# Stabilized Inverse Probability Weighting via Isotonic Calibration

**Lars van der Laan**                                                                    LVDLAAN@UW.EDU
*Department of Statistics, University of Washington, Seattle, WA 98195, USA*

**Ziming Lin**                                                                               ZMLIN@UW.EDU
*Department of Statistics, University of Washington, Seattle, WA 98195, USA*

**Marco Carone**                                                                        MCARONE@UW.EDU
*Department of Biostatistics, University of Washington, Seattle, WA 98195, USA*

**Alex Luedtke**                                                                        ALUEDTKE@UW.EDU
*Department of Statistics, University of Washington, Seattle, WA 98195, USA*

## Abstract

Inverse weighting with an estimated propensity score is widely used by estimation methods in causal inference to adjust for confounding bias. However, directly inverting propensity score estimates can lead to instability, bias, and excessive variability due to large inverse weights, especially when treatment overlap is limited. In this work, we propose a post-hoc calibration algorithm for inverse propensity weights that generates well-calibrated, stabilized weights from user-supplied, cross-fitted propensity score estimates. Our approach employs a variant of isotonic regression with a loss function specifically tailored to the inverse propensity weights. Through theoretical analysis and empirical studies, we demonstrate that isotonic calibration improves the performance of doubly robust estimators of the average treatment effect.

**Keywords:** inverse probability weighting, calibration, isotonic regression, propensity, balancing

## 1. Introduction

### 1.1. Background

Estimation and inference tasks in causal inference often involve inverse weighting by an estimate of the propensity score — the probability of receiving treatment given a set of covariates (Rosenbaum and Rubin, 1983). This includes, for example, prominent methods for inference of counterfactual means and average treatment effects such as inverse probability weighting (IPW) (Rosenbaum and Rubin, 1983) and doubly-robust estimation strategies like augmented inverse probability weighting (Robins et al., 1994; Bang and Robins, 2005), targeted maximum likelihood estimation (van der Laan and Rose, 2011), and double machine learning (Chernozhukov et al., 2018a). Despite their wide usage and theoretical motivation, these approaches are prone to instability, bias, and excessive variability induced by large inverse propensity weights, especially in settings with limited treatment overlap (Wang et al., 2006; Crump et al., 2009; Petersen et al., 2012; Zhou et al., 2020). As a result, there is a growing interest in developing robust methods to estimate inverse propensity weights, aiming to mitigate the influence of extreme values in these weights. To address this issue, various estimation strategies have been proposed, which can be broadly categorized into two groups: direct and inversion-based methods.

Inversion methods aim to construct an improved estimator of inverse propensity weights by applying post-hoc adjustments to an initial propensity score estimator before inverting it. One advantage of inversion methods is their convenience, as the initial propensity score estimator can be

obtained using any off-the-shelf regression algorithm. Notable inversion methods include propensity score truncation (or clipping) and propensity score trimming. Propensity score truncation constrains the estimated propensity scores within a range bounded away from 0 and 1 (Wang et al., 2006; Bembom and van der Laan, 2008; Cole and Hernán, 2008; Petersen et al., 2012; Gruber et al., 2022). Propensity score trimming is a related approach that involves excluding observations with extreme propensity score weights, i.e., assigning them zero weight (Imbens, 2004; Crump et al., 2009; Petersen et al., 2012). The choice of truncation or trimming threshold can be predetermined or data-adaptive. Pre-specified thresholds, however, may yield inconsistent estimates of propensity score weights, leading to biased estimations of treatment effects. To address this bias and improve the reliability of inferences, data-adaptive methods for threshold selection, such as collaborative targeted maximum likelihood estimation, have been proposed (Bembom and van der Laan, 2008; Ju et al., 2019; Gruber et al., 2022). While inversion methods can lead to improved inference, they are limited by the fact that even when the propensity score estimator is properly truncated or trimmed, it may not serve as a good estimator of the true inverse propensity weights.

Direct learning methods improve upon inversion methods by using supervised statistical learning tools to directly estimate the inverse propensity weights through the optimization of a specifically tailored objective function. Notable direct learning methods include entropy (Hainmueller, 2012) and covariate balancing weights (Imai and Ratkovic, 2014; Zubizarreta, 2015; Zhao, 2019; Wang and Zubizarreta, 2020), minimax linear estimation (Hirshberg et al., 2019; Hirshberg and Wager, 2021), and automatic debiased machine learning (Chernozhukov et al., 2018b, 2021, 2022a,b; Bruns-Smith et al., 2023). These approaches have been demonstrated, in simulations, to yield less biased and more stable parameter estimates with improved confidence interval coverage compared to inversion methods relying on propensity score estimation (Zhao, 2019; Chernozhukov et al., 2021). Even so, when estimated using aggressive machine learning algorithms, such as neural networks or gradient-boosted trees, these inverse weights may still exhibit poor calibration (Guo et al., 2017). Furthermore, the practical applicability of direct learning methods is hindered by their dependence on nonstandard loss functions, which are not readily available in standard software packages for machine learning. As a consequence of this limitation, direct methods may be employed in conjunction with parametric models, which may lack the necessary flexibility to accurately estimate inverse propensity weights. While some machine learning algorithms, such as xgboost gradient-boosted trees (Chen and Guestrin, 2016) and neural networks (Pedregosa et al., 2011), provide flexibility to accommodate customized loss functions, the extent of customization, scalability, and numerical efficiency can vary among commonly used learning algorithms and software choices.

## 1.2. Contributions of this work

In this work, we introduce a distribution-free approach for calibrating inverse propensity weights directly from user-supplied propensity score estimates. We propose a novel algorithm, isotonic calibrated inverse probability weighting (IC-IPW), which uses isotonic regression to transform cross-fitted propensity score estimates into well-calibrated inverse propensity weights. Departing from traditional inversion methods, IC-IPW learns an optimal monotone transformation of the propensity score estimator through a variant of isotonic regression that minimizes a loss function tailored to the inverse propensity score. This approach is computationally efficient, computable in linear time, and easily implemented using standard isotonic regression software. It retains the flexibility and simplicity of inversion-based estimation while providing robustness in settings with limited overlap.

To demonstrate the advantages of IC-IPW, we examine its use with augmented inverse probability weighted (AIPW) estimators for estimating the average treatment effect (ATE). We show that IC-IPW can relax the conditions required for achieving asymptotic linearity and nonparametric efficiency of AIPW, while also improving empirical performance in terms of bias and coverage.

Our work contributes to the recent studies of Gutman et al. (2022), Deshpande and Kuleshov (2023), and Ballinari and Bearth (2024) on the benefits of calibrated propensity score estimators for causal inference applications. In their empirical study, Gutman et al. (2022) found Platt's scaling to be beneficial for calibrating propensity score estimators in causal inference. Additionally, Deshpande and Kuleshov (2023) propose using propensity score calibration in IPW and AIPW estimators and provide theoretical guarantees for kernel density-based calibration using a log-likelihood loss. Similarly, Ballinari and Bearth (2024) proposed using propensity score calibration within the double machine learning framework to improve the finite-sample performance of ATE estimators. They examined the empirical performance of various probability calibration procedures, including Platt's and temperature scaling and a form of isotonic calibration that differs from ours. Distribution-free calibration guarantees for isotonic calibration of regression and conditional average treatment effect functions were established in Van Der Laan et al. (2023). We build on this work by establishing that distribution-free calibration guarantees for isotonic calibration can be extended to the inverse propensity score based on a tailored loss function.

Our work differs from prior studies of propensity score calibration in several key aspects. First, differing from prior work, we propose calibrating the inverse propensity score rather than the propensity score itself and introduce a measure for balance-based calibration of inverse propensity weights. This distinction is crucial because calibration methods developed for the propensity score may not yield well-calibrated inverse propensity weights. In this sense, our work is closely related to Deshpande and Kuleshov (2023), which considers calibrating propensity score estimates using the chi-squared divergence loss. While the authors do not explicitly consider calibration of the inverse propensity score, the chi-squared divergence loss can be reformulated as a loss for the inverse propensity score. Second, we propose a novel calibration procedure based on isotonic regression to construct calibrated inverse propensity weights, which is free from tuning parameters. Furthermore, we formally establish that our isotonic regression procedure provides distribution-free calibration guarantees that impose no assumptions on the functional form of the data-generating distribution. In contrast, Platt scaling approach of Gutman et al. (2022) relies on parametric assumptions, and the kernel-smoothing method of Deshpande and Kuleshov (2023) depends on smoothness assumptions and requires tuning of the kernel bandwidth parameter. Third, in the context of debiased machine learning, we combine cross-fitting and calibration in a novel, data-efficient manner. Specifically, we apply calibration once to cross-fitted propensity score estimators, unlike the approach of Ballinari and Bearth (2024), which repeats the calibration step across multiple data splits.

In our companion paper on calibrated debiased machine learning (van der Laan et al., 2024b), we establish the role of nuisance function calibration in constructing doubly robust asymptotically linear estimators, thereby providing not only doubly robust consistency but also facilitating doubly robust inference (e.g., confidence intervals and hypothesis tests). Notably, Theorem 4 in van der Laan et al. (2024b) shows that IPW estimators of linear functionals, such as the ATE, with weights obtained using our proposed IC-IPW procedure are asymptotically linear, even when the initial propensity score estimator is derived using flexible machine learning tools.

The outline of this manuscript is as follows. Our proposed IC-IPW algorithm is outlined in Section 2 and its theory is presented in Section 3. In Section 3.1, we give calibration and mean

square error guarantees for IC-IPW, and in Section 3.2 we theoretically illustrate the benefits of IC-IPW for ATE estimation using AIPW estimator. Finally, Section 4 presents simulation experiments to evaluate the estimator's performance under different levels of overlap.

## 2. Methodology: stabilized inverse weighting via isotonic calibration

### 2.1. Notation

Suppose we observe $n$ independent and identically distributed observations, $O_1, \ldots, O_n$, of the data structure $O := (W, A, Y)$ drawn from a probability distribution $P_0$. In this data structure, $W \in \mathbb{R}^d$ is a vector of baseline covariates, $A \in \mathcal{A} \subset \mathbb{R}$ is a discrete treatment assignment taking values in $\mathcal{A}$, and $Y \in \mathbb{R}$ is a bounded outcome. Without loss of generality, we assume $\{0, 1\} \subset \mathcal{A}$, where $\{0\}$ represents a reference treatment level and $\{1\}$ a treatment level of interest. For a given distribution $P$ and realization $(a, w)$ of $(A, W)$, we denote the (generalized) propensity score by $\pi_P(a \mid w) := P(A = a \mid W = w)$ and the outcome regression by $\mu_P(a, w) := E_P[Y \mid A = a, W = w]$. We also denote the inverse propensity score by $\alpha_P(a, w) := 1/\pi_P(a \mid w)$ and the conditional average treatment effect (CATE) function by $\tau_P(w) := \mu_P(1, w) - \mu_P(0, w)$. To simplify notation, we write $S_0$ for summaries $S_{P_0}$ of the true distribution $P_0$. Throughout this text, for a function $f$, we use $f^{-1}$ to denote the reciprocal function, defined pointwise as $f^{-1}(x) := \frac{1}{f(x)}$. All essential suprema are taken with respect to $P_0$ (or its marginal distributions).

### 2.2. Balance-based calibration for inverse propensity weights

For a treatment level $a_0 \in \mathcal{A}$, let $\pi_n(a_0 \mid \cdot)$ be an arbitrary estimator of $\pi_0(a_0 \mid \cdot)$, obtained using, for example, flexible statistical learning tools. Specifically, $\pi_n(a_0 \mid \cdot)$ could be derived by nonparametrically regressing the treatment indicator $1(A = a_0)$ on covariates. Alternatively, we could set $\pi_n(a_0 \mid \cdot) := \alpha_n^{-1}(a_0 \mid \cdot)$, where $\alpha_n(a_0 \mid \cdot)$ is an estimator of $\pi_0^{-1}(a_0 \mid \cdot)$ obtained using loss functions for inverse propensity weights (Zhao, 2019; Chernozhukov et al., 2021).

We say an estimator $\pi_n^{-1}(a_0 \mid \cdot)$ of the inverse propensity score function $\pi_0^{-1}(a_0 \mid \cdot)$ is *perfectly calibrated* for covariate balance if, for all functions $h : \mathbb{R} \to \mathbb{R}$ in $L^2(P_0)$,

$$\int (h \circ \pi_n)(a_0 \mid w) \left\{ \frac{\pi_0(a_0 \mid w)}{\pi_n(a_0 \mid w)} - 1 \right\} dP_0(w) = 0.$$

In words, a perfectly calibrated estimator $\pi_n^{-1}(a_0 \mid \cdot)$ achieves covariate balance, in the sense of Imai and Ratkovic (2014) and Ben-Michael et al. (2021), within subgroups defined by levels of estimated values of the propensity score. The estimator $\pi_n^{-1}(a_0 \mid \cdot)$ is perfectly calibrated for covariate balance if and only if $P_0$-almost surely $\pi_n(a_0 \mid W) = \gamma_0^{(a_0)}(W, \pi_n)$, where, for any function $\pi$, we define the *calibration function* $w \mapsto \gamma_0^{(a_0)}(w, \pi) := E_0[\pi_0(a_0 \mid W) \mid \pi(a_0 \mid W) = \pi(a_0 \mid w)]$.

As a measure of deviation from perfect calibration, we consider the following $\chi^2$-squared divergence measure (Pearson, 1900) of calibration error for inverse propensity weights:

$$\text{CAL}^{(a_0)}(\pi^{-1}) := \int \left\{ \frac{\gamma_0^{(a_0)}(w, \pi)}{\pi(a_0 \mid w)} - 1 \right\}^2 dP_0(w). \tag{1}$$

Intuitively, for an inverse propensity score estimator $\pi_n^{-1}(a_0 \mid \cdot)$ that is well-calibrated in the sense that $\text{CAL}^{(a_0)}(\pi_n^{-1})$ is small, a large inverse propensity weight should only be assigned to an individual when the true propensity score $\pi_0(a_0 \mid \cdot)$ is, on average, small among individuals with that

weight. Consequently, in scenarios with limited overlap, a well-calibrated estimator should avoid assigning excessively large weights to observations. Notably, Deshpande and Kuleshov (2023) established that calibration with respect to this $\chi^2$-divergence measure leads directly to smaller estimation errors for IPW and AIPW estimators of the ATE (see Lemma 3.3 and Theorem 3.5 of Deshpande and Kuleshov (2023)). Therefore, developing calibration procedures that guarantee inverse propensity score estimators with small calibration error under this measure is of great interest for causal inference.

We note that (1) differs from the $\ell^2$-integrated calibration error for the propensity score, commonly used in probability calibration (Gupta and Ramdas, 2021), which is given by

$$\int \{\gamma_0(a_0)(w, \pi) - \pi(a_0 \mid w)\}^2 \, dP_0(w). \tag{2}$$

Notably, a propensity score estimator $\pi_n$ that is well-calibrated with respect to this calibration measure may not lead to well-calibrated inverse propensity weights in the sense of (1). This discrepancy arises because the calibration error of (2) is insensitive to scenarios with limited overlap, where the inverse propensity score $\pi_n^{-1}(a_0 \mid w)$ may take large values. Distribution-free calibration guarantees for isotonic calibration of regression functions (such as the propensity score) with respect to this measure are a consequence of Van Der Laan et al. (2023).

## 2.3. Isotonic calibrated inverse probability weighting

In this section, we introduce our novel calibration algorithm, IC-IPW, which uses isotonic regression to construct inverse propensity weights calibrated according to the definition in (1).

IC-IPW is motivated by the observation that inversion strategies, such as propensity score truncation, commonly apply a nonincreasing monotone transformation to a propensity score estimator $\pi_n(a_0 \mid \cdot)$ to derive an inverse propensity score estimator. The simplest example of this is the estimator $\pi_n^{-1}(a_0 \mid \cdot)$, corresponding to the inversion map $x \mapsto x^{-1}$. Another option is the truncated estimator given by $\{(c_n \vee \pi_n(a_0 \mid \cdot)) \wedge (1 - c_n)\}^{-1}$, where $c_n \in (0, 1/2]$ is a possibly data-adaptive threshold, corresponding to the transformation $x \mapsto \{(c_n \vee x) \wedge (1 - c_n)\}^{-1}$. The optimal nonincreasing transformation $f_{n,0}^{(a_0)} : \mathbb{R} \to \mathbb{R}$ of $\pi_n(a_0 \mid \cdot)$, which minimizes the mean squared error for $(w, a) \mapsto 1(a = a_0)\pi_0^{-1}(a_0 \mid \cdot)$, is given by

$$f_{n,0}^{(a_0)} \in \underset{\theta \in \mathcal{F}_{\text{anti}}}{\operatorname{argmin}} \, E_0 \big[ 1(A = a_0)\{(\theta \circ \pi_n)(a_0 \mid W) - \pi_0^{-1}(a_0 \mid W)\}^2 \big]$$
$$= \underset{\theta \in \mathcal{F}_{\text{anti}}}{\operatorname{argmin}} \, E_0 \big[ 1(A = a_0)(\theta \circ \pi_n)(a_0 \mid W)^2 - 2(\theta \circ \pi_n)(a_0 \mid W) \big].$$

Notably, the mapping $(w, a) \mapsto f_{n,0}^{(a)} \circ \pi_n(a \mid w)$ is a mean-squared optimal estimator of the inverse propensity score $(w, a) \mapsto \pi_0^{-1}(a \mid w)$.

Our proposed IC-IPW estimator of $\pi_0^{-1}(a_0 \mid \cdot)$ is given by $\alpha_n^*(a_0 \mid \cdot) = (f_n^{(a_0)} \circ \pi_n)(a_0 \mid \cdot)$ for a data-dependent monotone nonincreasing transformation $f_n^{(a_0)} : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ obtained via antitonic empirical risk minimization as:

$$f_n^{(a_0)} \in \underset{\theta \in \mathcal{F}_{\text{anti}}}{\operatorname{argmin}} \sum_{i=1}^{n} \big[ 1(A_i = a_0)(\theta \circ \pi_n)^2(a_0 \mid W_i) - 2(\theta \circ \pi_n)(a_0 \mid W_i) \big], \tag{3}$$

where $\mathcal{F}_{\text{anti}} := \{\theta : \mathbb{R} \to \mathbb{R}; \ \theta \text{ is monotone nonincreasing}\}$ denotes the space of antitonic functions. We follow Groeneboom and Lopuhaa (1993) in taking the unique càdlàg piecewise constant solution to the antitonic regression problem, which only has jumps at observed values in $\{\pi_n(a_0 \,|\, W_i) : i \in [n]\}$. The loss function $(o, \alpha) \mapsto 1(a = a_0)\alpha(a \,|\, w)^2 - 2\alpha(a_0 \,|\, w)$, used in (3), has been applied for learning balancing weights (Zhao, 2019) as well as for estimating the inverse propensity score (Chernozhukov et al., 2018b). However, to the best of our knowledge, this loss function has not been previously utilized for isotonic regression or calibration purposes.

The function $f_n^{(a_0)}$ may take an infinite value when there exists a boundary observation $i(1) \in [n]$ with $\pi_n(a_0 \,|\, W_{i(1)}) = \min_{i \in [n]}\{\pi_n(a_0 \,|\, W_i)\}$ such that $A_{i(1)} \neq a_0$. This is related to the fact that isotonic regression can exhibit poor behavior at the boundary of the input space (Groeneboom and Lopuhaa, 1993). In the context of causal inference and debiased machine learning, infinite weights among observations $i$ with $A_i \neq a_0$ are usually not problematic, as typically only observations in the relevant treatment arm with $A_i = a_0$ need to be reweighted using these weights. That is, only estimates of $\{\frac{1(A_i = a_0)}{\pi_0(a_0 \,|\, W_i)} : i \in [n]\}$ are required. Nonetheless, we can address this issue by applying a boundary correction that adaptively truncates $f_n^{(a_0)}$ to lie within the range $[1, b_n^{(a_0)}]$, where the data-dependent truncation level is given by the largest finite value, $b_n^{(a_0)} = \max_{i \in [n]:A_i = a_0} f_n^{(a_0)}(\pi_n(a_0 \,|\, W_i))$. This truncation level preserves the value of $1(A_i = a_0)f_n^{(a_0)}(\pi_n(a_0 \,|\, W_i))$ for all $i \in [n]$, while ensuring that the weight estimates $f_n^{(a_0)}(\pi_n(a_0 \,|\, W_i))$ with $A_i \neq a_0$ remain finite. Alternatively, one can constrain the minimum number of observations in each constant segment of the antitonic solution, ensuring that at least one treated observation lies in each level set of $f_n^{(a_0)}$. We provide R and Python code implementing both the truncation and this constraint in Appendix B, leveraging a constrained implementation of isotonic regression in xgboost (Chen and Guestrin, 2016).

The first-order conditions characterizing the minimizing solution imply that the proposed estimator $\alpha_n^*(a_0 \,|\, \cdot)$ solves, for any map $h : \mathbb{R} \cup \{\infty\} \to \mathbb{R}$ with $h(\infty) := 0$, the covariate balancing equation:

$$\frac{1}{n}\sum_{i=1}^{n}(h \circ \alpha_n^*)(a_0 \,|\, W_i)\left\{1(A_i = a_0)\alpha_n^*(a_0 \,|\, W_i) - 1\right\} = 0. \tag{4}$$

Hence, the calibrated weight estimator $\alpha_n^*(a_0 \,|\, \cdot)$ achieves exact empirical covariate balance within levels of the estimated weights. Notably, by choosing $h$ as a level set indicator, we can show that $\frac{1}{N_t}\sum_{i \in I_t}1(A_i = a_0)\alpha_n^*(a_0 \,|\, W_i) = 1$ for all $t \in \mathbb{R}$, where $N_t = \sum_{i=1}^{n}1(\alpha_n^*(a_0 \,|\, W_i) = t)$ and $I_t = \{i : \alpha_n^*(a_0 \,|\, W_i) = t\}$. This indicates that the calibrated weight estimators are automatically stabilized among level sets of the estimated weights (Robins et al., 2000). As a consequence, the IC-IPW estimator is piecewise constant, where each unique value corresponds to the inverse of an empirical mean over a subset of $\{1(A_1 = a_0), \ldots, 1(A_n = a_0)\}$. Specifically, for each observation index $i_0 \in [n]$, there exist endpoints $\ell_0 \leq m_0$ that depend on the sample size $n$, such that

$$\alpha_n^*(a_0 \,|\, W_{i_0}) = \frac{m_0 - \ell_0 + 1}{\sum_{\ell_0 \leq r \leq m_0}1(A_{i(r)} = a_0)},$$

where $(i(r) : r \in [n])$ is an appropriate permutation of $[n]$.

We can show that the antitonic transformation satisfies $f_n^{(a_0)} = 1/g_n^{(a_0)}$, where $g_n^{(a_0)}$ is the isotonic calibrator of the propensity score $\pi_0(a_0 \,|\, \cdot)$ obtained via isotonic regression as:

$$g_n^{(a_0)} \in \operatorname*{argmin}_{\phi \in \mathcal{F}_{\text{iso}}}\sum_{i=1}^{n}\left\{1(A_i = a_0) - (\phi \circ \pi_n)(a_0 \,|\, X_i)\right\}^2, \tag{5}$$

where $\mathcal{F}_{\text{iso}} := \{\theta : \mathbb{R} \to \mathbb{R}; \ \theta \text{ is monotone nondecreasing}\}$ denotes the space of isotonic functions. Thus, the optimization problem given by (3) can be computed via isotonic probability calibration for $\pi_0(a_0 \mid \cdot)$ using $1(A = a_0)$ as the outcome (Niculescu-Mizil and Caruana, 2005).

It is important to emphasize that even in the case of a binary treatment, $A \in \{0, 1\}$, the calibration step in (5) must be performed independently for each treatment level $a_0 \in \mathcal{A}$ to achieve well-calibrated inverse propensity score estimates. Notably, Ballinari and Bearth (2024) assess an isotonic calibration method that does not follow this approach. Instead, they calibrate the propensity score estimator for the treatment arm, $\pi_n(1 \mid \cdot)$, using isotonic calibration, and then subtract 1 to derive a "calibrated" estimator for $\pi_n(0 \mid \cdot)$. This subtraction can lead to estimated weights in the control arm with infinite values, as the calibrated probabilities for observations in the control arm may be 0. Consequently, in the context of AIPW estimation, Ballinari and Bearth (2024) concluded that isotonic calibration of the propensity score estimator $\pi_n(1 \mid \cdot)$ cannot be used due to infinite weights. Our approach (Algorithm 1) avoids the issue of infinite weights in the AIPW estimator by calibrating the propensity score estimators $\pi_n(1 \mid \cdot)$ and $\pi_n(0 \mid \cdot)$ separately for the treatment and control arms. This finding highlights the advantage of calibration algorithms that target inverse scores directly, as focusing on propensity scores alone may not yield well-calibrated inverse scores.

---

**Algorithm 1** Cross-fitted Isotonic-Calibrated Inverse Probability Weighting

---

**Require:** dataset $\mathcal{D}_n = \{O_i : i \in [n]\}$, # of cross-fitting splits $k$, $a_0 \in \mathcal{A}$

1: partition $\mathcal{D}_n$ into datasets $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \ldots, \mathcal{C}^{(J)}$;

2: **for** $s = 1, 2, \ldots, J$ **do**

3:     let $j(i) = s$ for each $i \in \mathcal{C}^{(s)}$;

4:     get initial estimator $\pi_{n,s}$ of $\pi_0$ from $\mathcal{E}^{(s)} := \mathcal{D}_n \backslash \mathcal{C}^{(s)}$.

5: **end for**

6: set $f_n^{(a_0)} := 1/\{c_n^{(a_0)} \vee g_n^{(a_0)}\}$ where $c_n^{(a_0)} = \min_{i \in [n]: A_i = a_0} g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i))$ and $g_n^{(a_0)}$ is obtained using isotonic regression as

$$g_n^{(a_0)} \in \underset{\phi \in \mathcal{F}_{\text{iso}}}{\arg\min} \sum_{i=1}^{n} \left\{ 1(A_i = a_0) - (\phi \circ \pi_{n,j(i)})(a_0 \mid W_i) \right\}^2.$$

7: set $\alpha_{n,j}^*(a_0 \mid \cdot) := f_n^{(a_0)}(\pi_{n,j}(a_0 \mid \cdot))$ for each $j \in [J]$;

8: **return** weight functions $\{\alpha_{n,j}^*(a_0 \mid \cdot) : j \in [J]\}$ and estimates $\{\alpha_{n,j(i)}^*(a_0 \mid W_i) : i \in [n]\}$

---

In causal inference using debiased and targeted machine learning, cross-fitting inverse propensity weights is a common strategy to relax conditions on estimator complexity and mitigate overfitting (van der Laan et al., 2011; Chernozhukov et al., 2018a). Similarly, to ensure calibration guarantees under weak conditions, we recommend applying cross-fitting when fitting the initial propensity score estimator. Algorithm 1 outlines the procedure for IC-IPW with cross-fitting and an adaptive truncation-based boundary correction. Specifically, it involves partitioning the available data into $J$ folds, where each fold computes an initial propensity score estimator $\pi_{n,j}(a_0 \mid \cdot)$ using only the data from the complementary folds. The resulting out-of-fold estimates from the $k$ estimators are then combined to obtain $n$ propensity estimates for each observation. Akin to (5), these out-of-fold estimates can be used to learn a single nonincreasing transformation $f_n^{(a_0)} : \mathbb{R} \to \mathbb{R}$ such that $\{f_n^{(a_0)} \circ \pi_{n,j}(a_0 \mid \cdot) : j \in [J]\}$ are isotonic calibrated cross-fitted estimators. We note that

the isotonic fitting step of Algorithm 1 is deliberately not cross-fitted, so that all available data is used to calibrate the inverse weight estimates.

## 3. Theoretical properties

### 3.1. Calibration and estimation error guarantees

We will now demonstrate that the $\chi^2$-calibration error, as defined in Section 2.2, of our cross-fitted IC-IPW estimators proposed in Algorithm 1 tends to zero at a fast rate. Furthermore, we will show that IC-IPW generally improves the mean square error for the inverse propensity weights compared to the uncalibrated cross-fitted weights. Consequently, in the pursuit of obtaining calibrated weight estimates, IC-IPW typically does not harm the performance of the user-supplied weight estimators.

We will make use of the following conditions. For each $j \in [J]$, we denote the fold-specific IC-IPW estimator $\alpha_{n,j}^*$ by the mapping $(w, a) \mapsto f_n^{(a)} \circ \pi_{n,j}(a \mid w)$, where $f_n^{(a)}$ is the calibrator obtained from Algorithm 1.

**(C1)** *Estimator boundedness:* There exists $M < \infty$ such that, for all $j \in [J]$, $\operatorname{ess\,sup}_{a,w} |\pi_{n,j}(a \mid w)| < M$ with probability tending to one.

**(C2)** *Positivity:* There exists $0 < \eta < 1$ such that, for all $j \in [J]$, $\operatorname{ess\,sup}_{a,w} |\alpha_{n,j}^*(a \mid w)| < \eta^{-1}$ and $\operatorname{ess\,sup}_{a,w} |\pi_0^{-1}(a \mid w)| < \eta^{-1}$ with probability tending to one.

**(C3)** *Limited boundary observations:* $\#\{i \in [n] : A_i \neq a_0 \text{ and } \pi_{n,j(i)}(a_0 \mid W_i) < \min_{k \in [n]: A_k = a_0} \pi_{n,j(i)}(a_0 \mid W_k)\} = O_p(n^{1/3})$.

**(C4)** *Calibration function has finite total variation:* There exists an $M < \infty$ such that, for all $j \in [J]$, $n \in \mathbb{N}$, and $a_0 \in \mathcal{A}$, the function $g_{0,n,j}^{(a_0)} : \mathbb{R} \to \mathbb{R}$, defined by $g_{0,n,j}^{(a_0)} \circ \pi_{n,j} = \gamma_0^{(a_0)}(\cdot, \pi_{n,j})$, has a total variation norm that is almost surely bounded by $M$.

The following theorem establishes bounds for the calibration error of the IC-IPW estimator, and its proof follows from the empirical balancing property given by Equation (4).

**Theorem 1 (Calibration error)** *Under Conditions C1-C4, for each $a_0 \in \mathcal{A}$, it holds*

$$\frac{1}{J} \sum_{i=1}^{J} \operatorname{CAL}^{(a_0)}(\alpha_{n,j}^*) = O_p(n^{-2/3}).$$

Theorem 1 provides distribution-free calibration guarantees for the IC-IPW estimator under weak conditions. The calibration error tends to zero at the rate of $n^{-2/3}$ irrespective of the smoothness of the inverse propensity score or dimension of the covariate vector. Is it interesting to note that IP weight estimators that are consistent in mean squared error are typically asymptotically calibrated. However, particularly in settings with high-dimensional covariates or nonsmooth inverse propensity scores, the calibration error rate of such estimators can be arbitrarily slow.

Condition C1 is automatically satisfied when the initial propensity score estimator $\pi_{n,j}$ takes values in the range $[0, 1]$. Condition C2 is standard in causal inference and requires that all individuals have a positive probability of being assigned to either treatment or control. Condition C2

also ensures that the calibrated inverse propensity weights are uniformly bounded with probability approaching one, which holds as long as the truncation level $c_n$ in Alg. 1 remains uniformly bounded away from zero, a condition we expect to hold under C2. Condition C3 is a mild regularity condition that we expect to be satisfied under C2, since, for a fixed function $\pi(a_0 \mid \cdot)$, by properties of the minimum order statistic, $\min_{k \in [n]: A_k = a_0} \pi(a_0 \mid W_k)$ should converge in probability to $\operatorname{ess\,inf}_w \pi(a_0 \mid w)$ at a rate of $n^{-1}$ when $\pi(a_0 \mid W)$ is a continuous random variable, or equal $\operatorname{ess\,inf}_w \pi(a_0 \mid w)$ with probability tending to one when it is discrete. Condition C4 excludes cases in which the best possible predictor of the propensity score $\pi_0$, given only the initial propensity score estimator $\pi_{n,j}$, has pathological behavior, in that it has infinite variation norm as a (univariate) mapping of $\pi_{n,j}$. We stress here that isotonic regression is used only as a tool for calibration, and our theoretical guarantees do not require any monotonicity on components of the data-generating mechanism — for example, $\gamma_0(w, \pi_{n,j})$ need not be monotone as a function of $\pi_{n,j}(w)$.

The following theorem establishes that IC-IPW not only constructs calibrated weight estimates but can also reduce the mean squared error of the inverse propensity weights compared to uncalibrated cross-fitted weights. In the following, we define the optimal nondecreasing transformations of the initial cross-fitted propensity score estimators as:

$$\left\{ f_{n,0}^{(a_0)} : a_0 \in \mathcal{A} \right\} \in \operatorname*{argmin}_{\{f^{(a)}: a \in \mathcal{A}\} \subset \mathcal{F}_{\mathrm{anti}}} \sum_{j \in [J]} \int \left\{ f^{(a)} \circ \pi_{n,j}(a \mid w) - \pi_0^{-1}(a \mid w) \right\}^2 dP_0(a, w).$$

**Theorem 2 (Mean square error)** *Assume Conditions C1-C3 holds. As $n \to \infty$,*

$$\sum_{j \in [J]} \int \left\{ \alpha_{n,j}^*(a \mid w) - f_{n,0}^{(a)} \circ \pi_{n,j}(a \mid w) \right\}^2 dP_0(a, w) = O_p(n^{-2/3}).$$

*As a consequence, for any nondecreasing transformations $\{f^{(a)} \in \mathcal{F}_{\mathrm{anti}} : a \in \mathcal{A}\}$,*

$$\frac{1}{J} \sum_{j \in [J]} \| \alpha_{n,j}^* - \pi_0^{-1} \|^2 \leq O_p(n^{-2/3}) + \frac{1}{J} \sum_{j \in [J]} \int \left\{ f^{(a)} \circ \pi_{n,j}(a \mid w) - \pi_0^{-1}(a \mid w) \right\}^2 dP_0(a, w).$$

An important implication of Theorem 2 is that the data-dependent transformations $\{ f_n^{(a_0)} \in \mathcal{F}_{\mathrm{anti}} : a_0 \in \mathcal{A} \}$ are nearly mean square error optimal among all nondecreasing transformations of the propensity score estimators $\{ \pi_{n,j}(a_0 \mid \cdot) : a_0 \in \mathcal{A} \}$, up to an error term of $O_p(n^{-2/3})$. This result suggests that the IC-IPW estimator performs at least as well as, and potentially better than, $\pi_{n,j}^{-1}$ and its truncated variants.

### 3.2. Isotonic calibration and AIPW estimation of the ATE

In this section, we demonstrate the advantages of IC-IPW by examining isotonic-calibrated AIPW estimators for the ATE (Robins et al., 1994; Bang and Robins, 2005), showing how IC-IPW potentially relaxes the assumptions needed for asymptotic linearity and efficiency.

Let $\mu_{n,j}$ and $\pi_{n,j}$, for $j \in [J]$, be initial cross-fitted estimators for the outcome regression $\mu_0$ and the propensity score $\pi_0$. Let $\alpha_{n,j}^*$, also indexed by $j \in [J]$, denote the cross-fitted IC-IPW estimators of the inverse propensity score $\pi_0^{-1}$, obtained from $\pi_{n,1}, \ldots, \pi_{n,J}$ via Algorithm 1. Our cross-fitted IC-AIPW estimator for the ATE, $\psi_0 = E_0[\tau_0(W)]$, is defined as

$$\psi_n^* := \frac{1}{n} \sum_{i=1}^n \left[ \mu_{n,j(i)}(1, W_i) - \mu_{n,j(i)}(0, W_i) + \delta_{\pm}(A_i) \alpha_{n,j(i)}^*(A_i \mid W_i) \{ Y_i - \mu_{n,j(i)}(A_i, W_i) \} \right],$$

where $j(i) = s$ if observation $O_i$ belongs to the training fold $\mathcal{C}^{(s)}$, and $\delta_\pm(a) = 1(a = 1) - 1(a = 0)$ for $a \in \{0, 1\}$. Let $\varphi_0 : o = (w, a, y) \mapsto \mu_0(1, w) - \mu_0(0, w) - \psi_0 + \left\{\frac{1(a=1)}{\pi_0(1|w)} - \frac{1(a=0)}{\pi_0(0|w)}\right\}(y - \mu_0(a, w))$ denote the $P_0$-efficient influence function of the ATE parameter. The next theorem establishes that IC-AIPW provides valid inference for the ATE.

**(D1)** *Bounded estimator:* $\mu_{n,j}(A, W)$ is almost surely bounded with probability tending to one for each $j \in [J]$.

**(D2)** For each $a_0 \in \{0, 1\}$ and $j \in [J]$, the following hold:

(a) *Consistency of oracle transformation:* $\|f_{n,0}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, \cdot) - \pi_0^{-1}(a_0 \,|\, \cdot)\|_{P_0} = o_p(1)$.

(b) *Outcome regression rate:* $\|\mu_{n,j} - \mu_0\|_{P_0} = o_p(n^{-1/6})$.

(c) *Doubly robust rate:* $\|f_{n,0}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, \cdot) - \pi_0^{-1}(a_0 \,|\, \cdot)\|_{P_0} \times \|\mu_{n,j} - \mu_0\|_{P_0} = o_p(n^{-1/2})$.

**Theorem 3 (Efficiency for ATE)** *Under Conditions C1-C2 and D1-D2, the IC-AIPW estimator $\psi_n^*$ satisfies $\psi_n^* = \psi_0 + (P_n - P_0)\varphi_0 + o_p(n^{-1/2})$ and is, thus, regular, asymptotically linear, and nonparametric efficient. As a consequence, $\sqrt{n}(\psi_n^* - \psi_0) \xrightarrow{\mathsf{d}} \mathrm{N}(0, \sigma_0^2)$ with $\sigma_0^2 := \mathrm{Var}_0[\varphi_0(O)]$.*

Conditions D2a and D2b ensure the consistency of $\alpha_{n,j}^*$ and $\mu_{n,j}$ as estimators. Notably, Condition D2a only requires consistency of the best nondecreasing transformation of the initial inverse propensity score estimators $\{\pi_{n,j}^{-1} : j \in [J]\}$. On the other hand, D2b requires the outcome regression estimation rate to be faster than $n^{-1/6}$. Condition D2c imposes a doubly robust rate condition on the nuisance estimators, only requiring a rate for the best nondecreasing transformation of the initial inverse propensity score estimators. By Theorem 2, D2b and D2c together imply the standard doubly robust rate condition $\|\alpha_{n,j}^* - \alpha_0\|_{P_0}\|\mu_{n,j} - \mu_0\|_{P_0} = o_p(n^{-1/2})$ for each $j \in [J]$. Theorem 3 imposes weaker conditions compared to previous literature, as it allows the initial estimator of the inverse propensity score to be incorrectly specified by an arbitrary monotone transformation.

## 4. Experiments

We evaluate the empirical performance of IC-IPW in semi-synthetic experiments to estimate the ATE in scenarios with near positivity violations. Specifically, we assess the performance of AIPW with weights obtained via IC-IPW against several competing methods: naive inversion, propensity score trimming, Platt scaling, and direct learning approaches. Our analysis uses semi-synthetic data from the ACIC-2017 competition (Hahn et al., 2019), with covariates from the Infant Health and Development Program (Brooks-Gunn et al., 1992). Outcomes are generated from 32 distinct data-generating processes, focusing on those indexed from 17 to 24, which feature uncorrelated errors. Each process produces $M = 250$ replicated datasets, each containing $n = 4302$ samples.

For ATE estimation, we construct the IC-AIPW estimator using the calibrated inverse propensity weights established by Algorithm 1. We implement two variants of propensity score trimming: a deterministic method that truncates into $[0.01, 0.99]$, and an adaptive variant that learns the truncation threshold to minimize an empirical risk criterion. For indirect learning methods, we estimate cross-fitted propensity scores using gradient-boosted logistic regression in `xgboost`. The direct

learning method estimates the inverse propensity score directly using `xgboost` and the loss function $(o, \alpha) \mapsto 1(a = a_0)\alpha(a \mid w)^2 - 2\alpha(a_0 \mid w)$. As a parametric calibration method, we implement Platt scaling of the propensity scores as described in (Gutman et al., 2022). Moreover, we implement the targeted maximum likelihood estimation (TMLE) and TMLE with isotonic calibrated inverse probability weighting (IC-TMLE). Additional experimental details are given in Appendix A.3.

For each data-generating process, Table 1 and Table 2 (in Appendix A.3) display the Monte Carlo estimates of the bias, standard error (SE), and root-mean-square error (RMSE), as well as $95\%$ confidence interval coverage. In Table 1, for data-generating processes $(18, 20, 22, 24)$, where the overlap between the treatment arms is limited, we find that IC-IPW significantly reduces estimation bias and RMSE, while substantially improving the accuracy of $95\%$ coverage. For data-generating processes $(17, 19, 21, 23)$, Table 2 in Appendix A.3 shows that, when overlap is less of an issue, IC-IPW performs comparably to other estimation methods regarding bias, RMSE, and coverage, with the direct learning method generally performing best.

| Method | Setting 18 | | | | Setting 20 | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | SE | RMSE | Cov. | Bias | SE | RMSE | Cov. |
| Inversion | 0.17 | 0.072 | 0.18 | 0.25 | 0.17 | 0.13 | 0.22 | 0.56 |
| Direct Learning | 0.18 | 0.072 | 0.19 | 0.14 | 0.21 | 0.13 | 0.24 | 0.70 |
| Trimming | 0.23 | 0.039 | 0.24 | 0.00 | 0.24 | 0.12 | 0.26 | 0.21 |
| Adaptive Trim. | 0.17 | 0.072 | 0.18 | 0.25 | 0.17 | 0.13 | 0.22 | 0.56 |
| Platt Scaling | 0.088 | 0.047 | 0.10 | 0.68 | 0.087 | 0.11 | **0.14** | 0.87 |
| **IC-AIPW** | 0.045 | 0.056 | **0.072** | **0.95** | 0.035 | 0.17 | 0.18 | **0.92** |
| TMLE | 0.078 | 0.062 | 0.10 | 0.66 | 0.077 | 0.13 | 0.15 | 0.82 |
| **IC-TMLE** | **0.016** | 0.053 | 0.056 | 1.00 | **0.030** | 0.18 | 0.18 | **0.92** |
| Dropping | 0.20 | 0.077 | 0.21 | 0.21 | 0.20 | 0.13 | 0.24 | 0.53 |
| Method | Setting 22 | | | | Setting 24 | | | |
| | Bias | SE | RMSE | Cov. | Bias | SE | RMSE | Cov. |
| Inversion | 0.21 | 0.087 | 0.22 | 0.18 | 0.22 | 0.13 | 0.25 | 0.53 |
| Direct Learning | 0.13 | 0.081 | 0.14 | 0.62 | 0.18 | 0.13 | 0.22 | 0.81 |
| Trimming | 0.27 | 0.097 | 0.28 | 0.00 | 0.29 | 0.14 | 0.30 | 0.16 |
| Adaptive Trim. | 0.21 | 0.087 | 0.22 | 0.18 | 0.22 | 0.13 | 0.25 | 0.53 |
| Platt Scaling | 0.14 | 0.087 | 0.14 | 0.32 | 0.13 | 0.12 | **0.17** | 0.84 |
| **IC-AIPW** | 0.097 | 0.076 | 0.12 | 0.79 | 0.068 | 0.21 | 0.22 | 0.92 |
| TMLE | 0.12 | 0.063 | 0.14 | 0.42 | 0.12 | 0.14 | 0.18 | 0.74 |
| **IC-TMLE** | **0.039** | 0.061 | **0.073** | **0.95** | **0.005** | 0.22 | 0.22 | **0.93** |
| Dropping | 0.23 | 0.076 | 0.24 | 0.17 | 0.23 | 0.14 | 0.27 | 0.52 |

Table 1: Simulation results for estimating ATE using ACIC 2017 data with data-generating processes where overlap between treatment arms is limited. Our method is in bold.

## 5. Conclusion

We propose a $\chi^2$-divergence measure of calibration error for inverse propensity score estimators and introduce a post-hoc algorithm that minimizes calibration error using isotonic regression with

a tailored loss function for inverse propensity weights. We further emphasize the importance of developing methods that calibrate the inverse propensity score directly, as calibrating the propensity score may not yield well-calibrated inverse weights. While our primary focus is on isotonic regression for calibration, this approach can be extended to alternative methods, such as parametric scaling, histogram binning, kernel smoothing, and Venn-Abers calibration (Vovk and Petej, 2012; van der Laan and Alaa, 2024, 2025). Additionally, although we emphasize the calibration of inverse propensity weights, this method can also be applied to general inverse probability weights, including those used for handling missing data or censoring. Another interesting extension would be to study the use of the efficient plug-in learning framework (van der Laan et al., 2024a) to develop risk estimators that mitigate the poor boundary behavior of isotonic regression. Finally, as an adaptive histogram regressor, isotonic calibration performs automatic binning of propensity score estimates, which can be used to define subgroups with differing propensity scores for downstream analysis, such as propensity score matching (Xu and Otsu, 2022). Exploring these applications is an interesting direction for future work.

## Acknowledgments

## References

Daniele Ballinari and Nora Bearth. Improving the finite sample performance of double/debiased machine learning with propensity score calibration, 2024. URL https://arxiv.org/abs/2409.04874.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Richard E Barlow and Hugh D Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

Oliver Bembom and Mark J van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. 2008.

Eli Ben-Michael, Avi Feller, David A Hirshberg, and José R Zubizarreta. The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*, 2021.

Jeanne Brooks-Gunn, Fong-ruey Liaw, and Pamela Kato Klebanov. Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of pediatrics*, 120(3): 350–359, 1992.

David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. Augmented balancing weights as linear regression. *arXiv preprint arXiv:2304.14545*, 2023.

Tianqi Chen and Carlos 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.

V. Chernozhukov, Chetverikov D., M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21:C1–C68, 2018a. doi: 10.1111/ectj.12097.

Victor Chernozhukov, Whitney K Newey, and James Robins. Double/de-biased machine learning using regularized riesz representers. Technical report, cemmap working paper, 2018b.

Victor Chernozhukov, Whitney K Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv preprint arXiv:2104.14737*, 2021.

Victor Chernozhukov, Whitney Newey, Victor M Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022a.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022b.

Stephen R Cole and Miguel A Hernán. Constructing inverse probability weights for marginal structural models. *American journal of epidemiology*, 168(6):656–664, 2008.

Jeremy R Coyle, Nima S Hejazi, Ivana Malenica, Rachael V Phillips, and Oleg 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.

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

Shachi Deshpande and Volodymyr Kuleshov. Calibrated propensity scores for causal effect estimation. *arXiv preprint arXiv:2306.00382*, 2023.

Piet Groeneboom and HP Lopuhaa. Isotonic estimators of monotone densities and distribution functions: basic facts. *Statistica Neerlandica*, 47(3):175–183, 1993.

Susan Gruber, Rachael V Phillips, Hana Lee, and Mark J van der Laan. Data-adaptive selection of the propensity score truncation level for inverse-probability–weighted and targeted maximum likelihood estimators of marginal point treatment effects. *American Journal of Epidemiology*, 191(9):1640–1651, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Chirag Gupta and Aaditya K Ramdas. Distribution-free calibration guarantees for histogram binning without sample splitting. *arXiv preprint arXiv:2105.04656*, 2021.

Rom Gutman, Ehud Karavani, and Yishai Shimoni. Propensity score models are better when post-calibrated. *arXiv preprint arXiv:2211.01221*, 2022.

P Richard Hahn, Vincent Dorie, and Jared S Murray. Atlantic causal inference conference (acic) data analysis challenge 2017. *arXiv preprint arXiv:1905.09515*, 2019.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

David A Hirshberg and Stefan Wager. Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227, 2021.

David A Hirshberg, Arian Maleki, and Jose R Zubizarreta. Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*, 2019.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 243–263, 2014.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

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

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005.

Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J van der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Marginal structural models and causal inference in epidemiology. *Journal of the American statistical Association*, 89(427):846–866, 1994.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Lars van der Laan and Ahmed Alaa. Generalized venn and venn-abers calibration with applications in conformal prediction. *arXiv preprint arXiv:2502.05676*, 2025.

Lars van der Laan and Ahmed M. Alaa. Self-calibrating conformal prediction, 2024. URL https://arxiv.org/abs/2402.07307.

Lars Van Der Laan, Ernesto Ulloa-Pérez, Marco Carone, and Alex Luedtke. Causal isotonic calibration for heterogeneous treatment effects. In *International Conference on Machine Learning*, pages 34831–34854. PMLR, 2023.

Lars van der Laan, Marco Carone, and Alex Luedtke. Combining t-learning and dr-learning: a framework for oracle-efficient estimation of causal contrasts. *arXiv preprint arXiv:2402.01972*, 2024a.

Lars van der Laan, Alex Luedtke, and Marco Carone. Automatic doubly robust inference for linear functionals via calibrated debiased machine learning. *arXiv preprint arXiv:2411.02771*, 2024b.

Mark J van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

Mark J van der Laan, Sherri Rose, Wenjing Zheng, and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. *Targeted learning: causal inference for observational and experimental data*, pages 459–474, 2011.

Aad Van Der Vaart and Jon A Wellner. A local maximal inequality under uniform entropy. *Electronic Journal of Statistics*, 5(2011):192, 2011.

Adrianus Willem van der Vaart and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Yixin Wang and Jose R Zubizarreta. Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105, 2020.

Yue Wang, Maya L Petersen, David Bangsberg, and Mark J van der Laan. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. 2006.

Mengshan Xu and Taisuke Otsu. Isotonic propensity score matching. *arXiv preprint arXiv:2207.08868*, 2022.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. 2019.

Yunji Zhou, Roland A Matsouaka, and Laine Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical methods in medical research*, 29(12):3721–3756, 2020.

José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

## Appendix A. Simulation studies

### A.1. Implementation of the estimation

Code to replicate our simulation results are available on https://drive.google.com/drive/folders/1Lh4rm0vrMaaSZVnlBtgur5bvodMUzu4z.

In our simulation studies, we estimated the nuisance parameters (i.e., $\mu_0$ and $\pi_0$) using the Super Learner (van der Laan et al., 2007), an ensemble learning approach that utilizes cross-validation to select the optimal combination of pre-specified prediction methods. This was implemented by the R package sl3 (Coyle et al., 2021). For the library of prediction methods, we employed a stack of extreme gradient boosting (XGBoost) (Chen and Guestrin, 2016) learners with parameters maximum depth $\in \{1, 2, 3, 4, 5, 6\}$, eta $\in \{0.15, 0.2, 0.25, 0.3\}$, minimum child weight to be 5, and max number of iterations to be 20. To perform the isotonic calibration step, we used either the R function isoreg or a monotone 1-dimensional XGBoost with a customized loss function; both methods yielded similar results so we only report the results obtained with the first approach.

We perform a $(0.01, 0.99)$ cutoff for the fixed trimming method. For the adaptive trimming method, we employ a $(c_n, 1 - c_n)$ truncation with

$$c_n := \underset{c \in [0, 1/2]}{\operatorname{argmin}} \sum_{i=1}^{n} \alpha_{n,c}(A_i, W_i)^2 - 2\{\alpha_{n,c}(1, W_i) - \alpha_{n,c}(0, W_i)\};$$

and $\alpha_{n,c}(a, w) := \frac{2a - 1}{\max\{\min\{\pi_n(a|w), 1-c\}, c\}}$ is the truncated estimator of $\alpha_0$.

### A.2. Performance metrics

We estimate the performance metrics as follows. For a ATE estimator $\hat{\tau}$, we use $\hat{\tau}_m$ to denote its estimated value on the $m$-th instantiated dataset $\{o_{m,i}\}_{i=1}^{n}$. Also, we use $\tau_m$ to denote the true ATE of this data-generating process at each instantiated dataset, evaluated by averaging the true CATE effect values at this dataset given in the ACIC data. That is, for each $m \in M$, $\tau_m := \frac{1}{n} \sum_{i=1}^{n} \tau_0(w_{m,i})$, where $\tau_0$ is the true CATE function. The Monte Carlo Bias, Standard Error and RMSE are calculated as follows:

$$\operatorname{Bias}(\hat{\tau}) := \frac{1}{M} \sum_{m=1}^{M} (\hat{\tau}_m - \tau_m),$$

$$\operatorname{RMSE}(\hat{\tau}) := \sqrt{\frac{1}{M} \sum_{m=1}^{M} (\hat{\tau}_m - \tau_m)^2},$$

$$\operatorname{SE}(\hat{\tau}) := \sqrt{\operatorname{RMSE}(\hat{\tau})^2 - \operatorname{Bias}(\hat{\tau})^2};$$

whereas the Monte Cargo 95% coverage is defined by:

$$\operatorname{C}(\hat{\tau}) := \frac{1}{M} \sum_{m=1}^{M} 1\big(\tau_m \in [\hat{l}_m, \hat{u}_m]\big),$$

where the $\{(\hat{l}_m, \hat{u}_m)\}_{m=1}^{M}$'s are the estimated 95% confidence lower bounds and upper bounds. For CATE estimation, with a slight abuse of notations, for an arbitrary estimator $\hat{\tau}(w)$, we the Monte

Carlo RMSE is estimated as:

$$\text{RMSE}(\hat{\tau}) \coloneqq \sqrt{\frac{1}{nM} \sum_{m=1}^{M} \sum_{i=1}^{n} \left(\hat{\tau}(w_{m,i}) - \tau_0(w_{m,i})\right)^2}.$$

For an arbitrary estimator $\hat{\alpha}(w)$ of the inversed propensity weight function $\pi_0^{-1}(w)$, we estimate its RMSE by

$$\text{RMSE}(\hat{\alpha}) \coloneqq \sqrt{\frac{1}{nK} \sum_{k=1}^{K} \sum_{i=1}^{n} \left(\hat{\alpha}(w_{m,i}) - \pi_0^{-1}(w_{k,i})\right)^2},$$

where we recall the time for iteration is $K = 1000$ and evaluation is done on a out-of-fold dataset.

### A.3. Experimental results for additional settings

| Method | Setting 17 | | | | Setting 19 | | | |
|---|---|---|---|---|---|---|---|---|
| | Bias | SE | RMSE | Cov. | Bias | SE | RMSE | Cov. |
| Inversion | 0.0068 | 0.011 | 0.013 | 0.91 | 0.0082 | 0.051 | 0.052 | 0.96 |
| Direct Learning | **0.0030** | 0.011 | **0.011** | 0.98 | **0.0024** | 0.056 | 0.056 | 0.96 |
| Trimming | 0.0068 | 0.011 | 0.013 | 0.91 | 0.0082 | 0.051 | 0.052 | 0.96 |
| Adaptive Trim. | 0.0072 | 0.011 | 0.013 | 0.89 | 0.0091 | 0.051 | 0.052 | 0.96 |
| Platt Scaling | 0.0070 | 0.011 | 0.013 | 0.91 | 0.0083 | 0.051 | 0.052 | 0.96 |
| **IC-AIPW** | 0.0062 | 0.011 | 0.013 | 0.92 | 0.0068 | 0.052 | 0.052 | **0.95** |
| TMLE | 0.0058 | 0.011 | 0.012 | **0.93** | 0.0045 | 0.054 | 0.055 | 0.95 |
| **IC-TMLE** | 0.0071 | 0.010 | 0.012 | 0.90 | 0.0050 | 0.053 | 0.055 | 0.96 |
| Dropping | 0.0065 | 0.011 | 0.013 | 0.92 | 0.0060 | 0.052 | 0.053 | 0.94 |
| Method | Setting 21 | | | | Setting 23 | | | |
| | Bias | SE | RMSE | Cov. | Bias | SE | RMSE | Cov. |
| Inversion | 0.0067 | 0.013 | 0.015 | 1.00 | 0.0110 | 0.060 | **0.061** | **0.95** |
| Direct Learning | **0.0046** | 0.013 | **0.014** | 1.00 | **0.0051** | 0.061 | 0.062 | 0.97 |
| Trimming | 0.0067 | 0.013 | 0.015 | 1.00 | 0.0110 | 0.060 | 0.061 | **0.95** |
| Adaptive Trim. | 0.0070 | 0.013 | 0.015 | 1.00 | 0.0120 | 0.060 | 0.061 | 0.96 |
| Platt Scaling | 0.0069 | 0.013 | 0.015 | 1.00 | 0.0120 | 0.060 | 0.061 | 0.96 |
| **IC-AIPW** | 0.0063 | 0.013 | 0.015 | 1.00 | 0.0100 | 0.059 | 0.061 | **0.95** |
| TMLE | 0.0059 | 0.013 | 0.014 | 1.00 | 0.0078 | 0.060 | 0.062 | 0.96 |
| **IC-TMLE** | 0.0055 | 0.013 | 0.014 | 1.00 | 0.0082 | 0.060 | 0.062 | 0.95 |
| Dropping | 0.0060 | 0.013 | 0.015 | 1.00 | 0.0090 | 0.060 | 0.062 | 0.94 |

Table 2: Simulation results for estimating ATE using ACIC 2017 data with data-generating processes where overlap between treatment arms is good. Our method is in bold.

## Appendix B. Implementation of IC-IPW

The following subsections provide R and Python code for constructing calibrates inverse probability weights using IC-IPW. Here, we implement isotonic regression using xgboost, allowing control over the maximum tree depth and the minimum number of observations in each constant segment of the isotonic regression fit.

### B.1. R code

```
# Function: isoreg_with_xgboost
# Purpose: Fits isotonic regression using XGBoost.
# Inputs:
#   - x: A vector or matrix of predictor variables.
#   - y: A vector of response variables.
#   - max_depth: Maximum depth of the trees in XGBoost (default = 15).
#   - min_child_weight: Minimum sum of instance weights (Hessian)
#        needed in a child node (default = 20).
# Returns:
#   - A function that takes a new predictor variable x
# and returns the model's predicted values.
isoreg_with_xgboost <- function(x, y, max_depth = 15, min_child_weight = 20) {
  # Create an XGBoost DMatrix object from the data
  data <- xgboost::xgb.DMatrix(data = as.matrix(x), label = as.vector(y))

  # Set parameters for the monotonic XGBoost model
  params = list(max_depth = max_depth,
                min_child_weight = min_child_weight,
                monotone_constraints = 1,  # Enforce monotonic increase
                eta = 1, gamma = 0,
                lambda = 0)

  # Train the model with one boosting round
  iso_fit <- xgboost::xgb.train(params = params,
                                data = data,
                                nrounds = 1)

  # Prediction function for new data
  fun <- function(x) {
    data_pred <- xgboost::xgb.DMatrix(data = as.matrix(x))
    pred <- predict(iso_fit, data_pred)
    return(pred)
  }
  return(fun)
}
```

```
# Function: calibrate_inverse_weights
# Purpose: Calibrates inverse weights using isotonic regression
with XGBoost for two propensity scores.
# Inputs:
#   - A: Binary indicator variable.
#   - pi1: Cross-fitted (pooled out-of-fold)
          propensity score estimates for treatment group A = 1.
#   - pi0: Cross-fitted (pooled out-of-fold)
          propensity score estimates for control group A = 0.
# Returns:
#   - A list containing calibrated inverse weights for each group:
#       - alpha1_star: Calibrated inverse weights for A = 1.
#       - alpha0_star: Calibrated inverse weights for A = 0.
calibrate_inverse_weights <- function(A, pi1, pi0) {

  # Calibrate pi1 using monotonic XGBoost
  calibrator_pi1 <- isoreg_with_xgboost(pi1, A)
  pi1_star <- calibrator_pi1(pi1)

  # Set minimum truncation level for treated group
  c1 <- min(pi1_star[A == 1])
  pi1_star = pmax(pi1_star, c1)
  alpha1_star <- 1 / pi1_star

  # Calibrate pi0 using monotonic XGBoost
  calibrator_pi0 <- isoreg_with_xgboost(pi0, 1 - A)
  pi0_star <- calibrator_pi0(pi0)

  # Set minimum truncation level for control group
  c0 <- min(pi0_star[A == 0])
  pi0_star = pmax(pi0_star, c0)
  alpha0_star <- 1 / pi0_star

  # Return calibrated inverse weights for both groups
  return(list(alpha1_star = alpha1_star, alpha0_star = alpha0_star))
}
```

### B.2. Python code

```
import xgboost as xgb
import numpy as np
```

```python
# Function: isoreg_with_xgboost
# Purpose: Fits isotonic regression using XGBoost.
# Inputs:
#    - x: A vector or matrix of predictor variables.
#    - y: A vector of response variables.
#    - max_depth: Maximum depth of the trees in XGBoost (default = 15).
#    - min_child_weight: Minimum sum of instance weights in constant segment.
# Returns:
#    - A function that takes a new predictor variable x
and returns the model's predicted values.
def isoreg_with_xgboost(x, y, max_depth=15, min_child_weight=20):
    # Create an XGBoost DMatrix object from the data
    data = xgb.DMatrix(data=np.asarray(x), label=np.asarray(y))

    # Set parameters for the monotonic XGBoost model
    params = {
        'max_depth': max_depth,
        'min_child_weight': min_child_weight,
        'monotone_constraints': "(1)",  # Enforce monotonic increase
        'eta': 1,
        'gamma': 0,
        'lambda': 0
    }

    # Train the model with one boosting round
    iso_fit = xgb.train(params=params, dtrain=data, num_boost_round=1)

    # Prediction function for new data
    def predict_fn(x):
        data_pred = xgb.DMatrix(data=np.asarray(x))
        pred = iso_fit.predict(data_pred)
        return pred

    return predict_fn



def calibrate_inverse_weights(A, pi1, pi0):
    """
    Calibrates inverse weights using isotonic regression
    with XGBoost for two propensity scores.

    Args:
        A (np.array): Binary indicator variable.
        pi1 (np.array): Cross-fitted (pooled out-of-fold)
```

```
        propensity score estimates for treatment group A = 1.
    pi0 (np.array): Cross-fitted (pooled out-of-fold)
        propensity score estimates for control group A = 0.

Returns:
    dict: A dictionary containing calibrated inverse weights for each group:
        - alpha1_star: Calibrated inverse weights for A = 1.
        - alpha0_star: Calibrated inverse weights for A = 0.
"""

# Calibrate pi1 using monotonic XGBoost
calibrator_pi1 = isoreg_with_xgboost(pi1, A)
pi1_star = calibrator_pi1(pi1)

# Set minimum truncation level for treated group
c1 = np.min(pi1_star[A == 1])
pi1_star = np.maximum(pi1_star, c1)
alpha1_star = 1 / pi1_star

# Calibrate pi0 using monotonic XGBoost
calibrator_pi0 = isoreg_with_xgboost(pi0, 1 - A)
pi0_star = calibrator_pi0(pi0)

# Set minimum truncation level for control group
c0 = np.min(pi0_star[A == 0])
pi0_star = np.maximum(pi0_star, c0)
alpha0_star = 1 / pi0_star

# Return calibrated inverse weights for both groups
return {'alpha1_star': alpha1_star, 'alpha0_star': alpha0_star}
```

## Appendix C. Notation and Lemmas

### C.1. Notation

For a uniformly bounded function class $\mathcal{F}$, let $N(\epsilon, \mathcal{F}, L_2(P))$ denote the $\epsilon-$covering number (van der Vaart and Wellner, 1996) of $\mathcal{F}$ with respect to $L_2(P)$ and define the uniform entropy integral of $\mathcal{F}$ by

$$\mathcal{J}(\delta, \mathcal{F}) := \int_0^\delta \sup_{\mathrm{Q}} \sqrt{\log N(\epsilon, \mathcal{F}, L_2(\mathrm{Q}))} \, d\epsilon \,,$$

where the supremum is taken over all discrete probability distributions Q.

In the following, let $a_0 \in \mathcal{A}$ be arbitrary, and let $M > 0$, $C > 0$, and $\eta > 0$ be the constants associated with conditions C1 and C2. We redefine $\mathcal{F}_{\text{anti}} \subset \{\theta : [-M, M] \to \mathbb{R}; \; \theta \text{ is monotone nonincreasing}\}$ to denote the family of nonincreasing functions on $\mathcal{T}$ that are uniformly bounded by $\eta^{-1}$. For a function $f : [-M, M] \to \mathbb{R}$, we denote its variation norm by $\|f\|_{TV} := |f(-M)| + \int_{[-M,M]}|df|$. Let $\mathcal{F}_{TV}^{(a_0)}$ consist of all real-valued functions defined on $[-M, M]$ with a sup-norm bounded by 1 and a total variation norm uniformly bounded by $3M$, where $M$ is as described in condition C4. Additionally, define $\mathcal{F}_{n,\text{anti}}^{(j,a_0)} := \{\theta \circ \pi_{n,j}(a_0 \mid \cdot) : \theta \in \mathcal{F}_{\text{anti}}\}$ as the family of functions obtained by composing nonincreasing functions in $\mathcal{F}_{\text{anti}}$ with $\pi_{n,j}(a_0 \mid \cdot)$. Similarly, let $\mathcal{F}_{n,\text{TV}}^{(j,a_0)} := \{\theta \circ \pi_{n,j}(a_0 \mid \cdot); \; \theta \in \mathcal{F}_{TV}^{(a_0)}\}$ represent the family of functions obtained by composing functions in $\mathcal{F}_{TV}^{(a_0)}$ with $\pi_{n,j}$. Let $\mathcal{F}_{n,\text{Lip}}^{(j,a_0)}$ be the subset of $\{(w, a) \mapsto [1(a = a_0)\alpha(a \mid w) - 1][\pi(a \mid w)\alpha(a \mid w) - 1] : \pi \in \mathcal{F}_{n,\text{TV}}^{(j,a_0)}, \; \alpha \in \mathcal{F}_{n,\text{anti}}^{(j,a_0)}\}$ such that the essential supremum $\text{ess sup}_{w,a} |\pi(a \mid w)\alpha(a \mid w)| \leq (C + 1)$ for $C$ as given in condition C2. Note that $\mathcal{F}_{n,\text{Lip}}^{(j,a_0)}$ is uniformly bounded by $2(C + 1)$.

We will use the following empirical process notation: for a $P$−measurable function $f$, we denote $\int f(o)dP(o)$ by $Pf$. We also let $P_{n,j}$ for $j \in [J]$ denote the empirical distribution of $\mathcal{C}^{(j)}$ and, hence, write $P_{n,j}f := \frac{1}{\#|\mathcal{I}_j|} \sum_{i \in \mathcal{I}_j} f(O_i)$ with $\mathcal{I}_j$ indexing observations of $\mathcal{C}^{(j)} \subset \mathcal{D}_n$. For two quantities $x$ and $y$, we use the expression $x \lesssim y$ to mean that $x$ is upper bounded by $y$ times a universal constant that may only depend on global constants that appear in Conditions C1-C4.

## C.2. Lemmas

Recall from Algorithm 1 that $\alpha_{n,j}^* = f_n^{(a_0)} \circ \pi_{n,j}$ for each $j \in [J]$. For each $\theta : \mathbb{R} \to \mathbb{R}$, we denote the empirical risk by:

$$R_n^{(a_0)}(\theta) := \sum_{i=1}^{n} \left[ 1(A_i = a_0)(\theta \circ \pi_{n,j(i)})^2(a_0 \mid W_i) - 2(\theta \circ \pi_{n,j(i)})(a_0 \mid W_i) \right].$$

Let $\widetilde{f}_n^{(a_0)} := 1/g_n^{(a_0)}$ be the untruncated antitonic empirical risk minimizer in Algorithm 1, and, for each $\theta \in \mathcal{F}_{anti}$, define the truncated empirical risk $R_{n,\text{trunc}}^{(a_0)}(\theta)$ as

$$\sum_{i=1}^{n} 1\left\{ \widetilde{f}_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) < \infty \right\} \left[ 1(A_i = a_0)(\theta \circ \pi_{n,j(i)})^2(a_0 \mid W_i) - 2(\theta \circ \pi_{n,j(i)})(a_0 \mid W_i) \right].$$

**Lemma 4 (Near Risk Minimizer)** *Under C1-C3, $f_n^{(a_0)}$ minimizes the truncated empirical risk, satisfying $f_n^{(a_0)} \in \arg\min_{\theta \in \mathcal{F}_{\text{anti}}} R_{n,trunc}^{(a_0)}(\theta)$. Moreover, it is a near empirical risk minimizer of the untruncated empirical risk in that, for each $C < \infty$, we have*

$$R_n^{(a_0)}(f_n^{(a_0)}) - R_n^{(a_0)}(\theta) \leq CO_p(n^{-2/3}),$$

*for every $\theta \in \mathcal{F}_{\text{anti}}$ with $\|\theta\|_\infty < C$.*

**Proof** Since $\mathcal{F}_{\text{iso}}$ is a convex cone, the first-order conditions of the isotonic regression problem (Groeneboom and Lopuhaa, 1993) defining $g_n^{(a_0)}$ imply that $g_n^{(a_0)}$ is the empirical risk minimizer if and only if, for all $\theta \in \mathcal{F}_{\text{iso}}$, the following condition holds:

$$\frac{1}{n} \sum_{i=1}^{n} (\theta \circ \pi_{n,j(i)})(a_0 \mid W_i) \left\{ 1(A_i = a_0) - g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) \right\} \leq 0.$$

Now, for $i \in [n]$, note that the evaluation of the isotonic regression solution $g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i))$ is an empirical mean of $\{1(A_k = a_0) : k \in [n], g_n^{(a_0)}(\pi_{n,j(k)}(a_0 \mid W_k)) = g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i))\}$ (Groeneboom and Lopuhaa, 1993). Consequently, $g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) = 0$ implies that $1(A_i = a_0) = 0$, so the residual in the above expression is zero for boundary observations with a calibrated propensity score of $0$. Therefore, the previous display implies that

$$\frac{1}{n} \sum_{i=1}^{n} 1\{g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) > 0\}(\theta \circ \pi_{n,j(i)})(a_0 \mid W_i)\left\{1(A_i = a_0) - g_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) - 1\right\} \leq 0.$$

Noting that $\widetilde{f}_n^{(a_0)} = 1/g_n^{(a_0)}$ and $\theta$ is arbitrary, we conclude that $\widetilde{f}_n^{(a_0)}$ satisfies, for all $\theta \in \mathcal{F}_{\mathrm{iso}}$, the following condition:

$$\frac{1}{n} \sum_{i=1}^{n} 1\{\widetilde{f}_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) < \infty\}(\theta \circ \pi_{n,j(i)})(a_0 \mid W_i)\left\{1(A_i = a_0)\widetilde{f}_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) - 1\right\} \leq 0.$$

In view of these first-order conditions, it follows that $\widetilde{f}_n^{(a_0)}$ minimizes the empirical risk restricted to observations with finite weights, that is,

$$\widetilde{f}_n^{(a_0)} \in \operatorname*{argmin}_{\theta \in \mathcal{F}_{\mathrm{anti}}} R_{n,trunc}^{(a_0)}(\theta).$$

Hence, $R_{n,trunc}^{(a_0)}(\widetilde{f}_n^{(a_0)}) - R_{n,trunc}^{(a_0)}(\theta) \leq 0$ for all $\theta \in \mathcal{F}_{anti}$. Furthermore, since $f_n^{(a_0)} = \widetilde{f}_n^{(a_0)} \wedge b_n^{(a_0)}$ where $b_n^{(a_0)} = \max_{i \in [n]: A_i = a_0} f_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i))$ is the maximum finite value of $\widetilde{f}_n^{(a_0)}$, we know that $f_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) = \widetilde{f}_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i))$ for all $i$ such that $\widetilde{f}_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i)) < \infty$. As a consequence, the same inequality holds for the truncated minimizer $f_n^{(a_0)}$:

$$R_{n,trunc}^{(a_0)}(f_n^{(a_0)}) - R_{n,trunc}^{(a_0)}(\theta) \leq 0 \text{ for all } \theta \in \mathcal{F}_{anti}.$$

However, we know that infinite weight estimates can only occur at the boundary observations $\#\{i \in [n] : A_i \neq a_0 \text{ and } \pi_{n,j(i)}(a_0 \mid W_i) < \min_{k \in [n]: A_k = a_0} \pi_{n,j(i)}(a_0 \mid W_k)\} = 1/b_n^{(a_0)}$ and, by C3, there are at most $O_p(n^{1/3})$ such observations. By C2, $f_n^{(a_0)}(\pi_{n,j(i)}(a_0 \mid W_i))$ is almost surely uniformly bounded for each $i \in [n]$, and, therefore,

$$R_n^{(a_0)}(f_n^{(a_0)}) - R_n^{(a_0)}(\theta) \leq C O_p(n^{1/3}/n) = C O_p(n^{-2/3}),$$

for all $M < \infty$ and $\theta \in \mathcal{F}_{\mathrm{anti}}$ with $\|\theta\|_\infty < C$.

∎

**Lemma 5 (Empirical Calibration)** *For calibrated predictors $\alpha_{n,j}^*(a_0 \mid \cdot) = f_n^{(a_0)} \circ \pi_{n,j}(a_0 \mid \cdot)$ indexed by $j \in [J]$ obtained using Algorithm 1, we have*

$$\frac{1}{n} \sum_{i=1}^{n} \left[h \circ \alpha_{n,j(i)}^*(a_0 \mid W_i)\right]\left\{1(A_i = a_0)\alpha_{n,j(i)}^*(a_0 \mid W_i) - 1\right\} = M O_p(n^{-2/3}). \tag{6}$$

*for each $M < \infty$ and every $h \in \mathcal{F}_{\mathrm{anti}}$ with $\|h\|_\infty < M$.*

**Proof** This proof follows from a modification of the argument used to establish Lemma C.1 in Van Der Laan et al. (2023). For each $g : \mathbb{R} \to \mathbb{R}$, denote the empirical least squares risk by

$$R_n(g) := \sum_{i=1}^{n} \left\{ 1(A_i = a_0) - (g \circ \pi_{n,j(i)})(a_0 \,|\, W_i) \right\}^2.$$

Recall that $f_n^{(a_0)} = 1/(c_n^{(a_0)} \vee g_n^{(a_0)})$, where $g_n^{(a_0)}$ is obtained via

$$g_n^{(a_0)} \in \operatorname*{argmin}_{g \in \mathcal{F}_{\mathrm{iso}}} \sum_{i=1}^{n} \left\{ 1(A_i = a_0) - (g \circ \pi_{n,j(i)})(a_0 \,|\, W_i) \right\}^2 = \operatorname*{argmin}_{g \in \mathcal{F}_{\mathrm{iso}}} R_n(g).$$

The isotonic regression solution $(g_n^{(a_0)}$ can be expressed as a piecewise-constant function, satisfying

$$(g_n^{(a_0)} \circ \pi_{n,j})(a_0 \,|\, w) = b_0 + \sum_{t=1}^{T} b_t 1(\pi_{n,j}(a_0 \,|\, w) \ge u_t), \quad \text{for each } j \in [J],$$

with $T + 1 \in \mathbb{N}$ values determined by the coefficients $\{b_t\}_{t=0}^{T}$ and thresholds $\{u_t\}_{t=1}^{T}$ (Barlow and Brunk, 1972). By monotonicity of $g_n^{(a_0)}$, it must hold that $b_0 \in \mathbb{R}$ and $b_t > 0$ for all $t \in \{1, \ldots, T\}$.

For any jump point $u_t$ of $g_n^{(a_0)}$, we can always choose $\varepsilon \in \mathbb{R}$ with a small enough absolute value such that

$$\xi_n(\varepsilon) : x \mapsto g_n^{(a_0)}(x) + \varepsilon 1(x \ge u_t), \quad \text{for each } j \in [J],$$

is an isotonic function in $\mathcal{F}_{iso}$. In particular, this holds for any $\varepsilon$ such that $|\varepsilon| < \min\{b_t\}_{t=1}^{T}$.

Since $g_n^{(a_0)}$ is the minimizing isotonic solution and $R_n(\xi_n(0)) = R_n(g_n^{(a_0)})$, we have that $R_n(\xi_n(\varepsilon)) \ge R_n(g_n^{(a_0)})$ for all $\varepsilon$ sufficiently close to zero. Thus, since $\varepsilon \mapsto R_n(\xi_n(\varepsilon))$ is differentiable with a local minimum achieved at $\varepsilon = 0$, the derivative $\frac{d}{d\varepsilon} R_n(\xi_n(\varepsilon))\big|_{\varepsilon=0}$ is zero. Computing the derivative, we find that

$$\sum_{i=1}^{n} 1(\pi_{n,j(i)}(a_0 \,|\, W_i) \ge u_t) \left\{ 1(A_i = a_0) - (g_n^{(a_0)} \circ \pi_{n,j(i)})(a_0 \,|\, W_i) \right\} = 0.$$

Since the jump point $u_t$ can be chosen arbitrarily, we can show that the same equation holds for linear combinations of indicator functions of the form $a_{n,j}(w) = \sum_{t=0}^{T} b_t' 1(\pi_{n,j}(a_0 \,|\, w) \ge u_t)$. That is,

$$\sum_{i=1}^{n} a_{n,j(i)}(W_i) \left\{ 1(A_i = a_0) - (g_n^{(a_0)} \circ \pi_{n,j(i)})(a_0 \,|\, W_i) \right\} = 0.$$

Let $\widetilde{\alpha}_{n,j}^{*}(a_0 \,|\, \cdot) := 1/(g_n^{(a_0)} \circ \pi_{n,j})(a_0 \,|\, \cdot)$ denote the untruncated isotonic calibrated weights. Then, we have

$$\sum_{i=1}^{n} a_{n,j(i)}(W_i) \left\{ 1(A_i = a_0) - \{\widetilde{\alpha}_{n,j(i)}^{*}(a_0 \,|\, W_i)\}^{-1} \right\} = 0.$$

Next, observe that, for any $h : \mathbb{R} \cup \{\infty\} \to \mathbb{R}$, the function $h \circ \widetilde{\alpha}^*_{n,j}$ must take the form of $a_{n,j}$, since a transformation of a step function preserves its jump points. Hence, we have with probability tending to one,

$$\sum_{i=1}^{n} \left[ h \circ \widetilde{\alpha}^*_{n,j(i)}(a_0 \,|\, W_i) \right] \left\{ 1(A_i = a_0) - \{\widetilde{\alpha}^*_{n,j(i)}(a_0 \,|\, W_i)\}^{-1} \right\} = 0.$$

In addition, since $h : \mathbb{R} \cup \{\infty\} \to \mathbb{R}$ is arbitrary, we have, by Condition C2, with probability tending to one, for any $h : \mathbb{R} \cup \{\infty\} \to \mathbb{R}$ with $h(\infty) := 0$, that

$$\sum_{i=1}^{n} \left[ h \circ \widetilde{\alpha}^*_{n,j(i)}(a_0 \,|\, W_i) \right] \left\{ 1(A_i = a_0)\widetilde{\alpha}^*_{n,j(i)}(a_0 \,|\, W_i) - 1 \right\} = 0.$$

Next, observe, by definition of the truncation in Algorithm 1, that $\widetilde{\alpha}^*_{n,j(i)}(a_0 \,|\, W_i)$ can only differ from $\alpha^*_{n,j(i)}(a_0 \,|\, W_i)$ for $i \in [n]$ such that $A_i \neq a_0$ and $g_n^{(a_0)}(\pi_n(a_0 \,|\, W_i)) = 0$. Thus, by C3, the two functions can only differ at most at $O_p(n^{1/3})$ observations. It follows from C2 that

$$\sum_{i=1}^{n} \left[ h \circ \alpha^*_{n,j(i)}(a_0 \,|\, W_i) \right] \left\{ 1(A_i = a_0)\alpha^*_{n,j(i)}(a_0 \,|\, W_i) - 1 \right\} = MO_p(n^{1/3}),$$

and, therefore,

$$\frac{1}{n}\sum_{i=1}^{n} \left[ h \circ \alpha^*_{n,j(i)}(a_0 \,|\, W_i) \right] \left\{ 1(A_i = a_0)\alpha^*_{n,j(i)}(a_0 \,|\, W_i) - 1 \right\} = MO_p(n^{-2/3}),$$

■

**Lemma 6** *Under Condition C4, for all $j \in [J]$, the map $g^{*(a_0)}_{0,n,j}$, defined such that $g^{*(a_0)}_{0,n,j}(\pi_{n,j}(a_0 \,| W)) = \gamma^{(a_0)}_0(W, \alpha^*_{n,j})$, has its total variation, $P_0$-almost surely, bounded above by $3M$, where $M$ is as specified in Condition C4.*

**Proof** This proof follows from a modification of the argument used to establish Lemma C.3 in Van Der Laan et al. (2023). Recall that $\alpha^{*-1}_{n,j}(a_0 \mid w) = \{g_n^{(a_0)} \circ \pi_{n,j}(a_0 \mid w)\}^{-1}$, where $g_n^{(a_0)}$ is an isotonic step function defined in Algorithm 1. Therefore, $\gamma^{(a_0)}_0(\cdot, \alpha^{*-1}_{n,j}) = \gamma^{(a_0)}_0(\cdot, g_n^{(a_0)} \circ \pi_{n,j}(a_0 \mid \cdot))$.

Note that the map $g^{*(a_0)}_{0,n,j}$ such that $g^{*(a_0)}_{0,n,j}(\pi_{n,j}(a_0 \,| w)) = \gamma^{(a_0)}_0(w, \alpha^*_{n,j})$ is given by

$$t \mapsto E_0\left[ \pi_0(a_0 \,|\, W) \,|\, g_n^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, W) = g_n^{(a_0)}(t), \mathcal{D}_n \right].$$

Since $g_n^{(a_0)}$ is nondecreasing and piecewise constant, we have, for any $t \in \mathbb{R}$, that

$$E_0\left[ \pi_0(a_0 \,|\, W) \,|\, g_n^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, W) = g_n^{(a_0)}(t), \mathcal{D}_n \right] = E_0\left[ \pi_0(a_0 \,|\, W) \,|\, \pi_{n,j}(a_0 \,|\, W) \in B_t, \mathcal{D}_n \right]$$

for the set $B_t := \left\{ z \in \mathbb{R} : g_n^{(a_0)}(z) = g_n^{(a_0)}(t) \right\}$, where $B_t = \{z \in \mathbb{R} : a(t) \le z < b(t)\}$ for some endpoints $a(t), b(t) \in \mathbb{R}$. The law of total expectation further implies that

$$E_0[\pi_0(a_0 \,|\, W) \,|\, \pi_{n,j}(a_0 \,|\, W) \in B_t] = E_0[g_{0,n,j}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, W) \,|\, \pi_{n,j}(a_0 \,|\, W) \in B_t],$$

where $g_{0,n,j}^{(a_0)}$ in C4 is such that $g_{0,n,j}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, W) = \gamma_0(W, \pi_{n,j}(a_0 \,|\, W))$ $P_0$-almost surely. By Condition C4, the function $g_{0,n,j}^{(a_0)}$ is of bounded total variation. Heuristically, since $t \mapsto E[g_{0,n,j}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, W) \,|\, \pi_{n,j}(a_0 \,|\, W)] \in B_t$ is obtained by locally averaging $g_{0,n,j}^{(a_0)}$ within the bins $(B_t : t)$, its total variation should also be bounded. Up to notation, this proof follows from the proof of Lemma C.3 in Van Der Laan et al. (2023). In particular, their proof establishes that $t \mapsto E[g_{0,n,j}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, W) \,|\, \pi_{n,j}(a_0 \,|\, W) \in B_t]$ and, therefore also $g_{0,n,j}^{*(a_0)}$, has total variation norm bounded above by $3\|g_{0,n,j}^{(a_0)}\|_{TV}$ almost surely. The result then follows, noting that $\|g_{0,n,j}^{(a_0)}\|_{TV} \le M$ by C4. ∎

## Appendix D. Proofs of Theorems

### D.1. Proof of Theorem 1

For each $j \in [J]$ and any function $h : \mathbb{R} \to \mathbb{R}$, we have, by the law of iterated expectations, that

$$\int \left[ h \circ \alpha_{n,j}^*(a_0 \,|\, w) \right] \left\{ 1(A = a_0) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\} dP_0(w, a)$$
$$= \int \left[ h \circ \alpha_{n,j}^*(a_0 \,|\, w) \right] \left\{ \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\} dP_0(w).$$

Choosing $h$ such that $h(\alpha_{n,j}^*(w)) = \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(w) - 1$ in the above display, we find that

$$\int \left\{ \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\} \left\{ 1(A = a_0) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\} dP_0(w, a)$$
$$= \int \left\{ \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\}^2 dP_0(w). \tag{7}$$

By Lemma 5, we also have

$$\frac{1}{J} \sum_{j \in [J]} \sum_{i \in \mathcal{C}^{(j)}} \left\{ \gamma_0^{(a_0}(W_i, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(a_0 \,|\, W_i) - 1 \right\} \left\{ 1(A_i = a_0) \alpha_{n,j}^*(a_0 \,|\, W_i) - 1 \right\} = O_p(n^{-2/3}).$$

Combining this with Equation (7), we obtain the identity

$$\frac{1}{J} \sum_{j \in [J]} \mathrm{CAL}^{(a_0)}(\alpha_{n,j}^*)$$
$$= \frac{1}{J} \sum_{j \in [J]} \int \left\{ \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\}^2 dP_0(w)$$
$$= \frac{1}{J} \sum_{j \in [J]} \int \left[ f_{n,j}^{(a_0)}(w, a) \left\{ \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1}) \alpha_{n,j}^*(a_0 \,|\, w) - 1 \right\} \right] d(P_0 - P_{n,j})(w, a) + O_p(n^{-2/3}),$$
$$\tag{8}$$

where $(w, a) \mapsto f_{n,j}^{(a_0)}(w, a) := \{1(a = a_0)\alpha_{n,j}^*(a \,|\, w) - 1\}$.

We now use empirical process techniques to obtain a convergence rate for the left-hand side of the above display. Our proof follows along the lines of the proof of Theorem 3 in Van Der Laan et al. (2023). To this end, by definition of $f_n^{(a_0)}$ and C2, $w \mapsto \alpha_{n,j}^*(a_0 \,|\, w) = f_n^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, w)$ is a bounded monotone transformation of $w \mapsto \pi_{n,j}(a_0 \,|\, w)$ and falls in the random function class $\mathcal{F}_{n,\text{anti}}^{(j,a_0)}$. Under C4 and Lemma 6, $\pi_{n,j}(w) \mapsto \gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1})$ has total variation norm bounded by $3K$ for some constant $K > 0$. Hence, $\gamma_0^{(a_0)}(w, \alpha_{n,j}^*)$ falls almost surely in the random function class $\mathcal{F}_{n,\text{TV}}^{(j,a_0)}$. We conclude, by definition, that $(w, a) \mapsto f_{n,j}^{(a_0)}(w, a)\{\gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1})\alpha_n^*(a_0 \,|\, w) - 1\}$ falls in the function class $\mathcal{F}_{n,\text{Lip}}^{(j,a_0)}$.

Recall the positivity constant $\eta > 0$ from Condition C2 and note that $\text{ess sup}_{a,w}|f_{n,j}^{(a_0)}(w)| \leq \eta + 1$. Denote the localized function class $\mathcal{F}_{n,j}(\delta_n) = \{f \in \mathcal{F}_{n,\text{Lip}}^{(j,a_0)} : \|f\| \leq J(\eta+1)\delta_n\}$. Using Hölder's inequality and Jensen's inequality, Equation (8) implies

$$\delta_n^2 \leq \frac{1}{J} \sum_{j \in [J]} \sup_{f \in \mathcal{F}_{n,j}(\delta_n)} (P_0 - P_{n,j})f + O_p(n^{-2/3}). \tag{9}$$

where $\delta_n^2 := J^{-1} \sum_{j=1}^{J} \int \{\gamma_0^{(a_0)}(w, \alpha_{n,j}^{*-1})\alpha_{n,j}^*(a_0 \,|\, w) - 1\}^2 dP_0(w)$ is the fold-averaged calibration error. We denote the first term on the right-hand side of (9) by $\frac{1}{J} \sum_{j \in [J]} \phi_{n,j}(\delta_n)$, where

$$\phi_{n,j}(\delta) := \sup_{f \in \mathcal{F}_{n,j}(\delta)} (P_0 - P_{n,j})f.$$

and define the random variable

$$Z_n := \delta_n^2 - \frac{1}{J} \sum_{j \in [J]} \sup_{f \in \mathcal{F}_{n,j}(\delta_n)} (P_0 - P_{n,j})f,$$

where we note $Z_n = O_p(n^{-2/3})$ by (9).

Showing the asserted stochastic order, $\delta_n^2 = O_p(n^{-2/3})$, is equivalent to showing for all $\epsilon > 0$, we can find a $2^S$ sufficiently large so that $\limsup_{n\to\infty} P(n^{2/3}\delta_n^2 > 2^S) < \epsilon$. To this end, we just need to show $\lim_{n\to\infty} P(n^{2/3}\delta_n^2 > 2^S) \to 0$ as $S \to \infty$. Let $K_\varepsilon < \infty$ be such that $Z_n \leq K_\varepsilon n^{-2/3}$ with probability at least $1 - \varepsilon$. Define the event $A_s := \{n^{2/3}\delta_n^2 \in (2^s, 2^{s+1}]\}$ for each $s$. From a

peeling argument and Markov's inequality, we find

$$
\begin{aligned}
P\big(n^{2/3}\delta_n^2 > 2^S\big) &\le P\big(n^{2/3}\delta_n^2 > 2^S, Z_n \le K_\varepsilon n^{-2/3}\big) + \varepsilon \\
&\le \sum_{s=S}^{\infty} P\big(2^{s+1} \ge n^{2/3}\delta_n^2 > 2^s, Z_n \le K_\varepsilon n^{-2/3}\big) + \varepsilon \\
&\le \sum_{s=S}^{\infty} P\Big(A_s, \delta_n^2 \le \frac{1}{J}\sum_{j\in[J]}\phi_{n,j}(\delta_n) + Z_n, Z_n \le K_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\le \sum_{s=S}^{\infty} P\Big(A_s, \delta_n^2 \le \frac{1}{J}\sum_{j\in[J]}\phi_{n,j}(\delta_n) + K_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\le \sum_{s=S}^{\infty} P\Big(n^{-2/3}2^s < \delta_n^2 \le \frac{1}{J}\sum_{j\in[J]}\phi_{n,j}\big(n^{-1/3}2^{\frac{s+1}{2}}\big) + K_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\le \sum_{s=S}^{\infty} P\Big(n^{-2/3}2^s < \frac{1}{J}\sum_{j\in[J]}\phi_{n,j}\big(n^{-1/3}2^{\frac{s+1}{2}}\big) + K_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\le \sum_{s=S}^{\infty} \frac{J^{-1}\sum_{j\in[J]} E_0^n\big[\phi_{n,j}\big(n^{-1/3}2^{\frac{s+1}{2}}\big)\big] + K_\varepsilon n^{-2/3}}{n^{-2/3}2^s} + \varepsilon.
\end{aligned}
$$

In view of the above display, we aim to bound the local modulus of continuity, $E_0^n[\phi_{n,j}(\delta)]$ for a fixed $\delta > 0$. Using that $\mathcal{F}_{n,\mathrm{Lip}}^{(j,a_0)}$ is uniformly bounded and applying Theorem 2.1 in Van Der Vaart and Wellner (2011) conditional on $\mathcal{E}^{(j)}$, we find

$$
E_0^n[\phi_{n,j}(\delta)\,|\,\mathcal{E}^{(j)}] \lesssim n^{-1/2}\mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{Lip}}^{(j,a_0)}\big)\Big(1 + \frac{\mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{Lip}}^{(j,a_0)}\big)}{\sqrt{n}\delta^2}\Big).
$$

We note that, conditional on $\mathcal{E}^{(j)}$, $\mathcal{F}_{n,\mathrm{Lip}}^{(j,a_0)}$ is a multivariate Lipschitz transformation of $\mathcal{F}_{n,\mathrm{TV}}^{(j,a_0)}$ and $\mathcal{F}_{n,\mathrm{anti}}^{(j,a_0)}$. Therefore, by Theorem 2.10.20 of van der Vaart and Wellner (1996), we have that $\mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{Lip}}^{(j,a_0)}\big) \lesssim \mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{TV}}^{(j,a_0)}\big) + \mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{anti}}^{(j,a_0)}\big)$. Since functions of bounded total variation can be written as a difference of non-decreasing monotone functions, we have by the same theorem that $\mathcal{J}\big(\delta, \mathcal{F}_{\mathrm{TV}}^{(a_0)}\big) \lesssim \mathcal{J}(\delta, \mathcal{F}_{\mathrm{anti}})$. By the covering number bound for bounded monotone functions given in Theorem 2.7.5 of van der Vaart and Wellner (1996), we have $\mathcal{J}(\delta, \mathcal{F}_{\mathrm{anti}}) \lesssim \sqrt{\delta}$. We have by the change-of-variables formula

$$
\mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{anti}}^{(j,a_0)}\big) = \int_0^\delta \sup_Q \sqrt{N\big(\varepsilon, \mathcal{F}_{n,\mathrm{anti}}^{(j,a_0)}, \|\cdot\|_Q\big)}\, d\varepsilon = \int_0^\delta \sup_Q \sqrt{N\big(\varepsilon, \mathcal{F}_{\mathrm{anti}}, \|\cdot\|_{Q\circ\pi_{n,j}^{-1}}\big)}\, d\varepsilon = \mathcal{J}(\delta, \mathcal{F}_{\mathrm{anti}}),
$$
$$
\mathcal{J}\big(\delta, \mathcal{F}_{n,\mathrm{TV}}^{(j,a_0)}\big) = \int_0^\delta \sup_Q \sqrt{N\big(\varepsilon, \mathcal{F}_{n,\mathrm{TV}}^{(j,a_0)}, \|\cdot\|_Q\big)}\, d\varepsilon = \int_0^\delta \sup_Q \sqrt{N\big(\varepsilon, F_{n,\mathrm{TV}}, \|\cdot\|_{Q\circ\pi_{n,j}^{-1}}\big)}\, d\varepsilon = \mathcal{J}(\delta, \mathcal{F}_{n,\mathrm{TV}});
$$

where, with a slight abuse of notation, $Q \circ \pi_{n,j}^{-1}$ is the push-forward probability measure for the random variable $\pi_{n,j}(a_0 \mid W)$ conditional on $\mathcal{E}^{(j)}$.

166

We are now in a position to apply Theorem 3.2.5 of van der Vaart and Wellner (1996) to obtain a rate of convergence for the fold-averaged calibration error $\delta_n^2$. Returning to our previous bound for $E_0^n[\phi_{n,j}(\delta) \mid \mathcal{E}^{(j)}]$ and applying the law of total expectation, we conclude that

$$\frac{1}{J} \sum_{j=1}^{J} E_0^n[\phi_{n,j}(\delta)] \lesssim n^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{\text{anti}}) \left( 1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{\text{anti}})}{\sqrt{n}\delta^2} \right),$$

where the right-hand side is non-random. This gives us

$$\frac{J^{-1} \sum_{j\in[J]} E_0^n\left[\phi_{n,j}\left(n^{-1/3}2^{\frac{s+1}{2}}\right)\right]}{n^{-2/3}2^s} \lesssim 2^{-3s/4}$$

since $\mathcal{J}(\delta, \mathcal{F}_{\text{anti}}) \lesssim \sqrt{\delta}$. Our previous display implies that for any $n \in \mathbb{N}$,

$$P\left(n^{2/3}\delta_n^2 > 2^S\right) \lesssim \varepsilon + \sum_{s=S}^{\infty} \frac{1}{2^{3s/4}} \overset{S\to\infty}{\to} \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we conclude that $\mathbb{P}\left(n^{2/3}\delta_n^2 > 2^S\right) \overset{S\to\infty}{\to} 0$. Taking the limit on both sides with respect to $n$ gives us the desired result:

$$\delta_n^2 = \frac{1}{J} \sum_{j\in[J]} \int \left\{ \gamma_0^{(a)}(w, \alpha_{n,j}^*)\alpha_{n,j}^*(a \mid w) - 1 \right\}^2 dP_0(w) = O_p(n^{-2/3}).$$

## D.2. Proof of Theorem 2

The population risk function used to define the oracle transformations $\{ f_{n,0}^{(a_0)} : a_0 \in \mathcal{A} \}$ satisfies

$$\sum_{j\in[J]} \int \left\{ f^{(a)} \circ \pi_{n,j}(a \mid w) - \pi_0^{-1}(a \mid w) \right\}^2 dP_0(a, w)$$

$$= \sum_{a_0\in\mathcal{A}} \sum_{j\in[J]} \int 1(A = a_0)\left\{ f^{(a_0)} \circ \pi_{n,j}(a_0 \mid w) - \pi_0^{-1}(a_0 \mid w) \right\}^2 dP_0(a, w).$$

Since the oracle transformations are variation independent, we have for $a_0 \in \mathcal{A}$,

$$f_{n,0}^{(a_0)} \in \underset{f\in\mathcal{F}_{\text{anti}}}{\operatorname{argmin}} \sum_{j\in[J]} \int 1(A = a_0)\left\{ f \circ \pi_{n,j}(a_0 \mid w) - \pi_0^{-1}(a_0 \mid w) \right\}^2 dP_0(a, w)$$

$$= \underset{f\in\mathcal{F}_{\text{anti}}}{\operatorname{argmin}} \sum_{j\in[J]} \int \ell_{n,j}^{(a_0)}(a_0, w; f) dP_0(a, w) = \underset{f\in\mathcal{F}_{\text{anti}}}{\operatorname{argmin}} R^{(a_0)}(f),$$

where, for each $j \in [J]$, we define the fold-specific loss function and the population risk

$$(a, w, f) \mapsto \ell_{n,j}^{(a_0)}(a, w; f) := \left\{ 1(A = a_0)f \circ \pi_{n,j}(a_0 \mid w)^2 - 2f \circ \pi_{n,j}(a_0 \mid w) \right\};$$

$$f \mapsto R^{(a_0)}(f) := \sum_{j\in[J]} \int \ell_{n,j}^{(a_0)}(a_0, w; f) dP_0(a, w).$$

We claim that Condition C2 implies $\operatorname{ess\,sup}_{a,w} |f_{n,0}^{(a)} \circ \pi_{n,j}(a \,|\, w)| < \eta^{-1}$. This holds because, otherwise, we could truncate $f_{n,0}^{(a)}$ to lie in $[-1/\eta, 1/\eta]$ almost surely and obtain a monotone non-increasing function with smaller population risk. To see this, note that, for any $f \in \mathcal{F}_{anti}$,

$$
\sum_{j \in [J]} \int \ell_{n,j}^{(a_0)}(a_0, w; f) \, dP_0(a, w)
$$

$$
= \sum_{j \in [J]} \int \left\{ 1(|f| \le \eta^{-1}) + 1(|f| > \eta^{-1}) \right\} \ell_{n,j}^{(a_0)}(a_0, w; f) \, dP_0(a, w)
$$

$$
= \sum_{j \in [J]} \int 1(A = a_0) \, 1(|f| \le \eta^{-1}) \left\{ f \circ \pi_{n,j}(a \,|\, w) - \pi_0^{-1}(a_0 \,|\, w) \right\}^2 \, dP_0(a, w)
$$

$$
+ \sum_{j \in [J]} \int 1(A = a_0) \, 1(|f| > \eta^{-1}) \left\{ f \circ \pi_{n,j}(a \,|\, w) - \pi_0^{-1}(a_0 \,|\, w) \right\}^2 \, dP_0(a, w),
$$

where the global minimizer over $f$ of the left-hand side also minimizes the first term on the right-hand side by C2, and the second term on the right-hand side also takes its minimal value (zero) for any function $f$ taking values in $[-1/\eta, 1/\eta]$. Thus, we also have $f_{n,0}^{(a_0)} \in \mathcal{F}_{anti}^{1/\eta}$.

For $a_0 \in \mathcal{A}$, denote $(w, a_0) \mapsto \alpha_{0,j}^*(a_0 \,|\, w) := f_{n,0}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, w)$. By the law of iterated expectations, the excess risk satisfies for each $j \in [J]$:

$$
P_0 \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - P_0 \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \tag{10}
$$

$$
= \int \left[ \pi_0(a_0 \,|\, w) \left\{ (\alpha_{n,j}^*(a_0 \,|\, w))^2 - (\alpha_{0,j}^*(a_0 \,|\, w))^2 \right\} - 2 \left\{ \alpha_{n,j}^*(a_0 \,|\, w) - \alpha_{0,j}^*(a_0 \,|\, w) \right\} \right] dP_0(w)
$$

$$
= \int \left[ \alpha_{n,j}^*(a_0 \,|\, w) - \alpha_{0,j}^*(a_0 \,|\, w) \right] \left[ \pi_0(a_0 \,|\, w) \left\{ \alpha_{n,j}^*(a_0 \,|\, w) + \alpha_{0,j}^*(a_0 \,|\, w) \right\} - 2 \right] dP_0(w). \tag{11}
$$

Observing that $\mathcal{F}_{anti}$ is a convex cone, then for all $h \in \mathcal{F}_{anti}^{1/\eta}$, the path $\{t \mapsto (1 - t)f_{n,0}^{(a_0)} + th : t \in [0,1]\}$ lies entirely in $\mathcal{F}_{anti}$ and passes through $f_{n,0}^{(a_0)}$ at $t = 0$. Furthermore, since Condition C2 implies the derivative of the risk for every element in the path is almost surely bounded, from the mean value theorem for differentiation and dominated convergence theorem we obtain

$$
\lim_{t \downarrow 0} \frac{R^{(a_0)}\left(f_{n,0}^{(a_0)} + t(h - f_{n,0}^{(a_0)})\right) - R^{(a_0)}\left(f_{n,0}^{(a_0)}\right)}{t}
$$

$$
= (-2) \sum_{j \in [J]} \int 1(A = a_0) \left\{ \left(h - f_{n,0}^{(a_0)}\right) \circ \pi_{n,j}(a_0 \,|\, w) \right\} \left\{ f_{n,0}^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, w) - \pi_0^{-1}(a_0 \,|\, w) \right\} dP_0(a, w)
$$

$$
\ge 0.
$$

Thus, the minimizing solutions $\{\alpha_{0,j}^* : j \in [J]\}$ corresponding to the population risk satisfy, for all $h \in \mathcal{F}_{anti}^{1/\eta}$, the inequality:

$$
\sum_{j \in [J]} \int \left\{ h \circ \pi_{n,j}(a_0 \,|\, w) - \alpha_{0,j}^*(a_0 \,|\, w) \right\} \left\{ \pi_0(a_0 \,|\, w) \alpha_{0,j}^*(a_0 \,|\, w) - 1 \right\} dP_0(w) \le 0.
$$

Taking $h$ such that $\alpha_{n,j}^*(a \mid \cdot) : w \mapsto h \circ \pi_{n,j}(a \mid w)$ in the above display, Equation (11) implies that the excess risk satisfies the following lower bound:

$$
\sum_{j \in [J]} \left\{ P_0 \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - P_0 \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \right\}
$$

$$
= \sum_{j \in [J]} \int \pi_0(a_0 \mid w) \left[ \alpha_{n,j}^*(a_0 \mid w) - \alpha_{0,j}^*(a_0 \mid w) \right]^2 dP_0(w)
$$

$$
+ \sum_{j \in [J]} \int \left[ \alpha_{n,j}^*(a_0 \mid w) - \alpha_{0,j}^*(a_0 \mid w) \right] \left[ 2\pi_0(a_0 \mid w)\alpha_{0,j}^*(a_0 \mid w) - 2 \right] dP_0(w)
$$

$$
\geq \sum_{j \in [J]} \int \pi_0(a_0 \mid w) \left[ \alpha_{n,j}^*(a_0 \mid w) - \alpha_{0,j}^*(a_0 \mid w) \right]^2 dP_0(w).
$$

Now, we obtain an upper bound for the excess risk. By Lemma 4, the truncated empirical risk minimizer $f_n^{(a_0)}$ is a near empirical risk minimizer satisfying

$$
\sum_{j \in [J]} \left\{ P_{n,j} \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - P_{n,j} \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \right\} \leq O_p(n^{-2/3}).
$$

Thus,

$$
\sum_{j \in [J]} P_0 \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - \sum_{j \in [J]} P_0 \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)})
$$

$$
= \sum_{j \in [J]} \left\{ P_0 \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - P_{n,j} \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) \right\} + \sum_{j \in [J]} \left\{ P_n \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - P_{n,j} \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \right\}
$$

$$
- \sum_{j \in [J]} \left\{ P_0 \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) - P_{n,j} \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \right\}
$$

$$
\leq \sum_{j \in [J]} (P_0 - P_{n,j})\{\ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)})\} + O_p(n^{-2/3}),
$$

where we used that $\sum_{j \in [J]} \{P_{n,j}\ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - P_{n,j}\ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)})\} \leq O_p(n^{-2/3})$.

Putting the lower and upper bounds together, we obtain the inequality:

$$
\sum_{j \in [J]} \int \pi_0(a_0 \mid w) \left[ \alpha_{n,j}^*(a_0 \mid w) - \alpha_{0,j}^*(a_0 \mid w) \right]^2 dP_0(w)
$$

$$
\leq \sum_{j \in [J]} (P_0 - P_{n,j})\{\ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)})\} + O_p(n^{-2/3}). \tag{12}
$$

Note, for each $j \in [J]$, that

$$
\left\| \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \right\| \lesssim \left\| \alpha_{n,j}^*(a_0 \mid \cdot) - \alpha_{0,j}(a_0 \mid \cdot) \right\|,
$$

since $\alpha_{n,j}^*(a_0 \mid \cdot)$ and $\alpha_{0,j}(a_0 \mid \cdot)$ are uniformly bounded by $1/\eta$. Under C2, we have $\pi_0(a_0 \mid W) > \eta$ almost surely, thus,

$$
\left\| \ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)}) \right\|^2 \lesssim \left\| \alpha_{0,j}^*(a_0 \mid w) - \alpha_{0,j}(a_0 \mid \cdot) \right\|^2
$$

$$
\lesssim \eta^{-1} \int \pi_0(a_0 \mid w) \left[ \alpha_{n,j}^*(a_0 \mid w) - \alpha_{0,j}^*(a_0 \mid w) \right]^2 dP_0(w).
$$

Hence, (12) implies

$$\delta_n^2 \le \sum_{j\in[J]}(P_0 - P_{n,j})\{\ell_{n,j}^{(a_0)}(f_n^{(a_0)}) - \ell_{n,j}^{(a_0)}(f_{n,0}^{(a_0)})\} + Z_n \le \sum_{j\in[J]}\phi_{n,j}^{(a_0)}(\delta_n) + Z_n, \quad (13)$$

where $Z_n$ is some random variable satisfying $Z_n = O_p(n^{-2/3})$. Here, we define

$$\delta_n^2 := \sum_{j\in[J]}\int \pi_0(a_0\,|\,w)\left[\alpha_{n,j}^*(a_0\,|\,w) - \alpha_{0,j}^*(a_0\,|\,w)\right]^2 dP_0(w);$$

$$\phi_{n,j}^{(a_0)}(\delta) := \sup_{f\in\mathcal{F}_{n,j}^{(a_0)}:\|f\|\le C\delta}(P_0 - P_{n,j})f;$$

$$\mathcal{F}_{n,j}^{(a_0)} := \{\ell_{n,j}^{(a_0)}(\theta_1) - \ell_{n,j}^{(a_0)}(\theta_2) : \theta_1, \theta_2 \in \mathcal{F}_{\text{anti}}^{1/\eta}\};$$

and $C > 0$ is a sufficiently large constant depending only on $J$, $M$, and $\eta$.

Almost surely, we know that the function class $\mathcal{F}_{n,j}^{(a_0)}$ is uniformly bounded and is a Lipschitz transformation of $\mathcal{F}_{n,\text{anti}}^{(j,a_0),1/\eta} \times \mathcal{F}_{n,\text{anti}}^{(j,a_0),1/\eta}$, the product of function classes defined by truncating $\mathcal{F}_{n,\text{anti}}^{(j,a_0)}$ with $1/\eta$. By Theorem 2.10.20 of van der Vaart and Wellner (1996) and arguing as in the proof of Theorem 1, we have

$$\mathcal{J}\big(\delta, \mathcal{F}_{n,j}^{(a_0)}\big) \lesssim \mathcal{J}\big(\delta, \mathcal{F}_{n,\text{anti}}^{(j,a_0),1/\eta}\big) \lesssim \mathcal{J}\big(\delta, \mathcal{F}_{n,\text{anti}}^{(j,a_0)}\big) = \mathcal{J}\big(\delta, \mathcal{F}_{\text{anti}}\big) \lesssim \sqrt{\delta}.$$

Applying Theorem 2.1 in Van Der Vaart and Wellner (2011) conditional on $\mathcal{E}^{(j)}$, we then have for each $j \in [J]$,

$$E_0^n\big[\phi_{n,j}^{(a_0)}(\delta)\,|\,\mathcal{E}^{(j)}\big] \lesssim n^{-1/2}\mathcal{J}\big(\delta, \mathcal{F}_{n,j}^{(a_0)}\big)\Big(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{n,j}^{(a_0)})}{\sqrt{n}\delta^2}\Big).$$

Arguing similarly as in the proof of Theorem 1, since the right-hand side above is upper-bounded by a deterministic function of $\mathcal{J}(\delta, \mathcal{F}_{\text{anti}})$, we can show that the summation term satisfies

$$\sum_{j=1}^{J}E_0^n[\phi_{n,j}^{(a_0)}(\delta)] \lesssim n^{-1/2}\mathcal{J}\big(\delta, \mathcal{F}_{\text{anti}}\big)\Big(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{\text{anti}})}{\sqrt{n}\delta^2}\Big).$$

Proceeding exactly as in the proof of Theorem 1, we aim to show that $\lim_{n\to\infty}P(n^{1/3}\delta_n > 2^S) \to 0$ as $S \to \infty$. Let $M_\varepsilon < \infty$ be such that $|Z_n| \le M_\varepsilon n^{-2/3}$ with probability at least $1 - \varepsilon$. Denote the event $A_s := \{r_n^{-1}\delta_n \in (2^s, 2^{s+1}]\}$. By a standard peeling argument and Markov's

inequality, we have for any rate $\{r_n\}_{n \geq 0}$,

$$
\begin{aligned}
P\big(r_n^{-1}\delta_n > 2^S\big) &= P\big(r_n^{-1}\delta_n > 2^S, |Z_n| \leq M_\varepsilon n^{-2/3}\big) + \varepsilon \\
&= \sum_{s=S}^{\infty} P\big(2^{s+1} \geq r_n^{-1}\delta_n > 2^s, |Z_n| \leq M_\varepsilon n^{-2/3}\big) + \varepsilon \\
&= \sum_{s=S}^{\infty} P\Big(A_s, \delta_n^2 \leq \frac{1}{J}\sum_{j\in[J]} \phi_{n,j}^{(a_0)}(\delta_n) + Z_n, |Z_n| \leq M_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\leq \sum_{s=S}^{\infty} P\Big(r_n^2 2^{2s} < \delta_n^2 \leq \frac{1}{J}\sum_{j\in[J]} \phi_{n,j}^{(a_0)}(r_n 2^{(s+1)}) + M_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\leq \sum_{s=S}^{\infty} P\Big(r_n^2 2^{2s} < \frac{1}{J}\sum_{j\in[J]} \phi_{n,j}^{(a_0)}(r_n 2^{(s+1)}) + M_\varepsilon n^{-2/3}\Big) + \varepsilon \\
&\leq \sum_{s=S}^{\infty} \frac{J^{-1}\sum_{j\in[J]} E\big[\phi_{n,j}^{(a_0)}(r_n 2^{(s+1)})\big] + M_\varepsilon n^{-2/3}}{r_n^2 2^{2s}} + \varepsilon \\
&\leq \sum_{s=S}^{\infty} 2^{-2s} r_n^{-2} \left( n^{-1/2} \mathcal{J}\big(r_n 2^{(s+1)}, \mathcal{F}_{\text{anti}}\big)\left(1 + \frac{\mathcal{J}(r_n 2^{(s+1)}, \mathcal{F}_{\text{anti}})}{\sqrt{n} r_n^2 2^{2(s+1)}}\right) + M_\varepsilon n^{-2/3}\right) + \varepsilon \\
&\lesssim \sum_{s=S}^{\infty} 2^{-2s} r_n^{-2} \left[ n^{-1/2} \sqrt{r_n 2^{(s+1)}}\left(1 + \frac{\sqrt{r_n 2^{(s+1)}}}{\sqrt{n} r_n^2 2^{2(s+1)}}\right) + M_\varepsilon n^{-2/3}\right] + \varepsilon \\
&\lesssim \sum_{s=S}^{\infty} \left[ 2^{-2s/3} r_n^{-3/2} n^{-1/2} + 2^{-2s} r_n^{-3} n^{-1} + 2^{-2s} r_n^{-2} M_\varepsilon n^{-2/3}\right] + \varepsilon.
\end{aligned}
$$

Choosing $r_n = n^{-1/3}$, the above implies

$$
\lim_{n\to\infty} P\big(r_n^{-1}\delta_n > 2^S\big) \lesssim \sum_{s=S}^{\infty} \frac{1}{2^s} + \varepsilon \overset{S\to\infty}{\to} \varepsilon.
$$

Since $\varepsilon > 0$ was arbitrary, we conclude that for all $a_0 \in \mathcal{A}$, $\delta_n = O_p(n^{-1/3})$ and therefore

$$
\delta_n^2 = \sum_{j\in[J]} \int \pi_0(a_0 \,|\, w)\big[\alpha_{n,j}^*(a_0 \,|\, w) - \alpha_{0,j}^*(a_0 \,|\, w)\big]^2 dP_0(w) = O_p(n^{-2/3}).
$$

Summing the left-hand side over $a_0 \in \mathcal{A}$, we obtain

$$
\sum_{j\in[J]} \int \big[\alpha_{n,j}^*(a \,|\, w) - \alpha_{0,j}^*(a \,|\, w)\big]^2 dP_0(a, w) = O_p(n^{-2/3}),
$$

as desired. The second inequality follows from the above display after applying Minkowski's inequality and using that $\{\alpha_{0,j}^* : j \in [J]\}$ minimizes the fold-averaged mean square error distance from $\alpha_0$ over all monotone transformations of $\{\pi_{n,j}(a_0 \,|\, \cdot) : j \in [J], a_0 \in \mathcal{A}\}$.

## Appendix E. Proof of Theorem 3

Denote the pseudo-outcome function:

$$\chi_0 : o = (w, a, y) \mapsto \mu_0(1, w) - \mu_0(0, w) + \left\{ \frac{1(a = 1)}{\pi_0(1 \mid w)} - \frac{1(a = 0)}{\pi_0(0 \mid w)} \right\} \{y - \mu_0(a, w)\}. \quad (14)$$

Note $\psi_0 = E_0[\chi_0(O)]$. For $j \in [J]$, let $o \mapsto \chi_{n,j}^*(o)$ denote the fold-specific estimator of the pseudo-outcome $\chi_0$ corresponding to the nuisance estimators $\alpha_{n,j}^*$ and $\mu_{n,j}$ of $\pi_0^{-1}$ and $\mu_0$. We note that $\psi_n^* = \frac{1}{n} \sum_{i=1}^n \chi_{n,j(i)}^*(O_i)$.

Adding and subtracting, we have the von Mises bias expansion:

$$\psi_n^* - \psi_0 = \frac{1}{J} \sum_{j=1}^J P_{n,j} \chi_{n,j}^* - P_0 \chi_0$$

$$= (P_n - P_0)\chi_0 + \frac{1}{J} \sum_{j=1}^J (P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0) + \frac{1}{J} \sum_{j=1}^J P_0(\chi_{n,j}^* - \chi_0).$$

To establish the desired result, it suffices to show that $(P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0) = o_p(n^{-1/2})$ and $P_0(\chi_{n,j}^* - \chi_0) = o_p(n^{-1/2})$ for each $j \in [J]$. In the following, we implicitly condition on $\{O_1, \ldots, O_n\}$ in our expectations. To show the latter bound, note by applying the law of total expectation and Hölder's inequality we get for each $j \in [J]$,

$$P_0(\chi_{n,j}^* - \chi_0)$$
$$= \sum_{a_0 \in \{0,1\}} (-1)^{1+a_0} E_0 \big[ \mu_{n,j}(a_0, W) - \mu_0(a_0, W) + 1(A = a_0)\alpha_{n,j}^*(a_0 \mid W)\{Y - \mu_{n,j}(a_0, W)\}$$
$$\quad - 1(A = a_0)\pi_0^{-1}(a_0 \mid W)\{Y - \mu_0(a_0, W)\} \big]$$
$$= \sum_{a_0 \in \{0,1\}} (-1)^{1+a_0} E_0 \big[ \mu_{n,j}(a_0, W) - \mu_0(a_0, W) + 1(A = a_0)\alpha_{n,j}^*(a_0 \mid W)\{\mu_0(a_0, W) - \mu_{n,j}(a_0, W)\} \big]$$
$$= \sum_{a_0 \in \{0,1\}} (-1)^{1+a_0} E_0 \big[ \{1(A = a_0)\alpha_{n,j}^*(a_0 \mid W) - 1\}\{\mu_0(a_0, W) - \mu_{n,j}(a_0, W)\} \big]$$
$$= \sum_{a_0 \in \{0,1\}} (-1)^{1+a_0} E_0 \big[ 1(A = a_0)\{\alpha_{n,j}^*(a_0 \mid W) - \pi_0^{-1}(a_0 \mid W)\}\{\mu_0(a_0, W) - \mu_{n,j}(a_0, W)\} \big]$$
$$\leq \sum_{a_0 \in \{0,1\}} E_0 \big[ \pi_0(a_0 \mid W) \{\alpha_{n,j}^*(a_0 \mid W) - \pi_0^{-1}(a_0 \mid W)\}^2 \big] \|\mu_0(a_0, \cdot) - \mu_{n,j}(a_0, \cdot)\|$$
$$\lesssim \max_{a_0 \in \{0,1\}} \big\| \alpha_{n,j}^*(a \mid \cdot) - \pi_0^{-1}(a \mid \cdot) \big\|_{P_{0,a_0}} \|\mu_0(a_0, \cdot) - \mu_{n,j}(a_0, \cdot)\|;$$

By Theorem 2, we have

$$\sum_{j \in [J]} \|\alpha_{n,j}^*(a \mid \cdot) - \pi_0^{-1}(a \mid \cdot)\|_{P_0} \lesssim o_p(n^{-1/3}) + \|f_{n,0}^{(a)} \circ \pi_{n,j}(a \mid \cdot) - \pi_0^{-1}(a \mid \cdot)\|_{P_0},$$

where $f_{n,0}^{(a)}$ is the optimal nondecreasing calibration for the cross-fitted propensity score estimator. Using symmetry and D2b and D2c, we conclude that $P_0(\chi_{n,j}^* - \chi_0) = o_p(n^{-1/2})$, as desired.

Next, we will show that $E_0^n \left[ |(P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0)| \right] = o(n^{-1/2})$, which implies, by Markov's inequality, that $(P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0) = o_p(n^{-1/2})$. To this end, we denote the function class:

$$\Big\{ o \mapsto \mu_{n,j}(1, w) - \mu_{n,j}(0, w) + \big\{ 1(a_0 = 1)\theta \circ \pi_{n,j}(1 \,|\, w) - 1(a_0 = 0)\theta \circ \pi_{n,j}(0 \,|\, w) \big\} \{ y - \mu_n(a_0, w) \}$$

$$- \mu_0(1, w) + \mu_0(0, w) - \big\{ 1(a_0 = 1)\pi_0^{-1}(1 \,|\, w) - 1(a_0 = 0)\pi_0^{-1}(0 \,|\, w) \big\} \{ y - \mu_0(a_0, w) \} : \theta \in \mathcal{F}_{\text{anti}}^{1/\eta} \Big\}$$

by $\mathcal{G}_{n,j}$. Of note, $\mathcal{G}_{n,j}$ is deterministic conditional on the training set $\mathcal{E}^{(j)}$. With probability tending to one, we have by definition and C2, $\alpha_{n,j}^*(a_0 \,|\, \cdot) = f_n^{(a_0)} \circ \pi_{n,j}(a_0 \,|\, \cdot)$ where $f_n^{(a_0)} \in \mathcal{F}_{\text{anti}}^{1/\eta}$. Hence, $\chi_{n,j}^* - \chi_0 \in \mathcal{G}_{n,j}$ with probability tending to one. Without loss of generality, we assume the event $\chi_{n,j}^* - \chi_0 \in \mathcal{G}_{n,j}$ occurs. Since by Conditions D1 and D2b, we know that $\mu_0(a, w)$ is bounded with probability tending to one for $a \in \{0, 1\}$; combined with the analysis we did in the proof of Theorem 1, we know that the function class $\mathcal{G}_{n,j}$ is uniformly bounded with probability tending to one, and therefore a Lipschitz transformation of $\mathcal{F}_{n,\text{anti}}^{(j,a_0),1/\eta} \times \mathcal{F}_{n,\text{anti}}^{(j,a_0),1/\eta}$.

An argument identical to the proof of Theorem 2 shows that, under C1-C2, that $\mathcal{J}(\delta, \mathcal{G}_{n,j}) \leq K\sqrt{\delta}$ for some fixed constant $K > 0$. Let $\delta_n := \|\chi_{n,j}^* - \chi_0\|_{P_0}$; observe that by Minkowski's inequality and Hölder's inequality we have

$$\delta_n \leq \max_{a_0 \in \{0,1\}} \|\mu_{n,j}(a_0, \cdot) - \mu_0(a_0, \cdot)\|_{P_0} + \max_{a_0 \in \{0,1\}} \|\alpha_{n,j}^*(a_0 \,|\, \cdot) - \pi_0^{-1}(a_0 \,|\, \cdot)\|_{P_0} C_Y$$

$$+ \max_{a_0 \in \{0,1\}} \big\| \alpha_{n,j}^*(a_0 \,|\, \cdot)\mu_{n,j}(a_0, \cdot) - \pi_0^{-1}(a_0 \,|\, \cdot)\mu_{n,j}(a_0, \cdot) + \pi_0^{-1}(a_0 \,|\, \cdot)\mu_{n,j}(a_0, \cdot) - \pi_0^{-1}(a_0 \,|\, \cdot)\mu_0(a_0, \cdot) \big\|_{P_0}$$

$$\lesssim \max_{a_0 \in \{0,1\}} \|\alpha_{n,j}^*(a_0 \,|\, \cdot) - \pi_0^{-1}(a_0 \,|\, \cdot)\|_{P_0} + \max_{a_0 \in \{0,1\}} \|\mu_{n,j}(a_0 \,|\, \cdot) - \mu_0(a_0 \,|\, \cdot)\|_{P_0} = o_p(1),$$

where we used Conditions C2, D1-D2 and Theorem 2. Hence, we can find a deterministic sequence $\varepsilon_n \downarrow 0$ such that $\delta_n = o_p(\varepsilon_n)$. Without loss of generality, we can assume that $\delta_n \leq C\varepsilon_n$ for some fixed $C > 0$, since this occurs with probability tending to one. Working on this event, we have

$$E_0^n \big[ |(P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0)| \big] \lesssim E_0^n \Big[ \sup_{g \in \mathcal{G}_{n,j} : \|g\| \leq C\varepsilon_n} |(P_{n,j} - P_0)g| \Big].$$

Noting $\mathcal{G}_{n,j}$ is uniformly bounded and that $\mathcal{J}(\varepsilon_n, \mathcal{G}_{n,j}) \to 0$ since $\mathcal{J}(1, \mathcal{G}_{n,j}) < \infty$, we can apply Theorem 2.1 of Van Der Vaart and Wellner (2011) conditional on $\mathcal{E}^{(j)}$ to conclude that:

$$E_0^n \big[ |(P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0)| \,|\, \mathcal{E}^{(j)} \big] \lesssim n^{-1/2} \mathcal{J}(\varepsilon_n, \mathcal{G}_{n,j}) \Big( 1 + \frac{\mathcal{J}(\varepsilon_n, \mathcal{G}_{n,j})}{\sqrt{n}\delta^2} \Big) = o(n^{-1/2}),$$

where the right-hand side is nonrandom. Taking the expectation of both sides, we find

$$E_0^n \big[ |(P_{n,j} - P_0)(\chi_{n,j}^* - \chi_0)| \big] = o(n^{-1/2})$$

as desired. As the above bounds hold for all $j \in [J]$, we conclude $\psi_n^* - \psi_0 = (P_n - P_0)\chi_0 + o_p(n^{-1/2})$. Noting that $\chi_0 - P_0\chi_0$ is the efficient influence function of the ATE for a nonparametric model, we also conclude that $\psi_n^*$ is regular and efficient.