}{\bar{\pi}(a|W)}\right]+g_M.$$

Hence,

$$\sigma_M^2-\sigma_e^2=-E[\Gamma(W)]+\sum_{a=0}^{1}E\left[\frac{\Sigma(a,W)}{\bar{\pi}(a|W)}\right]+g_M. \tag{2.70}$$

Let $\Delta(a,w):=\mu(a,w)-\bar{\mu}(a,\bar{\pi}_1(w))$. Now, because

$$\text{Eff}(\mathcal{M}_{\hbar^*})=E\left[\left(\sqrt{\pi_{\bar{\theta}^*,\hbar^*}(W)(1-\pi_{\bar{\theta}^*,\hbar^*}(W))}\hbar^*(W)\right)^2\right] \tag{2.71}$$

we have that $\text{Eff}(\mathcal{M}_{\hbar^*})$ is equal to

$$\text{Eff}(\mathcal{M}_{\hbar^*})=E\left[\bar{\pi}_1(W)\bar{\pi}_0(W)\hbar^{*2}(W)\right]=\sum_{a=0}^{1}E\left[\frac{1-\bar{\pi}(a|W)}{\bar{\pi}(a|W)}\Delta^2(a,W)\right]+2E\left[\Delta(1,W)\Delta(0,W)\right].$$

Next, note that for a given $a\in\{0,1\}$, and for P-almost all w, $\Sigma(a,w)=E\left[\Delta^2(a,W)|\bar{\pi}_1(W)=\bar{\pi}_1(w)\right]$. Applying the law of total expectation on the right-hand side above and plugging this in yields that

$$\text{Eff}(\mathcal{M}_{\hbar^*})=\sum_{a=0}^{1}E\left[\left(\frac{1}{\bar{\pi}(a|W)}-1\right)\Sigma(a,W)\right]+2E\left\{\Delta(1,W)\Delta(0,W)\right\}.$$

Now, let $\xi(w) := \text{Cov}\left(\mu(1,W), \mu(0,W) | \bar{\pi}_1(W) = \bar{\pi}_1(w)\right)$, and note that $\Gamma(w) = \sum_{a=0}^{1} \Sigma(a,w) - 2\xi(w)$ for P-almost all w. Plugging this into the above yields that

$$\text{Eff}(\mathcal{M}_{\text{ñ}^*}) = -E\left[\Gamma(W)\right] - 2E\left[\xi(W)\right] + \sum_{a=0}^{1} E\left\{\frac{\Sigma(a,W)}{\bar{\pi}(a|W)}\right\} + 2E\left[\Delta(1,W)\Delta(0,W)\right]. \quad (2.72)$$

By the tower property, $E\left[\mu(a,W)\bar{\mu}(1-a,\bar{\pi}_1(W))\right] = E\left[\bar{\mu}(a,\bar{\pi}_1(W)\bar{\mu}(1-a,\bar{\pi}_1(W))\right]$ for $a \in \{0,1\}$. Hence,

$$E\left[\Delta(1,W)\Delta(0,W)\right] = E\left[\mu(1,W)\mu(0,W)\right] - E\left\{\bar{\mu}(0,\bar{\pi}_1(W))\bar{\mu}(1,\bar{\pi}_1(W))\right\}. \quad (2.73)$$

Moreover, the definition of conditional covariance and the tower property imply that

$$E\left[\xi(W)\right] = E[\mu(1,W)\mu(0,W)] - E\left\{\bar{\mu}(1,\bar{\pi}_1(W)\bar{\mu}(0,\bar{\pi}_1(W))\right\}. \quad (2.74)$$

Combining Equations 2.72, 2.73, and 2.74, we have that

$$\text{Eff}(\mathcal{M}_{\text{ñ}^*}) = -E[\Gamma(W)] + \sum_{a=0}^{1} E\left[\frac{\Sigma(a,W)}{\bar{\pi}(a|W)}\right].$$

Equation 2.70 and the above identity show that

$$\sigma_M^2 - \sigma_e^2 = \text{Eff}(\mathcal{M}_{\text{ñ}^*}) + g_M.$$

□

Chapter 3

CAUSAL ISOTONIC CALIBRATION

3.1 Introduction

Estimation of causal effects via both randomized experiments and observational studies is critical to understanding the effect of interventions and to improve decision making. Moreover, it is often the case that understanding treatment effect heterogeneity can provide more insights than overall population effects (Obermeyer and Emanuel 2016, Athey 2017). For instance, treatment effect heterogeneity can help elucidate how the mechanism of an intervention is acting at a more granular level, may help design policies targeted to sub-populations who can be most benefited to them (Imbens and Wooldridge 2009), and can also be used to investigate whether interventions can be implemented in populations that have different characteristics than the ones for which they were developed. These necessities have arisen in a wide range of fields, such as marketing (Devriendt et al. 2018), the social sciences (Imbens and Wooldridge 2009), and the health sciences (Kent et al. 2018).

In the health sciences, heterogeneous treatment effects (HTEs) are of high importance to understanding and quantifying how certain exposures or interventions affect the health of various sub-populations (Dahabreh et al. 2016, Lee et al. 2020). For instance, it can be beneficial to prioritize treatment to certain sub-populations when treatment resources are scarce, or to individualize treatment assignments when the treatment can have no effect (or even be harmful) in certain sub-populations (Dahabreh et al. 2016). As an example, treatment assignment based on risk scores has been used to provide clinical guidance in cardiovascular disease prevention (Lloyd-Jones et al. 2019), and to improve decision-making in oncology (Cucchiara et al. 2018, Collins and Varmus 2015). More recently, Q-learning has been developed to formulate treatment decision rules in personalized medicine (Clifton and

Laber 2020, Murphy 2003). Broadly, Q-learning can be used to derive optimal treatment strategies by maximizing a utility function. Implementation of Q-learning relies on estimation of regret functions based on HTEs. Therefore, accurate estimation of HTEs is of vital importance to enhance decision making and improving the delivery of precision medicine.

A wide range of statistical methods are available for assessing HTEs; recent examples include Wager and Athey (2018), Carnegie et al. (2019), Lee et al. (2020) and Nie and Wager (2021), among others. In particular, many methods scrutinize HTEs via conditional average treatment effects (CATEs); this is the case, for example, in Imbens and Wooldridge (2009) and Dominici et al. (2020). Under causal identification conditions, the CATE is the difference in the conditional mean of the counterfactual outcome corresponding to treatment versus control given covariates, which can be defined at a group or individual level. When interest lies in predicting treatment effect, the CATE can be viewed as the oracle predictor of the individual treatment effect that can still feasibly learned from data. Optimal treatment rules have been derived based on the sign of the CATE estimator (Murphy 2003, Robins 2004), with more recent works incorporating the use of data-adaptive CATE estimators (Luedtke and Van Der Laan 2016). Thus, due to its wide applicability and scientific relevance, CATE estimation has been of great interest in statistics and data science.

CATE estimation can be challenging for many reasons, including that the nuisance parameters involved in its estimation can be highly non-smooth, high-dimensional, or otherwise difficult to model correctly. Usually, CATE estimators (often referred to as learners) build upon estimators of the conditional mean outcome given covariates and treatment level, of the propensity score, or both. For instance, plug-in estimators such as those studied in Künzel et al. (2019) —also known as T-learners— are obtained by taking the difference between estimators of the conditional mean outcome given covariates and treatment level obtained separately for each treatment level. Potential pitfalls of plug-in estimators is that they rely on estimation of nuisance parameters that are at least as non-smooth or high-dimensional as the CATE, and are prone to the misspecification of the involved outcome regression models; these issues can result in the slow convergence or inconsistency of the CATE estimator.

Recently, doubly-robust (DR) methods have been developed as a means of preserving consistency under potentially inconsistent estimation of either the outcome regression or propensity score Kennedy (2020). Other available non-parametric nearest-neighbor methods, such as matching (Crump et al. 2008), may fail to handle a large number of covariates in practice (Wager and Athey 2018).

Even when estimation strategies are robust enough to avoid model misspecification, or if the smoothness conditions hold, these can often fail to produce useful predictions. For instance, it has been found that flexible statistical methods often produce biased predictions that overestimate (or underestimate) the true CATE in the extremes of the predicted values (van Klaveren et al. 2019). Additionally, due to the high variability in the tails of the estimated effects, an individuals or groups predicted risk could be severely overestimated or underestimated.

The implications of a biased method in the tails of the predicted values are profound if the CATE is used as a support tool to guide therapy, and range from harmful application of treatment to unfair algorithms (Van Calster et al. 2019). For example, it has been found that the pooled cohort equations (Goff et al. 2014) risk model used to predict cardiovascular disease tends to underpredict observed events in samples with lower socioeconomic status or with chronic inflammatory diseases (Lloyd-Jones et al. 2019). On the other end of the spectrum, if the CATE is overestimated, physicians may be overconfident and assign unnecessary treatment with the idea that the strength of the estimated CATE is sufficient evidence to provide treatment. Therefore, it is desirable that a method produces unbiased estimates across all stratums defined by its predicted values. In prediction settings, this property is commonly known as calibration.

Informally, a well-calibrated predictor guarantees that the average predicted value within a fixed bin is approximately equal to the conditional mean of the outcome given that the predictor lies inside the bin. In the machine learning literature, calibration has been widely used to enhance prediction models (Bella et al. 2010). In our context, if a CATE estimator is well calibrated, then for arbitrary subpopulations defined by quantiles of the estimated

CATE, we would observe that the true average treatment effect on those patients will be close to their average fitted values. That is, if well calibrated, CATE predictors can provide informative scores and treatment-assignments which, when utilized in decision-making, can lead to improved rewards over the status-quo. It is thus important to develop statistical methods for developing treatment-effect predictors that can provide useful treatment rules implied by the predictor under minimal assumptions. Hence, there is a need to develop statistical methods that can leverage the predictive power of data adaptive methods that estimate the CATE as well as provide useful decision rules under weak assumptions.

While calibration has been widely studied in predictive analytic settings, to our knowledge little work has been done specifically for calibrating CATE estimators. Recently, Leng and Dimmery (2021) proposed a calibration method for the CATE via maximum likelihood estimation. However, their calibration method relies on availability of experimental randomized data, which is often not the case. This gap in the literature is important to address because, regardless of whether an initial treatment effect score is a moderate or accurate approximation of the true CATE, its predictions can be of use if they are well calibrated. In this work, we propose to fill this gap by proposing a method that improves a given CATE estimator guaranteeing it is well calibrated while preserving its predictive power. In particular, we propose using cross-fitting with pseudo-outcomes to carry the calibration step, which, as far as we know, has also not been previously considered.

This work is structured as follows. In Section 3.2, we introduce general notation, formally define calibration and its measure. In Section 3.3 we provide an overview of current calibration methods, and, more specifically, describe isotonic regression, a commonly used calibration method which we use in our method. In Section 3.4 we outline our proposed method, causal isotonic calibration, and, in Section 3.5, state its theoretical properties in large and finite samples. Specifically, we show the convergence rate of the calibration measure and provide a bound of its predictive accuracy. In Section 3.6 we examine the performance of our method via a suite of simulation studies that consider a wide scope of scenarios, ranging from relatively simple CATE structures to more complex scenarios where the CATE is

challenging to estimate. We conclude with a discussion of our findings and provide guidance for applying our method in Section 3.7.

3.2 Statistical Setup

3.2.1 Notation & Definitions

Suppose we observe n independent identically distributed (iid) realizations of observations $O = \{W, A, Y\}$ drawn from a distribution P, where $W \in \mathcal{W} \subset \mathbb{R}^d$ is a vector of baseline covariates, $A \in \{0, 1\}$ is a binary indicator of treatment, and $Y \in \mathcal{Y} \subset \mathbb{R}$ is a numerical outcome variable. For instance, W can consist of demographic characteristics as well as medical history of an individual, A can denote the indicator that takes the value 1 if an individual is treated and 0 otherwise, and Y could be a binary indicator that encodes whether the treatment is successful. Let $\mathcal{D}_n := \{(W_i, A_i, Y_i)\}_{i=1}^n$ denote the observed dataset. Based on the potential outcomes framework (Rubin 1974), we denote by (Y_0, Y_1) the potential outcomes that would have been observed in the counterfactual case where treatment $A = 0$ and $A = 1$ were assigned, respectively.

Let $\tau : \mathcal{W} \to \mathbb{R}$ be a function that maps a realization w of the random variable W to a treatment-effect score $\tau(w)$. For now, we treat τ as a fixed function. Later, we will provide describe how τ could be obtained via sample splitting.

Let the individual treatment effect be equal to $Y_1 - Y_0$. Without loss of generality we assume that higher values of $Y_1 - Y_0$ correspond to better outcomes. We denote the true CATE by $\tau_0(w) := E[Y_1 - Y_0 \mid W = w]$. Next, we let $\pi_0(w) := P(A = 1|W = w)$ denote the probability of treatment for a fixed covariate w. Additionally, and given treatment level $a \in \{0, 1\}$ and covariate $w \in \mathcal{W}$, we let $\mu_0(a, w) := E[Y|A = a, W = w]$ denote the conditional mean of Y given a and w. Finally, we let $\gamma_0(\tau, w) := E[Y_1 - Y_0|\tau(W) = \tau(w)]$ denote the conditional mean of the individual treatment effect given that the treatment-effect score is equal to $\tau(w)$. Note that, by the tower property, $\gamma_0(\tau, w)$ is also equal to $E[\tau_0(W)|\tau(W) = \tau(w)]$. As such, integrals that only involve $\gamma_0(\tau, w)$ and additional functions

of $w \in \mathcal{W}$ can be taken with respect the marginal distribution of W, which we denote by P_W.

3.2.2 Measuring Calibration and the Calibration-Distortion Decomposition

Various definitions of calibration of risk predictors have been proposed in the literature — we refer the reader to Gupta and Ramdas (2021) and Gupta et al. (2020a) for a review. As such, there are different ways to assess how well calibrated a risk predictor is. In this section, we outline our definition of calibration as well as our rationale for our choice. Given a treatment effect score $\tau(w)$, the best predictor of the individual treatment effect in terms of mean-squared error is $\gamma_0(\tau, w) := E[Y_1 - Y_0|\tau(W) = \tau(w)]$. By the law of total expectation, the function $\gamma_0(\tau, w)$ has the property that, for a given interval $[a, b)$,

$$E[\tau_0(W)I(\gamma_0(\tau, W) \in [a, b))] = E\left[\gamma_0(\tau, W)I(\gamma_0(\tau, W) \in [a, b))\right]. \tag{3.1}$$

Then, given new observations $\mathcal{S} = \{W_i, A_i, Y_i\}_{i=1}^{n_s}$ drawn from P, $\gamma_0(\tau, w)$ will satisfy

$$\frac{1}{n}\sum_{i=1}^{n_s} \tau_0(W_i)I(a \leq \gamma_0(\tau, W_i) < b) \approx \frac{1}{n}\sum_{i=1}^{n_s} \gamma_0(\tau, W_i)I(a \leq \gamma_0(\tau, W_i) < b).$$

Thus, $\gamma_0(\tau, w)$ is approximately unbiased within bins of predictions $[a, b)$ when evaluated in the independent data set $\mathcal{S}$.

The above property in Equation 3.1 shows that $\gamma_0(\tau, \cdot)$ is perfectly calibrated in the interval $[a, b)$. Therefore, when a given risk predictor τ is such that $\tau(W) = \gamma_0(\tau, W)$ with P-probability one, it is said that τ is perfectly calibrated Gupta et al. (2020b).

In general, perfect calibration is an unrealistic property for a treatment-effect predictor to have. It thus desirable that $\tau(w)$ is as close to $\gamma_0(\tau, w)$ as possible, motivating our definition of calibration. This leads us to define the following calibration measure of τ:

$$CAL(\tau) := \int \left[\gamma_0(\tau, w) - \tau(w)\right]^2 dP_W(w). \tag{3.2}$$

The above calibration measure is known as the ℓ_2-expected calibration error for predictors of treatment heterogeneity Xu and Yadlowsky (2022) in treatment-effect settings, or as the the

ℓ_2-expected calibration error Gupta et al. (2020b) in prediction settings. Observe that the calibration measure in Equation 3.2 is 0 if τ is perfectly calibrated. Additionally, note that in Equation 3.2 our choice of measure is P_W; however, we could have used a different measure other than P_W. Such cases can occur, for instance, when there is a change of population that results in covariate shift and we are interested in measuring how well τ is calibrated in the new population.

Interestingly, the above calibration measure plays a role in the mean squared error decomposition between the treatment predictor and the true CATE. In particular,

$$\begin{aligned}\int \left[\tau_0(w) - \tau(w)\right]^2 dP_W(w) &= \int \left[\tau(w) - \gamma_0(\tau, w)\right]^2 dP_W(w) + \int \left[\tau_0(w) - \gamma_0(\tau, w)\right]^2 dP_W(w) \\ &= CAL(\tau) + DIS(\tau), \end{aligned} \tag{3.3}$$

where $DIS(\tau) := E\left[Var\{\tau_0(W)|\tau(W)\}\right]$. We refer to the above as the calibration-distortion decomposition of the mean-squared error. The first term on the right-hand side measures the calibration of the treatment effect predictor. To interpret the second term, we envision a scenario where a distorted message is passed between two parties. The goal is for the second party to discern the value of $\tau_0(w)$, where the value of $w \in \mathcal{W}$ is only known to the first party. The first party transmits w, which is then distorted through a function τ and received by the second party. The second party knows the functions τ and τ_0, and may use this information to try to discern $\tau_0(w)$. If τ is one-to-one, then $\tau_0(w)$ can be discerned by simply applying $\tau_0 \circ \tau^{-1}$ to the received message $\tau(w)$. More generally, if there exists a function f such that $\tau_0 = f \circ \tau$, then the second party can recover the value of $\tau_0(w)$. For example, f may be the identity function so that $\tau = \tau_0$. If no such function f exists, then it may not be possible for the second party to recover the value of $\tau_0(w)$. Instead, they may predict $\tau_0(w)$ based on $\tau(w)$ via $\gamma_0(\tau, w)$. Averaged across $W \sim P_W$, the mean-squared error of this approach is equal to $E\left[Var\{\tau_0(W)|\tau(W)\}\right]$, which corresponds to $DIS(\tau)$. Given this analogy to this distorted message scenario, we refer to the second term as a distortion term.

The calibration-distortion decomposition shows that, at a given level of distortion, better-calibrated treatment effect predictors will have lower mean-squared error for the true CATE

function. We will explore this fact later in this work when showing that, in addition to improving calibration, our proposed calibration procedure can improve the mean-squared error of many treatment effect predictors.

3.3 Review of Calibration Methods and Objective of this Work

A common approach to improving the calibration measure of τ is to apply an additional function $\theta : \mathbb{R} \to \mathbb{R}$, such that $\theta \circ \tau(w)$ is a mapping of the original predictions with the aim that $CAL(\theta \circ \tau)$ is smaller than $CAL(\tau)$. Such a function θ is known as a calibrator, because its goal is to calibrate τ. Calibration methods use the dataset $\mathcal{D}_n$ and the known treatment effect predictor τ to construct a calibrator θ_n .

Currently, among the most common calibration methods used in practice — see Huang et al. (2020) for an in-depth overview — are Platt's scaling (Platt et al. 1999), histogram binning Zadrozny and Elkan (2001), Bayesian binning into quantiles (Naeini et al. 2015), and isotonic calibration (Barlow and Brunk 1972). The aforementioned methods have been developed in standard regression problems , but not in the context of CATE calibration.

Depending on the setting, the aforementioned calibration methods have their strengths and limitations. Broadly, Platt's scaling is designed for binary outcomes and uses the estimated values of the treatment effect predictor to fit a parametric regression model of the form $\text{logit}P(Y = 1|\tau(W) = t) = \alpha + \beta t$ for $\alpha, \beta \in \mathbb{R}$. An advantage of Platt's scaling method is its straightforward implementation and familiarity. A disadvantage is that the parametric model may not be representative of the true distribution of the data. Histogram binning, also known as quantile binning, is also designed for binary data and consists in dividing the sorted values of the treatment effect predictor into a fixed number of bins. Given a prediction value, the method finds the bin containing the prediction and returns the empirical mean of the observed outcome. A major limitation of histogram binning is that it requires specification of the number of bins *a priori*. Bayesian binning into quantiles improves upon histogram binning by considering multiple binning models and their combinations. A drawback of this approach is that it requires pre-specification of the total number of binning models that are

combined as well as the prior distribution of each of the binning models. Finally, isotonic calibration consists in finding a monotone nondecreasing function that minimizes the empirical mean squared error of the observed outcome and the calibrated estimates. As such, it is also non-parametric and does not require any additional specification of tuning parameters. While the monotonicity assumption may be restrictive, it is not an unreasonable transformation of any risk predictor as one expects that the relationship between the predicted and true values is approximately monotone.

In this work, we describe and study isotonic calibration of treatment effect predictors, which to our knowledge has not previously been examined. We have selected this approach to leverage its theoretical properties, which we outline in the next sections and are shown in more in detail in the Appendix. Finally, it is worth noting that aside from its theoretical guarantees, isotonic calibration has the advantage of not requiring additional tuning parameters, facilitating its use when applying the method in practice.

3.4 Causal Isotonic Calibration of treatment effect predictors

In this section, we define our method, causal isotonic calibration, and, in Section 3.5, we show its theoretical properties. In order to provide more context of why these properties are important, we first outline general notions of what a treatment effect calibrator should satisfy.

3.4.1 Desiderata of treatment-effect calibrators

Let θ_n denote a calibrator of τ constructed based on a sample of size n. Naturally, the goal of the calibrator θ_n is that $CAL(\theta_n \circ \tau)$ shrinks to 0 at a fast rate as sample size increases, but it is also desirable that $\theta_n \circ \tau$ is as predictive as the initial treatment-effect score τ. That is, we want that θ_n to satisfy both of the following:

Property 1. The rate δ_n at which $CAL(\theta_n \circ \tau)$ converges to zero is sufficiently fast.

Property 2. The calibrated treatment-effect predictor $\theta_n \circ \tau$ should be comparably predictive for the treatment effect τ_0 as the initial predictor τ.

The "sufficient" rate in Property 1 will depend on the application of interest. Property 2 requires that the calibrator does not destroy the predictive-power of the initial predictor in the process of attaining Property 1, which would occur if the calibration term in the decomposition in (3.3) were made small at the cost of a dramatic inflation of the distortion term. It is worth noting that there exist cases where a calibrated risk score can be a poor predictor of the risk score. For instance, an estimator of the average treatment effect (ATE) has a very fast calibration rate, but poor predictive performance.

Example It is known that the marginal average treatment-effect (ATE) $\psi = E[Y_1 - Y_0]$ can be estimated at a rate of $n^{-1/2}$ using, for instance, an augmented inverse propensity weighted estimator (Robins et al. 1995) or a targeted maximum likelihood estimator (Van der Laan et al. 2011). Consider a calibrator θ_n that completely ignores the initial predictor τ and outputs τ_n, where τ_n is a constant function that equals a $n^{-1/2}$-consistent estimator for the ATE computed from the dataset $\mathcal{D}_n$. Since τ_n is a constant function, we have $\gamma_0(\tau_n, w) = E[Y_1 - Y_0]$ is exactly the ATE. Thus, by design,

$$CAL(\tau_n) = \int \left[\gamma_0(\tau_n, w) - \tau_n(w)\right]^2 dP(w) = \int \left\{E[Y_1 - Y_0] - \tau_n(w)\right\}^2 dP(w) = O_P(n^{-1}).$$

The calibrator τ_n attains a very fast calibration rate of n^{-1} and satisfies Property 1. However, the calibrated predictor is constant and thus has caused the distortion term to take its maximal possible value, namely $Var[\tau_0(W)]$. Therefore, estimators of the ATE can be calibrators that can satisfy Property 1, but violate Property 2.

3.4.2 Causal isotonic calibration

Inspired by isotonic calibration from the machine learning literature, we propose a calibration method for treatment effects which we refer to as *causal isotonic calibration*. In this

subsection, we define the method and, in the following section, we detail the sense in which it satisfies the two desiderata described earlier.

First, let $\mathcal{F}_{iso}$ be the space of real-valued monotone non-decreasing functions, that is, $\mathcal{F}_{iso} := \{\theta : \mathbb{R} \to \mathbb{R};\ \theta \text{ is monotone non-decreasing}\}$. Broadly, isotonic regression is a univariate regression method whose solution is a function that belongs to $\mathcal{F}_{iso}$ and is such that it minimizes the empirical risk based on a mean squared error loss. Section **??** in the Appendix describes isotonic regression in more detail, and shows an interesting theoretical property that is used in our results.

Then, let τ be a given treatment-effect predictor, and $\mathcal{D}_n = \{(W_i, A_i, Y_i)\}_{i=1}^n$ the calibration data. Additionally, for an observation o, let

$$D_0(o) := \mu_0(1, w) - \mu_0(0, w) + \left(\frac{a}{\pi_0(w)} - \frac{1-a}{1-\pi_0(w)}\right)[y - \mu_0(a, w)]\,.$$

We refer to $D_0(o)$ as the pseudo-outcome. The pseudo-outcome is serves as a surrogate for the CATE and has been used in previous methods that estimate τ_0 (see for instance, Kennedy 2020, Athey and Wager 2021, Luedtke and Van Der Laan 2016)

Using the dataset $\mathcal{D}_n$, our proposed method consists in calibrating τ with the pseudo-outcomes $\{D_0(O_i)\}_{i=1}^n$ as the outcomes of the isotonic regression, and $\{\tau(W_i)\}_{i=1}^n$ as the covariates. However, $D_0(O)$ depends on π_0 and μ_0, which are usually unknown. Hence, they need to be estimated prior to the calibration step. Estimation of both functions can be carried via flexible data-adaptive methods. Letting D_n denote the estimated pseudo-outcome function derived from $\mathcal{D}_n$, then, one could fit the calibrator θ_n by finding the minimizer $\theta \in \mathcal{F}_{iso}$ of the empirical risk function

$$\frac{1}{n}\sum_{i=1}^{n}[\theta \circ \tau(W_i) - D_n(O_i)]^2. \tag{3.4}$$

However, minimizing Equation 3.4 with D_n will require us to use $\mathcal{D}_n$ twice: once for obtaining the pseudo-outcomes D_n, and a second time to carry the calibration step. Using $\mathcal{D}_n$ twice has two main potential drawbacks. First, it could lead to over-fitting because the empirical risk function in Equation 3.4 can be a biased estimate of the population risk $E[\{\theta \circ \tau(W) -$

$D_0(O)\}^2]$; second, the empirical risk in (3.4) is challenging to study theoretically because of the simultaneous, and dependent, randomness of O_i and D_n.

There are two commonly-used approaches that one could take in order to avoid the aforementioned drawbacks. These approaches allow us to obtain an empirical risk function that is nearly unbiased for the population risk of interest. The first method consists in carrying sample splitting by randomly dividing the data $\mathcal{D}_n$ into mutually exclusive sets. The first set is used to estimate μ_0 and π_0, and the second to carry the calibration step. Drawbacks of sample splitting include that the sample sizes of each data set need to be selected, and that the whole sample is not used for calibration. The second approach is commonly known as cross-fitting and has the advantage that it uses the complete data to estimate μ_0 and π_0 as well as carry the calibration step. Due to these advantages, we propose to use cross-fitting, which we detail in what follows.

In order to carry cross-fitting in our setting, we first randomly divide the sample $\mathcal{D}_n$ into k mutually exclusive data sets of approximately equal size. For simplicity we suppose that n is divisible by k and that all of these sets are chosen to be of size exactly n/k. Let $s \in \{1, ..., k\}$, and denote by $\mathcal{T}_s$ each of the k disjoint data sets, and $\mathcal{T}_s^c$ their complement. Additionally, let $j : \{1, ...n\} \to \{1, ..., k\}$ denote a function that maps each index $i \in \{1, ...n\}$ to a given s, where s is such that $o_i \in \mathcal{T}_s$. Then, for each s, we obtain an estimate of D_0 using $\mathcal{T}_s^c$, and denote it by D_n^s. Then, given observation $o_i \in \mathcal{T}_s$, we let $D_n^{j(i)}(o_i)$ denote its evaluation, namely

$$D_n^{j(i)}(o) := \left(\frac{a}{\pi_n^{j(i)}(w)} - \frac{1-a}{1-\pi_n^{j(i)}(w)} \right) \left[y - \mu_n^{j(i)}(a, w) \right] + \mu_n^{j(i)}(1, w) - \mu_n^{j(i)}(0, w),$$

where $\mu_n^{j(i)}$ and $\pi_n^{j(i)}$ are the estimates of μ_0 and π_0 derived from $\mathcal{T}_{j(i)}^c$, respectively. We repeat the estimation procedure k times in order to obtain the evaluations $\{D_n^{j(i)}(o_i)\}_{i=1}^n$ for each observation in the sample.

After obtaining the cross-fitted estimates of D_0, we compute the risk function to carry the calibration step. For a fixed $\theta \in \mathcal{F}_{iso}$, let the empirical risk function based on $\mathcal{D}_n$, be

equal to

$$R_{n,\tau}(\theta) := \frac{1}{n}\sum_{i=1}^{n}(D_n^{j(i)}(o_i) - \theta \circ \tau(w_i))^2. \tag{3.5}$$

Letting $\theta_n^* \in \text{argmin}_{\theta \in \mathcal{F}_{iso}} R_{n,\tau}(\theta)$, the causal isotonic calibrator is defined as $\tau_n^* := \theta_n^* \circ \tau$. We note that the solution to Equation 3.5 is usually not unique. This is because once can fit an infinite number of monotone increasing functions that interpolate between the jump points of the resulting step function. To simplify our analysis, we assume that the solution is given by a step function whose jump points are contained in the set $\{\tau(w_i)\}_{i=1}^n$. With this simplifying assumption, the solution to the isotonic regression is unique.

Thus, the causal isotonic calibrator τ_n^* can be obtained by performing the isotonic regression of the outcomes $\{D_n^{j(i)}(o_i)\}_{i=1}^n$ on covariates $\{\tau(w_i)\}_{i=1}^n$, and can be done with widely-available software. Algorithm 2 in the Appendix outlines the steps to carry the causal isotonic calibration method.

3.4.3 Sample splitting when τ is unavailable

Throughout this work we have assumed that an initial CATE estimator τ is known. However, it may be the case that τ is unavailable. In such cases, we propose to carry sample splitting as an alternative. Specifically, given a data set, we can split into two mutually exclusive and exhaustive sets. The first is used to obtain an estimate of τ_0, and the second is used to calibrate this estimate. We refer to the calibrated estimate of τ_0 as $\tau_{n,s}^*$. The sample splitting approach implies a trade-off between the sample size allocated to derive the initial estimator of τ_0 and both the predictive accuracy of $\tau_{n,s}^*$ and how small $CAL(\tau_n^*)$ is. To better understand the relationship between the proportions of the sample allocated to obtain an initial estimate of τ_0 and to calibrate this estimate, we conducted a suite of simulation studies detailed in Section 3.6.

3.5 Theoretical properties of the causal isotonic calibrator

3.5.1 Large sample properties

In this section, we show that the causal isotonic calibration asymptotically satisfies the desired Properties 1 and 2. First, to ease notation and simplify our theoretical proofs, we assume that an additional data set $\mathcal{D}_m$ of size m is available to estimate μ_0 and π_0. This data set could be obtained through sample splitting the original set of observations, for instance. Additionally, we denote by $\mathcal{D}_{cal}$ the data set of size n_{cal} that is used for calibration. Then, we let

$$D_m(o) := \left(\frac{a}{\pi_m(w)} - \frac{1-a}{1-\pi_m(w)}\right)[y - \mu_m(a,w)] + \mu_m(1,w) - \mu_m(0,w),$$

the estimated pseudo-outcome evaluated at o, where μ_m and π_m are the estimated conditional means and propensity score obtained through $\mathcal{D}_m$. After obtaining the function D_m, we use the observations $\{D_m(O_i), \tau(W_i)\}_{i=1}^{n_{cal}}$ to carry isotonic regression and obtain τ_n^*. Finally, it is worth noting that all of our theoretical results can be generalized to the cross-fitting case.

Our theoretical results will hold under 5 conditions. Specifically, two conditions are related to the data generating mechanism, two pertain to the estimators of μ_0 and π_0, and the last assumption is related to the function class $\mathcal{F}_{iso}$. The conditions on the data generating mechanism are common in causal inference settings and require that the counterfactual outcomes as well as the propensity score are bounded almost surely. The conditions on the estimators can be satisfied by bounding the estimated regression functions and if we used sample splitting to calibrate.

Before stating the conditions we introduce additional definitions and notation, Let $\mathcal{F}_1 = \{\gamma_0(\theta \circ \tau, w) : \mathcal{W} \to \mathbb{R}; \theta \in \mathcal{F}_{iso}\}$ denote the set of functions defined by the conditional mean of the treatment effect conditioning on $\theta \circ \tau(W)$ where θ is an arbitrary isotonic calibrator. For a family of functions $\mathcal{F}$, let $N(\epsilon, \mathcal{F}, L_2(P))$ denote its ϵ- covering number (van der Vaart and Wellner 1996) and define the entropy integral of $\mathcal{F}$ by

$$\mathcal{J}(\delta, \mathcal{F}, L_2) := \int_0^\delta \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon.$$

where F is an arbitrary, measurable envelope function for the class $\mathcal{F}$. The conditions we require for our results to hold are as follows:

Condition 10. *Boundedness: $\mathcal{Y}$ —the support of Y— is a uniformly bounded subset of $\mathbb{R}$.*

Condition 11. *Positivity: There exist constants u and l such that $0 < l < u < 1$, we have that $l < \pi_0(W) < u$ almost surely when $W \sim P$.*

Condition 12. *Independence of π_m and μ_m: no data from $\mathcal{D}_{cal}$ was used to fit π_m and μ_m.*

Condition 13. *Bounded range of π_m and μ_m: there exist finite constants $\eta > 0$ and α such that, for every $n \in \mathbb{N}$, $1 - \eta > \pi_m(W) > \eta$ and $\alpha > |\mu_m(W)|$ with P-probability one.*

Condition 14. *Bounded entropy integral of $\mathcal{F}_1$: The entropy integral of $\mathcal{F}_1$ exists and there exists a universal constant $u > 0$ such that, for all $\delta > 0$, $\mathcal{J}(\delta, \mathcal{F}_1, L_2) \leq u\mathcal{J}(\delta, \mathcal{F}_{iso}, L_2)$.*

Condition 10 holds trivially when outcomes are binary, and in the case of continuous outcomes this assumption is not unreasonable because in practice outcomes of interest often satisfy known bounds. Condition 11 is standard in causal inference and requires that all individuals have a positive probability of being assigned to either treatment or control. Condition 12 follows as a direct consequence of the sample splitting approach, because the estimators are obtained from an independent sample from the data that is used to carry the calibration step. Condition 13 requires that the estimators of the conditional mean of the outcome and the propensity score are bounded. A bounded range of the estimators is not unreasonable, and can be enforced by imposing a threshold when fitting the estimators. Finally, Condition 14 requires that the size of the family of functions $\mathcal{F}_1$ is not much larger than $\mathcal{F}_{iso}$.

The following theorem establishes the calibration rate for a calibrated predictor using the proposed causal isotonic calibration method.

Theorem 3.1 (τ_n^* is well-calibrated)**.** *Suppose Conditions 10 to 14 hold. Then,*

$$CAL(\tau_n^*) = O_P\left(n_{cal}^{-2/3}\right) + O_P\left(\left[E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|\right]^2\right).$$

Note that the calibration rate attained by the causal isotonic regression method can be of the order of $n_{cal}^{-2/3}$. Our method can satisfy Property 1 at a rate of $n_{cal}^{-2/3}$ as long as $E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|^2$ is of the order of $n_{cal}^{-2/3}$, which requires that one or both of π_0 and μ_0 is estimated well in an appropriate sense. We now turn to establishing that isotonic calibration satisfies Property 2, namely that it maintains the predictive accuracy of the initial predictor τ. In what follows predictive accuracy is quantified in terms of mean-squared error. At first glance, the calibration-distortion decomposition appears to raise some concerns that isotonic calibration may distort τ so much that the ultimate predictive accuracy of τ_n^* may worsen the predictive performance of τ. This possibility seems especially concerning given that isotonic regression tends to return a step function, so that there could be many $w, w' \in \mathcal{W}$ such that $\tau(w) \neq \tau(w')$ but $\tau_n^*(w) = \tau_n^*(w')$. The following theorem alleviates this concern by establishing that the root-mean-squared error is, up to a remainder term that decays with sample size, no larger than the root-mean-squared error of the initial treatment effect predictor τ. A consequence of this theorem is that isotonic calibration does not distort τ so much as to destroy its predictive performance.

Theorem 3.2 (Isotonic calibration does not inflate mean-squared error too much)**.** *Suppose Conditions 10 to 14 hold. Let* $\tau_0^* = \text{argmin}_{\theta\circ\tau:\theta\in\mathcal{F}_{iso}} \left\|\tau_0 - \theta\circ\tau\right\|_{P_0}$ *be the best meta-isotonic approximation of the CATE given the initial predictor* τ*. Then,*

$$\left\|\tau_n^* - \tau_0^*\right\| = O_P\left(n_{cal}^{-1/3}\right) + O_P\left(E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|\right).$$

As a consequence,

$$\left\|\tau_n^* - \tau_0\right\| \leq \left\|\tau - \tau_0\right\| + O_P\left(n_{cal}^{-1/3}\right) + O_P\left(E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|\right).$$

It is worth noting that the big-oh in probability terms on the right-hand sides of the displays above are of the same form as the big-oh in probability bound on the square root of $CAL(\tau_n^*)$ implied by Theorem 3.1.

3.5.2 Finite sample properties of the isotonic calibrated estimator

Based on the properties of isotonic regression outlined in the Appendix, the resulting isotonic calibrated predictor τ_n^* has the property that for every function r,

$$\frac{1}{n_{cal}} \sum_{i=1}^{n_{cal}} (r \circ \tau_n^*(W_i)) \left[D_m(O_i) - \tau_n^*(W_i) \right] = 0 \tag{3.6}$$

P-almost surely. This property implies that the empirical errors $\{D_m(O_i) - \tau_n^*(W_i)\}_{i=1}^n$ are empirically uncorrelated with all transformations of the predictions $\{\tau_n^*(O_i)\}_{i=1}^n$. Two interesting consequences can be derived from this property. First, choosing $r \equiv 1$, we find that

$$\frac{1}{n_{cal}} \sum_{i=1}^{n_{cal}} \tau_n^*(W_i) = \frac{1}{n} \sum_{i=1}^{n_{cal}} D_m(O_i)$$

P-almost surely. Noting that the right hand side is the AIPW estimator for the marginal ATE, we can conclude that the empirical average of τ_n^* is an efficient estimator for the marginal ATE, regardless of the value of τ. Second, taking $r \circ \tau_n^* = 1(\tau_n^* = \tau')$ for some level τ', we find

$$\frac{1}{n_{cal}} \sum_{i=1}^{n_{cal}} 1(\tau_n^*(W_i) = \tau') \tau_n^*(W_i) = \frac{1}{n_{cal}} \sum_{i=1}^{n_{cal}} 1(\tau_n^*(W_i) = \tau') D_m(O_i),$$

which implies that

$$E_{P_n} \left[D_m(O) \mid \tau_n^*(W_i) = \tau' \right] = \tau'.$$

The above can be viewed as a form of finite sample exact calibration. Finally, taking $r \circ \tau_n^*(w) = \tau_n^*(w)$, we have that

$$\frac{1}{n_{cal}} \sum_{i=1}^{n_{cal}} \tau_n^*(W_i) [D_m(O_i) - \tau_n^*(W_i)] = 0$$

which implies that the estimated regression function is empirically uncorrelated with the residuals.

3.6 Simulation studies

We carried a suite of simulation studies in order to assess the performance and applicability of the asymptotic properties of the causal isotonic calibrator in finite samples.

3.6.1 Data Generating Mechanisms

We simulated data from eight different data generating mechanisms that varied the type of outcome, the complexity of the true CATE, and the number of covariates. In four scenarios the outcome was binary, and in the remaining four the outcome was continuous. Within each of the outcome types, we examined the case with 4 confounders ('low dimensional'), and a case with 100 confounders ('high dimensional'), where only 20 were confounders and the remaining 80 were noise. Within the low dimensional and the high dimensional cases, we considered a setting where the mean outcome —with the appropriate link function— is linear on covariates and contains interactions between treatment and covariates, and a more complicated model that involved nonlinear transformations of the covariates as well as interactions between treatment and covariates. The logit of the propensity score was always linear on covariates and included 4 or 20 confounders in the low dimensional and high dimensional settings, respectively. In all scenarios, the confounders were independent and uniformly distributed between -1 and 1. The total sample sizes —including the sample size allocated for calibration— we examined were 1,000, 2,000, 5,000, and 10,000.

In the binary outcome scenarios, we set $\text{logit}[\mu_0(a, w)] = g(a, w)$ and allowed for g to include interactions between the treatment indicator and the confounders. The setting with nonlinear transformations of W included terms with absolute values, squared-root transformations, and indicator functions with a fixed threshold. To simulate treatment assignment, we considered the case where the logit of the propensity score was linear on covariates and in all cases we included the same confounders used on the outcome model. Coefficients of the propensity score were selected such that the probabilities of treatment were bounded with values that ranged between 0.05 and 0.95 in the low dimensional cases and between 0.01 and 0.99 in the high dimensional settings. The continuous scenarios were analogous to the binary settings, except for the link function, where we used $\mu_0 = g$.

We will refer to the eight scenarios considered as scenarios 1 through 8. Scenario 1 denotes the low dimensional setting that considers a binary outcome whose mean does not

include nonlinear transformations of the covariates. Scenario 2 is similar to scenario 1, with the exception that it includes nonlinear transformations. Scenarios 3 and 4 correspond to the high dimensional binary outcome settings, where scenario 3 does not have nonlinearities and scenario 4 does. Scenarios 5 to 8 consider continuous outcomes. Scenarios 5 and 6 are low dimensional, the first does not nonlinearities whereas the second one does. Finally, scenarios 7 and 8 consider high dimensional settings with and without nonlinear transformations of the covariates, respectively. More details regarding the analytical form of the outcome mean functions and the propensity scores can be found in Section 3.D.1 of the Appendix.

3.6.2 Treatment effect estimators under consideration

We considered different initial CATE estimators that depended on the data generating mechanism. In the low dimensional settings, we fitted gradient boosted regression trees (Chen and Guestrin 2016) with different maximum depths ranging from 2 to 8, random forests (Breiman 2001), generalized linear model with elastic net regularization (Friedman et al. 2010), generalized additive models (Wood 2017), and multivariate adaptive regression splines (Friedman 1991). In the high dimensional settings, we considered two CATE estimators: generalized linear model with elastic net regularization, and a combination of variable screening with elastic net regularization followed by regression trees with gradient boosting with a maximum depth obtained via cross-validation. In all cases we defined the loss functions and mean models according to whether the outcome was continuous or binary. We used the sl3 R package (Coyle et al. 2021) to implement each of the CATE estimators. More detailed information on the CATE estimators can be found in Table 3.D.5.

3.6.3 Implementation of the causal isotonic calibrator

In our simulation study, we used the sample splitting approach outlined in Section 3.4.3 to estimate τ_0 and then subsequently calibrate this estimate. To analyze the impact of sample splitting, we examined different proportions of the data set to fit the initial CATE estimator. These proportions ranged from 0.2 to 0.8 by increments of 0.1. After estimating each of the

treatment effect predictors, we used cross-fitting to obtain the pseudo-outcomes. Following Algorithm 2, we carried isotonic regression of the estimated values of the CATE estimator as covariates and the cross-fitted pseudo-outcomes as the outcome. We used the R function isoreg() to carry out the isotonic regression step (R Core Team 2013).

During the cross-fitting procedure we estimated the two unknowns, $\mu_0(a, w)$ and $\pi_0(w)$, of the pseudo-outcomes using super learner (van der Laan et al. 2007) in the low dimensional case and elastic net penalty regression in the high dimensional case. Super learner is an ensemble learning approach that uses cross-validation to select a convex combination of a library of candidate prediction methods. We included the following models in the libraries of the propensity score as well as the outcome model: generalized linear model, generalized linear model with elastic net penalty, boosted random forest trees with depths of 2,4, and 6, and generalized additive models. In the high dimensional scenario, we directly estimated the mean model and the propensity score via a logistic regression (or linear when the outcome was continuous) with elastic net penalty. Note that all of our models for the outcome regression were misspecified in those scenarios with nonlinearities in the outcome model. However, across all scenarios, the propensity score estimator was a consistent estimator of the true propensity score. We imposed a threshold on the estimated propensity scores such that it took values between 0.01 and 0.99. More details on the candidate learners can be found on Table 3.D.5. As with the CATE estimators, we used the sl3 R package (Coyle et al. 2021) to estimate μ_0 and π_0.

3.6.4 Performance assessment of the causal isotonic calibrator

We compared the causal isotonic calibrator to its uncalibrated version via three metrics. To ensure a fair comparison, the uncalibrated estimators were fitted using the complete data set, rather than on a subset as was done for the CATE estimator that was calibrated using sample splitting.

We computed the following performance metrics: the calibration measure defined in Equation 3.1, the bias of our method within bins defined by extreme values of the predictions,

and the mean squared error compared to the true CATE. Given our theoretical results, compared to uncalibrated CATE estimators, we expected to see that our proposed method had a smaller calibration measure, comparable mean squared error, and lower bias within bins of predictions. The bias within bins defined by extreme predictions of the CATE is a particularly relevant measure because predictions in the extremes tend to produce confidence when assigning treatment. Moreover, because the calibration measure in (3.2) is based on an average over the complete distribution of W, it can tend to assign little weight to regions where the magnitude of the predicted CATE is particularly large. As such, it could in principle be that a treatment effect predictor is uncalibrated in the tails of its predictions, but well calibrated overall compared to its calibrated counterpart.

We estimated the calibration measure, the bias within bins of extreme predictions, and the mean squared error with an independent sample $\mathcal{V}$ of size $n_{\mathcal{V}} = 10^4$. Finally, in order to account for the randomness of τ, we repeated each simulation 300 times and reported the average of the estimated metrics. We estimated the performance metrics as follows.

With some abuse of notation, let $\hat{\tau}$ denote an arbitrary estimated treatment effect predictor, or its calibrated version. For each fitted $\hat{\tau}$ in the simulations, we computed its mean squared error by taking the empirical mean of the squared difference between the fitted values of the CATE estimators and τ_0:

$$\widehat{MSE}(\hat{\tau}) := \sum_{w_i \in \mathcal{V}} [\hat{\tau}(w_i) - \tau_0(w_i)]^2.$$

Next, we obtained the estimated calibration measure in two steps. In the first step, we estimated $\gamma_0(\hat{\tau}, w)$ using an independent data set of 100,000 observations and fitted gradient boosted regression trees with the fitted values of the treatment effect predictors as covariates and the true CATE as the outcome. For each simulation setting and CATE estimator, the depths of each of the regression trees were obtained using cross-validation in a separate simulation. Let $\widehat{\gamma_0}(\hat{\tau}, w)$ denote the estimated function. In the second step, we used the

sample $\mathcal{V}$ to estimate the calibration measure as

$$\widehat{CAL}(\tau) := \sum_{i:w_i \in \mathcal{V}} \left[\tau_0(w_i) - \hat{\tau}(w_i)\right] \left[\widehat{\gamma_0}(\hat{\tau}, w_i) - \hat{\tau}(w_i)\right].$$

The above measure has the advantage of having less bias with respect to $CAL(\hat{\tau})$ than the plug-in estimator $\sum_{i:w_i \in \mathcal{V}} \left[\widehat{\gamma_0}(\hat{\tau}, w_i) - \hat{\tau}(w_i)\right]^2$ (Xu and Yadlowsky 2022).

In the case where $\hat{\tau}$ was obtained via isotonic calibration, we validated the aforementioned approach of estimating $CAL(\hat{\tau})$. Because the isotonic calibrated estimator is a step function, it is possible to estimate $\gamma_0(\hat{\tau}, w)$ computed the empirical mean of τ_0 at each interval defined by the jump points of $\hat{\tau}$. Using the aforementioned approach and a large data set of 100,000 observations, we estimated $\gamma_0(\hat{\tau}, w)$. We computed the difference of the estimated calibration measures via gradient boosted regression trees and the empirical mean approach. The results of this validation study in scenarios 1 and 5 are reported in Figure 3.D.1, which shows that the calibration measure using the empirical mean approach versus our estimated calibration measure are close to 0. Similar results were attained in the remaining scenarios.

The third measure we computed to assess the performance of the estimators was to compare the average estimated values within bins of predictions against the true CATE. That is, for fixed interval $[a, b)$, we computed

$$\sum_{\{i:w_i \in \mathcal{V}\}} \tau_0(w_i) I(a \leq \hat{\tau}(w_i) < b), \text{ and } \sum_{\{i:w_i \in \mathcal{V}\}} \hat{\tau}(w_i) I(a \leq \hat{\tau}(w_i) < b),$$

using the external validation set $\mathcal{V}$. We chose two intervals $[a, b)$ based on the observed predictions $\{\hat{\tau}(w_i)\}_{w_i \in \mathcal{V}}$. The first interval was defined as the values between the minimum and the 10th percentile of the predictions, and the second interval consisted of the of the 90th percentile and maximum. We standardized the above sums by dividing by the total number of observations in the corresponding intervals. We contrasted the two resulting averages by computing the difference. In the continuous scenarios, we reported the relative difference dividing by the average of the true CATE within the bins of predictions.

3.6.5 Simulation results

Overview of results

The overall improvement and performance with regard to the calibration measure depended mostly on the initial CATE estimators and the proportion allocated to conduct the calibration step. However, the general trend across all settings was that gradient boosted regression trees and random forests were outperformed by our method in terms of the calibration measure, but these regression methods generally had better performance than their calibrated counterparts in terms of mean squared error. The rest of the CATE estimators (glm, glmnet, gam, and splines) performed similarly calibration-wise, and mostly outperformed our approach in terms of mean squared error. Finally, regarding the proportion allocated to calibrate, we saw that, generally, the best performance of our method was achieved by splitting the data in half to fit the initial estimator and using the remaining 50% to calibrate. As we will see in the following sections, the performance of our method when allocating 80% or 20% of the sample to fit the initial CATE depended on the specific scenario.

In terms of the bias within bins of predictions, our method generally outperformed the uncalibrated methods in large sample sizes, and, depending on the scenario, performed similarly or better when sample size was small. In order to summarize our results, we reported the difference in the average predictions only when using 50% of the sample to calibrate because we saw that our method had the best overall performance using that proportion. Additionally, we only report the performance of three CATE estimators: glmnet, splines, and random forests. We focus on these methods for two reasons. First, these methods vary in their level of flexibility, with random forests being the most flexible and glmnet the least. Second, these methods performed relatively well with respect to our calibration measure.

Binary outcome scenario results

In the low dimensional settings, the results varied depending on whether nonlinear terms were present or not. For instance, causal isotonic calibration was overall better calibrated

than most methods when no nonlinearities were present, but had similar or slightly worse calibration when nonlinearities were present (Figures 3.D.2, 3.D.3, and 3.D.4), except for boosted trees whose calibration was worse than their calibrated counterparts (Figure 3.D.5). Overall, our method had the lowest calibration measure when sample size was large and the proportion to calibrate was 0.5. In small sample sizes, our method's calibration performance improved as the proportion allocated to calibrate was higher. However, when sample size was low, our proposed method has a similar or worse calibration compared to most of the other methods. This also held when comparing the bias within bins of predictions, where it is visibly higher in our method in small sample sizes, and decreases to outperforms most of the methods when sample size is large (Table 3.D.1). Regarding the predictive performance of our method, we observed that, in the absence of nonlinearities, the causal isotonic calibrator performed similarly to most of the non-calibrated methods in mean squared error, and slightly worse when the proportion to fit the CATE estimator was 0.8 (Figures 3.D.6 and 3.D.7). On the other hand, in presence of nonlinearities, using 80% of the sample to fit the CATE estimator yielded the best mean squared error performance (Figures 3.D.8 and 3.D.9); however, our method was generally outperformed by the non-calibrated methods.

In the high dimensional settings, when no nonlinearities were present, penalized regression outperformed our method in mean squared error, was better calibrated in small samples, and similarly calibrated when sample size was large (Figures 3.D.10 and 3.D.11). On the contrary, using penalized regression to select variables followed by boosted trees was overall less calibrated compared to our method. We observed the same result when contrasting the bias within bins of predictions (Table 3.D.2). When nonlinearities were present, our method had similar or better calibration than glmnet in large sample sizes and when the proportion to calibrate was 80% or 50%. Additionally, the bias within bins of predictions between glmnet and our method was very similar. Boosted trees were less calibrated overall and had higher bias, but had lower mean squared error than the calibrated version, (Figures 3.D.12 and 3.D.13, Table 3.D.2).

Continuous outcome scenario results

In the low dimensional settings, as the proportion to fit the initial CATE estimator decreased, the calibration of our method improved. Generally, the causal isotonic method was generally better calibrated than tree-based methods (Figures 3.D.14 and 3.D.15). On the other hand, the uncalibrated methods that were not tree-based were better calibrated when sample size was small, and had similar or worse performance as sample size increased (Figures 3.D.16 and 3.D.17). Our method achieved its best predictive performance when the proportion to fit the CATE estimator was 0.5. With this proportion and in the absence of nonlinearities, as sample size increased our method performed similarly to or outperformed the majority of the methods (Figures 3.D.18 and 3.D.19). When nonlinear terms were present, the difference in mean squared error converged to 0, except with splines and tree based methods, which had better overall predictive performance (Figures 3.D.20 and 3.D.21). Finally, contrasting the average within bins of predictions we saw that our method outperformed most of the methods in the absence of nonlinearities, but when nonlinear terms were present our method performed slightly worse in small sample sizes and better in larger samples (Table 3.D.3).

In the large-dimensional scenarios the results did not vary greatly depending on whether nonlinearities were included in the model. In both high dimensional settings, our method had substantially better calibration compared to gradient boosted trees with prior variable screening from glmnet (Figures 3.D.22 and 3.D.23). In contrast, glmnet had had better or very similar calibration compared to its isotonic calibrated version. Additionally, in moderate to large sample sizes, our method outperformed the uncalibrated methods when comparing the bias within bins of predictions (Table 3.D.4). Finally, the predictive performance of glmnet and our method was similar, and boosted trees had better performance when nonlinearities were present, but worse performance when no nonlinearities were present and in large samples (Figures 3.D.24 and 3.D.25).

3.7 Discussion

In this work, we propose causal isotonic calibration to calibrate any CATE estimator τ. Our proposed method guarantees that, under minimal assumptions, the calibration measure defined in Equation 3.2 converges to 0. This property holds regardless of whether τ is a good approximation of the CATE function. Additionally, our method does not lose the initial predictive power of τ because its mean squared error is upper bounded by mean squared error of τ plus additional terms that converge to zero at an $n^{-1/3}$ rate. To our knowledge, our method is the first in the literature to calibrate CATE estimators without requiring trial data, which was needed by the method in Leng and Dimmery (2021).

While our method guarantees that the calibration measure converges to zero, it does rely on the assumption that $\mu_0(a, w)$ or $\pi_0(w)$ are consistently estimated. However, if flexible data-adaptive estimators are used, this condition can often be met. Additionally, if calibrating treatment effect predictors based on trial data, the assumption on the consistency of the estimators is met directly because π_0 is known. Hence, our method can be readily applied to calibrate treatment effect predictors and better understand treatment effect heterogeneity in clinical trials.

Simulation results showed that the causal isotonic calibrator loses some predictive power overall, but it is in general better or as calibrated as its counterparts. We believe that the main drawback that causes this performance is due to the sample splitting. A way to circumvent this disadvantage would be to split the data in half, and use one half to estimate the CATE and the other to calibrate it, then repeat this procedure inverting the roles of the data. Then, the final isotonic calibrator will be an average of the two fitted calibrators. By Jensen's inequality, this approach will necessarily yield an estimator with lower mean-squared error than either of the two calibrated sample splitting based estimators used to define it. However, at this point it is unclear whether the resulting estimator would in fact be calibrated.

Aside from the overall assessment of our method in the simulations, we found two inter-

esting results through these studies. The first is that the tree-based methods were generally poorly calibrated. We found that our method generally improved the calibration of these partitioning methods without a major loss in predictive power. This is of particular high relevance because recent methods to estimate the CATE have relied on regression trees because of their flexibility (Athey and Imbens 2016) and interpretability (Lee et al. 2020). Particularly, because of the monotonicity constraint, our calibration method will not change the ordering of the treatment rules originally derived from tree-based methods, but rather will provide a more calibrated estimate of the CATE in the corresponding terminal nodes (leaves). A second interesting finding was that glmnet was very well calibrated in the high dimensional settings, even when the model was misspecificied due to the nonlinearities in the outcome model. While the simulations were constructed differently than in van Klaveren et al. (2019), authors of that study also found that penalized regression were well calibrated — which was measured in the same way as in our study— and suggest that researchers should fit penalized methods in the presence of treatment-covariate interactions. On the other hand, Leng and Dimmery (2021) suggest that regularization induced bias and model misspecification can worsen the regularized method's calibration in linear settings. However, Leng and Dimmery (2021) did not report the same measures of calibration as we do, so our results are not directly comparable. More research should be carried to investigate the calibration properties of penalized regression methods.

APPENDIX

3.A Algorithms

Input: Input data set $\mathcal{D}_n$, k = number of splits for cross-fitting, initial τ

1: Partition $\mathcal{D}_n$ into $\mathcal{T}_1, ...\mathcal{T}_k$ data sets.
2: **for** s = 1,...,k **do**
3: Let $j(i) = s$ for all $i \in \mathcal{T}_k$
4: Using $\mathcal{T}_k^c$, obtain an estimate D_n^s pseudo-outcome function D_0^s
5: Obtain the fitted pseudo-outcomes $\{D_n^{j(i)}(o_i)\}_{i\in\mathcal{T}_k}$
6: **end for**
7: For each observation in $\mathcal{D}_n$ obtain predictions from τ
8: With $\mathcal{D}_n$, obtain $\tau_n^* = \theta_n^* \circ \tau$ where $\theta_n^* \in \text{argmin}_{\theta\in\mathcal{F}_{iso}} \frac{1}{n}\sum_{i=1}^n [D_n^{j(i)}(o_i) - \theta \circ \tau(w_i)]]^2$

Output: $\hat{\tau}_n^*$

Algorithm 2: Causal isotonic calibration

3.B Calibration via isotonic regression

For simplicity and because calibration has been developed mostly for prediction settings, in this section we assume we are working on a standard regression problem where an iid sample consisting of feature-outcome pairs (W, Y) is observed and we want to estimate $q : w \mapsto E[Y|W = w]$ regression setting and that we are calibrating a fixed function $g : \mathcal{W} \to \mathbb{R}$ that is an estimate of q.

Based on a realization of the data set $\mathcal{D}_n = \{w_i, y_i\}_{i=1}^n$, isotonic calibration consists in finding a real-valued function θ such that it minimizes an empirical loss function $\frac{1}{n}\sum_{i=1}^n [\theta \circ g(w_i) - y_i]^2$ with the restriction that the θ is a monotone non-decreasing function. The

isotonic calibrator is such that

$$\theta_n^* = \underset{\theta \in \mathcal{F}_{iso}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [\theta \circ g(w_i) - y_i]^2. \tag{3.7}$$

Recall that the solution is unique because of our simplifying assumption that the jump points can only take values $\{g(w_i)\}_{i=1}^n$. Then, it can be shown that θ_n^* is a monotone non-decreasing function that is piecewise constant. That is, for a given $w \in \mathcal{W}$, $\theta_n^* \circ g(w)$ can be expressed as $\theta_n^* \circ g(w) = a_0 + \sum_{j=1}^{J} a_j 1(g(w) \geq u_j)$ for coefficients $\{a_j\}_{j=0}^J$ and jump-points $\{u_j\}_{j=1}^J$.

The following lemma outlines important properties of θ_n^* that will allow us to show the theoretical results of our proposed method.

Lemma 3.1. *Let g denote a $\mathcal{W} \mapsto \mathbb{R}$ function , $\mathcal{D}_n$ the observed calibration data, and let θ_n^* be its isotonic calibrator defined by*

$$\theta_n^* = \underset{\theta \in \mathcal{F}_{iso}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} [\theta \circ g(w_i) - y_i]^2. \tag{3.8}$$

Then, for any real-valued function r, we have that

$$\frac{1}{n} \sum_{i=1}^{n} [r \circ \theta_n^* \circ g(w_i)] \, [\theta_n^* \circ g(w_i) - y_i] = 0. \tag{3.9}$$

Moreover, letting $\{u_j\}_{j=1}^J$ denote the jump points of θ_n^* it is also true that

$$\frac{1}{n} \sum_{i=1}^{n} I(u_j \leq \theta_n^* \circ g(w_i) < u_{j+1}) [\theta_n^* \circ g(w_i)] = \frac{1}{n} \sum_{i=1}^{n} I(u_j \leq \theta_n^* \circ g(w_i) < u_{j+1}) y_i. \tag{3.10}$$

Proof. To simplify notation, let $\eta_n^* : w \mapsto \theta_n^* \circ g(w)$ denote the mapping from $\mathcal{W}$ to its corresponding value of the calibrated g through θ_n^*. Additionally, let $R_n(\theta) := \frac{1}{n} \sum_{i=1}^{n} [\theta \circ g(w_i) - y_i]^2$.

It is known that $\eta_n^*(w) = a_0 + \sum_{j=1}^{J} a_j 1(g(w) \geq u_j)$, where $a_0 \in \mathbb{R}$, $\{a_j\}_{j=1}^J$ are positive coefficients, and jump points $\{u_j\}_{j=1}^J$. Fix an arbitrary jump point $\bar{u}_j$, and let $\xi_n : \mathbb{R}^2 \to \mathbb{R}$ denote the function $\xi_n(\varepsilon, h) = \theta_n^*(h) + \varepsilon 1(h \geq \bar{u}_j)$. Note that $\delta > 0$ can be chosen to be sufficiently small that, for all $|\varepsilon| \leq \delta$, $h \mapsto \xi_n(\varepsilon, h)$ is non-decreasing — for instance

if $\delta = \min\{a_j\}_{j=1}^J$, then the desired property holds. In a slight abuse of notation, we let $R_n(\xi_n(\varepsilon)) := \frac{1}{n}\sum_{i=1}^n [\xi_n(\varepsilon, g(w_i)) - y_i]^2$ and $R_n(\xi_n(-\varepsilon)) := \frac{1}{n}\sum_{i=1}^n [\xi_n(-\varepsilon, g(w_i)) - y_i]^2$.

Now, because $\eta_n^*(w)$ is a risk minimizer, for all $\varepsilon \geq 0$, both of the following hold:

$$R_n(\xi_n(\varepsilon)) - R_n(\eta_n^*) \geq 0,$$
$$R_n(\xi_n(-\varepsilon)) - R_n(\eta_n^*) \geq 0.$$

Moreover, when $\varepsilon = 0$, $R_n(\xi_n(0)) - R_n(\eta_n^*) = 0$. Therefore, if ε is sufficiently close to 0, the derivative with respect to ε of $R_n(\xi_n(\varepsilon)) - R_n(\eta_n^*)$ must be non-negative, and $R_n(\xi_n(-\varepsilon)) - R_n(\eta_n^*)$ must be non-positive. Hence, both of the following hold:

$$\frac{d}{d\varepsilon}[R_n(\xi_n(\varepsilon)) - R_n(\theta_n)]\Big|_{\varepsilon=0} \geq 0,$$
$$\frac{d}{d\varepsilon}[R_n(\xi_n(-\varepsilon)) - R_n(\theta_n)]\Big|_{\varepsilon=0} \leq 0.$$

This in turn implies that both of the following hold:

$$\frac{2}{n}\sum_{i=1}^n 1(g(w_i) \geq \bar{u}_j)\,[\eta_n^*(w_i) - y_i] \geq 0,$$
$$\frac{2}{n}\sum_{i=1}^n 1(g(w_i) \geq \bar{u}_j)\,[\eta_n^*(w_i) - y_i] \leq 0.$$

Hence,

$$\frac{1}{n}\sum_{i=1}^n 1(g(w_i) \geq \bar{u}_j)\,[\eta_n^*(w_i) - y_i] = 0.$$

Because the jump point $\bar{u_j}$ was arbitrary, we have that for all functions of the form $s(w) = b_0 + \sum_{j=1}^J b_j 1(g(w) \geq u_j)$ with coefficients $\{b_j\}_{j=0}^J$, one can show that

$$\frac{1}{n}\sum_{i=1}^n s(w_i)\,[\eta_n^*(w_i) - y_i] = 0$$

by taking linear combinations of $1(g(w) \geq u_j)$ and noting that the score equations are linear. Then, for any real-valued function r and letting $s_n = r \circ \eta_n^*$, we have that

$$\frac{1}{n}\sum_{i=1}^n r \circ \eta_n^*(w_i)\,[\eta_n^*(w_i) - y_i] = 0,$$

which shows the main result of this lemma. Finally, noting that

$$\frac{1}{n}\sum_{i=1}^{n} 1(u_j \leq g(w_i) < u_{j+1}) \left[\eta_n^*(w_i) - y_i\right] = \frac{1}{n}\sum_{i=1}^{n} 1(g(w_i) \geq u_j) \left[\eta_n^*(w_i) - y_i\right] - \frac{1}{n}\sum_{i=1}^{n} 1(g(w_i) \geq u_{j+1}) \left[\eta_n^*(w_i) - y_i\right],$$

implies that, for consecutive jump-points $u_j < u_{j+1}$,

$$\frac{1}{n}\sum_{i=1}^{n} 1(u_j \leq g(w_i) < u_{j+1}) \left[\eta_n^*(w_i) - y_i\right] = 0.$$

□

3.C Proofs of Theorems 3.1 and 3.2

In this section we provide proofs of the theoretical results stated in the main paper. First, we provide an overview of the notation and definitions used in the sections, then, we outline the conditions that the theorems rely on, and, finally, we show the corresponding proofs.

Review of notation & additional definitions

Let $O = (W, A, Y)$ be a random variable with distribution P and support $\mathcal{O}$, where $W \in \mathcal{W} \subset \mathbb{R}^d$ is a vector of baseline covariates, $A \in \{0, 1\}$ is a binary indicator of treatment, and $Y \in \mathcal{Y} \subset \mathbb{R}$ is a numerical outcome variable. All expectations are taken with respect to the measure P. Recall that (Y_0, Y_1) denote the potential outcomes that would have been observed in the counterfactual case where treatment $A = 0$ and $A = 1$ were assigned, respectively.

We denote the true CATE by $\tau_0(w) = E[Y_1 - Y_0 \mid W = w]$. Next, we let $\pi_0(w) = P(A = 1|W = w)$ denote the probability of treatment for a fixed covariate w. Additionally, and given treatment level $a \in \{0, 1\}$ and covariate $w \in \mathcal{W}$, we let $\mu_0(a, w) = E[Y|A = a, W = w]$ denote the conditional mean of Y given $(A, W) = (a, w)$. Let $\tau : \mathcal{W} \to \mathbb{R}$ be a function that maps a realization w of the random variable W to a treatment-effect prediction $\tau(w)$. Finally, recall that $\gamma_0(\tau, w) = E[Y_1 - Y_0|\tau(W) = \tau(w)]$ denotes the conditional mean of the ITE given that the treatment-effect prediction is equal to $\tau(w)$.

In our upcoming arguments we will generally condition on the data set $\mathcal{D}_m := \{A_i, W_i, Y_i\}_{i=1}^m$. To streamline presentation, we will make this conditioning implicit in the notation. The m observations of this sequence will be used to obtain μ_m and π_m, which are estimates of μ_0 and π_0, respectively.

For a given observation o, recall that the pseudo-outcome is defined as

$$D_0(o) = \mu_0(1, w) - \mu_0(0, w) + \left(\frac{a}{\pi_0(w)} - \frac{1-a}{1-\pi_0(w)}\right)[y - \mu_0(a, w)]. \tag{3.11}$$

Note that $E[D_0(O)|W] = \tau_0(w)$. Next, denote by

$$D_m(o) := \left(\frac{a}{\pi_m(w)} - \frac{1-a}{1-\pi_m(w)}\right)[y - \mu_m(a, w)] + \mu_m(1, w) - \mu_m(0, w).$$

Additionally, let $\mathcal{D}_{cal} = (W_i, A_i, Y_i)_{i=1}^n$ the sample of n iid observations that will be used to derive the causal isotonic calibrator, which we denote by τ_n^*.

Recall that $\mathcal{F}_{iso} = \{\theta : \mathbb{R} \to \mathbb{R}; \theta \text{ is monotone non-decreasing}\}$ is defined as the family of bounded monotone non-decreasing functions, and $\mathcal{F}_1 = \{\gamma_0(\theta \circ \tau, w) : \mathcal{W} \to \mathbb{R}; \theta \in \mathcal{F}_{iso}\}$ is the family of functions defined as the conditional expectation of the difference of the potential outcomes conditioning on the event $\{\theta \circ \tau(W) = \theta \circ \tau(w)\}$. Additionally, let $\mathcal{F}_2 := \{\theta \circ \tau : \mathcal{W} \to \mathbb{R}; \theta \in \mathcal{F}_{iso}\}$ denote the family of functions defined by the composition of a given function in θ and τ, and let $\mathcal{F}_{3,m} := \{[\tau_2 - \tau_1][D_m - \tau_2] : \mathcal{O} \to \mathbb{R}; \tau_1 \in \mathcal{F}_1, \tau_2 \in \mathcal{F}_2\}$ denote the family of functions given by the product of $[\tau_2(w) - \tau_1(w)]$ and $[D_m(o) - \tau_2(w)]$, where $\tau_2 \in \mathcal{F}_1$, $\tau_1 \in \mathcal{F}_2$, and D_m is the estimated pseudo-outcome. Finally, for a family of functions $\mathcal{F}$, let $N(\epsilon, \mathcal{F}, L_2(P))$ denote its ϵ- covering number (van der Vaart and Wellner 1996) and define the entropy integral of $\mathcal{F}$ by

$$\mathcal{J}(\delta, \mathcal{F}, L_2) := \int_0^\delta \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon.$$

where F is an arbitrary, measurable envelope function for the class $\mathcal{F}$.

In the results below, we will use the following empirical process notation, for a $P-$measurable function f we denote $\int f(o)dP(o)$ by Pf and, letting P_n denote the empirical distribution, we let $P_n f$ be equal to $\frac{1}{n}\sum_{i=1}^n f(O_i)$, where each O_i is an observation from $\mathcal{D}_{cal}$. Additionally, we let $\|f\|_P^2 = Pf^2$, for notational simplicity we omit the dependency in P and use $\|f\|^2$

instead of $\|f\|_P^2$. The following section restates the sufficient conditions for the proofs of the results in our work to hold.

Conditions

Condition 1. *Boundedness: $\mathcal{Y}$ —the support of Y— is a uniformly bounded subset of $\mathbb{R}$.*

Condition 2. *Positivity: There exist constants u and l such that $0 < l < u < 1$, we have that $l < \pi_0(W) < u$ almost surely when $W \sim P$.*

Condition 3. *Independence of π_m and μ_m: no data from $\mathcal{D}_{cal}$ was used to fit π_m and μ_m.*

Condition 4. *Bounded range of π_m and μ_m: there exist finite constants $\eta > 0$ and α such that, for every $n \in \mathbb{N}$, $1 - \eta > \pi_m(W) > \eta$ and $\alpha > |\mu_m(W)|$ with P-probability one.*

Condition 5. *Bounded entropy integral of $\mathcal{F}_1$: The entropy integral of $\mathcal{F}_1$ exists and there exists a universal constant $u > 0$ such that, for all $\delta > 0$, $\mathcal{J}(\delta, \mathcal{F}_1, L_2) \leq u\mathcal{J}(\delta, \mathcal{F}_{iso}, L_2)$.*

Before outlining our theoretical results, we briefly describe how Conditions 1,2, and 4 imply that the function classes $\mathcal{F}_{iso}$, $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_{3,m}$ are bounded.

First, note that the isotonic calibrator is given by $\theta_n^* = \text{argmin}_{\theta:\theta\in\mathcal{F}_{iso}} \frac{1}{n}\sum_{i=1}^n[\theta \circ \tau(w_i) - D_m(o_i)]^2$. Because of Conditions 1, 2, and 4, we know that $D_m(o)$ is bounded uniformly over all observations $o \in \mathcal{O}$. That is, there exists a finite constant B such that $B := \sup_{m\in\mathbb{N},o\in\mathcal{O}} D_m(o)$. Hence, we can redefine $\mathcal{F}_{iso}$ to be equal to $\mathcal{F}_{iso} := \{\theta : \mathbb{R} \to \mathbb{R}; \theta \text{ is bounded monotone n}$ and the resulting calibrator would remain the same. Thus, by construction, all functions that belong to $\mathcal{F}_{iso}$ can be bounded by a constant $C := \sup_{x\in\mathbb{R},\theta\in\mathcal{F}_{iso}} \theta(x)$. Moreover, because $\mathcal{F}_{iso}$ is bounded, it directly implies $\mathcal{F}_2$ is bounded. Next, because of Condition 1, there exists a constant M which does not depend on θ nor τ and is such that $M > E[Y_1 - Y_0|\theta \circ \tau]$, implying that $\mathcal{F}_1$ is bounded. Finally, because $\mathcal{F}_1$, $\mathcal{F}_2$, D_m, and the potential outcomes are uniformly bounded, the function class $\mathcal{F}_{3,m}$ is also uniformly bounded.

Finally, for two quantities x and y, we use the expression $x \lessapprox y$ to denote that x is upper bounded by y times a constant that does not depend on any specific parameter that

is indexing x nor y. For instance, Condition 5 implies that for all $\delta > 0$, $\mathcal{J}(\delta, \mathcal{F}_1, L_2) \lessapprox \mathcal{J}(\delta, \mathcal{F}_{iso}, L_2)$.

Proof of Theorem 3.1

Theorem 3.1. *Suppose Conditions 1 to 5 hold. Then,*

$$CAL(\tau_n^*) = O_P\left(n^{-2/3}\right) + O_P\left(\left[E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|\right]^2\right).$$

Proof. First, note that

$$\begin{aligned}
&E\left\{\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]\left[D_0(O) - \tau_n^*(W)\right]\right\} = E\left\{E\left\{\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]\left[D_0(o) - \tau_n^*(W)|W\right]\right\}\right\} \\
=&E\left\{\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]\left[\tau_0(W) - \tau_n^*(W)\right]\right\} = E\left\{E\left\{\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]\left[\tau_0(W) - \tau_n^*(W)\right]|\tau_n^*(W)\right\}\right\} \\
=&E\left\{\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]\right\} = E\left\{\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right]^2\right\}.
\end{aligned}$$

The above equality implies that

$$\int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}^2 dP(w) = \int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}\left\{D_0(o) - \tau_n^*(w)\right\} dP(o).$$

Adding and subtracting $D_m(o)$ to the above expression gives

$$\int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}\left\{D_0(o) - D_m(o)\right\} dP(o) + \int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}\left\{D_m(o) - \tau_n^*(w)\right\} dP(o). \tag{3.12}$$

Note that by Lemma 3.1, τ_n^* has the property that it satisfies the score equations of the form,

$$\frac{1}{n}\sum_{i=1}^{n} r(\tau_n^*(W_i))\left[D_m(O_i) - \tau_n^*(W_i)\right] = 0$$

for all real-valued functions r. Setting $r(\tau') \equiv E[Y_1 - Y_0 \mid \tau_n^*(W) = \tau'] - \tau'$, we conclude that

$$\int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}\left\{D_m(o) - \tau_n^*(w)\right\} dP_n(o) = 0.$$

Subtracting the above score equation from the second term in Equation 3.12, we obtain the basic inequality (which is an equality),

$$\begin{aligned}
\int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}^2 dP(w) \leq &\int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}\left\{D_0(o) - D_m(o)\right\} dP(o) \\
&+ \int \left\{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\right\}\left\{D_m(o) - \tau_n^*(w)\right\} d(P - P_n)(o).
\end{aligned}$$

Rewriting the above basic inequality using norm and empirical process notation yields:

$$\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|^2 \leq \underbrace{P\left\{\gamma_0(\tau_n^*,\cdot)-\tau_n^*\right\}\left\{D_0-D_m\right\}}_{\text{(I)}} + \underbrace{(P-P_n)\left\{\gamma_0(\tau_n^*,\cdot)-\tau_n^*\right\}\left\{D_m-\tau_n^*\right\}}_{\text{(II)}} \tag{3.13}$$

In order to show the desired result, we will find bounds for both of the above terms. We can bound term (I) using the law of iterated conditional expectations and the Cauchy-Schwartz inequality. First, note that

$$\begin{aligned} P\left\{\gamma_0(\tau_n^*,\cdot)-\tau_n^*\right\}\left\{D_0-D_m\right\} &= \int \left\{\gamma_0(\tau_n^*,w)-\tau_n^*(w)\right\}\left\{E[D_0 \mid W=w]-E[D_m \mid W=w]\right\} dP(w) \\ &\leq \|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|\,\|E[D_0 \mid W]-E[D_m \mid W]\|\,. \end{aligned} \tag{3.14}$$

Next, we express $\|E[D_0 \mid W]-E[D_m \mid W]\|$ in terms of $\|\pi_m-\pi_0\|$ and $\|\mu_m-\mu_0\|$. Recalling that $E[D_0(O) \mid W=w]=\tau_0(w)$, we have that

$$\begin{aligned} &E[D_m(O) \mid W=w]-E[D_0(O) \mid W=w] \\ &= \mu_m(1,w)-\mu_0(1,w)-[\mu_m(0,w)-\mu_0(0,w)] \\ &\quad + \frac{\pi_0(w)}{\pi_m(w)}\left[\mu_0(1,w)-\mu_m(1,w)\right] - \frac{1-\pi_0(w)}{1-\pi_m(w)}\left[\mu_0(0,w)-\mu_m(0,w)\right] \\ &= \left[\frac{\pi_0(w)-\pi_m(w)}{\pi_m(w)}\right]\left[\mu_0(1,w)-\mu_m(1,w)\right] \\ &\quad + \left[\frac{\pi_m(w)-\pi_0(w)}{1-\pi_m(w)}\right]\left[\mu_0(0,w)-\mu_m(0,w)\right]. \end{aligned}$$

By Assumption 2, $P(1-\eta>\pi_m(W)>\eta)=1$ for some η. The latter assumption combined with Cauchy-Schwartz gives

$$\begin{aligned} \|E[D_0 \mid W]-E[D_m \mid W]\| \lessapprox\; &\|[\pi_m(W)-\pi_0(W)][\mu_0(0,W)-\mu_m(0,W)]\| \\ &+ \|[\pi_m(W)-\pi_0][\mu_0(1,W)-\mu_m(1,W)]\|\,. \end{aligned}$$

We also have by Assumption 2, that for any P-measurable function $h:\mathcal{W}\to\mathbb{R}$,

$$\begin{aligned} &\int h(w)^2[\mu_0(1,w)-\mu_m(1,w)]^2 dP(w) = \int h(w)^2[\mu_0(1,w)-\mu_m(1,w)]^2\frac{\pi_0(w)}{\pi_0(w)}dP(w) \\ &= \int h(w)^2[\mu_0(a,w)-\mu_m(a,w)]^2\frac{a}{\pi_0(w)}dP(a,w) \leq \frac{1}{l}\int h(w)^2[(\mu_0(a,w)-\mu_m(a,w))]^2 dP(a,w). \end{aligned}$$

The same argument and bound works for $\int h(w)^2[\mu_0(0,w)-\mu_m(0,w)]^2 dP(w)$. Setting $h(w) = [\pi_m(w) - \pi_0(w)]$, we conclude

$$\|E[D_m \mid W] - E[D_0 \mid W]\| \lessapprox \|(\pi_m(W) - \pi_0(W))(\mu_0(A,W) - \mu_m(A,W))\|. \tag{3.15}$$

Combining Equations 3.13, 3.14 and 3.15 yield the following bound for term (I),

$$P\left\{\gamma_0(\tau_n^*,\cdot) - \tau_n^*\right\}\left\{D_0 - D_m\right\} \lessapprox \|\gamma_0(\tau_n^*,\cdot) - \tau_n^*\| \, \|\left\{\pi_m - \pi_0\right\}\left\{\mu_0 - \mu_m\right\}\|. \tag{3.16}$$

Next, we upper bound the empirical process of term (II). First, recalling our definition of the constants $B = \sup_{m\in\mathbb{N},o\in\mathcal{O}} D_m(o)$ and $C = \sup_{x\in\mathbb{R},\theta\in\mathcal{F}_{iso}} \theta(x)$, let $K := B + C$. Set $\delta_n := \|\gamma_0(\tau_n^*,\cdot) - \tau_n^*\|$, which is a random rate. For a given rate δ, let

$$\begin{aligned} S_n(\delta) &:= \sup_{\theta_1,\theta_2\in\mathcal{F}_{iso}:\|\gamma_0(\theta_2\circ\tau,\cdot)-\theta_1\circ\tau\|\le\delta K} (P - P_n)\left\{\gamma_0(\theta_2\circ\tau,\cdot) - \theta_1\circ\tau\right\}\left\{D_m - \theta_1\circ\tau\right\} \\ &= \sup_{f\in\mathcal{F}_{3,m}:\|f\|\le\delta K} (P - P_n) f \end{aligned}$$

Note that $\tau_n^* = \theta_n^* \circ \tau$ for some $\theta_n^* \in \mathcal{F}_{iso}$. As a consequence, we have the following upper bound,

$$(P - P_n)\left\{\gamma_0(\tau_n^*, W) - \tau_n^*\right\}\left\{D_m - \tau_n^*\right\} \le S_n(\delta_n). \tag{3.17}$$

Due to the randomness in δ_n, the above cannot be further upper-bounded immediately. To bound the term above we will take a deterministic δ, and upper bound the modulus of continuity $\phi_n(\delta) := E\left\{S_n(\delta)\right\}$. Upper bounding $\phi_n(\delta)$ for a fixed δ suffices to obtain a rate of convergence for $\|\gamma_0(\tau_n^*,\cdot) - \tau_n^*\|$. To bound the above term, we will use empirical process techniques based on the function classes $\mathcal{F}_{iso}$, $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_{3,m}$. First, recall our definitions of the function classes: $\mathcal{F}_{iso} = \{\theta : \mathbb{R} \to \mathbb{R}; \theta \text{ is bounded monotone non-decreasing}\}$, $\mathcal{F}_1 = \{E[Y_1 - Y_0|\theta\circ\tau] : \mathcal{W} \to \mathbb{R}; \theta \in \mathcal{F}_{iso}\}$, $\mathcal{F}_2 := \{\theta\circ\tau : \mathcal{W}\to\mathbb{R}; \theta\in\mathcal{F}_{iso}\}$, and $\mathcal{F}_{3,m} := \{[\tau_2 - \tau_1][D_m - \tau_2] : \mathcal{O}\to\mathbb{R}; \tau_1\in\mathcal{F}_1, \tau_2\in\mathcal{F}_2\}$. Note that since μ_m and π_m are by assumption obtained from a independent data set $\mathcal{D}_m$, $\mathcal{F}_{3,m}$ is a deterministic function class conditional on $\mathcal{D}_m$.

We now aim to determine the uniform entropy integral $\mathcal{J}(\delta,\mathcal{F}) = \int_0^\delta \sup_Q \sqrt{N(\varepsilon, \mathcal{F}\,\|\cdot\|_Q)}d\varepsilon$ for each of the above function classes. Note that, because D_m is fixed, $\mathcal{F}_{3,m}$ is a multivariate

Lipschitz transformation of $\mathcal{F}_1$ and $\mathcal{F}_2$, and therefore by Theorem 2.10.20 in van der Vaart and Wellner (1996), we have $\mathcal{J}(\delta, \mathcal{F}_{3,m}) \lessapprox \mathcal{J}(\delta, \mathcal{F}_1) + \mathcal{J}(\delta, \mathcal{F}_2)$. Additionally, by Assumption 5, $\mathcal{J}(\delta, \mathcal{F}_1) \lessapprox \mathcal{J}(\delta, \mathcal{F}_{iso})$. The same upper bound holds for $\mathcal{F}_2$ by noting that

$$\mathcal{J}(\delta, \mathcal{F}_2) = \int_0^\delta \sup_Q \sqrt{N(\varepsilon, \mathcal{F}_2, \|\cdot\|_Q)} d\varepsilon = \int_0^\delta \sup_Q \sqrt{N(\varepsilon, \mathcal{F}_{iso}, \|\cdot\|_{Q\circ\tau^{-1}})} d\varepsilon = \mathcal{J}(\delta, \mathcal{F}_{iso}),$$

where $Q \circ \tau^{-1}$ is the push forward probability measure for the random variable $\tau(W)$. We are now ready to bound $\phi_n(\delta)$. Applying Theorem 2.10.20 in van der Vaart and Wellner (1996), we obtain

$$\begin{aligned} E\left[\sup_{f\in\mathcal{F}_{3,m}:\|f\|\leq\delta K} (P - P_n)f \middle| \mathcal{S}_m\right] &\lessapprox n^{-1/2}\mathcal{J}(\delta, \mathcal{F}_{3,m})\left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{3,m})}{\sqrt{n}\delta^2}\right) \\ &\lessapprox n^{-1/2}\mathcal{J}(\delta, \mathcal{F}_{iso})\left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{n}\delta^2}\right). \end{aligned}$$

Because the right hand side of the above bound is not random, taking a second expectation yields

$$\phi_n(\delta) = E\left[\sup_{f\in\mathcal{F}_{3,m}:\|f\|\leq\delta K} (P - P_n)f\right] \lessapprox n^{-1/2}\mathcal{J}(\delta, \mathcal{F}_{iso})\left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{n}\delta^2}\right). \tag{3.18}$$

We are now ready to proceed with the main argument that gives a rate of convergence for $\delta_n = \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|$. First, note that combining Equations 3.13, 3.16 and 3.17 imply that the event

$$\left\{\|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 \leq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(\delta_n)\right\}$$

occurs with probability one. Next, we proceed with a standard peeling argument. Let ε_n be some deterministic sequence to be chosen later, and let A_s be equal to the event $\{2^{s+1}\varepsilon_n \geq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \geq 2^s\varepsilon_n\}$. Then, we have that

$$\begin{aligned} &P\left(\|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \geq 2^S\varepsilon_n\right) = \sum_{s=S}^\infty P\left(2^{s+1}\varepsilon_n \geq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \geq 2^s\varepsilon_n\right) = \sum_{s=S}^\infty P(A_s) \\ &= \sum_{s=S}^\infty P\left(A_s, \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 \leq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(\delta_n)\right) \end{aligned} \tag{3.19}$$

Note that because $\delta_n = \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|$, in all the events in the above sum we have that $S_n(\delta_n) \leq S_n(2^{s+1}\epsilon_n)$. Next, manipulating the inequalities in the above events, one can show that

$$\begin{aligned}
&\left\{A_s, \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 \leq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(\delta_n)\right\} \subseteq \\
&\left\{A_s, \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 \leq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(2^{s+1}\epsilon_n)\right\} \subseteq \\
&\left\{2^{2s}\epsilon_n^2 \leq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 \leq 2^{s+1}\epsilon_n \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(2^{s+1}\epsilon_n)\right\} \subseteq \\
&\left\{2^{2s}\epsilon_n^2 \leq 2^{s+1}\epsilon_n \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(2^{s+1}\epsilon_n)\right\}
\end{aligned}$$

which implies that the sum in Equation 3.19 is upper bounded by

$$\sum_{s=S}^{\infty} P\left(2^{2s}\varepsilon_n^2 \leq 2^{s+1}\varepsilon_n \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(2^{s+1}\epsilon_n)\right).$$

Applying Markov's inequality and Equation 3.18, we find

$$\begin{aligned}
&\sum_{s=S}^{\infty} P\left(2^{2s}\varepsilon_n^2 \leq 2^{s+1}\varepsilon_n \, \|(\pi_m - \pi_0)(\mu_m - \mu_0)\| + S_n(2^{s+1}\epsilon_n)\right) \\
&\leq \sum_{s=S}^{\infty} \frac{2^{s+1}\varepsilon_n E\left\{\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right\} + \phi_n(2^{s+1}\varepsilon_n)}{2^{2s}\varepsilon_n^2} \\
&\lessapprox \sum_{s=S}^{\infty} \left\{ \frac{E\left\{\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right\}}{2^{s-1}\varepsilon_n} + \frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})\left(1 + \frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})}{\sqrt{n}2^{2s+1}\varepsilon_n^2}\right)}{2^{2s}\sqrt{n}\varepsilon_n^2} \right\}.
\end{aligned}$$

By well-known bounds on the uniform entropy integral for the class of isotonic functions, we have that $\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso}) = 2^{s/2+1/2}\sqrt{\varepsilon_n}$. Using this fact, it can be shown that

$$\frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})}{2^{2s}\sqrt{n}\varepsilon_n^2} \lessapprox \frac{1}{2^s}\frac{\mathcal{J}(\varepsilon_n, \mathcal{F}_{iso})}{\sqrt{n}\varepsilon_n^2}$$

It follows that

$$\frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})\left(1 + \frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})}{\sqrt{n}2^{2s+1}\varepsilon_n^2}\right)}{2^{2s}\sqrt{n}\varepsilon_n^2} \lessapprox 2^{-s}\frac{\mathcal{J}(\varepsilon_n, \mathcal{F}_{iso})\left(1 + \frac{\mathcal{J}(\varepsilon_n, \mathcal{F}_{iso})}{\sqrt{n}\varepsilon_n^2}\right)}{\sqrt{n}\varepsilon_n^2}.$$

We now choose ε_n such that both $\mathcal{J}(\varepsilon_n, \mathcal{F}_{iso}) \lessapprox \sqrt{n}\varepsilon_n^2$ and $E\,\|(\pi_m - \pi_0)(\mu_m - \mu_0)\| \lessapprox \varepsilon_n$, by setting

$$\varepsilon_n = \max\left\{n^{-1/3}, E\,\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right\}.$$

Then we have

$$P\left(\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|\geq 2^S\varepsilon_n\right)\lessapprox\sum_{s=S}^{\infty}\frac{1}{2^s}\xrightarrow[S\to\infty]{}0.$$

As a consequence of the above, for every $\varepsilon>0$, we can find a constant 2^S sufficiently large such that $P\left(\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|\geq 2^S\varepsilon_n\right)<\varepsilon$. In other words, we have shown that $\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|=O_P(\varepsilon_n)$ for our choice of $\varepsilon_n=\max\left\{n^{-1/3},\|(\pi_m-\pi_0)(\mu_m-\mu_0)\|\right\}$. Noting that, $CAL(\tau_n^*)^{1/2}=\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|$, this shows that $CAL(\tau_n^*)^{1/2}=O_P(\varepsilon_n)$ or, equivalently, that $CAL(\tau_n^*)=O_P(\varepsilon_n^2)$. □

Proof of Theorem 3.2

Theorem 3.2. *Suppose Conditions 1 to 5 hold. Let* $\tau_0^*=\text{argmin}_{\theta\circ\tau:\theta\in\mathcal{F}_{iso}}\|\tau_0-\theta\circ\tau\|$ *be the best isotonic approximation of the CATE given the initial predictor* τ*. Then,*

$$\|\tau_0^*-\tau_n^*\|=O_P\left(n^{-2/3}\right)+O_P\left(E\left\|(\pi_m-\pi_0)(\mu_m-\mu_0)\right\|\right).$$

As a consequence,

$$\|\tau_0-\tau_n^*\|\leq\|\tau_0-\tau\|+O_P\left(n^{-1/3}\right)+O_P\left(E\left\|(\pi_m-\pi_0)(\mu_m-\mu_0)\right\|\right).$$

Proof. We may write $\tau_n^*=\theta_n^*\circ\tau$ for some $\theta_n^*\in\mathcal{F}_{iso}$. We know that θ_n^* is the minimizer in $\mathcal{F}_{iso}$ of the empirical risk function

$$R_n(\theta):\theta\mapsto\frac{1}{n}\sum_{i=1}^{n}(D_m(O_i)-\theta\circ\tau(W_i))^2.$$

Consider the one-sided path $\varepsilon\mapsto\theta_n^*+\varepsilon(\theta-\theta_n^*):\varepsilon\in[0,1]$ through θ_n^* for $\theta\in\mathcal{F}_{iso}$. Note that $\mathcal{F}_{iso}$ is a convex space. Then, we have that

$$-\frac{2}{n}\sum_{i=1}^{n}(\theta-\theta_n^*)\circ\tau(W_i)(D_m(O_i)-\theta_n^*\circ\tau(W_i))=\lim_{\varepsilon\downarrow 0}\frac{R_n(\theta_n^*+\varepsilon(\theta-\theta_n^*))-R_n(\theta_n^*)}{\varepsilon}\geq 0 \quad (3.20)$$

for all $\theta\in\mathcal{F}_{iso}$. Let $\theta_0=\text{argmin}_{\theta\in\mathcal{F}_{iso}}\|\theta\circ\tau-\tau_0\|_P$ be the population level isotonic risk minimizer. Taking $\theta=\theta_0$ in Equation (3.20), we obtain the inequality,

$$\frac{1}{n}\sum_{i=1}^{n}\left\{(\theta_0-\theta_n^*)\circ\tau(W_i)\right\}\left\{D_m(O_i)-\theta_n^*\circ\tau(W_i)\right\}\leq 0. \quad (3.21)$$

Rearranging terms and adding and subtracting $P_n \{\theta_n^* \circ \tau\} D_0$ in the above inequality implies that

$$P_n \{(\theta_0 - \theta_n^*) \circ \tau\} \{D_m - D_0\} \leq P_n \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\}.$$

Adding and subtracting $P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\}$ yields

$$P_n \{(\theta_0 - \theta_n^*) \circ \tau\} \{D_m - D_0\} - (P_n - P) \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\} \leq P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\}. \tag{3.22}$$

Next, adding and subtracting $P \{\theta_0 \circ \tau\} \{(\theta_0 - \theta_n^*) \circ \tau\}$ we have that

$$\begin{aligned} P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\} &= P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - E[D_0(O) \mid W]\} \\ &= P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - \tau_0\} \\ &= P \{(\theta_0 - \theta_n^*) \circ \tau\} \{(\theta_n^* - \theta_0) \circ \tau\} + P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_0 \circ \tau - \tau_0\} \\ &= - \|(\theta_0 - \theta_n^*) \circ \tau\|^2 + P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_0 \circ \tau - \tau_0\}, \end{aligned} \tag{3.23}$$

where in the first to second equality we used the fact that $E[D_0 \mid W = w] = \tau_0(w)$. Next, note that θ_0 is the risk minimizer of the population risk function $\theta \mapsto E_P \{\tau_0(W) - \theta \circ \tau(W)\}^2$ over $\mathcal{F}_{iso}$. As a consequence, the same argument used to derive Equation 3.21 establishes the analagous inequality:

$$P \{(\theta - \theta_0) \circ \tau\} \{\tau_0 - \theta_0 \circ \tau\} \leq 0$$

for any $\theta \in \mathcal{F}_{iso}$. Taking $\theta = \theta_n^*$, we find

$$P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_0 \circ \tau - \tau_0\} \leq 0. \tag{3.24}$$

Combining Equations 3.23 and 3.24, we find

$$P \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\} \leq - \|(\theta_0 - \theta_n^*) \circ \tau\|_P^2. \tag{3.25}$$

Finally, combining Equations 3.22 and 3.25, we obtain the following inequality

$$\|(\theta_0 - \theta_n^*) \circ \tau\|_P^2 \leq -P_n \{(\theta_0 - \theta_n^*) \circ \tau\} \{D_m - D_0\} + (P_n - P) \{(\theta_0 - \theta_n^*) \circ \tau\} \{\theta_n^* \circ \tau - D_0\}.$$

Adding and subtracting, $P\{(\theta_0 - \theta_n^*) \circ \tau\}\{D_m - D_0\}$ we finally obtain the key basic inequality,

$$\|(\theta_0 - \theta_n^*) \circ \tau\|^2 \leq P\{(\theta_0 - \theta_n^*) \circ \tau\}\{D_0 - D_m\} + (P - P_n)\{(\theta_0 - \theta_n^*) \circ \tau\}\{D_m - D_0\} + (P_n - P)\{(\theta_0 - \theta_n^*) \circ \tau\}\{\theta_n^* \circ \tau - D_0\}. \quad (3.26)$$

The above basic inequality is very similar to the key basic inequality given in Equation 3.13 in the proof of Theorem 3.1. Thus, we employ similar techniques to show its convergence rate.

First, using a similar approach to derive (3.16), one can upper bound the first term of the RHS in Equation 3.26 as

$$P\{(\theta_0 - \theta_n^*) \circ \tau\}\{D_0 - D_m\} \leq \|(\theta_0 - \theta_n^*) \circ \tau\| \, \|\{\pi_m - \pi_0\}\{\mu_m - \mu_0\}\|.$$

The second term in the RHS of (3.26) can be examined as follows. Recall the definitions of $\mathcal{F}_{iso} = \{\theta : \mathbb{R} \to \mathbb{R}; \theta \text{ is monotone non-decreasing and bounded}\}$, $\mathcal{F}_2 = \{\theta \circ \tau; \theta \in \mathcal{F}_{iso}\}$, and that $B := \sup_{m \in \mathbb{N}, o \in \mathcal{O}} D_m(o)$. Further, let $\mathcal{F}_{4,m} := \{(\tau_1 - \tau_2)(D_m - D_0); \tau_1, \tau_2 \in \mathcal{F}_2\}$, and let $Q := \sup_{o \in \mathcal{O}} D_0(o)$ — which exists because of Conditions 1 and 2. Additionally, let $R := Q + B$. Next, define for a fixed $\delta \in \mathbb{R}$,

$$\begin{aligned} Z_{1,n}(\delta) &:= \sup_{\theta_1, \theta_2 \in \mathcal{F}_{iso} : \|(\theta_1 - \theta_2) \circ \tau\| \leq \delta R} (P - P_n)\{(\theta_1 - \theta_2) \circ \tau\}\{D_m - D_0\} \\ &= \sup_{f \in \mathcal{F}_{4,m} : \|f\| \leq \delta R} (P - P_n) f. \end{aligned}$$

Let $\delta_{1,n} := \|(\theta_0 - \theta_n^*) \circ \tau\|$, then, we have that

$$(P - P_n)\{(\theta_0 - \theta_n^*) \circ \tau\}\{D_m - D_0\} \leq Z_{1,n}(\delta_{1,n}).$$

Finally, let $\psi_{1,n}(\delta) := E\{Z_{1,n}(\delta)\}$ be its corresponding modulus of continuity. Then, noting that $\mathcal{F}_{4,m}$ is a Lipschitz transformation of the function classes $\mathcal{F}_1$ and $\mathcal{F}_2$, we have by Theorem 2.10.20 in van der Vaart and Wellner (1996) and the results outlined in Theorem 3.1 that, for every $\delta \in \mathbb{R}$,

$$\psi_{1,n}(\delta) \lessapprox n^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{n}\delta^2}\right).$$

Finally, the third term in Equation 3.26 can be studied as follows. Let $\mathcal{F}_5 := \{(\tau_1 - \tau_2)(\tau_2 - D_0); \tau_1, \tau_2 \in \mathcal{F}_2\}$, and for a given δ, let

$$\begin{aligned} Z_{2,n}(\delta) &:= \sup_{\theta_1,\theta_2 \in \mathcal{F}_{iso}: \|(\theta_1-\theta_2)\circ\tau\| \leq \delta G} (P - P_n) \{(\theta_1 - \theta_2) \circ \tau\} \{\theta_2 - D_0\} \\ &= \sup_{f \in \mathcal{F}_5: \|f\| \leq \delta G} (P - P_n) f. \end{aligned}$$

where $G := Q + C$, and $\phi_{2,n} := E[Z_{2,n}]$. Note that $\mathcal{F}_5$ is a Lipschitz transformation of $\mathcal{F}_2$, which has the same entropy number as $\mathcal{F}_1$. Hence, by Condition 5 and Theorem 2.10.20 in van der Vaart and Wellner (1996), we have that

$$\psi_{2,n}(\delta) \lessapprox n^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{n}\delta^2}\right).$$

Then, using a similar peeling argument as in Theorem 3.1, one can show that

$$\begin{aligned} P\left(\|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \geq 2^S \varepsilon_n\right) &\leq \sum_{s=S}^{\infty} \frac{2^{s+1}\varepsilon_n E\left\{\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right\} + \phi_{1,n}(2^{s+1}\varepsilon_n) + \phi_{2,n}(2^{s+1}\varepsilon_n)}{2^{2s}\varepsilon_n^2} \\ &\lessapprox \sum_{s=S}^{\infty} \left\{ \frac{E\left\{\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|\right\}}{2^{s-1}\varepsilon_n} + 2\frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso}) \left(1 + \frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})}{\sqrt{n}2^{2s+1}\varepsilon_n^2}\right)}{2^{2s}\sqrt{n}\varepsilon_n^2} \right\}. \end{aligned}$$

Then, by the same arguments used in Theorem 3.1 and a similar choice of ϵ_n, we can establish that

$$\|(\theta_0 - \theta_n^*) \circ \tau\| = O_P\left(n^{-1/3}\right) + O_P\left(E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|\right).$$

By the triangle inequality and the fact that $\tau_0^* = \text{argmin}_{\theta\circ\tau:\theta\in\mathcal{F}_{iso}} \|\tau_0 - \theta \circ \tau\|$ implies $\|\tau_0 - \tau_0^*\| \leq \|\tau_0 - \tau\|$, we find

$$\|\tau_0 - \tau_n^*\| \leq \|\tau_0 - \tau_0^*\| + \|\tau_0^* - \tau_n^*\| \leq \|\tau_0 - \tau\| + \|\tau_0^* - \tau_n^*\|.$$

Combining this with the main bound gives

$$\|\tau_0 - \tau_n^*\| \leq \|\tau_0 - \tau\| + O_P\left(n^{-1/3}\right) + O_P\left(E\left\|(\pi_m - \pi_0)(\mu_m - \mu_0)\right\|\right).$$

□

3.D Tables and Figures

Total Sample Size			1000		2000		5000		10000	
Scenario	CATE estimator	Cal	Lower Quantile	Upper Decile	Lower Quantile	Upper Decile	Lower Quantile	Upper Decile	Lower Quantile	Upper Decile
1	splines	yes	-0.09	0.08	-0.06	0.05	-0.02	0.02	-0.01	0.01
		no	-0.03	-0.03	-0.03	0.01	-0.01	0.00	0.02	-0.01
	glmnet	yes	-0.08	0.1	-0.06	0.06	-0.02	0.04	-0.02	0.03
		no	-0.02	-0.05	0.05	-0.09	0.02	-0.06	0.02	-0.07
	random forest	yes	-0.08	0.09	-0.05	0.04	-0.02	0.03	-0.01	0.01
		no	0.14	-0.1	0.06	0.00	0.01	0.01	0.02	0.01
2	splines	yes	-0.07	0.1	-0.05	0.06	-0.03	0.04	-0.02	0.01
		no	-0.04	0.00	-0.08	0.01	-0.04	-0.05	-0.02	-0.02
	glmnet	yes	-0.1	0.1	-0.06	0.06	-0.04	0.04	-0.02	0.02
		no	0.03	0.02	0.01	-0.04	0.00	-0.03	0.02	-0.04
	random forest	yes	-0.08	0.09	-0.05	0.06	-0.03	0.04	-0.02	0.02
		no	-0.04	0.00	-0.1	0.02	-0.09	0.03	-0.1	0.04

Table 3.D.1: Difference between the average of the causal isotonic calibrator's predictions and the true CATE values within bins defined by predictions. The results correspond to the low dimensional with binary outcome scenarios. Scenario 1 has no nonlinearities in the outcome model whereas scenario 2 does. The average predictions are taken in observations whose predicted values lie within specific deciles. Values in the lower decile column show the results when the interval is defined by the minimum and the 10% percentile of the predictions, and the upper decile columns display the results when the interval is defined between the 90% percentile and the maximum. The column *Cal* indicates whether the method is calibrated or not. From left to right, each column shows the differences of the average predictions as sample size increases. We see that the differences between the causal isotonic method and the true CATE are closer to 0 as sample size increases, indicating a good performance in large samples. The uncalibrated methods also improve, but not as substantially.

Total Sample Size			1000		2000		5000		10000	
Scenario	CATE estimator	Cal	Lower Decile	Upper Decile	Lower Decile	Upper Decile	Lower Decile	Upper Quantile	Lower Quantile	Upper Decile
3	glmnet	yes	-0.06	0.11	-0.03	0.08	-0.01	0.06	-0.01	0.04
		no	0.02	-0.02	0.03	-0.01	0.01	0.02	0.02	0.02
	glmnet + btrees	yes	-0.06	0.09	-0.03	0.05	-0.01	0.02	0.00	0.01
		no	0.07	-0.12	0.12	-0.09	0.17	-0.14	0.2	-0.15
4	glmnet	yes	-0.06	0.06	-0.03	0.05	-0.02	0.03	-0.01	0.02
		no	0.00	-0.03	0.00	-0.03	-0.01	0.02	-0.01	0.02
	glmnet + btrees	yes	-0.08	0.06	-0.04	0.04	-0.02	0.01	-0.01	0.00
		no	-0.09	0.03	0.01	0.00	0.06	-0.05	0.1	-0.08

Table 3.D.2: Difference between the average of the causal isotonic calibrator's predictions and the true CATE values within bins defined by predictions. These results correspond to the high dimensional with binary outcome scenarios. Scenario 3 has no nonlinearities in the outcome model whereas scenario 4 does. The average predictions are taken in observations whose predicted values lie within specific deciles. Values in the lower decile column show the results when the interval is defined by the minimum and the 10% percentile of the predictions, and the upper decile columns display the results when the interval is defined between the 90% percentile and the maximum. The column *Cal* indicates whether the method is calibrated or not. From left to right, each column shows the differences of the average predictions as sample size increases. Differences between the causal isotonic method and the true CATE decrease as sample size increases. On the contrary, the boosted trees approach does not improve as sample size increases. Glmnet performes well overall. Abbreviations: glmnet: generlaized linear model with elastic net penalty, btrees: gradient boosted trees.

Total Sample Size			1000		2000		5000		10000	
Scenario	CATE Estimator	CAL	Lower Decile	Upper Quantile	Lower Decile	Upper Quantile	Lower Decile	Upper Quantile	Lower Decile	Upper Quantile
5	splines	yes	0.05	0.05	0.03	0.03	0.02	0.02	0.01	0.01
		no	-0.1	-0.02	0.01	0.02	-0.02	0.00	-0.03	0.02
	glmnet	yes	-0.07	0.05	-0.06	0.02	-0.02	0.01	-0.03	0.01
		no	0.07	0.09	0.07	0.07	0.05	0.05	0.09	0.1
	random forest	yes	0.05	0.04	0.03	0.03	0.01	0.01	0.01	0.01
		no	-0.46	-0.21	-0.27	-0.11	-0.12	-0.07	-0.06	-0.05
6	splines	yes	0.03	-0.09	0.02	-0.08	0.01	-0.03	0.01	-0.03
		no	0.02	0.06	0.01	0.08	0.00	-0.01	0.01	-0.02
	glmnet	yes	0.03	-0.04	0.02	-0.02	0.01	-0.01	0.01	-0.01
		no	0.00	0.00	-0.01	-0.01	-0.02	-0.02	-0.02	-0.02
	random forest	yes	0.04	-0.09	0.03	-0.07	0.01	-0.03	0.01	-0.03
		no	0.00	0.08	0.03	-0.02	0.03	-0.05	0.04	-0.07

Table 3.D.3: Relative difference between the average of the causal isotonic calibrator's predictions and the true CATE values within bins defined by predictions. These results correspond to the low dimensional with continuous outcome scenarios. Scenario 5 has no nonlinearities in the outcome model whereas scenario 6 does. The average predictions are taken in observations whose predicted values lie within specific deciles. Values in the lower decile column show the results when the interval is defined by the minimum and the 10% percentile of the predictions, and the upper decile columns display the results when the interval is defined between the 90% percentile and the maximum. The column *Cal* indicates whether the method is calibrated or not. From left to right, each column shows the relative differences of the average predictions as sample size increases. When sample size is large, see that the relative differences between the causal isotonic method and the true CATE are lower then the uncalibrated methods. However, when sample size is lower, our method performed similarly or worse.

Total Sample Size			1000		2000		5000		10000	
Scenario	CATE estimator	Cal	Lower Decile	Upper Decile	Lower Decile	Upper Decile	Lower Decile	Upper Quantile	Lower Quantile	Upper Decile
7	glmnet	yes	-0.2	0.11	-0.17	0.06	-0.09	0.04	-0.05	0.01
		no	0.14	0.15	0.09	0.09	0.17	0.15	0.21	0.2
	glmnet + btrees	yes	0.31	0.05	0.00	0.02	0.02	-0.01	-0.02	-0.01
		no	-0.43	-0.22	-0.43	-0.21	-0.35	-0.16	-0.33	-0.15
8	glmnet	yes	-0.15	0.07	-0.09	0.05	-0.04	0.03	-0.02	0.02
		no	-0.22	-0.23	-0.2	-0.19	-0.19	-0.19	-0.18	-0.18
	glmnet + btrees	yes	4.69	0.05	0.26	0.02	0.04	0.00	0.02	-0.01
		no	-0.14	-0.12	-0.23	-0.14	-0.33	-0.13	-0.35	-0.12

Table 3.D.4: Relative difference between the average of the causal isotonic calibrator's predictions and the true CATE values within bins defined by predictions.. These results correspond to the high dimensional with continuous outcome scenarios. Scenario 7 has no nonlinearities in the outcome model whereas scenario 8 does. The average predictions are taken in observations whose predicted values lie within specific deciles. Values in the lower decile column show the results when the interval is defined by the minimum and the 10% percentile of the predictions, and the upper decile columns display the results when the interval is defined between the 90% percentile and the maximum. The column *Cal* indicates whether the method is calibrated or not. From left to right, each column shows the relative differences of the average predictions as sample size increases. In moderate to large sample sizes, we see that our method generally outperforms its uncalibrated counterparts. Abbreviations: glmnet: generalized linear model with elastic net penalty, btrees: gradient boosted trees.

Scenario	Outcome	Number of confounders	CATE Estimators	Pseudo-outcome μ_n component	Pseudo-outcome π_n component
1, 2	Binary	4	logistic regression, logistic regression with elastic net regularization, generalized additive models, regression trees with gradient boosting with depths of 2,3, 5, 6, and 8, random forests, multivariate adaptive splines	Super learner with generalized linear model, generalized linear model with elastic net penalty, boosted random forest trees with depths of 2,4, and 6, and generalized additive model	Super learner with generalized linear model, generalized linear model with elastic net penalty, boosted random forest trees with depths of 2,4, and 6, and generalized additive model
3,4	Binary	100	logistic regression with elastic net regularization, variable screening followed by regression trees with gradient boosting, random forests	linear regression with elastic net regularization	logistic regression with elastic net regularization

Table 3.D.5: Information on the CATE and pseudo-outcome estimators used in each scenario

Scenario	Outcome	Number of confounders	CATE Estimators	Pseudo-outcome μ_n component	Pseudo-outcome π_n component
5,6	Continuous	4	linear regression, linear regression with elastic net regularization, generalized additive models, regression trees with gradient boosting, with depths of 2,3,5,6, and 8, random forests, multivariate adaptive splines	Super learner with glm, generalized linear model with elastic net penalty, boosted random forest with depths of 2, 4, and 6, and gam	Super learner with glm, glm with elastic net penalty, boosted random forest trees with depths of 2,4, and 6, and gam
7,8	Continuous	100	linear regression with elastic net regularization, variable screening followed by regression trees with gradient boosting, random forests	linear regression with elastic net regularization	logistic regression with elastic net regularization

Table 3.D.6: Information on the CATE and pseudo-outcome estimators used in each scenario

3.D.1 Data generating distribution in simulation studies

Below are the formulas for each of the simulated data sets. The only difference between the continuous and binary settings are the link functions of the outcome model.

Binary outcome

- Scenario 1:
  - $\text{logit}(P(Y = 1|A = a, W = w)) = 1.5 + 1.5a + 3aw_1 - 2.5(1 - a)w_2 + 2.5aw_3 + 1.5(1 - a)w_4$
  - $\text{logit}(\pi_0(w)) = -0.25 - w_1 + .5w_2 - w_3 + 0.5w_4$
- Scenario 2:
  - $\text{logit}(P(Y = 1|A = a, W = w)) = -1.5 + 1.5a + 2a|w_1||w_2| - 2.5(1 - a)|w_2|w_3 + 2.5w_3 + 2.5(1 - a)\sqrt{|w_4|} - 1.5aI(w_2 < .5) + 1.5(1 - a)I(w_4 < 0)$
  - $\text{logit}(\pi_0(w)) = -0.25 - w_1 + .5w_2 - w_3 + 0.5w_4$
- Scenario 3:
  - $\text{logit}(P(Y = 1|A = a, W = w)) = -0.5 + 3.5a + 3aw_1 + 6.5(1 - a)w_2 + 1.5aw_3 + 4(1 - a)w_4 + 2.5aw_5 - 6(1 - a)w_6 + 1aw_7 + 4.5(1 - a)w_8 + aw_9 + 2.5(1 - a)w_{10} + 1.5w_{11} - 2.5w_{12} + w_{13} - 1.5w_{14} + 3w_{15} - 2w_16 + 3w_{17} - w_{18} + 1.5w_{19} - 2w_{20}$
  - $\text{logit}(\pi_0(w)) = .2 - .5w_1 - 0.5w_2 - .5w_3 + .5w_4 - .5w_5 + 0.5w_6 - .5w_7 - .5w_8 - .5w_9 - .2w_{10} + .5w_{11} - w_{12} + w_{13} - 1.5w_{14} + w_{15} - w_{16} + 2w_{17} - w_{18} + 1.5w_{19} - w_{20}$
- Scenario 4:
  - $\text{logit}(P(Y = 1|A = a, W = w)) = 1.5 + 1.5a - 2.5(1 - a)|w_1| - 3.5(1 - a)|w_2||w_3| - 2.5(1 - a)|w_3| + 2(1 - a) - 2|w_4| + 3.5(1 - a)I(w_5 < 0.5) + 3(1 - a)I(w_6 >$

$0.5)|w_7|+1.5aw_7+3.5aI(w_8>0.5)+2.5aw_9+2aI(w_{10}<.2)+1.5a|w_{11}|-2.5aw_12+w_13-1.5|w_14|+2.5aI(w_{15}>.2)|w_14|-2a|w_{16}|w_{17}+3w_{17}-w_{18}+1.5w_{19}-w_{20}$

- $\text{logit}(\pi_0(w)) = .2-.4w_1+0.4w_2-.4w_3+.5w_4-.5w_5+0.4w_6-.4w_7+.5w_8-.5w_9+.4w_{10}-.4w_{11}+.5w_{12}-.5w_{13}+.4w_{14}-.4w_{15}+0.5w_{16}-.5w_{17}+0.4w_{18}-0.4w_{19}+0.5w_{20}$

Continuous outcome

- Scenario 5:
  - $E[Y|A=a,W=w] = 1.5+1.5a+3aw_1-2.5(1-a)w_2+2.5aw_3+1.5(1-a)w_4$
  - $\text{logit}(\pi_0(w)) = -0.25-w_1+.5w_2-w_3+0.5w_4$
- Scenario 6:
  - $E[Y|A=a,W=w] == -1.5+1.5a+2a|w_1||w_2|-2.5(1-a)|w_2|w_3+2.5w_3+2.5(1-a)\sqrt{|w_4|}-1.5aI(w_2<.5)+1.5(1-a)I(w_4<0)$
  - $\text{logit}(\pi_0(w)) = -0.25-w_1+.5w_2-w_3+0.5w_4$
- Scenario 7:
  - $E[Y|A=a,W=w] = -0.5+3.5a+3aw_1+6.5(1-a)w_2+1.5aw_3+4(1-a)w_4+2.5aw_5-6(1-a)w_6+1aw_7+4.5(1-a)w_8+aw_9+2.5(1-a)w_{10}+1.5w_{11}-2.5w_{12}+w_{13}-1.5w_{14}+3w_{15}-2w_16+3w_{17}-w_{18}+1.5w_{19}-2w_{20}$
  - $\text{logit}(\pi_0(w)) = .2-.5w_1-0.5w_2-.5w_3+.5w_4-.5w_5+0.5w_6-.5w_7-.5w_8-.5w_9-.2w_{10}+.5w_{11}-w_{12}+w_{13}-1.5w_{14}+w_{15}-w_{16}+2w_{17}-w_{18}+1.5w_{19}-w_{20}$
- Scenario 8:
  - $E[Y|A=a,W=w] = 1.5+1.5a-2.5(1-a)|w_1|-3.5(1-a)|w_2||w_3|-2.5(1-a)|w_3|+2(1-a)-2|w_4|+3.5(1-a)I(w_5<0.5)+3(1-a)I(w_6>0.5)|w_7|+$

$1.5aw_7 + 3.5aI(w_8 > 0.5) + 2.5aw_9 + 2aI(w_{10} < .2) + 1.5a|w_{11}| - 2.5aw_12 + w_13 - 1.5|w_14| + 2.5aI(w_{15} > .2)|w_14| - 2a|w_{16}|w_{17} + 3w_{17} - w_{18} + 1.5w_{19} - w_{20}$

– $\text{logit}(\pi_0(w)) = .2 - .4w_1 + 0.4w_2 - .4w_3 + .5w_4 - .5w_5 + 0.4w_6 - .4w_7 + .5w_8 - .5w_9 + .4w_{10} - .4w_{11} + .5w_{12} - .5w_{13} + .4w_{14} - .4w_{15} + 0.5w_{16} - .5w_{17} + 0.4w_{18} - 0.4w_{19} + 0.5w_{20}$

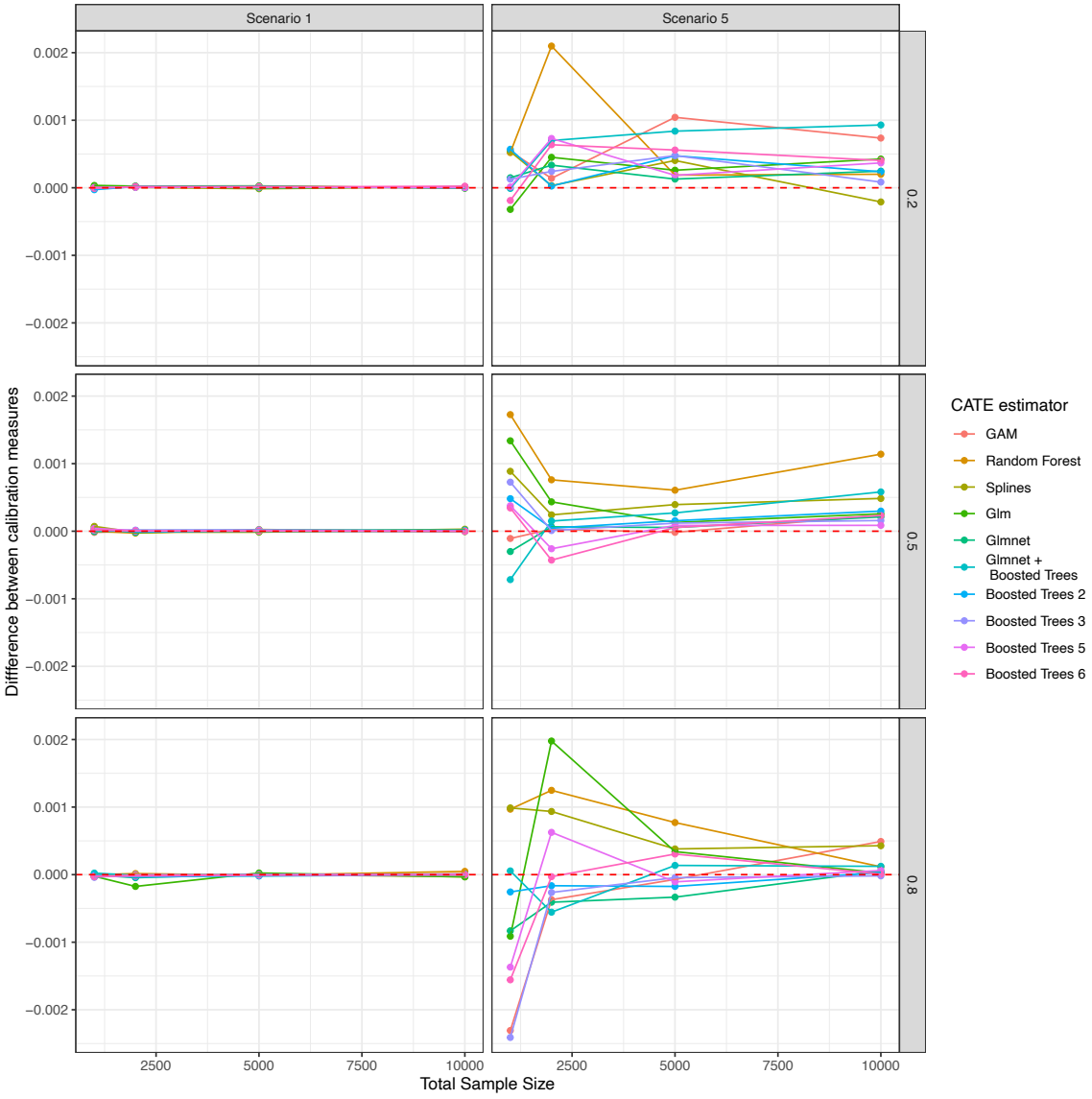

Figure 3.D.1: Difference in calibration measured between our gradient boosted approach and taking empirical means to estimate $\gamma_0(\hat{\tau}, w)$. The panels on the left and right indicate the difference in scenarios 1 and 5, respectively. Each row indicates a proportion of sample size used to derive the initial CATE estimator. The absolute difference between the calibration measures is no larger than 0.002. The numbers in the legend indicate the maximum depth of the gradient boosted trees. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

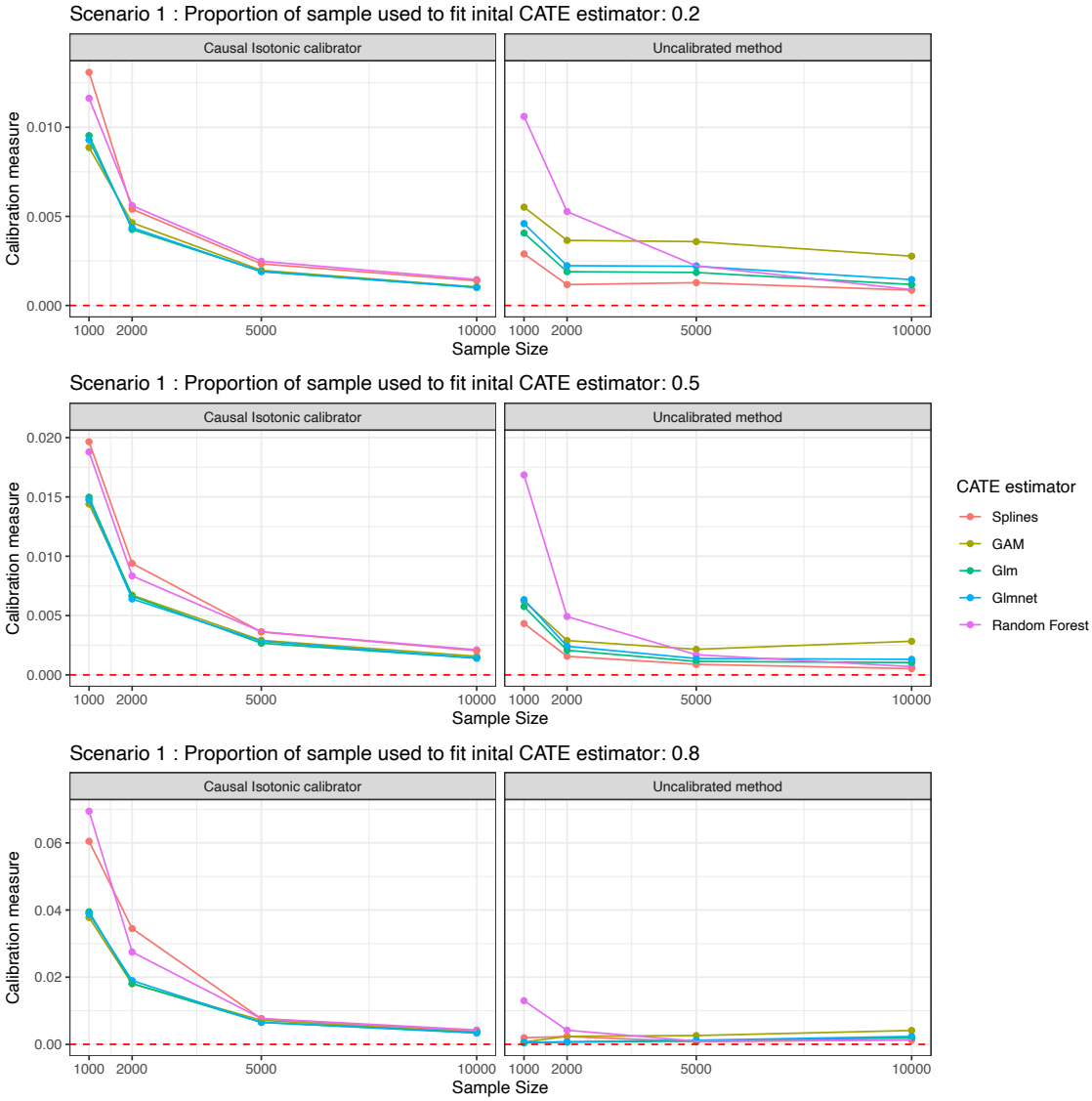

Figure 3.D.2: Calibration measures in scenario 1 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

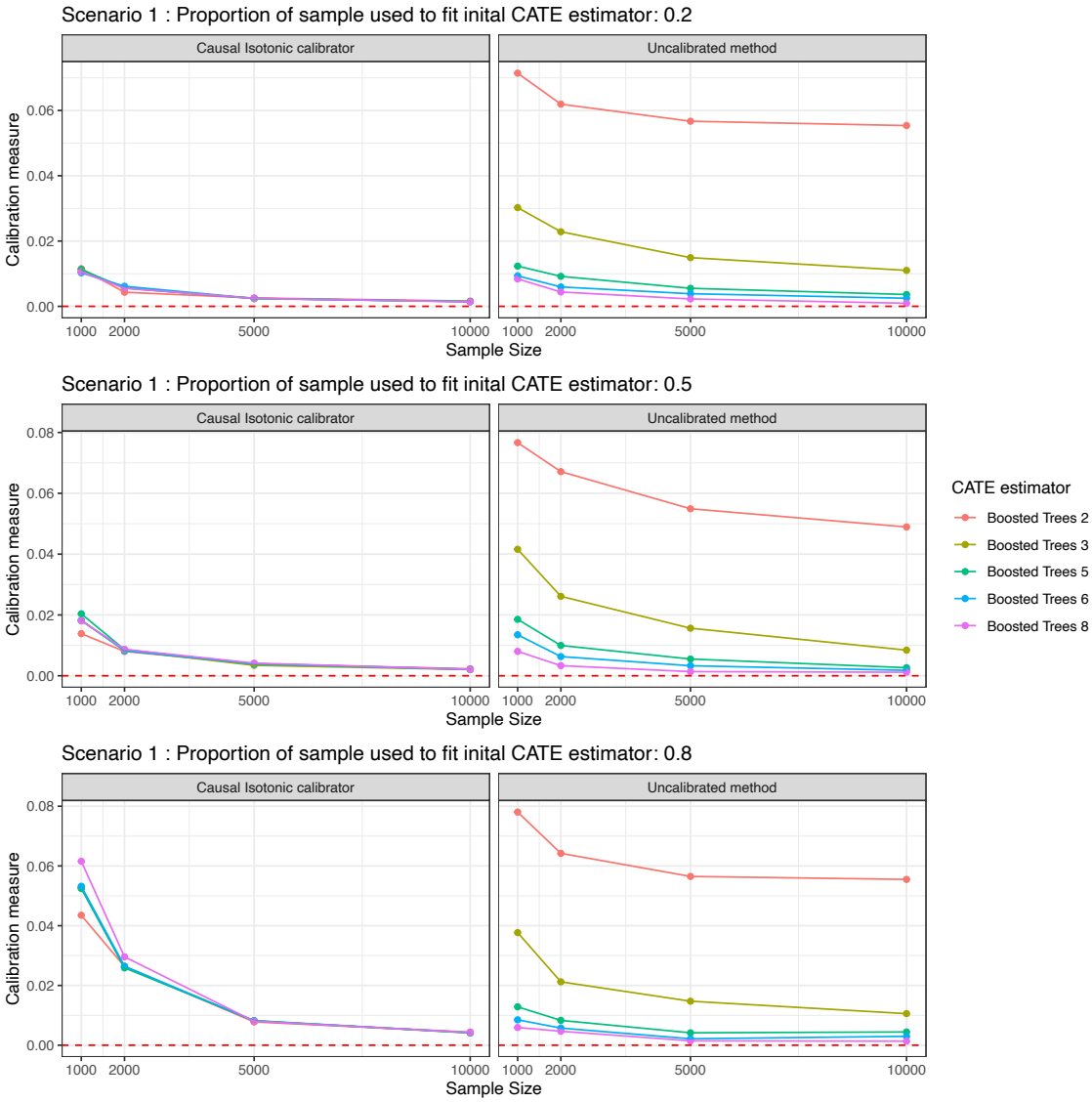

Figure 3.D.3: Calibration measures in scenario 1 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

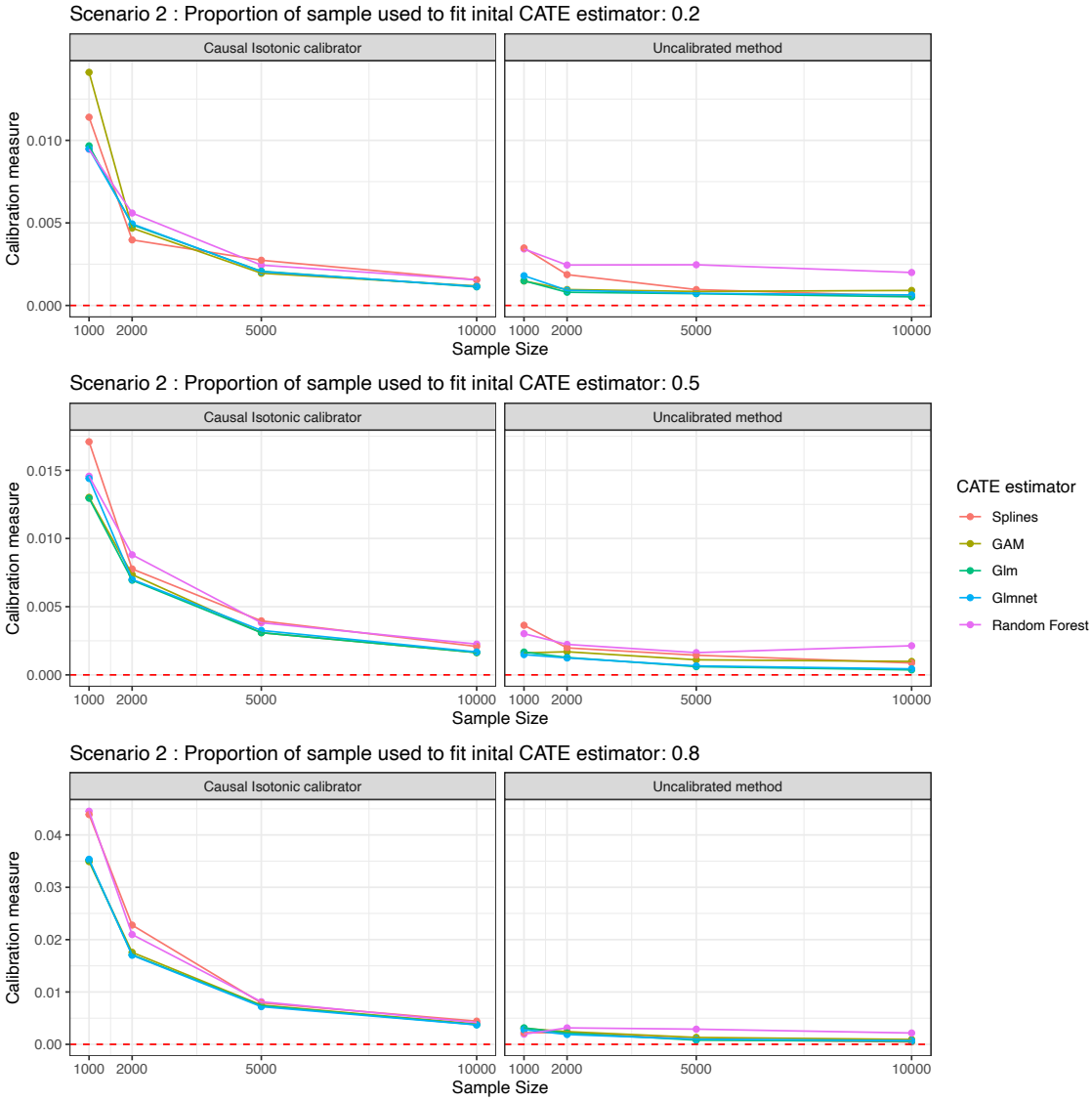

Figure 3.D.4: Calibration measures in scenario 2 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

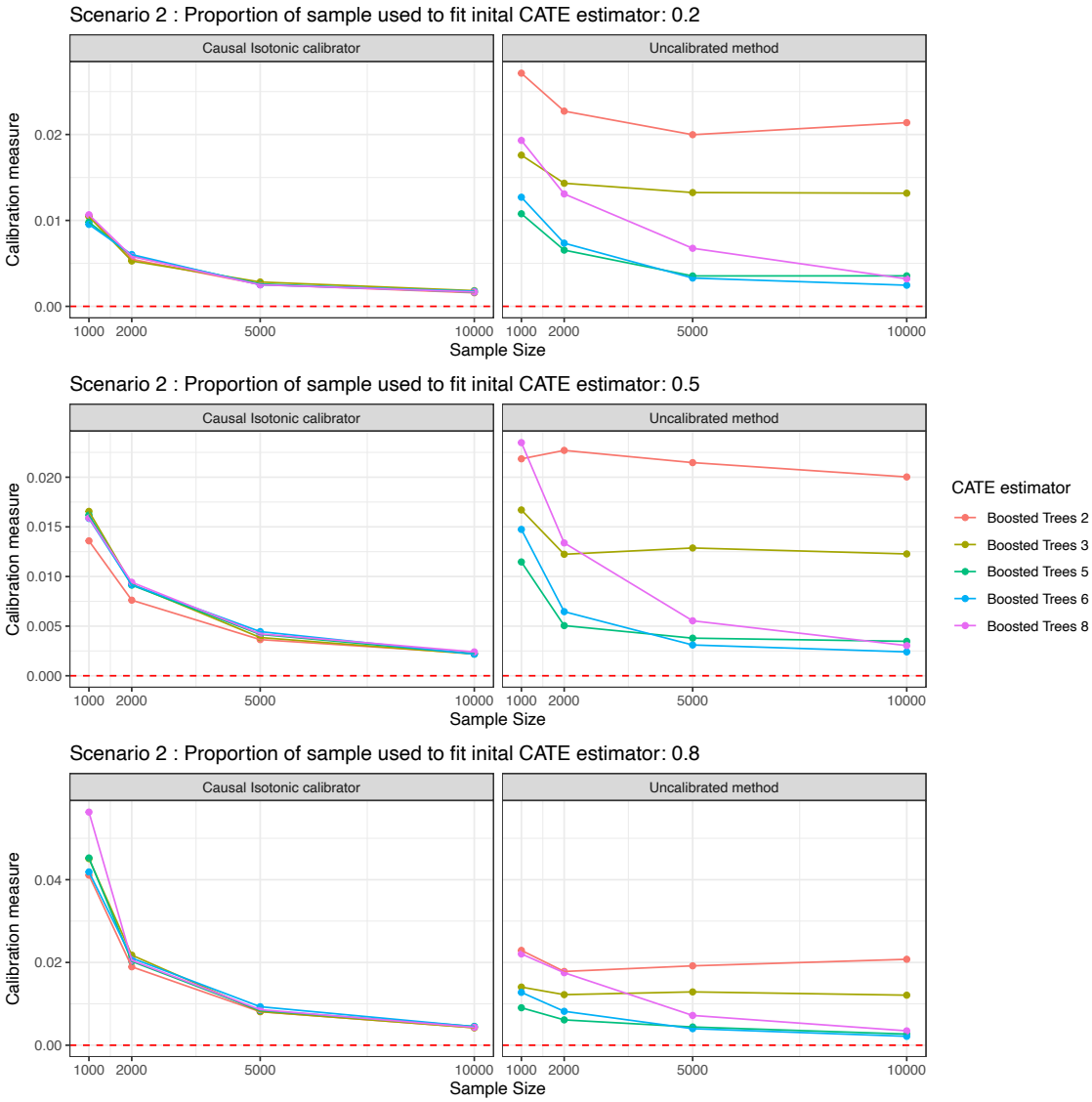

Figure 3.D.5: Calibration measures in scenario 2 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

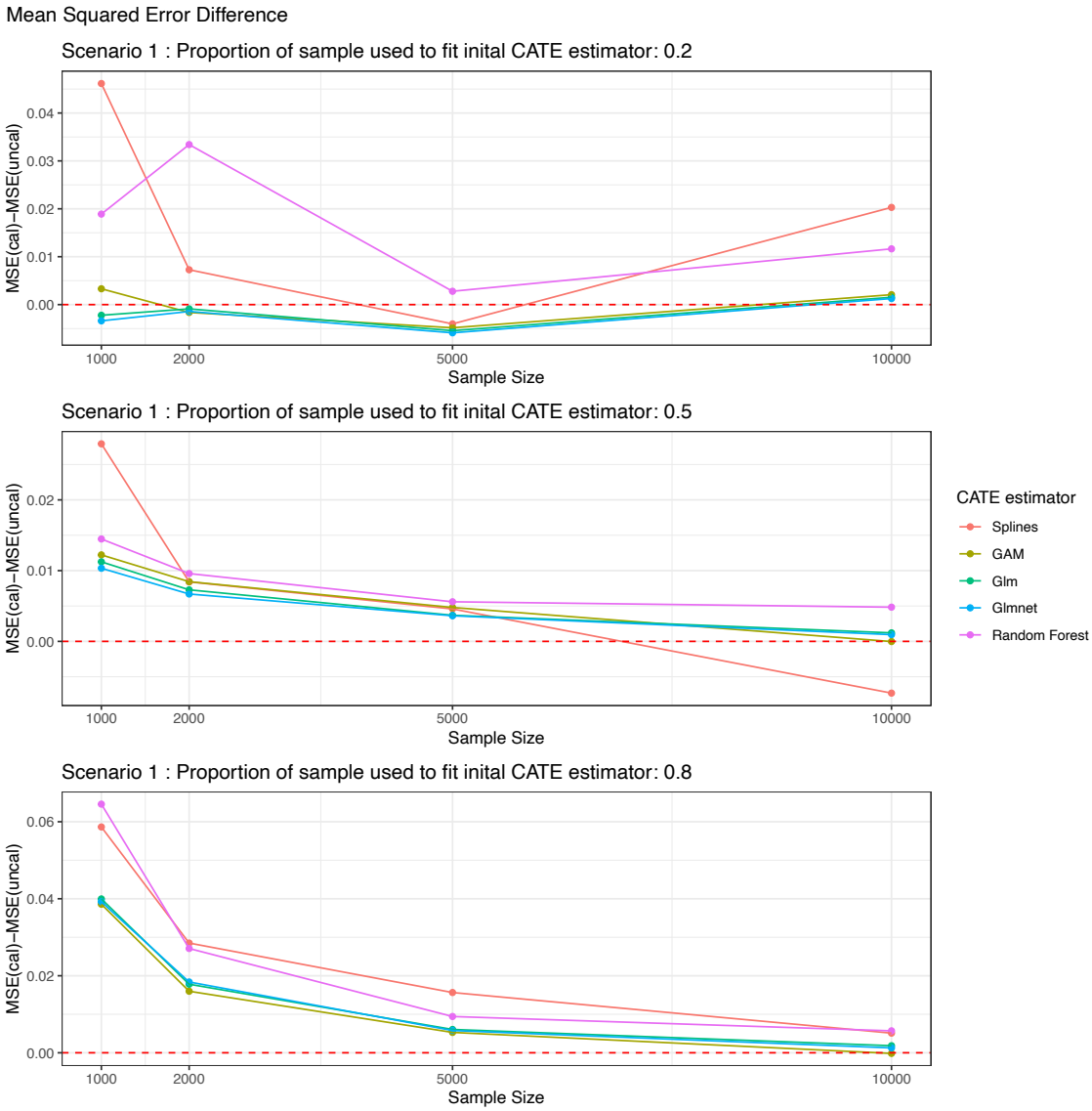

Figure 3.D.6: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in Scenario 1. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

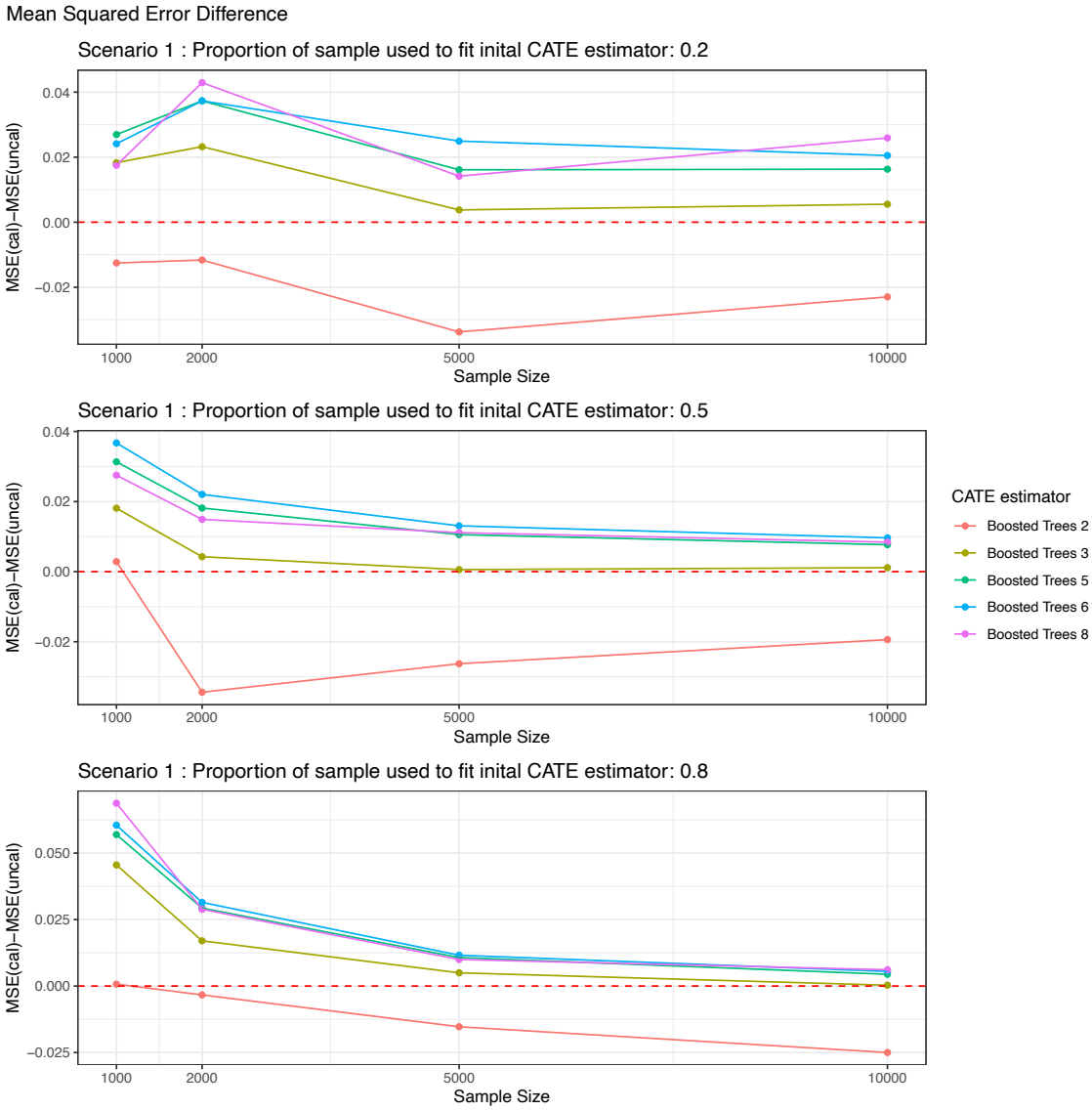

Figure 3.D.7: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 1. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

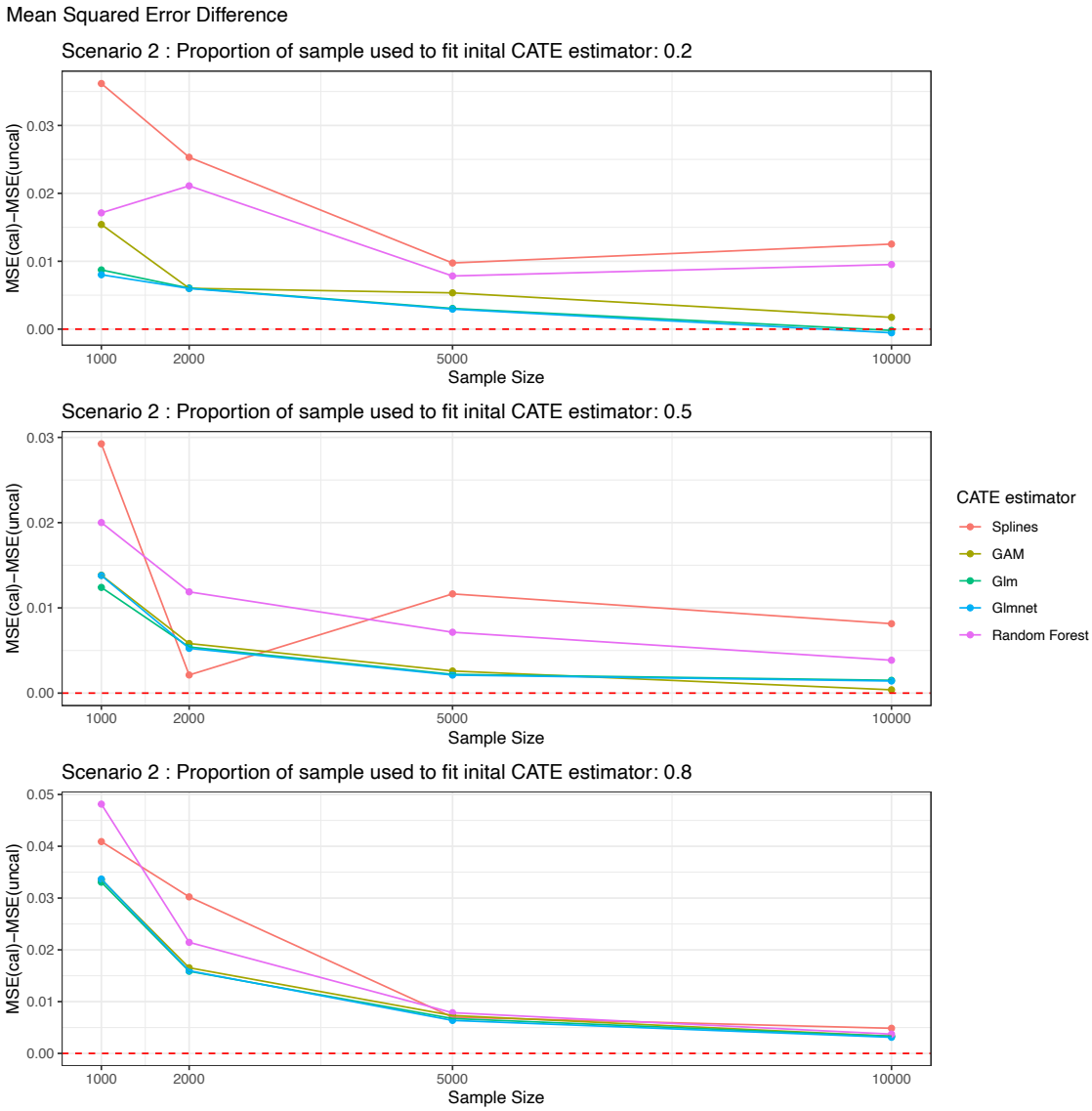

Figure 3.D.8: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 2. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

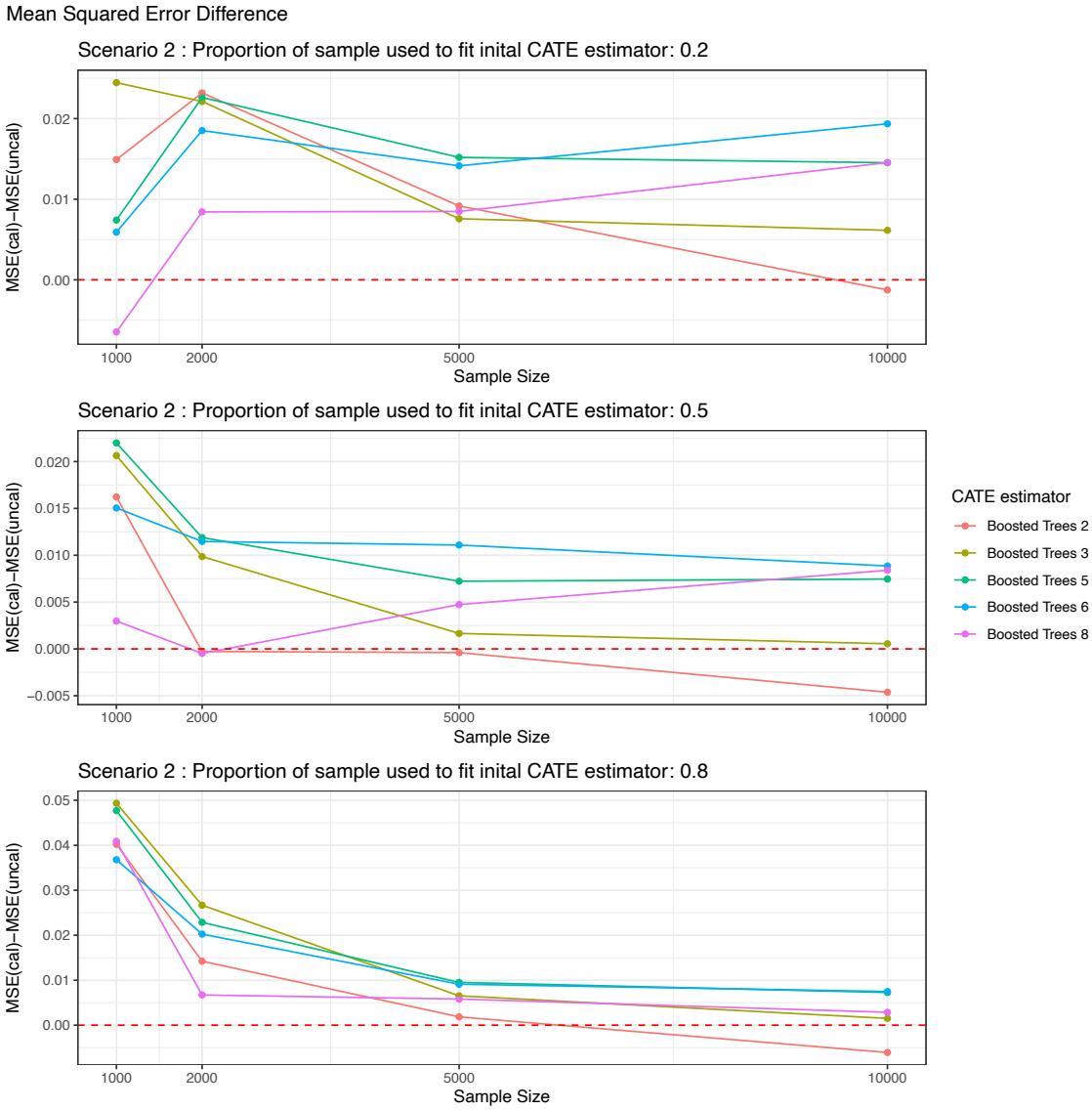

Figure 3.D.9: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 2. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

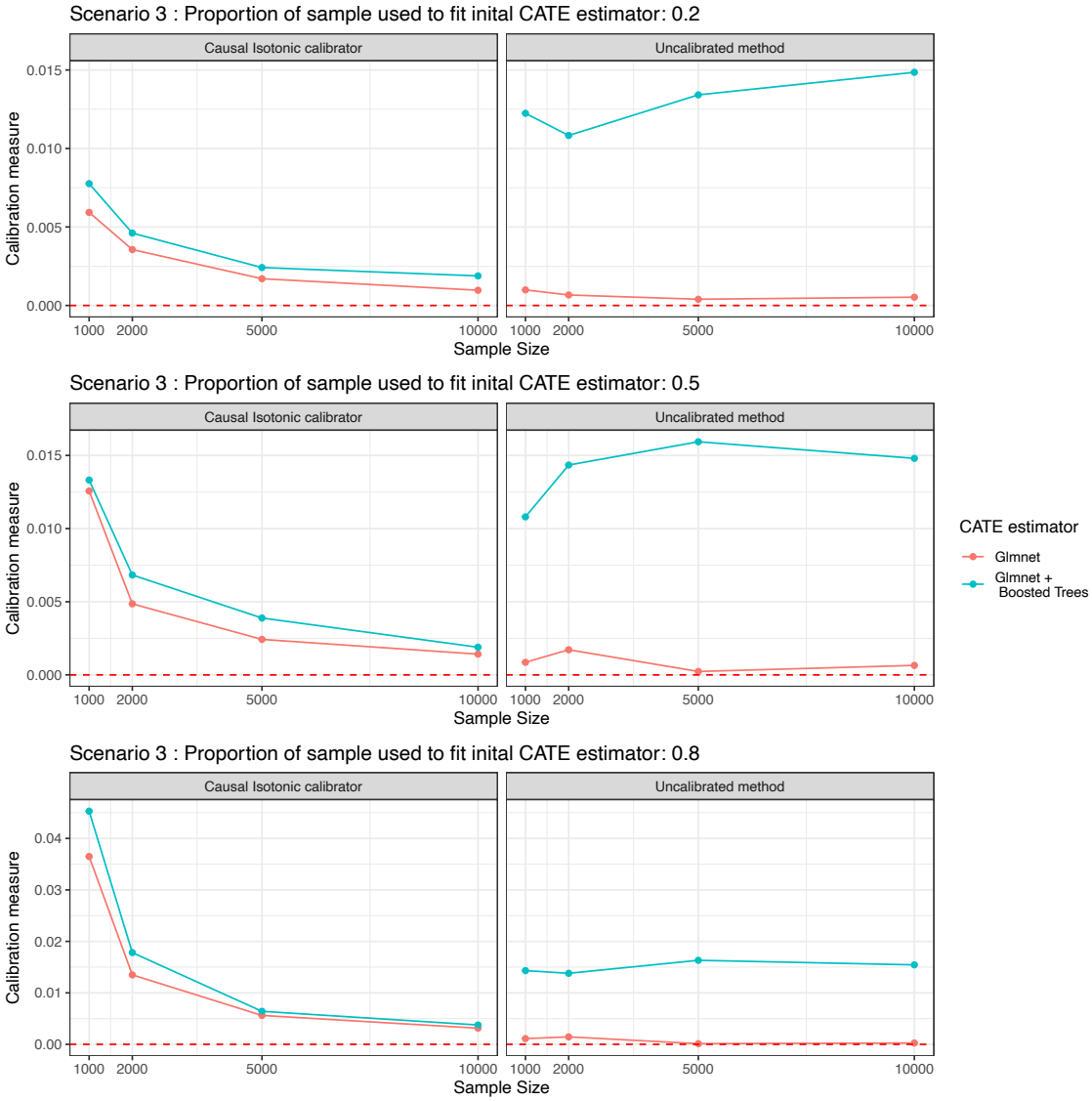

Figure 3.D.10: Calibration measures in scenario 3 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

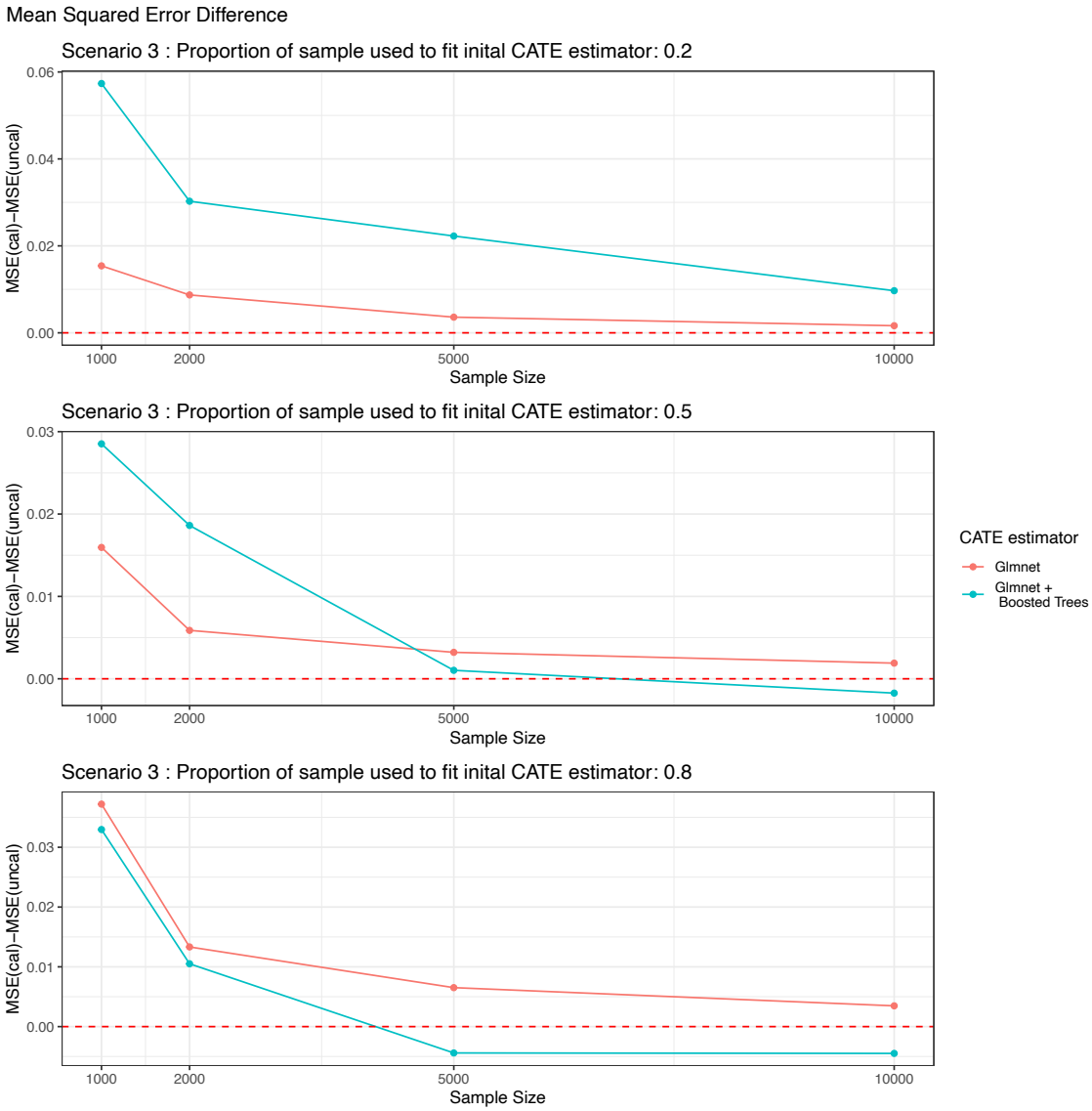

Figure 3.D.11: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 3. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: Glmnet: Generalized linear model with elastic net penalty.

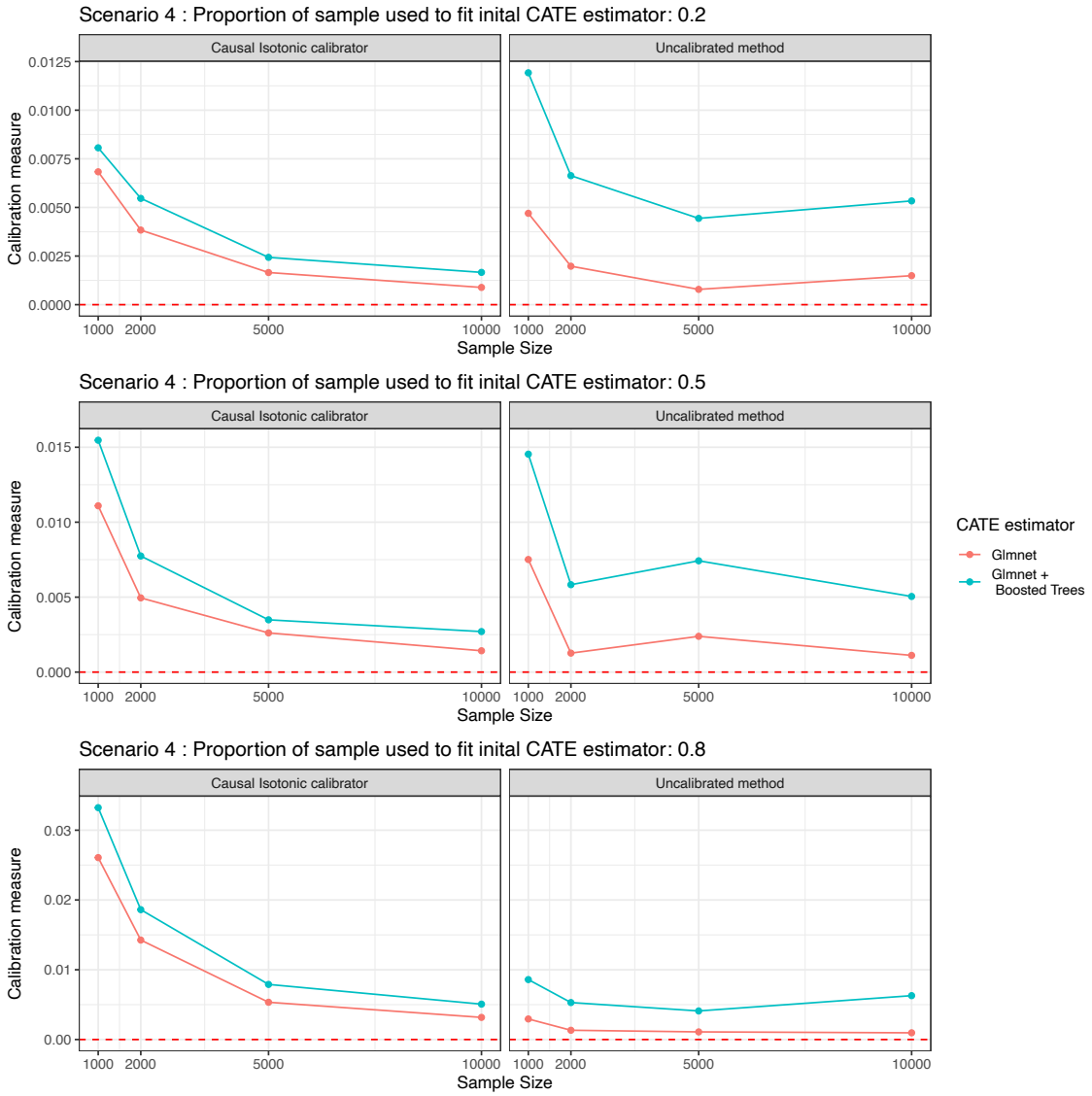

Figure 3.D.12: Calibration measures in scenario 4 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: Glmnet: Generalized linear model with elastic net penalty.

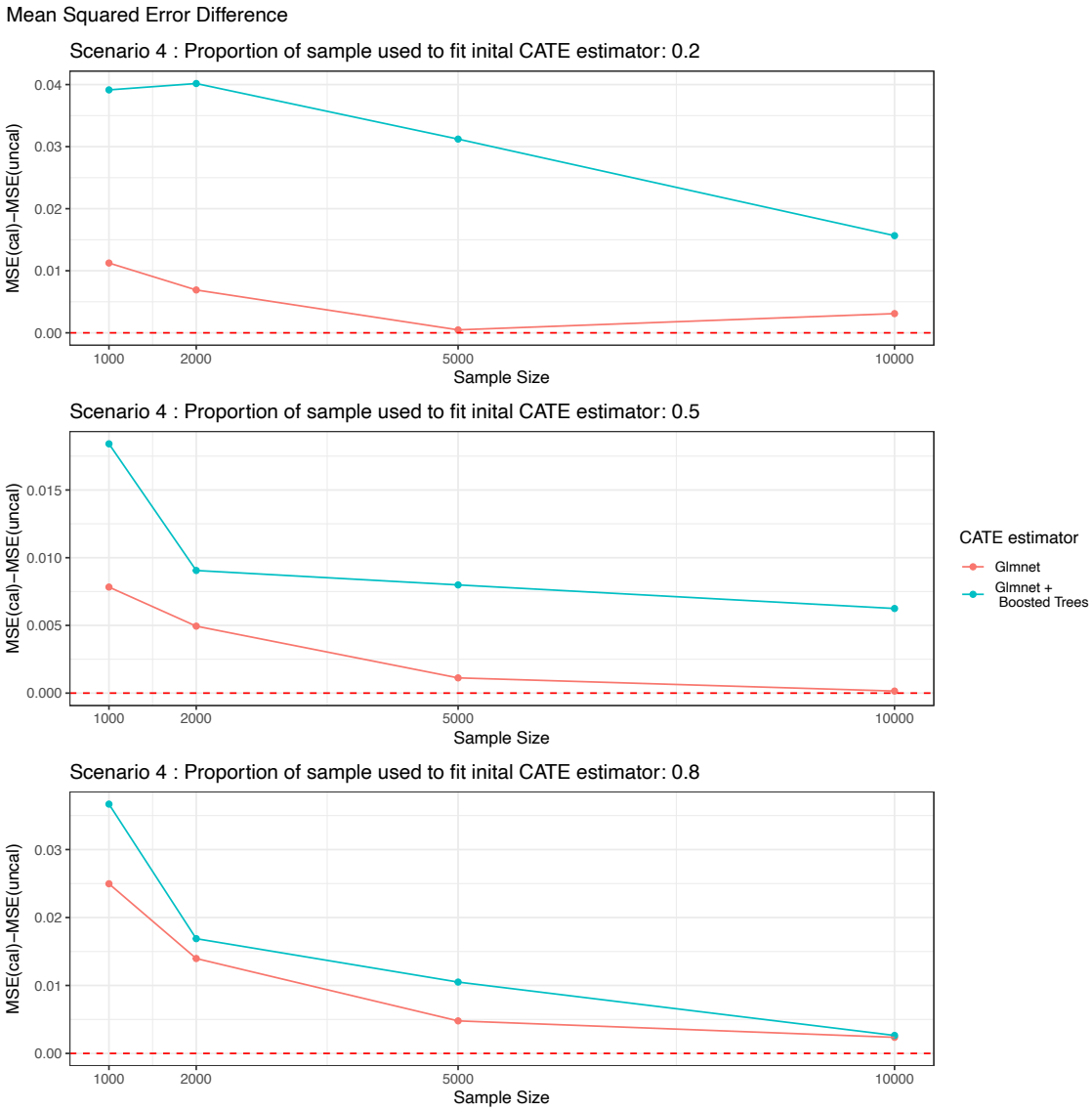

Figure 3.D.13: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 4. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: Glmnet: Generalized linear model with elastic net penalty.

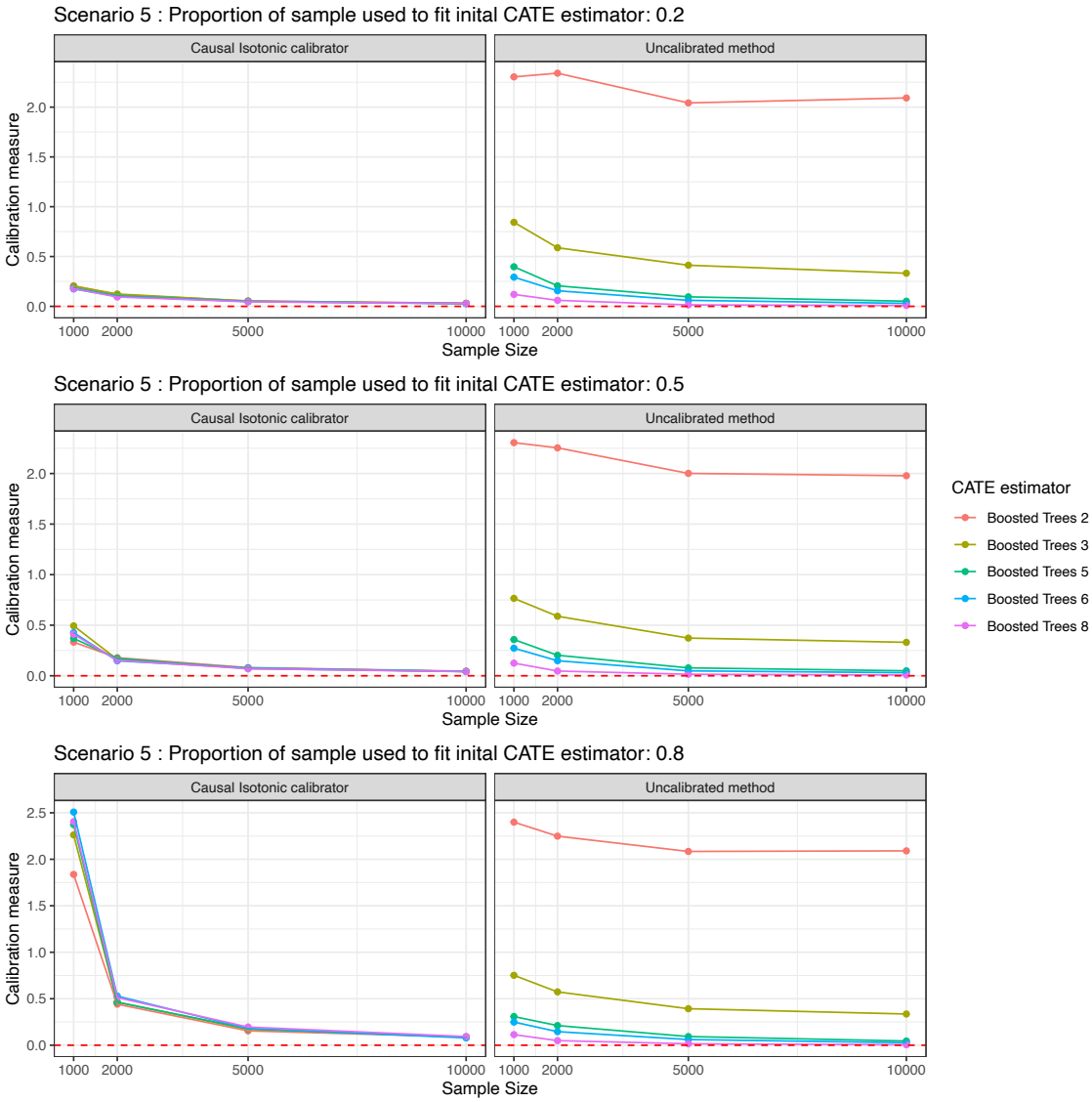

Figure 3.D.14: Calibration measures in scenario 5 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

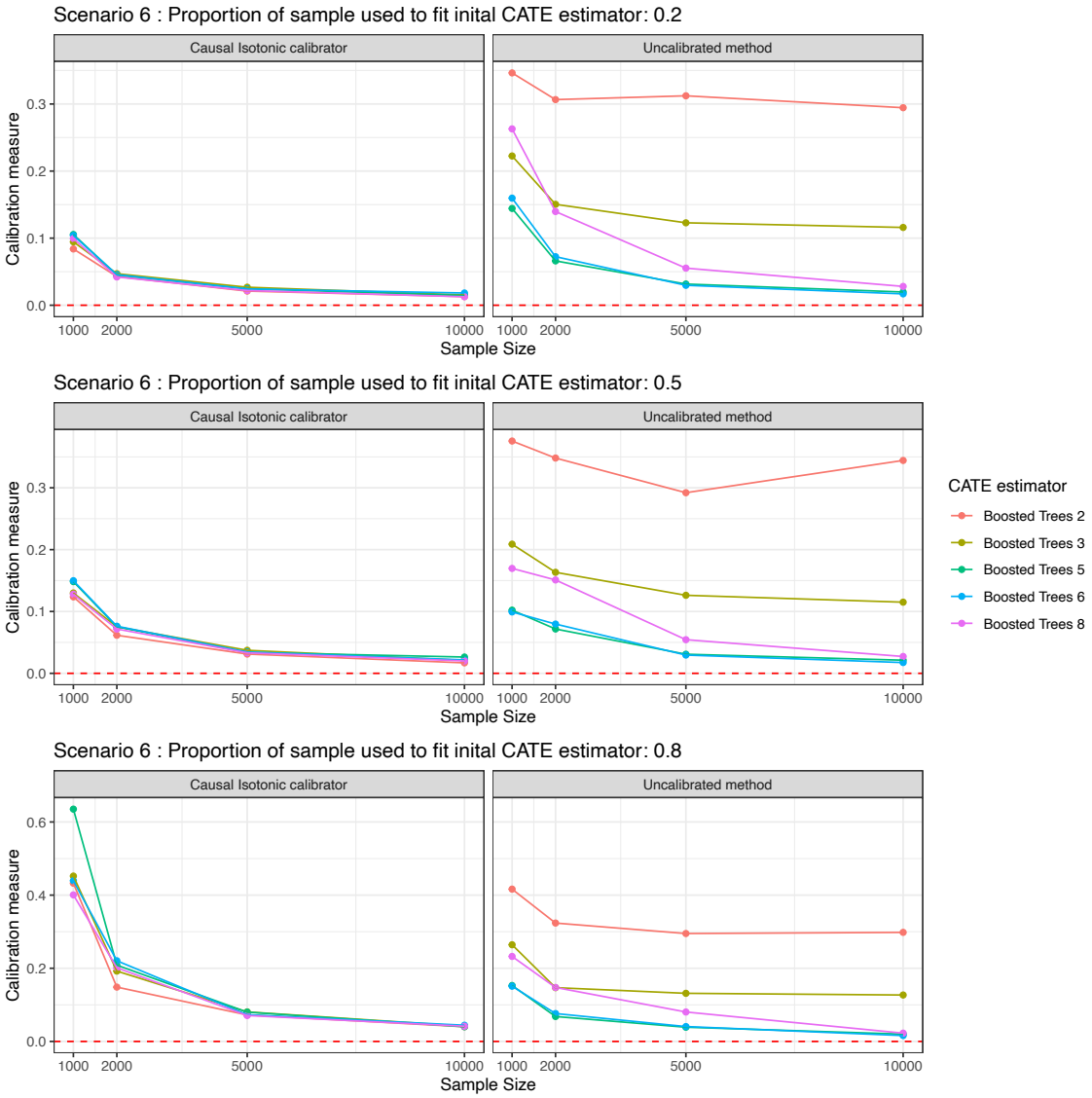

Figure 3.D.15: Calibration measures in scenario 6 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

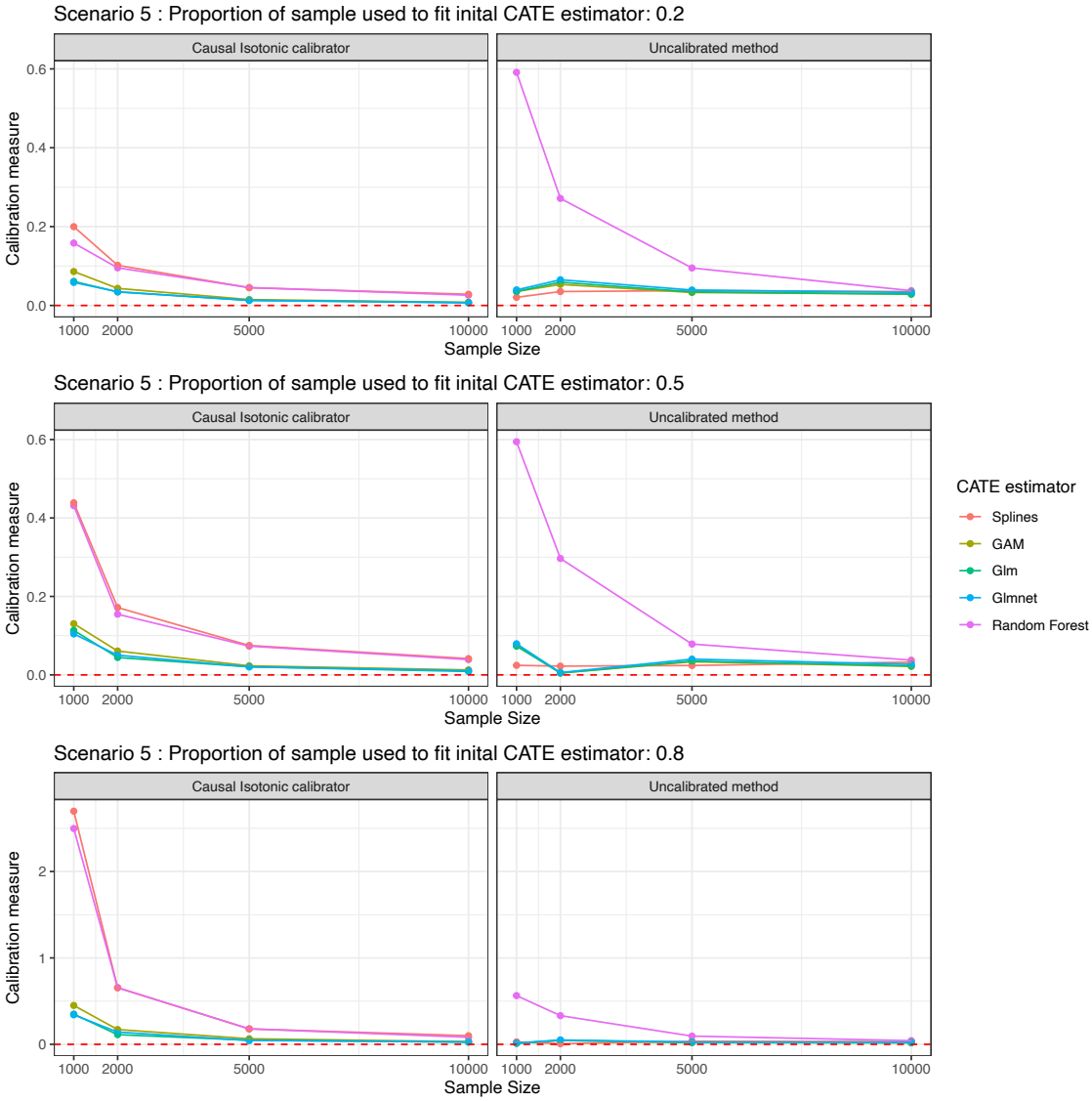

Figure 3.D.16: Calibration measures in scenario 5 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

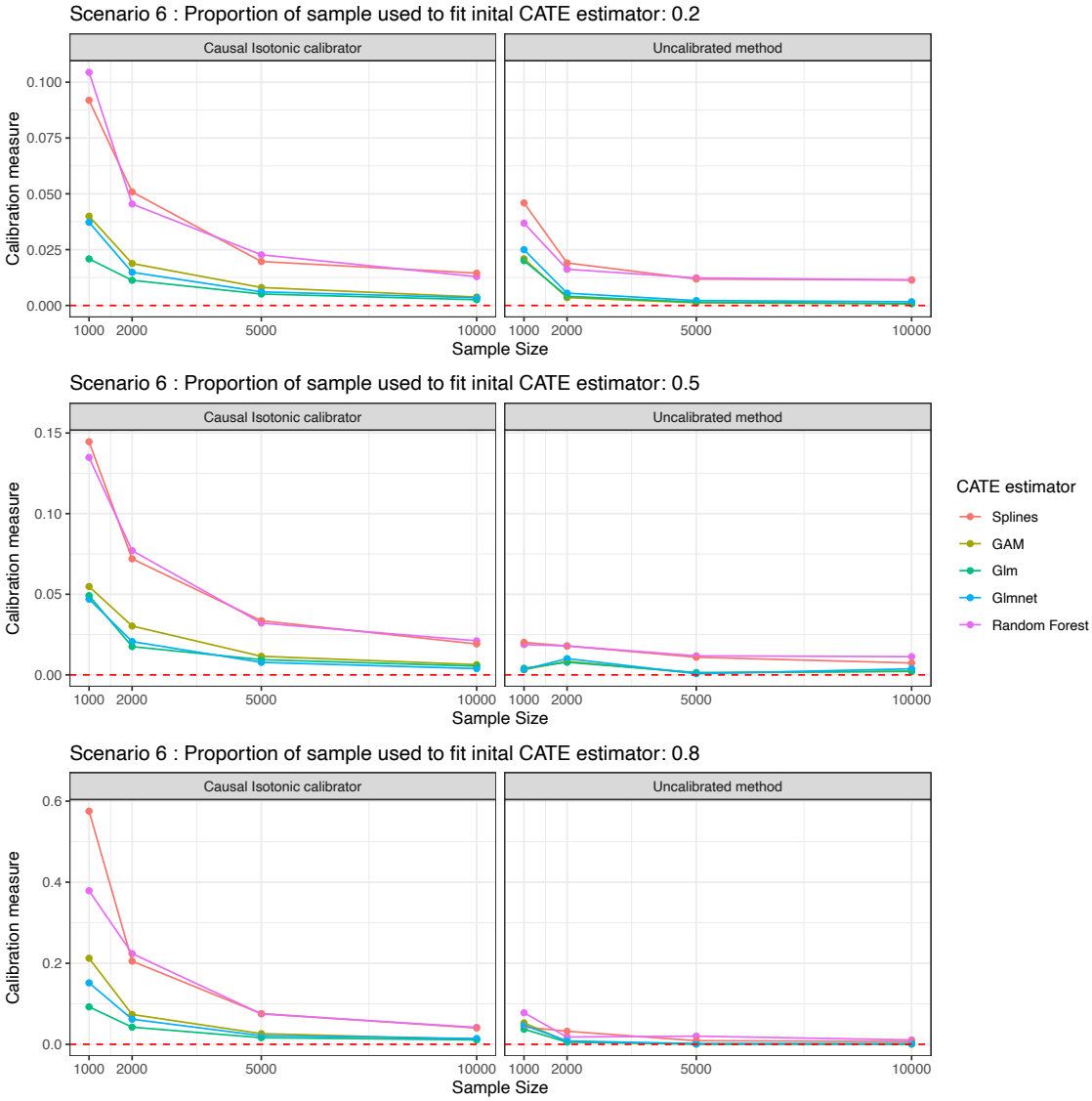

Figure 3.D.17: Calibration measures in scenario 6 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

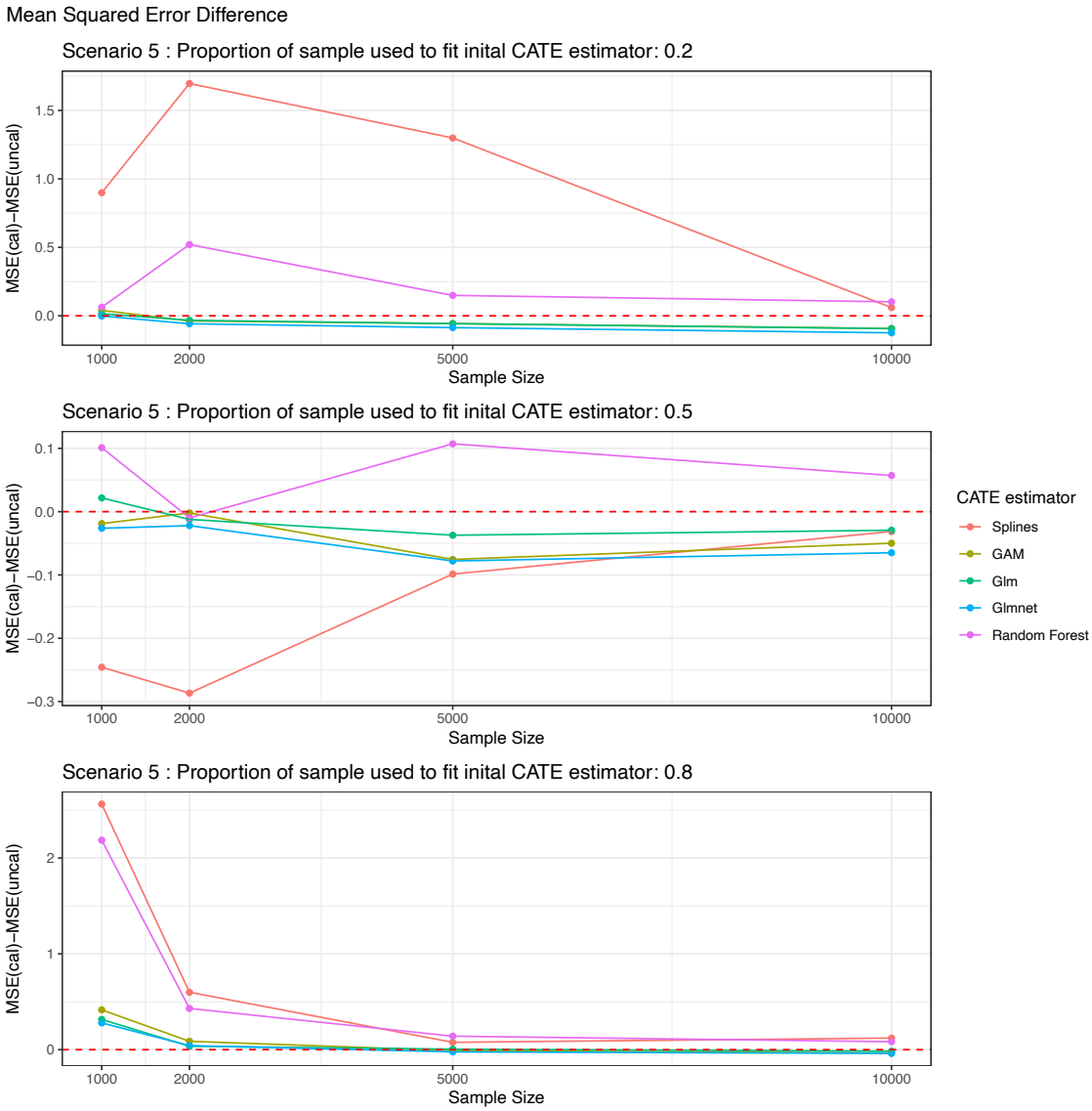

Figure 3.D.18: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 5. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty

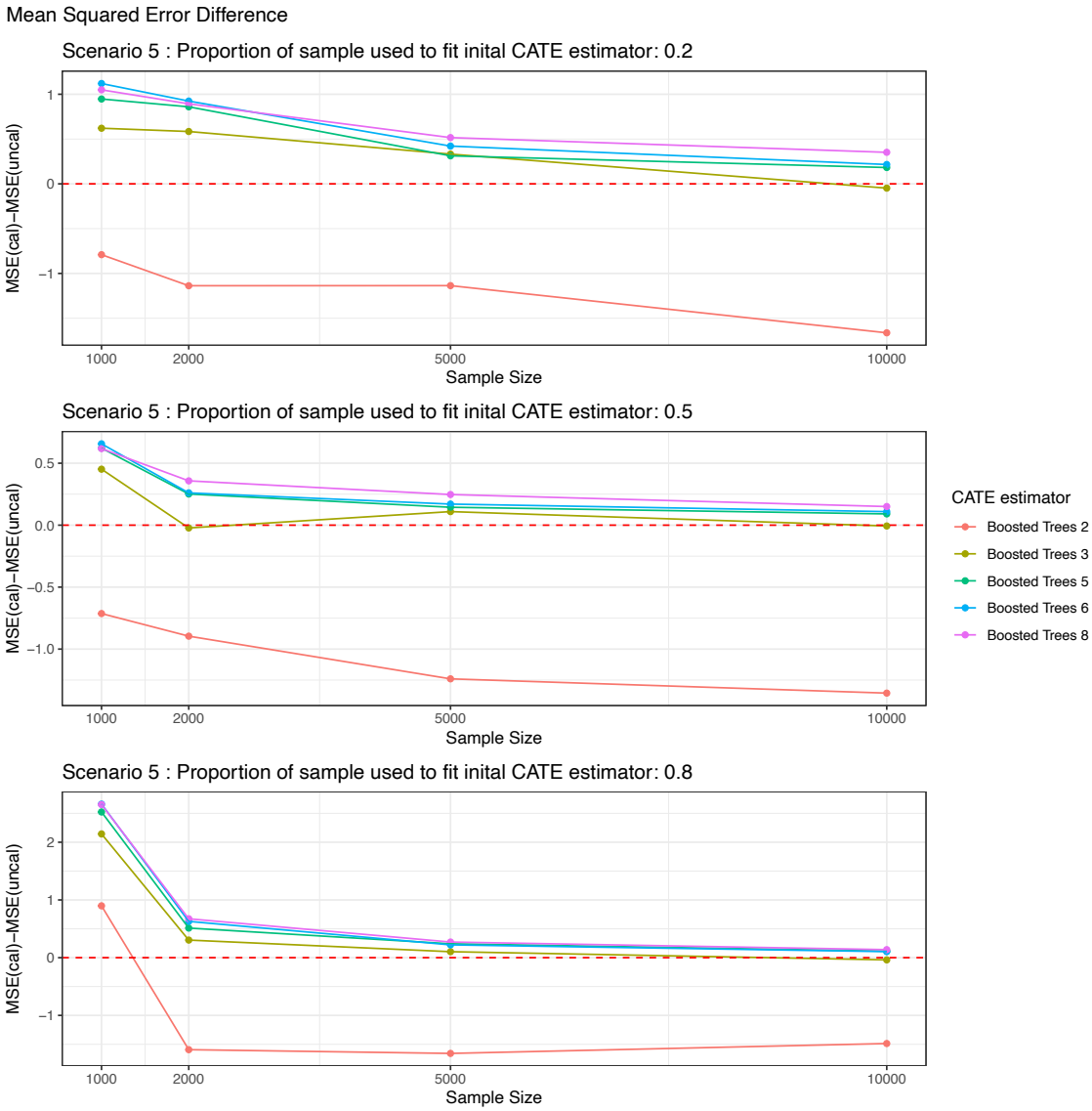

Figure 3.D.19: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 5. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

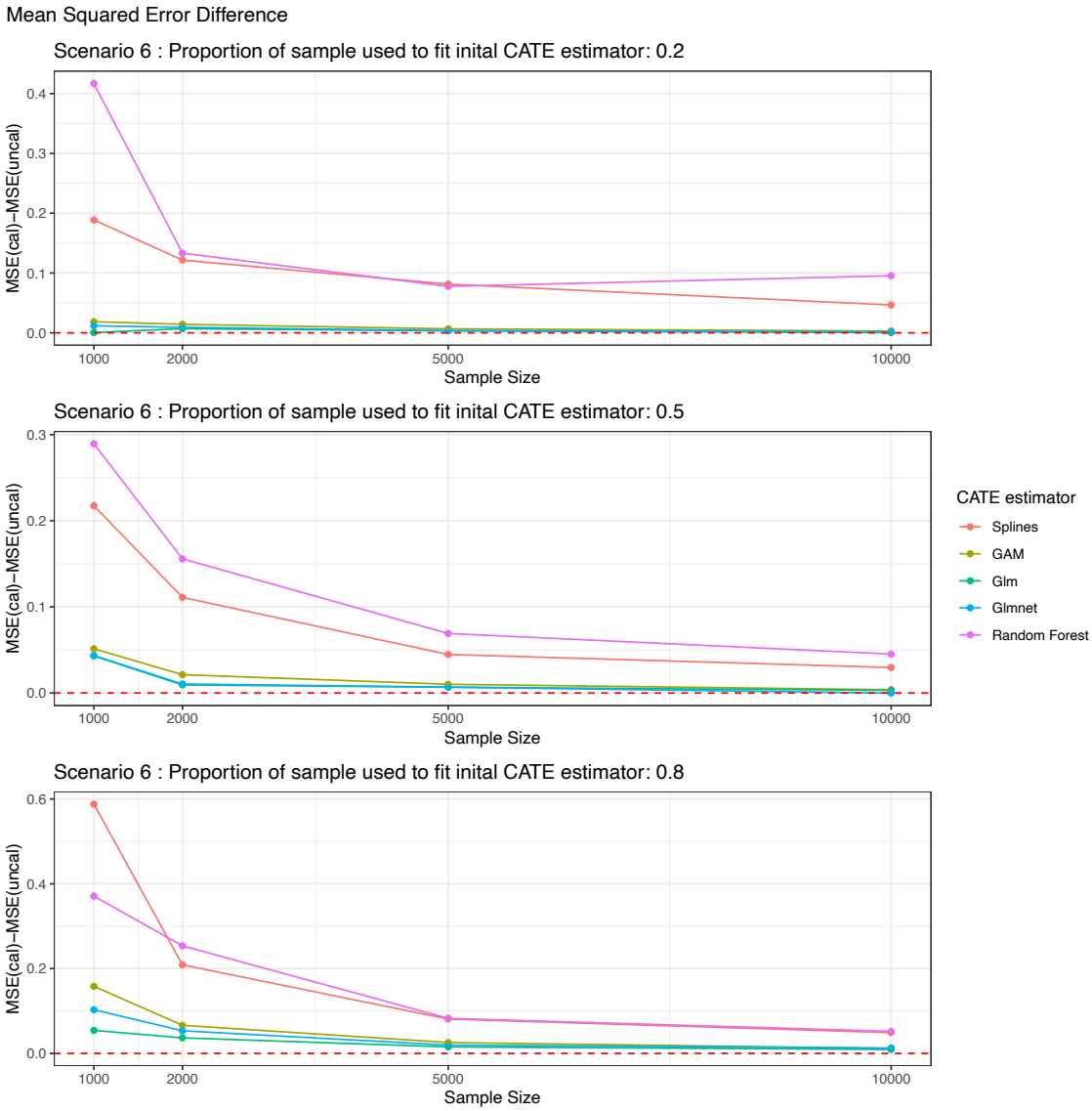

Figure 3.D.20: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 6. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

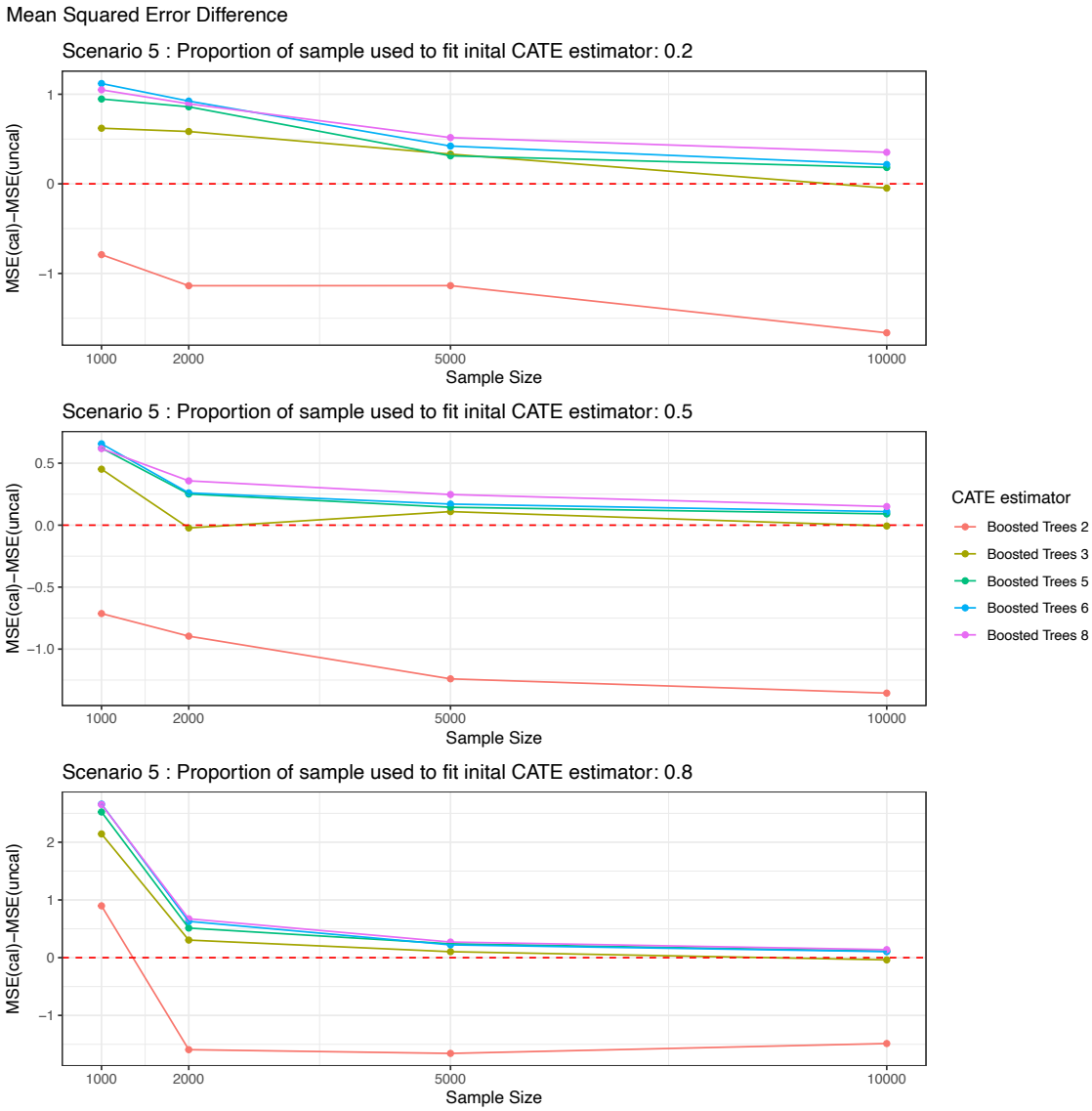

Figure 3.D.21: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 6. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. The numbers in the legend indicate the maximum depth of the gradient boosted trees.

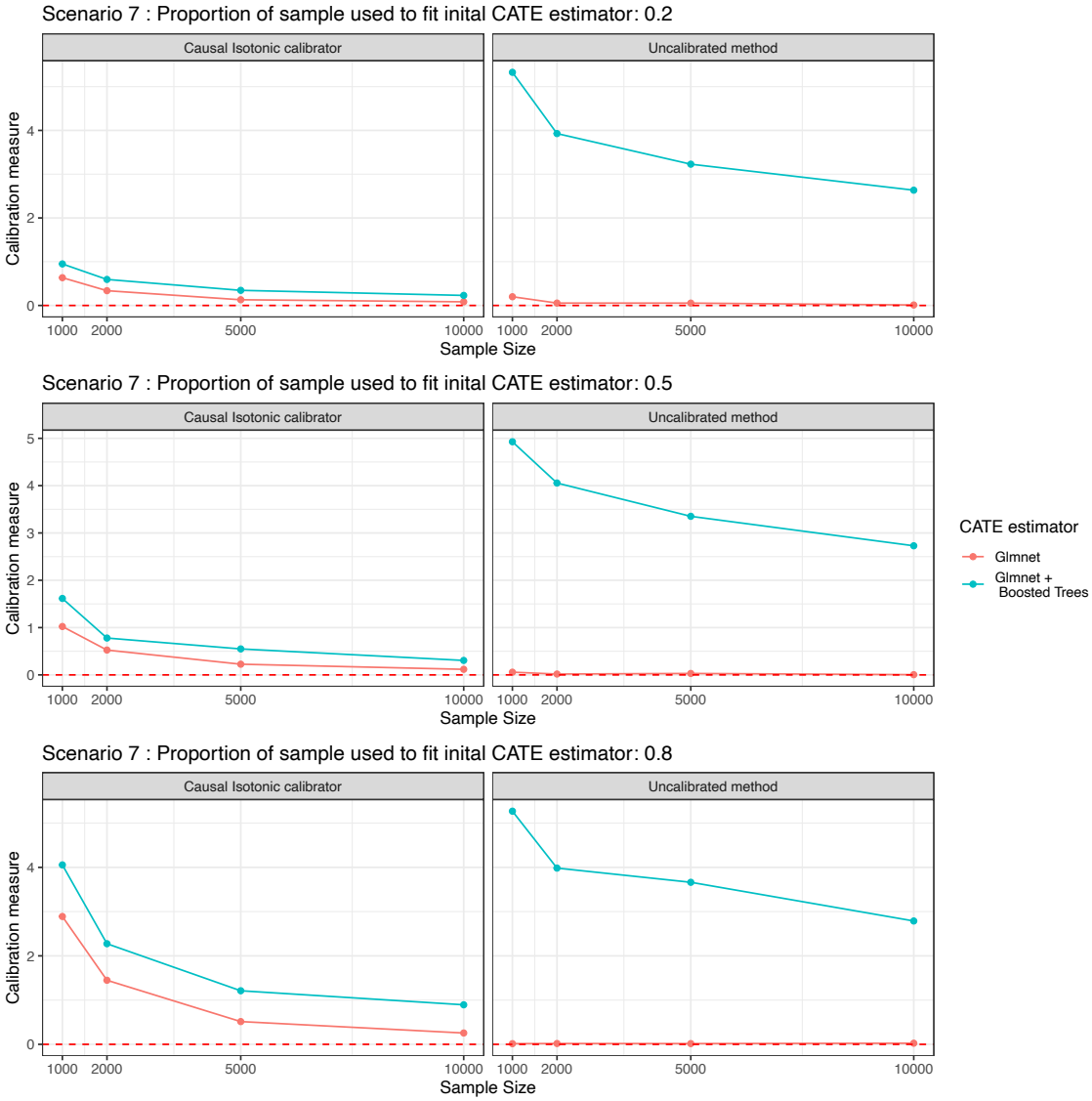

Figure 3.D.22: Calibration measures in scenario 7 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

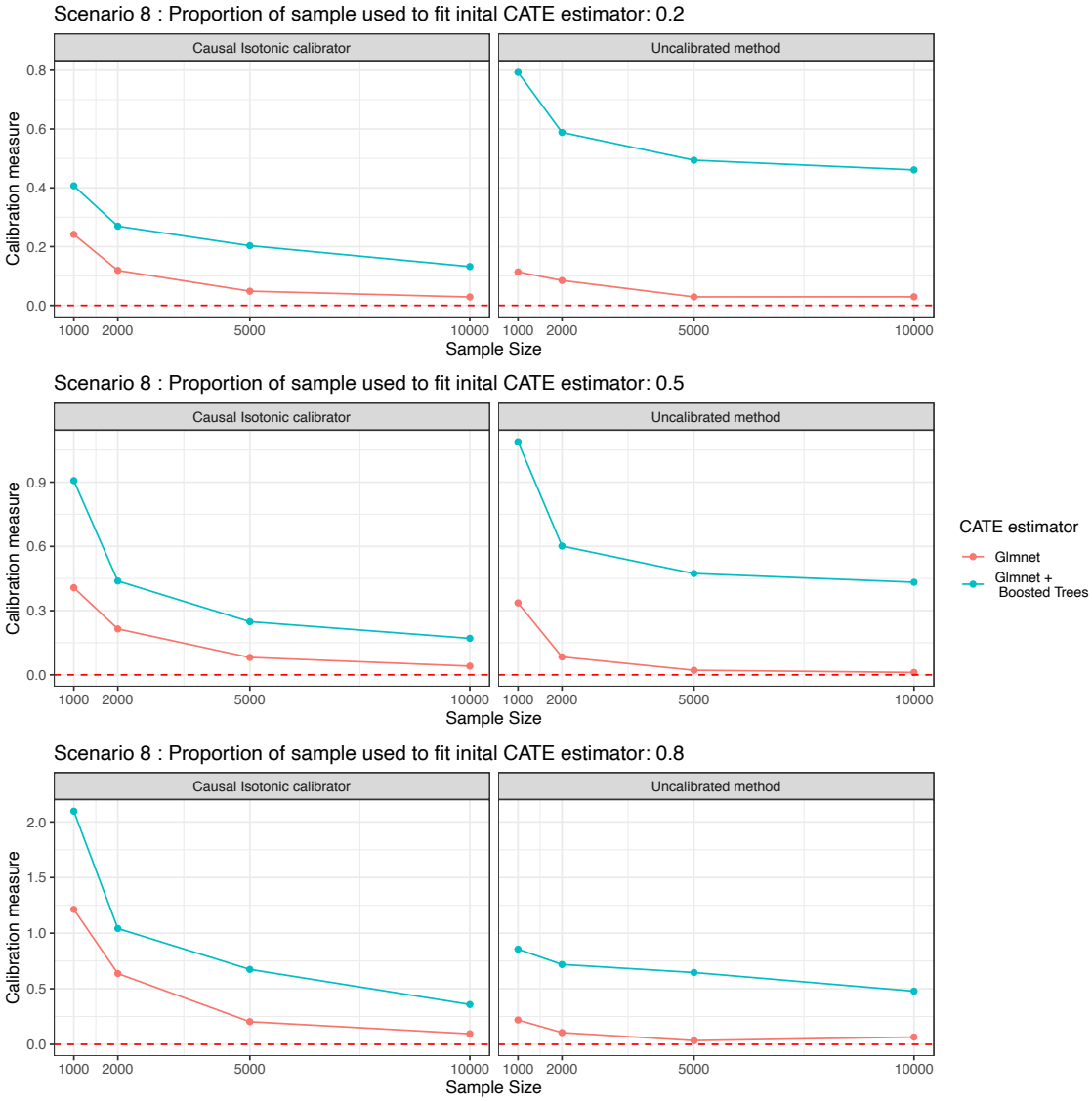

Figure 3.D.23: Calibration measures in scenario 8 across different proportions to divide the data between fitting the CATE estimator and carrying calibration. The panel on the left shows the calibration measure using our method, and the panel on the right shows the same measure using the uncalibrated estimator. Abbreviations: GAM: Generalized Additive Models, Glm: Generalized Linear Model, Glmnet: Generalized linear model with elastic net penalty.

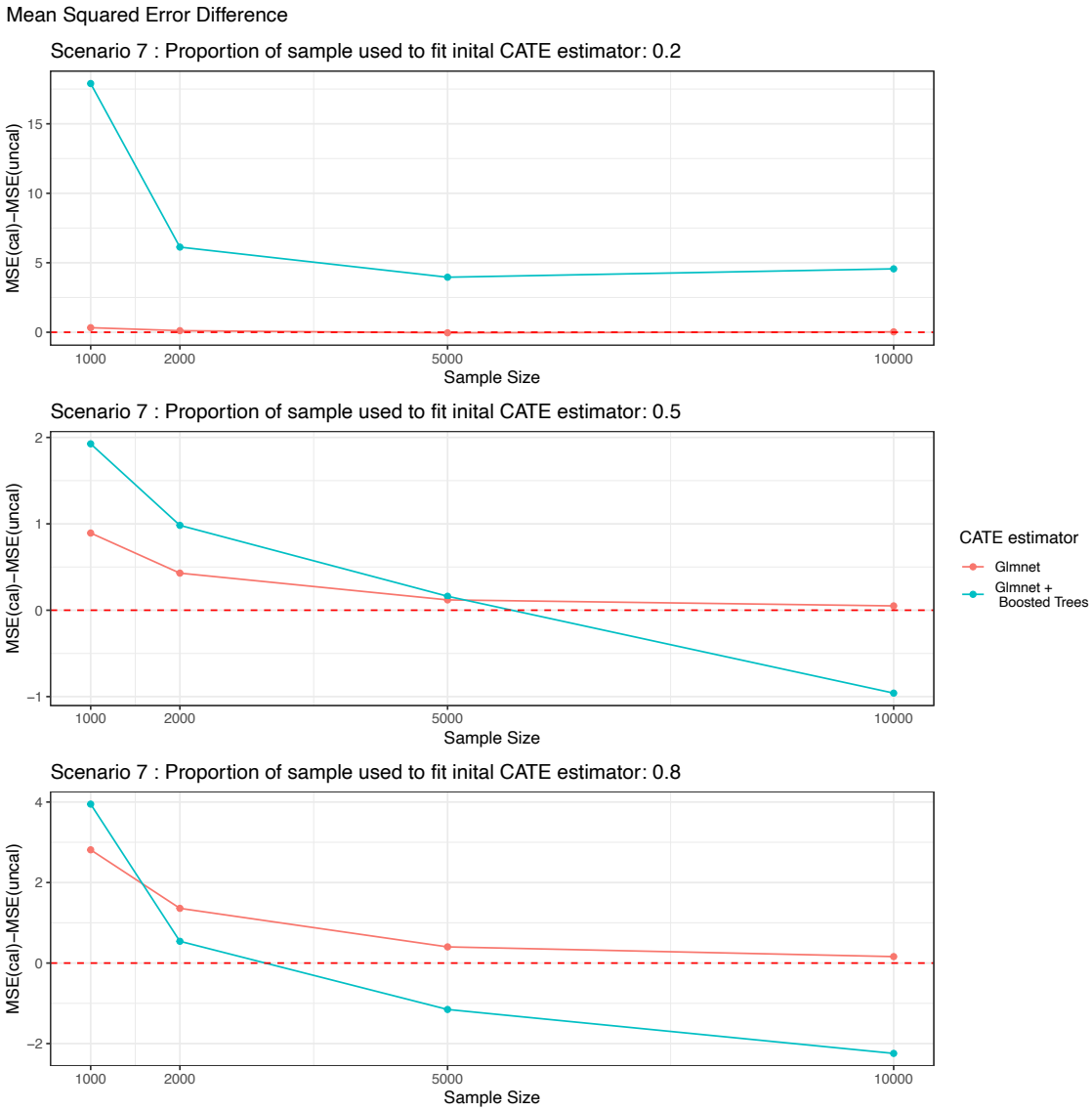

Figure 3.D.24: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 7. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: Glmnet: Generalized linear model with elastic net penalty.

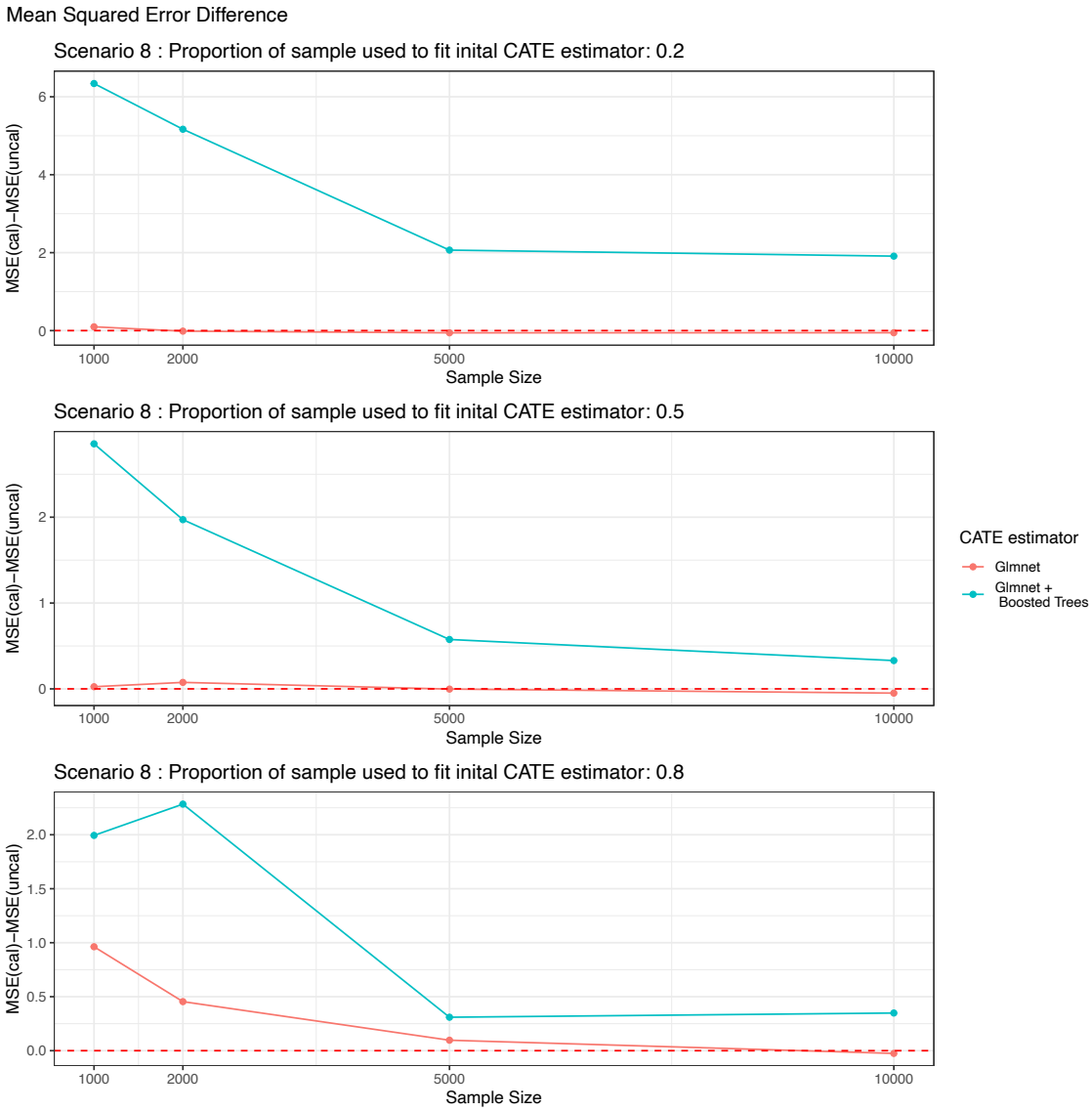

Figure 3.D.25: Difference in mean squared error between the calibrated (denoted by MSE(cal)) and uncalibrated predictors (denoted by MSE(uncal)) in scenario 8. A mean squared error difference above 0 (shown by the dotted red line) indicates a worse performance of the causal isotonic calibrator. Abbreviations: Glmnet: Generalized linear model with elastic net penalty.

Chapter 4

HIV-1 RISK PREDICTION IN KWAZULU-NATAL VIA SUPER LEARNER

4.1 Introduction

The HIV-1 epidemic in South Africa has shown a decreasing trend with an estimated decline in incidence from 3.94 to 2.25 events per 100 person-years between 2012 and 2017 (Vandormael et al. 2019). However, compared to other Sub-Saharan regions, Eastern and Southern Africa remain most heavily affected by HIV-1 (UNAIDS 2021). Specifically, HIV-1 prevalence in South Africa is estimated to be 12.8% (7.1 million seroconversions) (Commission 2017), with variation across provinces and demographic characteristics. In particular, among all provinces, KwaZulu-Natal has the highest prevalence estimated at 18% (Commission 2017). Additionally, women in the country experience a higher risk of HIV-1 than men, with women aged between 15 and 24 years experiencing the highest HIV-1 incidence of any age or sex cohort (Vandormael et al. 2019). Therefore, there is a need to keep expanding and improving evidence-based HIV-1 interventions targeted to key populations in the region.

Interventions to prevent and combat HIV-1 are more effective when leveraging information of individuals at increased risk of infection (Alistar et al. 2014, WHO, Ayton et al. 2020a, Cohen et al. 2016, Baral et al. 2019, Gilkey et al. 2019). For example, offering high-risk individuals different HIV-1 treatment or preventive services has been shown to be more cost-effective than uniform approaches (Alistar et al. 2014, Anderson et al. 2014). Moreover, clinical guidelines for HIV-1 pre-exposure prophylaxis (PrEP) include risk prediction tools to identify appropriate candidates (Gilkey et al. 2019). In research settings, information on HIV-1 risk acquisition can be used as a screening tool to improve recruitment (Balkus et al. 2016). Additionally, when information on the propensity of HIV-1 risk acquisition can be

obtained through a risk prediction model, the score can be used as an additional covariate to: improve the precision of the estimated intervention effect (Colantuoni and Rosenblum 2015, Benkeser et al. 2020, Steingrimsson et al. 2017); reduce bias due to informative follow-up (Benkeser et al. 2019); or better understand treatment effect heterogeneity (Gabriel et al. 2020). Hence, identifying individuals at high risk can be instrumental to the delivery and enhancement of HIV-1 prevention strategies.

Risk score prediction models help identify individuals at high risk by leveraging available information from different sources, such as demographic registries (Balzer et al. 2020), electronic health record data (Krakower et al. 2019), or participant information obtained from HIV-1 prevention trials (Balkus et al. 2016, Kahle et al. 2013, Wand et al. 2018). Leveraging individual-level information, risk prediction methods have been developed for various high-risk groups including: serodiscordant couples (Kahle et al. 2013), men-who-have-sex-with-men (Wahome et al. 2018), and Sub-Saharan women (Balkus et al. 2016, Balzer et al. 2020, Kahle et al. 2013, Wand et al. 2018). Specifically, Balkus et al. (2016) developed a model in South Africa, Uganda and Zimbabwe, and Wand et al. (2018) developed a risk predictor for women in KwaZulu-Natal, South Africa.

While previous risk scores for women in Sub-Saharan Africa have had good predictive performance (Balkus et al. 2018), open questions remain regarding their performance and potential for improvement. For instance, risk predictors from Wand et al. (2018) and Balkus et al. (2016) were fit to predict the risk of seroconversion at one year and have not been assessed across different time windows. Additionally, the benefit of incorporating machine learning tools in HIV-1 risk prediction models has not been fully explored, yet its potential benefit is promising (Balzer et al. 2020). Moreover, there has only been limited study as to whether there are covariates that are particularly predictive of HIV-1 risk. To address these gaps, we combined data from six HIV-1 trials conducted in Kwazulu-Natal to implement and assess the performance of an ensemble risk prediction method known as the super learner (van der Laan et al. 2007). Additionally, we estimated its predictive accuracy across different time windows while accounting for censoring of HIV-1 diagnosis times. Finally, we examined

the impact on the predictive accuracy of the model across different sets of baseline variables.

4.2 Methods

4.2.1 Trial information and primary objectives

We combined information from women who were randomized to the control and treatment arms in the following five phase II/III HIV-1 prevention trials conducted in Kwazulu-Natal: MDP301 (McCormack et al. 2010), HPTN035 (Karim et al. 2011), Carraguard (Skoler-Karpoff et al. 2008), MIRA (Padian et al. 2007), VOICE (Marrazzo et al. 2015), and from women who participated in ASPIRE (Baeten et al. 2016). Enrollment criteria were broadly similar across trials: women older than 18 years old, HIV-1 negative at enrollment, sexually active, not pregnant and with intention to maintain this status, and residing in or around the study area for a minimum of 1 year. In all trials, participants were screened for HIV-1 and sexually transmitted diseases at enrollment.

Study analyses were prespecified before commencement of data analysis (see supplementary Statistical Analysis Plan). Our primary objectives consisted in developing and assessing a risk score prediction model based on the super learner approach. We implemented the super learner to predict risk of HIV-1 infection across varying time windows since trial enrollment. Specifically, we predicted the risk of HIV-1 infection within 6, 12, 18, and 24 months of enrollment, and also the risk of HIV-1 infection within 24 months of enrollment conditional on being HIV-1 seronegative at month 12.

4.2.2 Fitting and assessing the risk model

Super learner is an ensemble learning approach that uses cross-validation to select a weighted combination of a library of candidate prediction methods (van der Laan et al. 2007). The candidate prediction methods considered may be parametric, such as an accelerated failure time model (Wei 1992), semiparametric, such as a Cox model (Cox 1972), or more flexible, such as a survival forest (Ishwaran et al. 2008). Table S3 in the Supplement lists the library

of candidate prediction methods that we used. The convex combination selected by super learner has been shown to perform about as well or better than the best candidate prediction method in the library using theoretical arguments (van der Laan et al. 2007), simulation studies (Polley et al. 2011), and cross-validated performance on real datasets (Balzer et al. 2020, Petersen et al. 2015).

To avoid overly optimistic assessments of risk score performance, we assessed the super learner via internal validation (cross-validation) as well as external validation. These assessments involved evaluating the performance of super learner risk scores on data that were neither used to fit the candidate algorithms included in the super learner nor to learn an optimal weighted combination of these algorithms. We used the VOICE trial data for internal validation and the ASPIRE trial data for external validation. More specifically, for our cross-validated assessments of performance we divided the VOICE trial into 10 folds, and a total of ten different super learner risk scores were fit, with each fit corresponding to a different choice of validation fold. For each fit, data from the 9 non-validation VOICE folds, along with the data from all trials except ASPIRE, were used as data to fit the super learner risk score. Predicted survival probabilities were then computed for each VOICE trial participant in the validation fold. These predictions were then compared to these participants' actual HIV-1 infection outcomes to estimate the cross-validated area under the curve (AUC) performance metric. The computed AUC metric accounted for the fact that some participants had insufficient follow-up to assess HIV-1 infection status through certain times (see Supplementary Appendix 4.A.3 for details). The 10 fold-specific performance estimates were then averaged to compute the final cross-validated AUC measure, and corresponding standard errors were obtained. For external validation, we combined VOICE with the training data to fit the super learner once and to predict the survival probabilities of interest for each ASPIRE trial participant. These predictions were then compared to the actual ASPIRE HIV-1 infection outcomes to compute the estimate the AUC of the super learner risk score. The main rationale behind choosing VOICE and ASPIRE for internal and external validation, respectively, was that these trials were conducted more recently compared to the

others. Section 4.A.2 in the supplementary materials contains detailed information on how we fit the super learner risk score and assessed its performance.

Finally, we carried two post-hoc analyses with the goals of assessing the predictive performance outside the AUC measure, and to obtain more accurate comparisons with previous HIV-1 risk prediction models. First, we computed the sensitivity and specificity values for different cutoffs of the estimated risks of HIV-1 infection within 12 and 24 months of enrollment. The cutoffs were fixed in order to attain specific sensitivity values based on the validation data. The resulting cutoff values were then used to compute the estimated sensitivity and specificity on the test data. Second, to have a more accurate comparison of our approach to previously reported risk scores, we assessed the performance of a forward step-wise Cox regression model. We chose this model because it is similar or has been used in previous studies whose scope is similar to ours (Balkus et al. 2016, Balzer et al. 2020, Wand et al. 2018).

4.2.3 Variable selection

Variable collection was not uniform across trials. To avoid the need for imputation methods to account for systematic missingness when validating our learned risk scores, we predicted HIV-1 risk based on the covariates that were measured in both our internal and external validation trials, namely VOICE and ASPIRE. This resulted in a total of 14 variables being included as predictors. These 14 variables were: age, education level, married, married or cohabiting, parity, number of set acts in the last 7 days, number of sex partners in the last 3 months, condom in last sex encounter, family planning method, partner's circumcision status, syphilis, gonorrhea, chlamydia status, and any sexually transmitted infection. Variables excluded from the super learner were: cohabiting, partner's age, partner has another partner, number of sex acts in the last 14 days, and trichomonas. Missing predictors were imputed with the `missForest R` package (Stekhoven and Bhlmann 2011), which relies on fitting random forests on the observed data to predict the missing values. As alluded to earlier, some variables that were measured in VOICE and ASPIRE were not measured in some of

the other trials. Therefore, these covariates had to be imputed for all individuals in these other trials, not just those with happenstance missingness. Note that because the variables used to validate the super learner are available in both the external and internal validation sets, our assessment of the performance of the super learner is not contingent on our choice of imputation method. Table S1 in the Supplement contains further information on the 14 variables used to develop the risk scores and indicates which of those variables were not measured in some of the trials.

To complement our primary objectives, we examined the impact of removing baseline variables when fitting the super learner. The variables we chose to remove were motivated by variables that were unmeasured in three recently completed HIV-1 prevention trials. Specifically, we restricted our variable set to the subset of those from our primary analyses that were also collected in several recent trials, namely (i) HVTN 702 (Gray et al. 2021) and HVTN 705 (Laher et al. 2020), and (ii) HVTN 703 (Corey et al. 2021). For instance, HVTN 703 did not measure contraceptive method, so we assessed the impact of excluding that variable when fitting the risk prediction model. Because HVTN 702 and HVTN 705 collected the same information that was obtainable in our trial data, we combined the two variable sets. Hereafter, we refer to these two variable sets as HVTN 702/705 and HVTN 703. To match the variables collected in HVTN 702/705, we excluded family planning method. To match the variables collected in HVTN 703, we further excluded married, married or cohabiting, parity, number of sex acts in the last 7 days, and partner's circumcision status. Finally, to obtain information on the marginal importance of the available baseline variables, we compared the predictive accuracy of a univariate Cox model adjusting for age versus bivariate Cox models that included age plus an additional covariate.

4.3 Results

4.3.1 Population

Baseline characteristics of the trial participants in the training (Carraguard, HPTN035, MDP302, and MIRA), internal (VOICE) and external validation (ASPIRE) are shown in Table 4.1. Across the three different datasets, most women were between 20 and 29 years old (>50%), reported having received no education or primary education at most (>70%), and were not married nor cohabiting (>46%). Additionally, the majority of the participants reported having more than one sex act in the prior 7 days from enrollment (between 69% to 87%). Information on sex partners was missing in a substantial part of the training data. In the internal and external validation data, the majority of participants (>80%) self-reported one sex partner in the last 3 months.

The training data consisted of 6,286 trial participants, from which 517 (8%) seroconversions occurred during 7,974 person-years of follow-up. The internal validation set from the VOICE trial consisted of 2,726 participants, where 241 (9%) seroconversions were observed in 2,973 person-years of follow up. In the ASPIRE external validation set, we observed 797 individuals where 87 (11%) seroconversions were measured during 1,432 person-years of follow-up.

Participant follow-up was heterogeneous across trials. The average follow-up was 1.27, 1.09, and 1.79 person-years on the training, internal validation, and external validation, respectively. The trials in the training data had a follow-up that ranged from 23% to 92% (0% to 41%) of participants with complete follow-up information at 12 months (24 months). VOICE and ASPIRE trial data had 58% and 77% (2% and 59%) of their observations with complete follow-up at 12 months (24 months), respectively. Table S2 in the Supplement provides further information on participant follow-up through different time points.

4.3.2 Predictive performance of the super learner risk score

The estimated internal validation AUC when predicting HIV-1 infection at 24 months was 0.57 (95% CI: 0.51-0.64), and the external validation AUC at 24 months was 0.67 (0.61-0.73). Restricting to variables available in HVTN 702/705 yielded similar estimated AUCs, and the super learner based on variables collected in the HVTN 703 trial had a slightly higher internal validation AUC of 0.60 (0.55-0.65) and a similar external validation AUC of 0.66 (0.59-0.71). Figure 4.1 shows the estimated AUCs across the different variable sets.

When predicting HIV-1 infection between 0 to 6, 12, and 18 months, the estimated internal validation AUC ranged between 0.62 and 0.66 depending on the time point, and the estimated external validation AUC remained close to 0.67 (Figure 4.2). The internal validation AUC was higher when predicting risk prior to 6, 12, or 18 months than when predicting risk to 24 months. Similar predictive performance was obtained when using all variables, HVTN 703 variables, or HVTN 702/705 variables (see Figure S1 in the Supplement).

The AUC of the super learner when conditioning on seronegative status at 12 months to predict HIV-1 infection at 24 months was 0.50 (0.41-0.64) in the internal validation set, suggesting predictive performance no better than that obtainable by random chance. Performance improved noticeably in the external validation set, with an estimated AUC of 0.66 (0.60-0.72). Using variables exclusively from the HVTN 703 trial increased the AUC to 0.58 (0.49-0.68) and 0.64 (0.54 -0.73) in the internal and external validation set, respectively. Similarly, the risk score based on variables from the HVTN 702/705 trial yielded an estimated AUC of 0.59 (0.51-0.67) and 0.65 (0.53-0.75).

Sensitivity and specificity values at different cutoffs are reported on table S4 in the Supplement. Cutoffs selected to attain high sensitivity values chosen from the internal validation tend to yield a higher sensitivity in the test set. For instance, the cutoff obtained to achieve 80% sensitivity in VOICE resulted in 90% sensitivity in the ASPIRE data. The opposite occurred with low sensitivity values. Overall, high cut-points produced high sensitivity but resulted in low specificity in both the VOICE and ASPIRE data. For example, in the valida-

tion data, predicting HIV-1 before 24 months with 80% sensitivity resulted in 33% specificity. It is worth noting that to estimate the sensitivity and specificity we used a similar approach to the one used to estimate the AUC. Hence, our sensitivity and specificity estimates account for the right censoring of the data.

In our last exploratory analysis, we estimated that the AUC in the univariate model with age was 0.61 (0.54-0.70) in the internal validation set, and 0.57 (0.52-0.63) in the external validation set. The variable married/cohabiting had the largest increase when included jointly with age with an internal validation AUC of 0.63 (0.57-0.72) and an external validation AUC of 0.58 (0.52-0.63).

To better understand which candidate prediction methods contributed most to the final risk score, we inspected the weighted combination used by super learner. Across all our objectives, we observed that the Cox proportional hazards and elastic net methods tended to have higher weights both in internal validation and external validation (Supplementary Figures S2 and S3). Of these methods, the most weight tended to be given to elastic net with main terms and Cox proportional hazards with stepwise forward regression. Of the more flexible approaches, namely survival forest and the generalized additive model, the survival forest tended to be given more weight. Moreover, averaging across folds, Cox proportional hazards forward stepwise model and glmnet with main terms had the highest weights. We observed similar results when fitting the super learner on the training data and validation data combined to predict on the test data (Supplementary Figure S3).

4.4 Discussion

In this work, we fitted and assessed the super learner based on information from women who enrolled in six HIV-1 prevention trials carried out in KwaZulu-Natal, South Africa. We obtained estimates of the AUC across different risk time windows, while accounting for right censoring. Additionally, we examined the impact of removing specific variables, which may help elucidate their utility to predict risk of seroconversion.

Our findings showed modest predictive performance of the super learner, with estimated

AUCs ranging between 0.57 to 0.68 depending on the time window and variable set. Overall, the predictive accuracy as well as estimated sentivity and specificity of the super learner is comparable to previous risk predictors reported in the literature (Balkus et al. 2016, Balzer et al. 2020, Wand et al. 2018). While the performance of the Cox model was similar, super learner has the added value that it can have good predictive performance even in cases where the true data generating mechanism is not well approximated by a Cox model, but is well approximated by another model in the library. The modest AUC in previous HIV risk scores and that of the super learner suggests a fundamental limit on risk scores based on baseline covariates used so far.

The internal validation AUC based on the VOICE trial was consistently lower than the external validation AUC obtained from the ASPIRE trial data (Figure 4.2). This difference could be partly explained by the fact that we used slightly more information to fit the super learner on ASPIRE. Across different time points, we found little variation in the AUC, however there was a decrease in the validation AUC at 24 months. Several factors could be attributed to the observed difference in the validation and test AUCs at 24 months. First, there is a lower amount of participants with follow-up information (2% at 24 months) in the VOICE trial (see Table S2) compared to the ASPIRE trial (59% at 24 months). Second, VOICE participants have different demographic and sex behavioral variables compared to women in ASPIRE and the training data. Specifically, on Table S1, we see that participants in the VOICE trial had different frequencies in married or lives with partner, condom use in the last sex encounter, and parity. To further investigate this difference, we carried a principal component analysis (PCA) with the variables that were measured across all trials. Based on the first two components of the PCA analysis, we did not find any substantial separation across participants in the three data sets, but rather that participants in ASPIRE set are located in a sub-region of the training and validation data (Figure S4 in the Supplement). Hence, while our results suggest that participants in VOICE could have different behaviors associated to HIV risk acquisition at 2 years, our PCA analysis was not able to uncover any substantial difference in their baseline characteristics compared to the population from

ASPIRE.

Results of the super learner across the different variable sets were insufficient to detect meaningful differences on its performance. While fitting the super learner with different variables changed the point estimates of the AUC, the confidence intervals overlapped (Figure S1 in the Supplement). In terms of point estimates, we found that the HVTN 703 variable set performed slightly better compared to using all the variables. Bivariate analyses showed that information on cohabiting yielded the largest increase in AUC when adjusting for age. This result is comparable to the finding in Wand et al. (2018) that single or cohabiting information was the most important factor for HIV-1 seroconversion. Thus, we believe that researchers should continue to incorporate cohabiting information when developing risk prediction models, as it is relatively easy to collect and less sensitive to subject reported biases as other risk behavior variables. Additionally, future work to examine variable importance using inferential techniques can provide more evidence on this matter, such as that proposed in Williamson et al. (2020).

One limitation of the risk score we developed is its generalizability to populations other than the one used to fit the risk score. While we fitted the super learner with information from 6 trials, its performance is questionable outside KwaZulu-Natal, which is the province in South Africa that gave rise to all of the data used in our analyses.

Moreover, because the number of adolescents in VOICE and ASPIRE is low (9% and 8%, respectively – Table 4.1), it is also unclear whether the learned risk score will perform well in these populations. To better understand its applicability and limitations outside trial populations, future work — along the lines of that done in Ayton et al. (2020b)— should be carried out to evaluate and develop risk scores for broader high-risk populations, such as adolescent women between 15 and 19 years (Ayton et al. 2020a). Finally, incorporating variables that were not available in our datasets such as socioeconomic information or partner's HIV-1 status could also help improve the performance of our risk score.

Although improvements in relation to those detailed above are possible, our risk score has the benefit of having already been fit on a large database of existing trial data and validated

on data from two separate trials which were carried out in a region with high HIV-1 risk, where efforts for prevention work are of high importance. Hence, this risk score can be readily applied to clinical research to support trial recruitment or target HIV-1 prevention strategies.

Tables and Figures

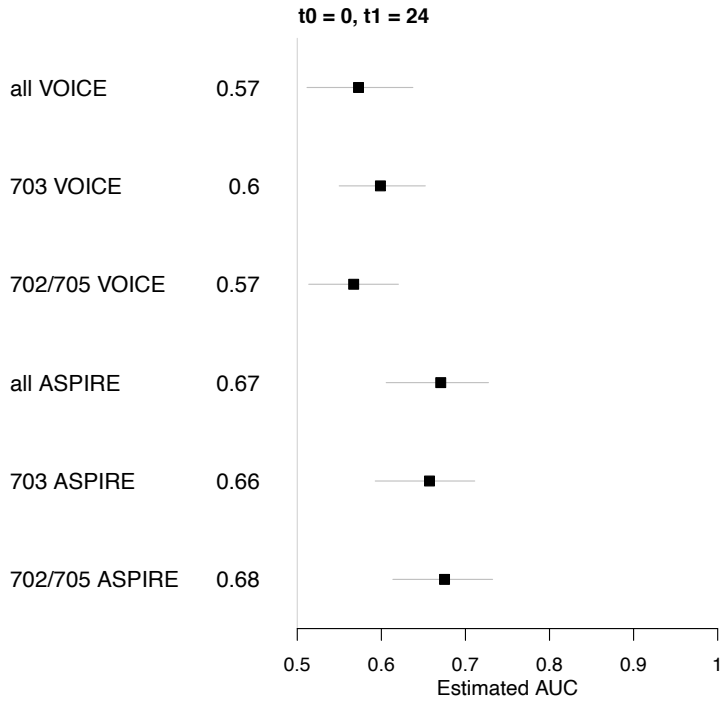

Figure 4.1: Estimated AUC in the internal (VOICE) and external (ASPIRE) validation data for $t_0 = 0$, $t_1 = 24$ months and different sets of variables that were used to train the super learner. For each of the internal and external validation data, the 95% confidence intervals for the AUC across the variable sets overlap.

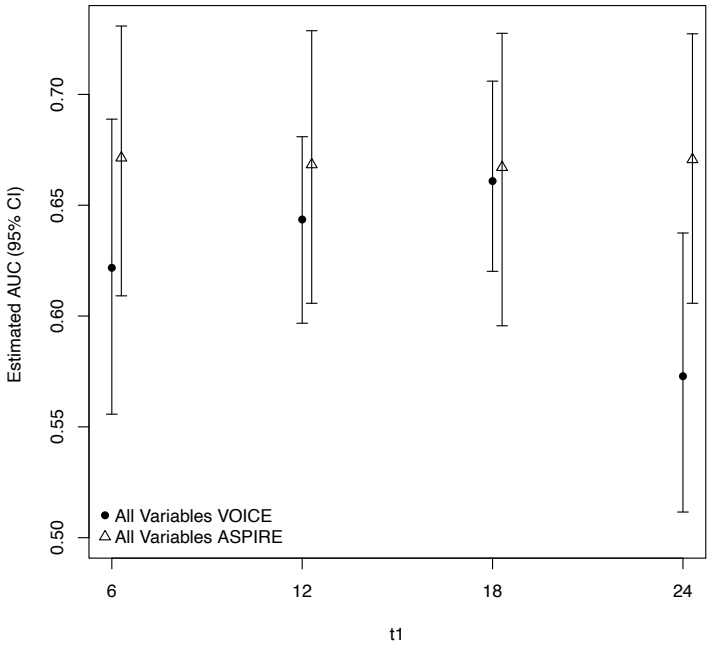

Figure 4.2: Internal (VOICE) and external (ASPIRE) validation AUC across different time points. The internal validation AUC ranges between 0.62 and 0.66, declining to 0.57 at 24 months. The external validation AUC has less variability across time, and was consistently higher than the internal validation set, with values close to 0.67.

Variable measured at baseline	Category	Training Set N = 6,286	Internal Validation N = 2,726	External Validation N = 797
Age (years)	<20	688 (11%)	240 (9%)	63 (8%)
	20-29	3,084 (49%)	1,940 (71%)	544 (68%)
	30-39	1,636 (26%)	534 (20%)	146 (18%)
	40+	878 (14%)	12 (<1%)	44 (6%)
Education level	None	2,780 (44%)	1,287 (47%)	390 (49%)
	Primary	1,863 (30%)	1,247 (46%)	373 (47%)
	Secondary or higher	149 (2%)	191 (7%)	34 (4%)
	(Missing)	1,494 (24%)	1 (<1%)	0 (0%)
Married or lives with partner	No	2,867 (46%)	2,167 (79%)	741 (93%)
	Yes	1,105 (18%)	559 (21%)	55 (7%)
	(Missing)	2,314 (37%)	0 (0%)	1 (<1%)
Parity	None	435 (7%)	396 (15%)	8 (1%)
	1 Children	1,544 (25%)	1,344 (49%)	373 (47%)
	2+ Children	2,004 (32%)	986 (36%)	321 (40%)
	(Missing)	2,303 (37%)	0 (0%)	95 (12%)
Number of vaginal sex encounters in the past 7 days	0	816 (13%)	576 (21%)	244 (31%)
	1	1,530 (24%)	447 (16%)	227 (28%)
	2	1,443 (23%)	584 (21%)	150 (19%)
	3+	2,132 (34%)	876 (32%)	165 (21%)
	(Missing)	365 (6%)	243 (9%)	11 (1%)
Number of sex partners in the last 3 months	0	3 (<1%)	0 (0%)	0 (0%)
	1	2,352 (37%)	2,155 (79%)	660 (83%)
	2+	135 (2%)	564 (21%)	137 (17%)
	(Missing)	3,828 (61%)	7 (<1%)	0 (0%)
Used condom in last sex encounter	Yes	3,594 (57%)	2,202 (81%)	531 (67%)
	No	2,682 (43%)	296 (11%)	266 (33%)
	(Missing)	10 (<1%)	228 (8%)	0 (0%)
Family planning method	None	1,312 (21%)	0 (0%)	0 (0%)
	Male/Female condom	1,456 (23%)	0 (0%)	0 (0%)
	IUD or Sterilization	608 (10%)	115 (4%)	39 (5%)
	Oral Contraceptive	381 (6%)	511 (19%)	155 (19%)
	Injectables	2,529 (40%)	2,100 (77%)	601 (75%)
	(Missing)	0 (0%)	0 (0%)	2 (<1%)
Primary partner circumcised	Yes	345 (5%)	730 (27%)	304 (38%)
	No	746 (12%)	1,617 (59%)	430 (54%)
	Don't know	358 (6%)	375 (14%)	61 (8%)
	(Missing)	4,837 (77%)	4 (<1%)	2 (<1%)
Chlamydia	Positive	552 (9%)	407 (15%)	129 (16%)
	Negative	5,695 (91%)	2,316 (85%)	668 (84%)
	(Missing)	39 (1%)	3 (<1%)	0 (0%)
Syphillis	Positive	148 (2%)	34 (1%)	5 (1%)
	Negative	4,542 (72%)	2,692 (99%)	792 (99%)
	(Missing)	1,596 (25%)	0 (0%)	0 (0%)
Gonnorhea	Positive	138 (2%)	89 (3%)	32 (4%)
	Negative	6,109 (97%)	2,635 (97%)	765 (96%)
	(Missing)	39 (1%)	2 (<1%)	0 (0%)
Any curable STI †	Yes	1,001 (16%)	478 (18%)	189 (24%)
	No	5,271 (84%)	2,248 (82%)	608 (76%)
	(Missing)	14 (<1%)	0 (0%)	0 (0%)

Table 4.1: Demographic information in the training, internal and external validation sets. † Diagnosed with at least one STI of: chlamydia, gonorrhoea or syphilis

APPENDIX

4.A Further Methodological Details

4.A.1 Notation and Assumptions

Throughout we use T to denote the time of HIV-1 diagnosis relative to enrollment, C to denote the earlier of loss to follow-up or trial termination, and W to denote measured baseline covariates. The risk scores studied in this work aim to estimate functions of the form $\psi_0(W) := P_0(T \leq t_1 \mid T > t_0, W)$, where t_0 and t_1 are fixed times. For instance, when estimating HIV-1 risk at $t_1 = 24$ months if seronegative at $t_0 = 12$ months, our target of inference is $\psi_0(W) = P_0(T \leq 24 \mid T > 12, W)$.

In practice, for each participant we observe $Z := (L, W, Y, \Delta)$ with support $\mathcal{Z}$, where $Y := \min\{T, C\}$, $\Delta = I\{Y = T\}$, and L denotes trial and site indicators that are not included in the baseline covariates W. We assume that we observe n i.i.d. observations $Z_1, \ldots, Z_n$ drawn from P_0. Further, in order to identify $\psi_0(W)$ via the distribution of the observed data, we require that the censoring time is conditionally independent of the HIV-1 diagnosis time given baseline covariates, that is,

$$C \perp\!\!\!\perp T \mid W, \tag{4.1}$$

and also that there is a nonzero probability of being censored through the end of the observation period, that is,

$$P\{C \geq t_1 \mid W\} > 0 \text{ with probability one over } W \sim P_0. \tag{4.2}$$

4.A.2 Super Learner Risk Score Model

Using the notation introduced in Supplementary Appendix 4.A.1, we now provide a more detailed description of the super learner algorithm introduced in the main text. Super learner

estimates the conditional probability of HIV-1 infection through a convex combination of each candidate models. Letting $\hat{\psi}_1, ..., \hat{\psi}_K$ be the fitted risks from each of the K models, the risk predicted by the super learner can be expressed as $\sum_{k=1}^{K} \alpha_k \hat{\psi}_k(w)$ where $\{\alpha_1, ...\alpha_k\}$ are obtained to minimize a given loss function. For this study, we minimized the risk of the cross-entropy loss function. We decided to use the cross-entropy loss because HIV-1 infection is a rare outcome, and the minimum squared error loss may not perform well. Given t_0 and t_1, the cross-entropy loss function for participant i is defined as

$$(W_i, T_i) \mapsto -I\{T_i \in (t_0, t_1]\} \log \left[\sum_{k=1}^{K} \alpha_k \hat{\psi}_{k,s(i)}(W_i) \right] - I\{T_i > t_1\} \log \left[1 - \sum_{k=1}^{K} \alpha_k \hat{\psi}_{k,s(i)}(W_i) \right].$$

Because the survival outcomes are not observed in all trial participants, we did not use the above loss directly. However, the risk function corresponding to the above loss is estimable using existing methods. In particular, one can estimate the risk function given by

$$E\left[I\{T \in (t_0, t_1]\} \{ -\log \left(\sum_{k=1}^{K} \alpha_k \hat{\psi}_k(W) \right) \} \right] + E\left[I\{T > t_1\} \{ -\log \left(1 - \sum_{k=1}^{K} \alpha_k \hat{\psi}_k(W) \right) \} \right] \tag{4.3}$$

$$= E[I\{T \in (t_0, t_1]\} \operatorname{wgt}^{(1)}(W)] + E[I\{T > t_1\} \operatorname{wgt}^{(2)}(W)],$$

where $\operatorname{wgt}^{(1)}(w) : \mathcal{W} \mapsto -\log \left(\sum_{k=1}^{K} \alpha_k \hat{\psi}_k(w) \right)$, and $\operatorname{wgt}^{(2)}(w) : \mathcal{W} \mapsto -\log \left(1 - \sum_{k=1}^{K} \alpha_k \hat{\psi}_k(w) \right)$. The above risk function represents a sum of two weighted survival probabilities, where the weights depend on covariates W. These weighted survival probabilities are estimable using the observed data as follows.

First, note the expression in Equation 4.3 can be re-expressed as

$$E\left[I\{T \leq t_1\} \operatorname{wgt}^{(1)}(W) \right] - E\left[I\{T \leq t_0\} \operatorname{wgt}^{(1)}(W) \right] + E\left[I\{T > t_1\} \operatorname{wgt}^{(2)}(W) \right]$$

Estimation for each quantity in the above sum is very similar, so we outline our approach for the first term. Before proceeding it is also worth noting that, because of the cross-validation approach, we can condition on the training set used to obtain the k learners $\{\hat{\psi}_j\}_{j=1}^{k}$ and treat them as fixed functions. To ease notation, we will condition on these data sets implicitly.

We can express the parameter of interest as a mapping of a given distribution P:

$$\tilde{\Psi} : P \mapsto E_P \left[E_P[I\{T \leq t_1\}|W] \, \text{wgt}^{(1)}(W) \right] = E_P \left[f_P(w) \, \text{wgt}^{(1)}(W) \right]$$

where, $f_P(w) := E_P[I\{T \leq t_1\}|W = w]$. Our goal is to estimate $\tilde{\Psi}(P_0)$. We do so in two steps. First, we apply the `survtmle` `R` package to carry Targeted Maximum Likelihood Estimation Van der Laan et al. (2011), Benkeser et al. (2018) of the cumulative incidence $\Psi(P_0)$, where

$$\Psi : P \mapsto E_P \left[E_P[I\{T \leq t_1\}|W\}] \right] = E_P[f_P(W)].$$

Then, we alter the resulting estimated influence function to obtain our final estimate of $\tilde{\Psi}$. It is worth noting that because of the assumptions outlined in Section 4.A.1, both parameters are identifiable using the observed data (Benkeser et al. 2018, Luedtke et al. 2017).

Letting $D_P : \mathcal{Z} \to \mathbb{R}$ be equal to the efficient influence function of Ψ at P relative to a nonparametric model, it can be shown that the efficient influence function of $\tilde{\Psi}$ at P relative to this same model is equal to $\tilde{D}_P(z) : z \mapsto \text{wgt}^{(1)}(w)[D(z) + \Psi(P)] - \tilde{\Psi}(P)$. Via this relationship, we can obtain the estimated marginal probability of $\Psi(P_0)$ (denoted by $\hat{\Psi}(P_n^*)$), and the evaluation of the estimated efficient influence function at each observation $\{\hat{D}_{P_n^*}(Z_i)\}_{i=1}^n$, where P_n^* is estimate of P_0 returned by TMLE. We use a one-step estimator of $\tilde{\Psi}(P_0)$, which is given by

$$\hat{\tilde{\psi}} := \frac{1}{n} \sum_{i=1}^{n} \text{wgt}^{(1)}(W_i)[\hat{D}_{P_n^*}(Z_i) + \hat{\Psi}(P_n^*)].$$

With the estimated survival probabilities and the fit from the candidate learners, we obtained the optimal weights that minimize the empirical risk function via the `nloptr` `R` package (Ypma et al. 2020). Each time we fit the super learner, we carried 10-fold cross-validation to obtain the optimal α weights for each model in the super learner — a detailed description of this process can be found in Algorithm 3. After fitting the super learner, we obtained the fitted survival probability for each participant in the hold-out dataset. Performing 10-fold cross-validation allowed us to obtain independent predictions for each participant in the VOICE dataset without overfitting. Similarly, for external validation, we

combined the VOICE trial data with the training data to obtain the optimal α weights via 10-fold cross-validation. Afterwards, we predicted the survival probabilities for each ASPIRE trial participant. Finally, it is worth noting that the estimated marginal probabilities of survival to obtain the α weights were obtained separately from each training fold, to avoid over-fitting. Similarly, variable imputation was also carried out within each training fold. In the HVTN 702/705 and HVTN 703 variable analyses, we restricted the imputation procedure to use the same variables to fit the super learner.

4.A.3 Area Under the Curve

We used the AUC to assess the performance of the super learner. For given t_0 and t_1 and independent observations Z_1 and Z_2, the AUC is defined as

$$\begin{aligned} & E\left[I\{\hat{\psi}(W_1) > \hat{\psi}(W_2)\} + 0.5I\{\hat{\psi}(W_1) = \hat{\psi}(W_2)\} \middle| T_1 \in (t_0, t_1], T_2 > t_1\right] \\ & = \frac{E\left[I\{T_1 \in (t_0, t_1]\} I\{T_2 > t_1\}\left(I\{\hat{\psi}(W_1) > \hat{\psi}(W_2)\} + 0.5I\{\hat{\psi}(W_1) = \hat{\psi}(W_2)\}\right)\right]}{P(t_0 \leq T_1 \leq t_1) P(T_2 > t_1)}. \end{aligned} \quad (4.4)$$

We estimated the AUC in a similar fashion as the risk function, that is, by plugging the estimated quantities into the empirical mean based on Equation 4.4. To estimate the numerator, we used the `survtmle R` package to obtain the estimated the weighted probabilities. The denominator consists of marginal probabilities which were also obtained using TMLE via the same package the `survtmle R` package. Finally, we derived the 95% confidence intervals (CI) for the AUC using 1,000 bootstrapped samples. Specifically, we re-sampled the predictions from the validation and test sets and computed the estimated AUC. Algorithms 4 and 5 outline the steps we carried out to fit the super learner and obtain the estimated AUC in the internal and external validation sets.

4.A.4 Algorithms

Let T = Training Data, V = VOICE Data, $E = T \cup V$, and A = ASPIRE data. Additionally, for a given observation Z_i, let $\hat{D}_i^{(1)}$ and $D_i^{(2)}$ be equal to the sum of the point estimate

of the marginal survival probability, respectively either $P(T \in (t_0, t_1])$ or $P(T > t_1)$, plus the corresponding individual i-specific efficient influence function estimate output by `survtmle`.

Input: Input dataset B, K algorithms to fit the super learner

1: Use B to impute B via `missForest` Package
2: With B use `survtmle` to obtain $\hat{D}_i^{(1)}$ and $\hat{D}_i^{(2)}, \forall i \in B$
3: Partition B into $B_1, ...B_{10}$ datasets.
4: **for** s = 1,...,10 **do**
5: **for** k = 1,..., K **do**
6: $\star$ Fit algorithm k on B_{-s}, call this $\hat{\psi}_{k,s}$
7: Let $s(i) = s$ for all $i \in B_{-s}$
8: **end for**
9: **end for**
10: Using $\{\hat{\psi}_{k,s(i)}(W_i), \hat{D}_i^{(1)}, \hat{D}_i^{(2)} : i \in B, k = \{1, 2, \ldots, K\}\}$ obtain optimal $\vec{\alpha}$ using the `nolptr` package
11: Repeat steps 2 to 11 five times to obtain $\vec{\alpha}^* = \frac{1}{5}\sum_{i=1}^{5} \vec{\alpha}_i$
12: **for** k = 1,....K **do**
13: $\star$ Fit algorithm k on B to obtain $\hat{\psi}_k$
14: **end for**

Output: $w \mapsto \sum_{k=1}^{K} \alpha_k^* \hat{\psi}_k(w)$,$\alpha^*$

Algorithm 3: Super Learner Algorithm. Steps marked with $\star$ may require another layer of CV for data adaptive algorithms, e.g. `glmnet`

Input: V, T

1: With V use `survtmle` package to obtain $\hat{D}_i^{(1)}$ and $\hat{D}_i^{(2)}, \forall i \in V$

2: Partition V into $V_1, ...V_{10}$ datasets

3: **for** s = 1,...,10 **do**

4: Use $\{T, V_{-s}\}$ to obtain $\hat{\psi}_s$ using super learner

5: Let $s(i) = s$ for all $i \in V_s$

6: **end for**

7: Compute $\widehat{AUC}$ using $\{\hat{\psi}_{s(i)}(W_i), \hat{D}_i^{(1)}, \hat{D}_i^{(2)} : i \in V\}$

Output: $\widehat{AUC}$ for V

Algorithm 4: Internal Validation

Input: $E = T \cup V$, A

1: With A use `survtmle` package to obtain $\hat{D}_i^{(1)}$ and $\hat{D}_i^{(2)}, \forall i \in \{A\}$

2: Use E obtain $\hat{\psi}$ using super learner

3: Compute AUC using $\{\hat{\psi}(W_i), \hat{D}_i^{(1)}, \hat{D}_i^{(2)} : i \in A\}$

Output: $\widehat{AUC}$ for A

Algorithm 5: External Validation

4.B Tables and Figures

	Carraguard	HPTN035	MDP301	MIRA	VOICE	ASPIRE
	N = 1485	1054	2392	1485	2750	803
	(15%)	(11%)	(24%)	(15%)	(28%)	(8%)
Age	✓	✓	✓	✓	✓	✓
Education Level		✓	✓	✓	✓	✓
Married	✓	✓		✓	✓	✓
Married or Cohabiting	✓	✓		✓	✓	✓
Parity	✓	✓		✓	✓	✓

Num Sex Acts: 7 days	✓	✓	✓	✓	✓	✓
~~Num Sex Acts: 14 days~~	✓					
Num Sex Partners: 3 mo.	✓	✓			✓	✓
Condom, Last Encounter	✓	✓	✓	✓	✓	✓
~~Age at Sexual Debut~~				✓		
Contraceptive	✓	✓	✓	✓	✓	✓
~~Cohabitating~~		✓		✓	✓	
~~Age of Partner~~	✓				✓	
Partner circumcised	✓				✓	✓
~~Partner has other Partner~~	✓				✓	
~~Screening: Trichomonas~~	✓		✓	✓		✓
Screening: Syphilis		✓	✓	✓	✓	✓
Screening: Gonorrhea	✓	✓	✓	✓	✓	✓
Screening: Chlamydia	✓	✓	✓	✓	✓	✓
Screening: Any STI*	✓	✓	✓	✓	✓	✓

Table S1: Summary of baseline covariate availability by trial. For validation purposes, we only consider baseline covariates available in both VOICE and ASPIRE in our prediction models – variables not available in both of these trials are denoted with a ~~strikethrough~~.

* Chlamydia, Gonorrhea or Syphilis, except Syphilis for Carraguard.

	Training				Internal Validation	External Validation
Trial	Carraguard N = 1,485 (15%)	HPTN035 N= 1,054 (11%)	MDP301 N= 2,392 (23%)	MIRA N = 1,485 (15%)	VOICE N = 2,726 (28%)	ASPIRE N = 797 (8%)
Month 6	0.88	0.95	0.84	0.97	0.91	0.92
Month 12	0.65	0.92	0.23	0.89	0.58	0.77
Month 18	0.38	0.65	0.00	0.48	0.18	0.65
Month 24	0.07	0.41	0.00	0.24	0.02	0.59

Table S2: Proportion of individuals with follow-up through months 6, 12, 18, and 24.

	Method		R function
"Classical" algorithms	Cox Model, Main Terms	(Cox 1972)	`coxph`
	Cox Model, Main Terms, Forward Stepwise	(Peduzzi et al. 1980)	`stepAIC`
	Cox Model, Interactions, Forward Stepwise	(Peduzzi et al. 1980)	`stepAIC`
	Acelerated Failure Time Model, Main Terms	(Wei 1992)	`survreg`
	Acelerated Failure Time Model, Interactions	(Wei 1992)	`survreg`
Data adaptive algorithms	Cox Model, Elastic Net Penalty	(Tibshirani 1997)	`glmnet`
	Cox Model, Interactions, Elastic Net Penalty	(Tibshirani 1997)	`glmnet`
	Survival Forest	(Ishwaran et al. 2008)	`rsfc`
	Generalized Additive Model, Cox	(Hastie and Tibshirani 1995)	`gam`

Table S3: Candidate prediction methods used in the super-learner library. *Interactions* denotes the inclusion of all interaction terms in the model.

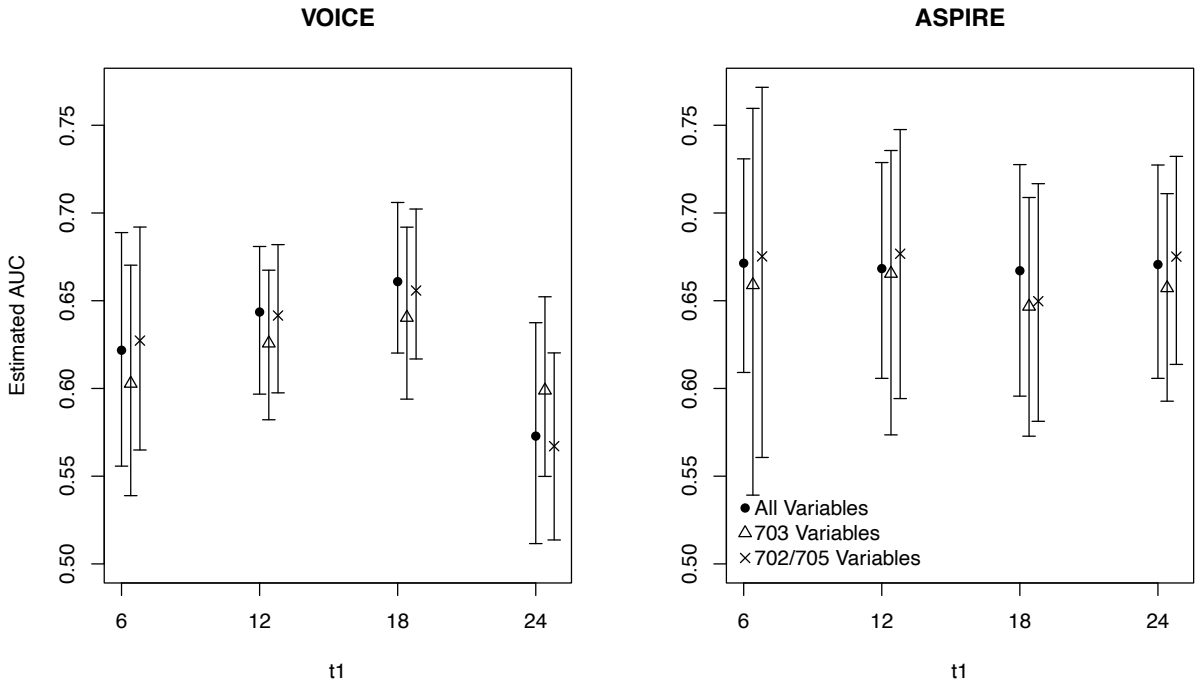

Figure S1: Internal (left panel, VOICE) and external validation (right panel, ASPIRE) AUC as we fix $t_0 = 0$ and vary t_1 by 6 month increments and across different variable sets. The AUCs at $t = 18$ and $t = 6$ months are slightly higher than the rest in the validation and test set, respectively. However, the confidence intervals at these times typically overlap, across all variable sets. In VOICE, the AUC is lowest when t_1 is equal to 24 months, and there is no clear trend in which variable set yields the highest AUC. In ASPIRE, except when $t_1 = 18$, the AUC with highest values are those from the super learner fitted with variables from the HVTN 702/705 trials, and the lowest AUCs are those from the HVTN 703 variables. Additionally, for a fixed variable set, the AUC does not seem to vary as t_1 varies.

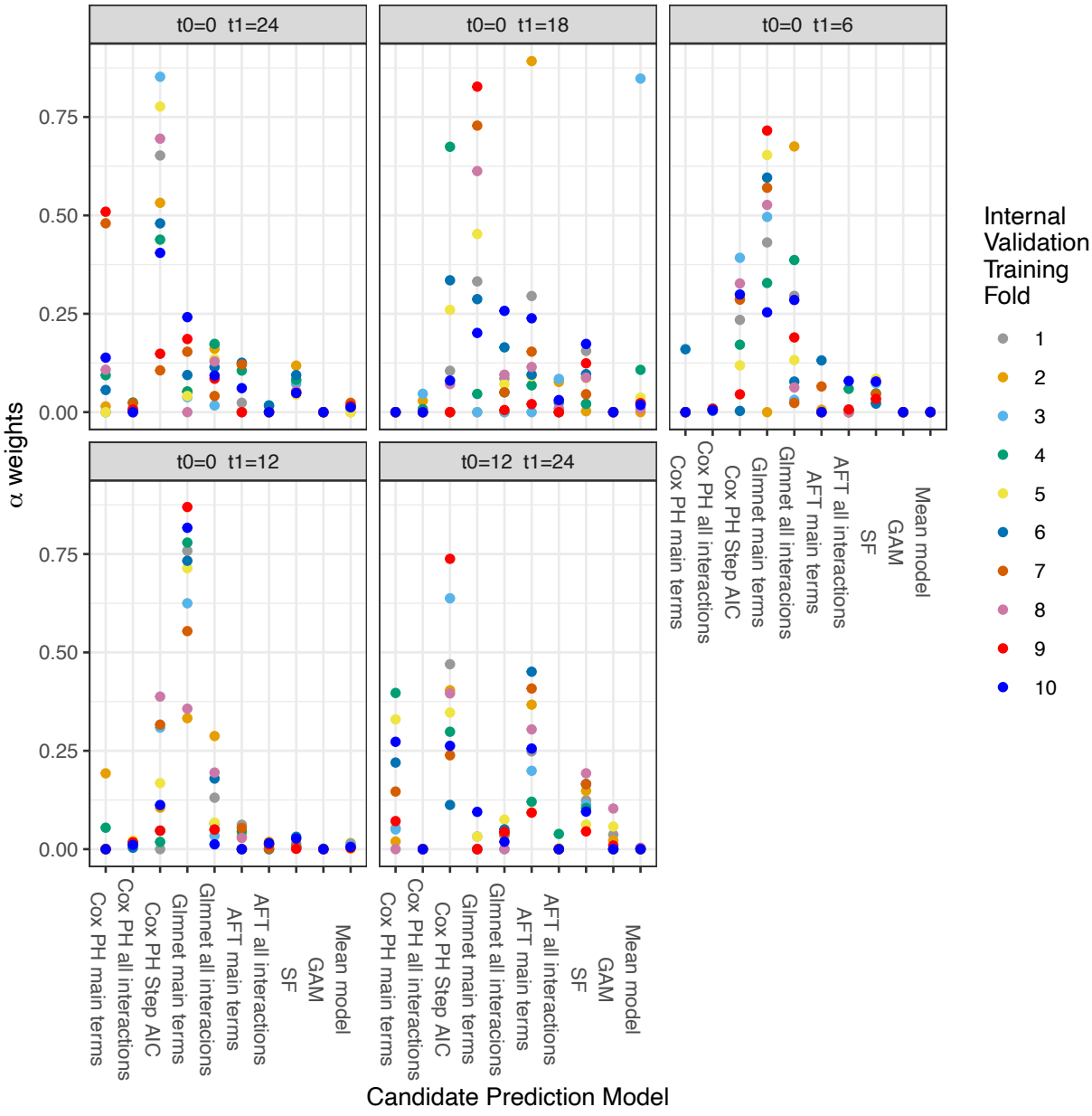

Figure S2: α-weights assigned by the super learner on each training fold during internal validation. Each panel shows the weights depending on the window (t0,t1) of months over which risk prediction is performed. Main terms refer to those models that adjust for all available covariates and all interactions adjust for all possible interactions. Abbreviations: PH, proportional hazards; Glmnet: Generalized linear model with elastic net penalty; AFT, accelerated failure time; SF, survival forest; GAM, generalized additive model.

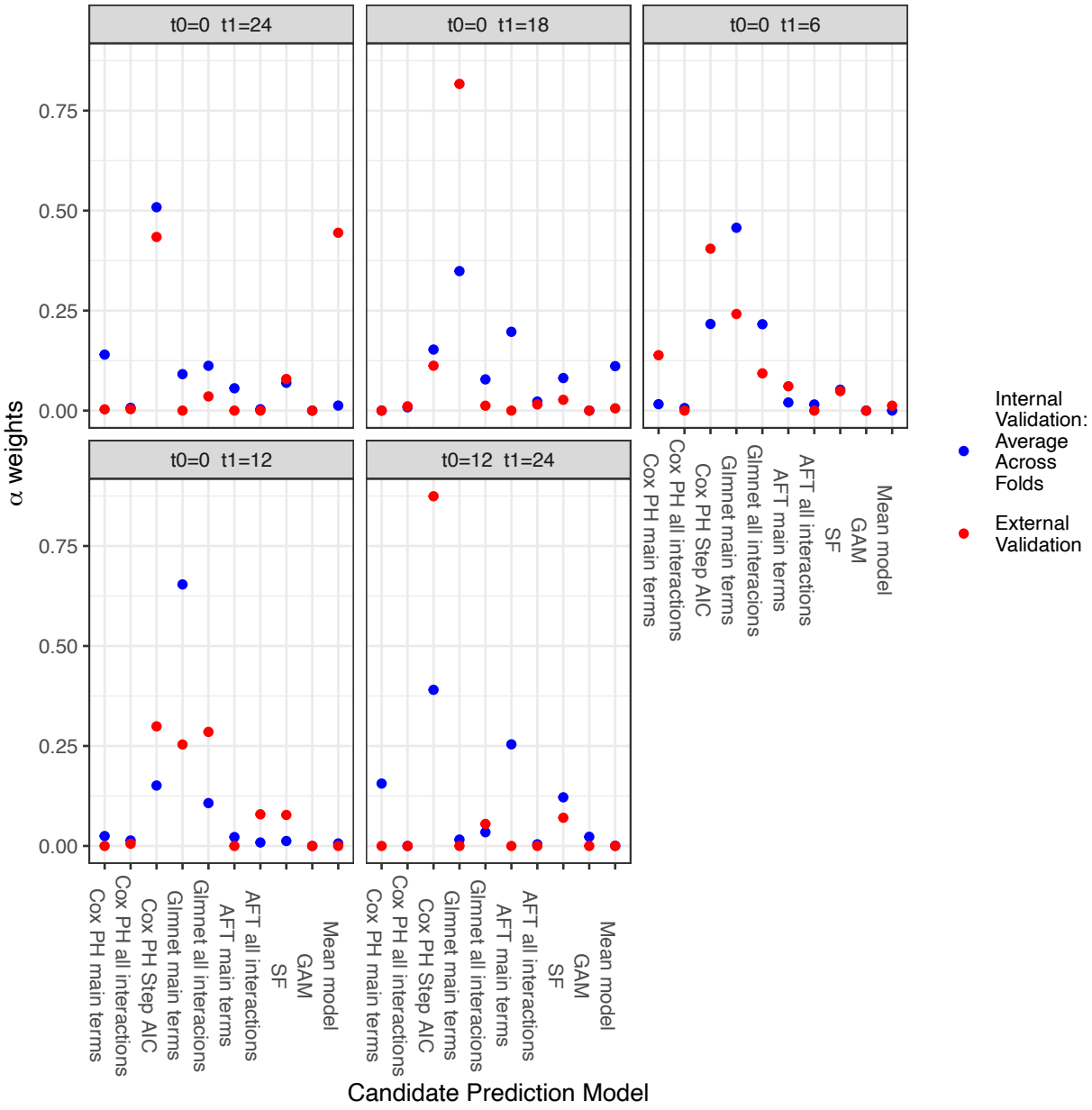

Figure S3: Average of α-weights assigned by the super learner across cross validation folds and when fitting the super learner on all the validation and training data. Each panel shows the weights depending on the window (t0,t1) of months over which risk prediction is performed. Main terms refer to those models that adjust for all available covariates and all interactions adjust for all possible interactions. Abbreviations: CV, cross validation; PH, proportional hazards; Glmnet: Generalized linear model with elastic net penalty; AFT, accelerated failure time; SF, survival forest; GAM, generalized additive model.

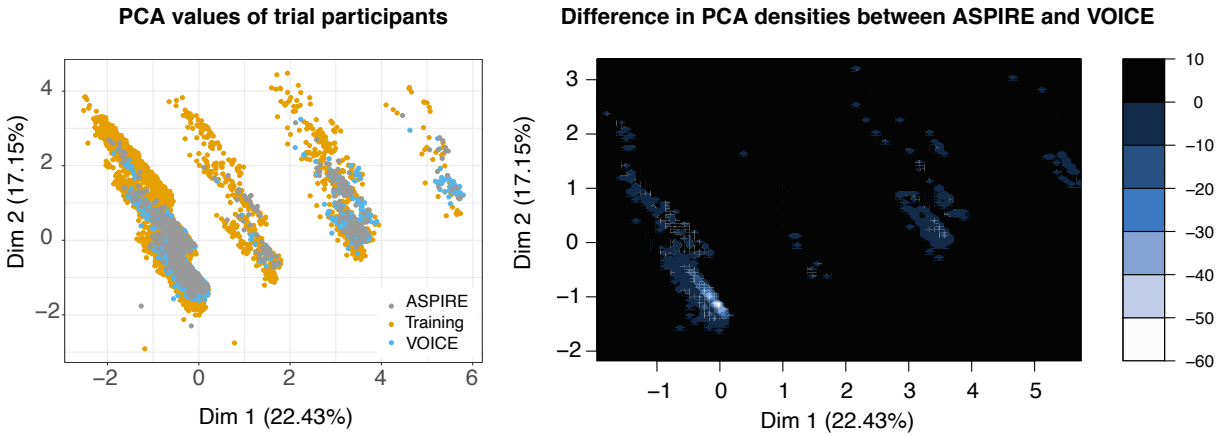

Figure S4: Principal Component Analysis plots of all trial participants using variables measured across all trials (age, condom use in last sex encounter, any curable STI, gonnorhea, chlamydia, family planning method, number of vaginal sex encounters in the past 7 days). On the left panel, we see all participants clustered in four groups and that there is substantial overlap between VOICE and ASPIRE observations. On the right panel, we see the difference between PCA densities between ASPIRE and VOICE. The most notable difference lies in the lower left region of the plot, where the density of VOICE participants is higher. This difference is mainly driven by family planning method and condom use in last sex encounter.

	Performance of the super learner when estimating risk of HIV-1 infection within 24 months of enrollment					Performance of the super learner when estimating risk of HIV-1 infection within 12 months of enrollment				
Sensitivity VOICE	Threshold predicted risk	Quantile predicted risk	Specificity VOICE	Sensitivity ASPIRE	Specificity ASPIRE	Threshold predicted risk	Quantile predicted risk	Specificity VOICE	Sensitivity ASPIRE	Specificity ASPIRE
0.90	0.1	0.07	0.24	0.99	0.12	0.05	0.03	0.25	0.92	0.19
0.80	0.12	0.07	0.33	0.9	0.25	0.06	0.03	0.34	0.87	0.27
0.70	0.14	0.08	0.46	0.71	0.5	0.07	0.03	0.43	0.81	0.38
0.60	0.15	0.08	0.57	0.53	0.72	0.08	0.03	0.6	0.71	0.56
0.50	0.17	0.08	0.70	0.4	0.84	0.08	0.03	0.67	0.63	0.64
0.40	0.18	0.09	0.78	0.25	0.89	0.09	0.03	0.76	0.54	0.74
0.30	0.2	0.09	0.84	0.14	0.94	0.10	0.04	0.85	0.42	0.83
0.20	0.23	0.10	0.90	0.06	0.98	0.12	0.04	0.90	0.25	0.89
0.10	0.25	0.10	0.95	0.00	1.00	0.13	0.04	0.95	0.19	0.93

Table S4: Estimated sensitivity and specificity for various cutoffs when predicting HIV-1 risk at 12 and 24 months since enrollment. The risk thresholds were fixed to attain different sensitivity values in the validation data. The sensitivity values in VOICE are shown on the left-most column. The actual threshold and their corresponding quantile are shown next. The next three columns show the attained sensitivity in VOICE, and the resulting sensitivity and specificity in ASPIRE using the same threshold. The right section of the table shows the same information, but when predicting risk between 0 and 12 months since enrollment.

BIBLIOGRAPHY

Consolidated Guidelines on HIV Prevention, Diagnosis, Treatment and Care for Key Populations. Technical report.

A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. econometrica, 74(1):235–267, 2006.

A. Abadie and G. W. Imbens. Matching on the estimated propensity score. Econometrica, 84(2):781–807, 2016.

S. S. Alistar, P. M. Grant, and E. Bendavid. Comparative effectiveness and cost-effectiveness of antiretroviral therapy and pre-exposure prophylaxis for HIV prevention in South Africa. BMC medicine, 12(1):1–11, 2014.

S.-J. Anderson, P. Cherutich, N. Kilonzo, I. Cremin, D. Fecht, D. Kimanga, M. Harper, R. L. Masha, P. B. Ngongo, W. Maina, et al. Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. The Lancet, 384(9939):249–256, 2014.

E. Andreou and B. J. Werker. An alternative asymptotic analysis of residual-based statistics. Review of Economics and Statistics, 94(1):88–99, 2012.

S. Athey. Beyond prediction: Using big data for policy problems. Science, 355(6324):483–485, 2017.

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

S. Athey and S. Wager. Policy learning with observational data. Econometrica, 89(1): 133–161, 2021.

P. C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate behavioral research, 46(3):399–424, 2011.

S. G. Ayton, M. Pavlicova, and Q. A. Karim. Identification of adolescent girls and young women for targeted HIV prevention: a new risk scoring tool in KwaZulu-Natal, South Africa. Scientific reports, 10(1):1–10, 2020a.

S. G. Ayton, M. Pavlicova, H. Tamir, and Q. A. Karim. Development of a prognostic tool exploring female adolescent risk for HIV prevention and PrEP in rural South Africa, a generalised epidemic setting. Sexually transmitted infections, 96(1):47–54, 2020b.

J. M. Baeten, T. Palanee-Phillips, E. R. Brown, K. Schwartz, L. E. Soto-Torres, V. Govender, N. M. Mgodi, F. Matovu Kiweewa, G. Nair, F. Mhlanga, et al. Use of a vaginal ring containing dapivirine for HIV-1 prevention in women. New England Journal of Medicine, 375(22):2121–2132, 2016.

J. E. Balkus, E. Brown, T. Palanee, G. Nair, Z. Gafoor, J. Zhang, B. A. Richardson, Z. M. Chirenje, J. M. Marrazzo, and J. M. Baeten. An empiric HIV risk scoring tool to predict HIV-1 acquisition in African women. Journal of acquired immune deficiency syndromes (1999), 72(3):333, 2016.

J. E. Balkus, E. R. Brown, T. Palanee-Phillips, F. M. Kiweewa, N. Mgodi, L. Naidoo, Y. Mbilizi, N. Jeenarain, Z. Gaffoor, S. Siva, et al. Performance of a validated risk score to predict HIV-1 acquisition among African women participating in a trial of the dapivirine vaginal ring. Journal of acquired immune deficiency syndromes (1999), 77(1):e8, 2018.

L. B. Balzer, D. V. Havlir, M. R. Kamya, G. Chamie, E. D. Charlebois, T. D. Clark, C. A. Koss, D. Kwarisiima, J. Ayieko, N. Sang, et al. Machine learning to identify persons

at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clinical Infectious Diseases, 71(9):2326–2333, 2020.

S. Baral, A. Rao, P. Sullivan, N. Phaswana-Mafuya, D. Diouf, G. Millett, H. Musyoki, E. Geng, and S. Mishra. The disconnect between individual-level and population-level HIV prevention benefits of antiretroviral treatment. The lancet HIV, 6(9):e632–e638, 2019.

R. E. Barlow and H. D. Brunk. The isotonic regression problem and its dual. Journal of the American Statistical Association, 67(337):140–147, 1972.

A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. Calibration of machine learning models. In Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pages 128–146. IGI Global, 2010.

D. Benkeser and M. van der Laan. The highly adaptive lasso estimator. In 2016 IEEE international conference on data science and advanced analytics (DSAA), pages 689–696. IEEE, 2016.

D. Benkeser, M. Carone, and P. B. Gilbert. Improved estimation of the cumulative incidence of rare outcomes. Statistics in medicine, 37(2):280–293, 2018.

D. Benkeser, P. B. Gilbert, and M. Carone. Estimating and testing vaccine sieve effects using machine learning. Journal of the American Statistical Association, 2019.

D. Benkeser, I. Díaz, A. Luedtke, J. Segal, D. Scharfstein, and M. Rosenblum. Improving precision and power in randomized trials for covid-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. Biometrics, 2020.

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

M. A. Brookhart, S. Schneeweiss, K. J. Rothman, R. J. Glynn, J. Avorn, and T. Stürmer.

Variable selection for propensity score models. American journal of epidemiology, 163(12): 1149–1156, 2006.

N. Carnegie, V. Dorie, and J. L. Hill. Examining treatment effect heterogeneity using bart. Observational Studies, 5(2):52–70, 2019.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

J. Clifton and E. Laber. Q-learning: theory and applications. Annual Review of Statistics and Its Application, 7:279–301, 2020.

M. S. Cohen, Y. Q. Chen, M. McCauley, T. Gamble, M. C. Hosseinipour, N. Kumarasamy, J. G. Hakim, J. Kumwenda, B. Grinsztejn, J. H. Pilotto, et al. Antiretroviral therapy for the prevention of HIV-1 transmission. New England Journal of Medicine, 375(9):830–839, 2016.

E. Colantuoni and M. Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. Statistics in medicine, 34(18):2602–2617, 2015.

F. S. Collins and H. Varmus. A new initiative on precision medicine. New England journal of medicine, 372(9):793–795, 2015.

S. A. N. A. Commission. South Africa's National Strategic Plan for HIV, TB and STIs 2017-2022. Technical report, 2017.

L. Corey, P. B. Gilbert, M. Juraska, D. C. Montefiori, L. Morris, S. T. Karuna, S. Edupuganti, N. M. Mgodi, A. C. deCamp, E. Rudnicki, Y. Huang, P. Gonzales, R. Cabello, C. Orrell, J. R. Lama, F. Laher, E. M. Lazarus, J. Sanchez, I. Frank, J. Hinojosa, M. E. Sobieszczyk, K. E. Marshall, P. G. Mukwekwerere, J. Makhema, L. R. Baden, J. I. Mullins, C. Williamson, J. Hural, M. J. McElrath, C. Bentley, S. Takuva, M. M.

Gomez Lorenzo, D. N. Burns, N. Espy, A. K. Randhawa, N. Kochar, E. Piwowar-Manning, D. J. Donnell, N. Sista, P. Andrew, J. G. Kublin, G. Gray, J. E. Ledgerwood, J. R. Mascola, and M. S. Cohen. Two Randomized Trials of Neutralizing Antibodies to Prevent HIV-1 Acquisition. New England Journal of Medicine, 384(11):1003–1014, 2021. doi: 10.1056/NEJMoa2031738. URL https://doi.org/10.1056/NEJMoa2031738. PMID: 33730454.

D. R. Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, 1972.

J. R. Coyle, N. S. Hejazi, I. Malenica, R. V. Phillips, and O. Sofrygin. sl3: Modern pipelines for machine learning and Super Learning. https://github.com/tlverse/sl3, 2021. URL https://doi.org/10.5281/zenodo.1342293. R package version 1.4.2.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics, 90(3):389–405, 2008.

V. Cucchiara, M. R. Cooperberg, M. DallEra, D. W. Lin, F. Montorsi, J. A. Schalken, and C. P. Evans. Genomic markers in prostate cancer decision making. European urology, 73 (4):572–582, 2018.

I. J. Dahabreh, R. Hayward, and D. M. Kent. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. International journal of epidemiology, 45(6):2184–2193, 2016.

F. Devriendt, D. Moldovan, and W. Verbeke. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big data, 6(1):13–41, 2018.

F. Dominici, F. J. Bargagli-Stoffi, and F. Mealli. From controlled to undisciplined data: estimating causal effects in the era of data science using a potential outcome framework. arXiv preprint arXiv:2012.06865, 2020.

A. Duerr, Y. Huang, S. Buchbinder, R. W. Coombs, J. Sanchez, C. del Rio, M. Casapia, S. Santiago, P. Gilbert, L. Corey, et al. Extended follow-up confirms early vaccine-enhanced risk of hiv acquisition and demonstrates waning effect over time among participants in a randomized trial of recombinant adenovirus hiv vaccine (step study). The Journal of infectious diseases, 206(2):258–266, 2012.

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

J. H. Friedman. Multivariate adaptive regression splines. The annals of statistics, 19(1): 1–67, 1991.

E. E. Gabriel, M. C. Sachs, D. A. Follmann, and T. M.-L. Andersson. A unified evaluation of differential vaccine efficacy. Biometrics, 76(4):1053–1063, 2020.

P. Gaenssler and E. Haeusler. On martingale central limit theory. In Dependence in probability and statistics, pages 303–334. Springer, 1986.

M. B. Gilkey, J. L. Marcus, J. M. Garrell, V. E. Powell, K. M. Maloney, and D. S. Krakower. Using HIV risk prediction tools to identify candidates for pre-exposure prophylaxis: perspectives from patients and primary care providers. AIDS patient care and STDs, 33(8): 372–378, 2019.

D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. Dagostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. Odonnell, et al. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. Journal of the American College of Cardiology, 63(25 Part B):2935–2959, 2014.

G. E. Gray, L.-G. Bekker, F. Laher, M. Malahleha, M. Allen, Z. Moodie, N. Grunenberg, Y. Huang, D. Grove, B. Prigmore, J. J. Kee, D. Benkeser, J. Hural, C. Innes, E. Lazarus,

G. Meintjes, N. Naicker, D. Kalonji, M. Nchabeleng, M. Sebe, N. Singh, P. Kotze, S. Kassim, T. Dubula, V. Naicker, W. Brumskine, C. N. Ncayiya, A. M. Ward, N. Garrett, G. Kistnasami, Z. Gaffoor, P. Selepe, P. B. Makhoba, M. P. Mathebula, P. Mda, T. Adonis, K. S. Mapetla, B. Modibedi, T. Philip, G. Kobane, C. Bentley, S. Ramirez, S. Takuva, M. Jones, M. Sikhosana, M. Atujuna, M. Andrasik, N. S. Hejazi, A. Puren, L. Wiesner, S. Phogat, C. Diaz Granados, M. Koutsoukos, O. Van Der Meeren, S. W. Barnett, N. Kanesa-Thasan, J. G. Kublin, M. J. McElrath, P. B. Gilbert, H. Janes, and L. Corey. Vaccine Efficacy of ALVAC-HIV and Bivalent Subtype C gp120MF59 in Adults. New England Journal of Medicine, 384(12):1089–1100, 2021. doi: 10.1056/NEJMoa2031499. URL `https://doi.org/10.1056/NEJMoa2031499`. PMID: 33761206.

C. Gupta and A. K. Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. arXiv preprint arXiv:2105.04656, 2021.

C. Gupta, A. Podkopaev, and A. Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 3711–3723. Curran Associates, Inc., 2020a. URL `https://proceedings.neurips.cc/paper/2020/file/26d88423fc6da243ffddf161ca712757-Paper.pdf`.

C. Gupta, A. Podkopaev, and A. Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. Advances in Neural Information Processing Systems, 33:3711–3723, 2020b.

T. Hastie and R. Tibshirani. Generalized additive models for medical research. Statistical methods in medical research, 4(3):187–196, 1995.

Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association, 27(4):621–633, 2020.

G. W. Imbens and J. M. Wooldridge. Recent developments in the econometrics of program evaluation. Journal of economic literature, 47(1):5–86, 2009.

H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer. Random survival forests. The annals of applied statistics, 2(3):841–860, 2008.

E. M. Kahle, J. P. Hughes, J. R. Lingappa, G. John-Stewart, C. Celum, E. Nakku-Joloba, S. Njuguna, N. Mugo, E. Bukusi, R. Manongi, et al. An empiric risk scoring tool for identifying high-risk heterosexual HIV-1 serodiscordant couples for targeted HIV-1 prevention. Journal of acquired immune deficiency syndromes (1999), 62(3):339, 2013.

S. S. A. Karim, B. A. Richardson, G. Ramjee, I. F. Hoffman, Z. M. Chirenje, T. Taha, M. Kapina, L. Maslankowski, M. A. Coletti, A. Profy, et al. Safety and effectiveness of BufferGel and 0.5% PRO2000 gel for the prevention of HIV infection in women. AIDS (London, England), 25(7):957, 2011.

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

D. M. Kent, E. Steyerberg, and D. van Klaveren. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. Bmj, 363, 2018.

D. S. Krakower, S. Gruber, K. Hsu, J. T. Menchaca, J. C. Maro, B. A. Kruskal, I. B. Wilson, K. H. Mayer, and M. Klompas. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. The Lancet HIV, 6(10):e696–e704, 2019.

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

F. Laher, L.-G. Bekker, N. Garrett, E. M. Lazarus, and G. E. Gray. Review of preventative HIV vaccine clinical trials in South Africa. Archives of virology, pages 1–14, 2020.

K. Lee, F. J. Bargagli-Stoffi, and F. Dominici. Causal rule ensemble: Interpretable inference of heterogeneous treatment effects. arXiv preprint arXiv:2009.09036, 2020.

Y. Leng and D. Dimmery. Calibration of heterogeneous treatment effects in random experiments. Available at SSRN 3875850, 2021.

Z. Lin, P. Ding, and F. Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. arXiv preprint arXiv:2112.13506, 2021.

D. M. Lloyd-Jones, L. T. Braun, C. E. Ndumele, S. C. Smith Jr, L. S. Sperling, S. S. Virani, and R. S. Blumenthal. Use of risk assessment tools to guide decision-making in the primary prevention of atherosclerotic cardiovascular disease: a special report from the american heart association and american college of cardiology. Circulation, 139(25):e1162–e1177, 2019.

A. R. Luedtke and M. J. Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. Annals of statistics, 44(2):713, 2016.

A. R. Luedtke, O. Sofrygin, M. J. van der Laan, and M. Carone. Sequential double robustness in right-censored longitudinal models. arXiv preprint arXiv:1705.02459, 2017.

J. M. Marrazzo, G. Ramjee, B. A. Richardson, K. Gomez, N. Mgodi, G. Nair, T. Palanee, C. Nakabiito, A. Van Der Straten, L. Noguchi, et al. Tenofovir-based preexposure prophylaxis for HIV infection among African women. New England Journal of Medicine, 372(6): 509–518, 2015.

S. McCormack, G. Ramjee, A. Kamali, H. Rees, A. M. Crook, M. Gafos, U. Jentsch, R. Pool, M. Chisembele, S. Kapiga, et al. PRO2000 vaginal gel for prevention of HIV-1 infection (Microbicides Development Programme 301): a phase 3, randomised, double-blind, parallel-group trial. The Lancet, 376(9749):1329–1337, 2010.

S. A. Murphy. Optimal dynamic treatment regimes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(2):331–355, 2003.

M. P. Naeini, G. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. Handbook of econometrics, 4:2111–2245, 1994.

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

Z. Obermeyer and E. J. Emanuel. Predicting the futurebig data, machine learning, and clinical medicine. The New England journal of medicine, 375(13):1216, 2016.

N. S. Padian, A. van der Straten, G. Ramjee, T. Chipato, G. de Bruyn, K. Blanchard, S. Shiboski, E. T. Montgomery, H. Fancher, H. Cheng, et al. Diaphragm and lubricant gel for prevention of HIV acquisition in southern African women: a randomised controlled trial. The Lancet, 370(9583):251–261, 2007.

P. Peduzzi, R. Hardy, and T. R. Holford. A stepwise variable selection procedure for nonlinear regression models. Biometrics, pages 511–516, 1980.

M. L. Petersen, E. LeDell, J. Schwab, V. Sarovar, R. Gross, N. Reynolds, J. E. Haberer, K. Goggin, C. Golin, J. Arnsten, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. Journal of acquired immune deficiency syndromes (1999), 69(1):109, 2015.

J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3):61–74, 1999.

E. C. Polley, S. Rose, and M. J. Van der Laan. Super learning. In Targeted learning, pages 43–66. Springer, 2011.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for

Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

J. M. Robins. Optimal structural nested models for optimal sequential decisions. In Proceedings of the second seattle Symposium in Biostatistics, pages 189–326. Springer, 2004.

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

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the american statistical association, 90(429):106–121, 1995.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

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

J. S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. Journal of Statistical Software, Forthcoming, 2008.

S. Skoler-Karpoff, G. Ramjee, K. Ahmed, L. Altini, M. G. Plagianos, B. Friedland, S. Govender, A. De Kock, N. Cassim, T. Palanee, et al. Efficacy of Carraguard for prevention of HIV infection in women in South Africa: a randomised, double-blind, placebo-controlled trial. The Lancet, 372(9654):1977–1987, 2008.

J. A. Steingrimsson, D. F. Hanley, and M. Rosenblum. Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions. Contemporary clinical trials, 54:18–24, 2017.

D. J. Stekhoven and P. Bhlmann. MissForestnon-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1):112–118, 10 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093/bioinformatics/btr597.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1):1, 2010.

R. Tibshirani. The lasso method for variable selection in the Cox model. Statistics in medicine, 16(4):385–395, 1997.

UNAIDS. End Inequalities. End AIDS. Global AIDS Strategy 2021- 2026. Technical report, 2021.

B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, and E. W. Steyerberg. Calibration: the achilles heel of predictive analytics. BMC medicine, 17(1):1–7, 2019.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. The International Journal of Biostatistics, 2(1), 2006.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. Statistical applications in genetics and molecular biology, 6(1), 2007.

M. J. Van der Laan, S. Rose, et al. Targeted learning: causal inference for observational and experimental data, volume 4. Springer, 2011.

A. W. van der Vaart. Asymptotic statistics, volume 3. Cambridge university press, 2000.

A. W. van der Vaart and J. Wellner. Weak convergence and empirical processes: with applications to statistics. Springer Science & Business Media, 1996.

D. van Klaveren, T. A. Balan, E. W. Steyerberg, and D. M. Kent. Models with interactions overestimated heterogeneity of treatment effects and were prone to treatment mistargeting. Journal of clinical epidemiology, 114:72–83, 2019.

A. Vandormael, A. Akullian, M. Siedner, T. de Oliveira, T. Bärnighausen, and F. Tanser. Declines in HIV incidence among men and women in a South African population-based cohort. Nature communications, 10(1):1–10, 2019.

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

E. Wahome, A. N. Thiongo, G. Mwashigadi, O. Chirro, K. Mohamed, E. Gichuru, J. Mwambi, M. A. Price, S. M. Graham, and E. J. Sanders. An empiric risk score to guide PrEP targeting among MSM in coastal Kenya. AIDS and Behavior, 22(1):35–44, 2018.

H. Wand, T. Reddy, S. Naidoo, S. Moonsamy, S. Siva, N. S. Morar, and G. Ramjee. A simple risk prediction algorithm for HIV transmission: results from HIV prevention trials in KwaZulu Natal, South Africa (2002–2012). AIDS and Behavior, 22(1):325–336, 2018.

L.-J. Wei. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in medicine, 11(14-15):1871–1879, 1992.

B. Williamson, P. Gilbert, M. Carone, and N. Simon. Nonparametric variable importance assessment using machine learning techniques. Biometrics, 2020. doi: 10.1111/biom.13392.

S. N. Wood. Introducing gams. In Generalized additive models, pages 161–194. Chapman and Hall/CRC, 2017.

Y. Xu and S. Yadlowsky. Calibration error for heterogeneous treatment effects. arXiv preprint arXiv:2203.13364, 2022.

J. Ypma, S. G. Johnson, H. W. Borchers, D. Eddelbuettel, B. Ripley, K. Hornik, J. Chiquet, and A. Adler. nloptr: The NLopt nonlinear-optimization package. 2020. URL `https://CRAN.R-project.org/package=nloptr`. R package version 1.2.2.2.

B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Icml, volume 1, pages 609–616. Citeseer, 2001.