How Much Weak Overlap Can Doubly Robust T-Statistics Handle?

Jacob Dorn*

February 19, 2025

Abstract

I study inference on the average causal effect under weak overlap. It is known that when the propensity score density is large near zero, no regular root-n-consistent estimators exist and standard estimators may fail to be asymptotically normal. As a result, previous approaches to causal inference under weak overlap have relied on nonstandard estimators, nonstandard confidence intervals, or have settled for targeting average effects within a nonstandard population. I show that statistical inference in this setting need not be so difficult: standard Wald confidence intervals for the standard doubly robust estimator are valid for the standard average causal effect, provided the estimator uses an appropriate trimming or clipping strategy to control extreme estimated propensities. The key is to clip at a rate decaying slowly enough to obtain asymptotic normality, but quickly enough that the bias introduced by clipping is second-order. I show that Wald confidence interval validity for clipped AIPW under weak overlap requires unusually stringent nuisance error conditions, but these conditions are achievable under appropriate smoothness conditions. The procedure also calls for a sequence of trimming or clipping thresholds, so I propose rules of thumb for the threshold choice. In simulations, clipped AIPW achieves near-nominal inference in large samples, but with only 1,000 or 10,000 observations, I find that the associated confidence intervals can slightly overreject the true null hypothesis. In an empirical application, the clipped AIPW estimator that targets the standard average treatment effect yields similar precision to the heuristic 10% fixed-trimming approach that changes the target sample.

1 Introduction

I study inference for causal effects in the presence of very weak overlap. By very weak overlap, I mean that propensity scores can be common enough near zero to introduce an infinite semiparametric efficiency bound, but not so common that the average effect is rendered unidentified.

*I am grateful for suggestions from Kevin Guo, Edward Kennedy, Samir Khan, Michal Kolesár, Lihua Lei, Xinwei Ma, Ulrich Müller, Yuya Sasaki, Yulong Wang, and Larry Wasserman. Artificial intelligence was used to suggest changes in the editing process. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

The framework I consider is almost standard, with the exception that I will permit very weak overlap. The econometrician observes n independent and identically distributed observations of covariates $X \in \mathbb{R}^d$, outcome $Y \in \mathbb{R}$, and binary treatment $D \in \{0, 1\}$ from some distribution P. For simplicity, I assume the econometrician would like to estimate the average potential outcome $\psi = \mathbb{E}[\mathbb{E}[Y \mid X, D = 1]]$; my results will carry through for average treatment effects and multiple discrete treatments. Two common strategies for estimating ψ are the Inverse Propensity Weighting (IPW) estimator $\hat{\psi}^{IPW} = \frac{1}{n}\sum \frac{D_i Y_i}{\hat{e}(X_i)}$, where $\hat{e}$ estimates the propensity function $e(X) = P(D = 1 \mid X)$, and the Augmented IPW (AIPW) estimator $\hat{\psi}^{AIPW} = \frac{1}{n}\sum \hat{\mu}_i + \frac{D_i}{\hat{e}(X_i)}(Y_i - \hat{\mu}(X_i))$, where $\hat{\mu}$ also estimates the outcome regression function $\mu(X) = E[Y \mid X, D = 1]$.

Traditional analysis of inverse propensity estimators proceeds under a strict overlap assumption that $e(X)$ is bounded away from zero, or at least that $E[1/e(X)]$ exists. If $E[1/e(X)]$ fails to exist, then IPW and AIPW estimators frequently divide by small numbers, the semiparametric bound is infinite, and IPW and AIPW may fail to be asymptotically normal even if the propensity function is known (Khan and Tamer, 2010; Ma and Wang, 2020; Heiler and Kazak, 2021).

Two standard approaches for making progress with IPW under weak overlap are to "clip" (Winsorize) or "trim" (drop from the dataset) observations with propensity scores below a given threshold. However, this strategy introduces a bias-variance tradeoff. On the one hand, if the threshold is set at a sufficiently large value, the IPW estimator will have an approximately normal sampling distribution, but the distribution will be centered at the wrong estimand (Khan and Tamer, 2010; Ma and Wang, 2020). On the other hand, if the threshold is set at a sufficiently small value, the IPW will have limited bias, but the large inverse propensity scores may introduce a nonstandard asymptotic distribution that requires a nonstandard statistical inference strategy (Ma and Wang, 2020; Heiler and Kazak, 2021). Empirical practice has generally favored the use of fixed trimming thresholds that introduce bias even asymptotically (Crump et al., 2009).

The main contribution of this paper is to show that replacing the IPW estimator with the AIPW estimator can solve the bias-variance tradeoff that applies to clipping or trimming for IPW. In particular, I show that Wald confidence intervals constructed using AIPW can achieve well-calibrated coverage for the target causal effect under even very weak overlap. I consider the AIPW estimator with a sequence of clipping thresholds b_n,

$$\hat{\psi}^{AIPW}_{clip}(b_n) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(X_i) + \frac{D_i(Y_i - \hat{\mu}(X_i))}{\max\{\hat{e}(X_i), b_n\}} = \frac{1}{n}\sum_{i=1}^{n} \phi(Z_i \mid b_n, \hat{\eta}), \tag{1}$$

where $\hat{\eta}$ is an estimate of the nuisance functions $\eta = (e(\cdot), \mu(\cdot))$. Results for the trimmed AIPW estimator

follow by analogous arguments. I consider the standard Wald confidence interval

$$\left[\hat{\psi}_{clip}^{AIPW}(b_n) + z_{\alpha/2}\hat{\sigma}_n, \hat{\psi}_{clip}^{AIPW}(b_n) + z_{1-\alpha/2}\hat{\sigma}_n\right],$$

with the standard error estimate

$$\hat{\sigma}_n = n^{-1/2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\phi(Z_i \mid b_n, \hat{\eta})^2 - \hat{\psi}_{clip}^{AIPW}(b_n)^2}.$$

I provide sufficient regularity conditions under which this standard confidence interval covers the standard causal estimand ψ with probability tending to $1-\alpha$. The results are uniform, potentially including families of distributions with overlap so weak that ψ nearly fails to be identified at all.

The key feature of the clipped AIPW estimator is that it is Neyman orthogonal with respect to nuisance function errors. Under Neyman orthogonality, small errors in the propensity and outcome nuisance functions near the true values have a squared effect on bias. Clipping can be viewed as an intentional error in the estimated propensity function that increases bias in order to reduce variance and achieve asymptotic normality. AIPW orthogonalizes the clipping step's added bias, while IPW does not achieve such a debiasing. I formalize this intuition to prove that the clipped AIPW estimator's t-statistics can be well-calibrated under appropriate regularity conditions. While the formal result requires some care to handle division by arbitrarily small numbers, the main technical contribution is to characterize sufficient regularity conditions for the Neyman orthogonality intuition to carry through.

I provide guarantees over a uniform version of Ma and Wang (2020)'s model family. I assume that there is a lower bound on the CDF of the propensity score of the form $P(e(X) \leq \pi) \leq C\pi^{\gamma_0 - 1}$, where $\gamma_0 > 1$ corresponds to a distribution tail index. $\gamma_0 = 2$ corresponds to a uniform density bound under which appropriate moments for IPW inference may barely fail to exist. When γ_0 is allowed to be below 2, the density of propensities around zero can be infinite and the unclipped AIPW estimator may fail to be asymptotically normal. As γ_0 tends to 1, increasingly weak overlap is allowed; when $\gamma_0 = 1$, the average treatment effect may cease to be identified.

These results for statistical inference require stronger convergence rate assumptions than are required under strict overlap. Under strict overlap, a sufficient condition for valid inference with the AIPW estimator is that the convergence rate $r_{\mu,n}$ of the outcome regression estimator $\hat{\mu}(\cdot)$ and the convergence rate $r_{e,n}$ of the propensity score estimator $\hat{e}(\cdot)$ satisfy the product-of-errors condition $n^{1/2}r_{\mu,n}r_{e,n} \to 0$. Under weak overlap, I require the stronger condition $n^{1/2}r_{\mu,n}r_{e,n}^{\min\{1,\gamma_0/2\}} \to 0$. This characterization creates an asymmetry between the contribution of the outcome and propensity score rates: if the outcome regression estimator

achieves a parametric consistency rate of $n^{-1/2}$, then AIPW can accommodate an arbitrarily heavy inverse propensity tail so long as the propensity score is consistent. Conversely, even if the propensity score is estimated at a parametric rate, the outcome regression estimator may need to achieve a consistency rate as fast as $n^{-1/4}$ in order to accommodate very weak overlap. I show that these stronger rate requirements are sharp, in the sense that if $n^{1/2} r_{\mu,n} r_{e,n} \to 0$ but $n^{1/2} r_{\mu,n} r_{e,n}^{\min\{1,\gamma_0/2\}} \to \infty$, then there exists a distribution for which the nuisance rates $r_{\mu,n}$ and $r_{e,n}$ are insufficient to achieve valid Wald confidence intervals.

Further, I show that a given outcome regression rate may be more difficult to achieve under weak overlap. I focus on Nadaraya-Watson regression under Hölder continuity restrictions on the conditional outcome mean function. I show that the weak overlap tail parameter γ_0 can play a role equivalent to adding $d/(\gamma_0 - 1)$ dimensions under traditional outcome regression. The problem is that under weak overlap, certain regions of the covariate space can have small propensities, and therefore can have few treated observations to use in regression. As a result, a given outcome regression rate may require unusually strong smoothness assumptions.

This paper leverages these theoretical results to provide a precise answer to how much weak overlap doubly robust t-statistics can handle. Under a Lipschitz-continuity restriction on the conditional mean outcome and β_e-order Hölder continuity of the propensity function, clipped AIPW can handle a tail parameter as small as $\frac{2(d+1)+d^2/\beta_e}{d+2}$. In one dimension, an infinitely-differentiable propensity function allows for reliable Wald confidence intervals if $\gamma_0 > \frac{5}{3}$. Under outcome Lipschitz-continuity, any $\beta_e > d^2/2$ allows Wald confidence intervals to be valid under some level of overlap weakness that keeps unadjusted IPW from being asymptotically normal.

I leverage these new theoretical results to provide new guidance for empirical work. I provide several rules of thumb for the choice of clipping threshold. In my favored regime, the econometrician is willing to posit a minimal consistency rate for one of the two nuisance function estimates. Such a minimal rate is often implied by theoretical justifications for a given nuisance estimator. I provide simple rules of thumb that estimate the clipping threshold that imposes the laxest possible requirement on the other nuisance estimator in order to guarantee Wald confidence interval validity. A third rule of thumb calculates a clipping threshold based on imposing an equal minimal consistency rate on both the outcome and propensity nuisance estimates. This final rule of thumb is theoretically attractive, but practically inconvenient due to the relative difficulty of outcome regression under weak overlap.

I also provide some intuition for clipped or trimmed AIPW with parametric nuisance estimates. When both the outcome and propensity functions are estimated at consistent parametric rates, my existing analysis applies and Wald confidence interval validity follows. If neither is consistent, then AIPW and IPW are inconsistent. The more interesting cases involve when only one nuisance is estimated consistently. In this

case, clipped AIPW remains asymptotically normal. However, the behavior diverges in a similar manner to Ma et al. (2023)'s analysis of trimmed AIPW with tailored debiasing. With only a consistent propensity estimate, clipped AIPW will have first-order bias like clipped IPW. However, with a consistent parametric outcome estimate, clipped AIPW should be asymptotically normal around the standard causal effect.

In simulations, I find that clipped AIPW achieves the promised properties asymptotically. I consider a setting of very weak overlap with nonparametric local linear outcome regression and propensity estimates. Unadjusted IPW and AIPW estimators perform poorly, with large errors and non-normal asymptotic t-statistic distributions. Clipped IPW displays its known asymptotic normality with first-order bias. Clipped AIPW achieves smaller bias than clipped IPW. With access to 1,000 or 10,000 observations, I find that p-values based on clipped AIPW t-statistics exhibit moderate overrejection. In large samples with 100,000 observations, a Kolmogorov-Smirnov test based on 5,000 simulations is unable to reject a null hypothesis that clipped AIPW p-values on the true effect are exactly uniformly distributed.

I apply the clipped AIPW estimator to data on right heart catheterization. I consider the setting of Connors et al. (1996), which has become a canonical setting with weak overlap, including providing the empirical application for Crump et al. (2009)'s paper proposing a 10% trimming rule of thumb. I compare clipped AIPW with a rule-of-thumb clipping threshold to an AIPW estimator applied to the 10% trimmed sample. I find that by including observations with small estimated propensities, the clipped AIPW strategy increases the estimated harm of the procedure by 0.17 standard errors, while increasing the estimated standard error by 5.1%. These results show that targeting the full-population treatment effect does not need to introduce a major efficiency loss.

Related Literature. Weak overlap is a common phenomenon in practice and in theory. The dominant response to weak overlap in inverse propensity score practice is trimming: dropping samples with small propensity estimates in order to estimate average effects within a more precise population (Currie and Walker, 2011; Bailey and Goodman-Bacon, 2015; Galiani et al., 2005), typically following the 10% rule of thumb from Crump et al. (2009). Other work for estimating causal effects includes proposals to reweight towards higher-precision populations (Yang and Ding, 2018; Li et al., 2018) or clipping strategies to Winsorize weights above (Lee et al., 2011; Ionides, 2008).[1] D'Amour et al. (2021) argue that weak overlap is likely to be prevalent in modern settings with high-dimensional covariates. Imbens (2004) argues that changing the target estimand may be necessary in the absence of sufficient precision.

The theoretical literature so far has either proposed a nonstandard causal estimator, targeted a non-standard causal estimand, or required nonstandard techniques to construct confidence intervals. Khan and

[1] Awkwardly, the epidemiological literature sometimes refers to the Winsorization strategy as "trimming." My results hold for both dropping or Winsorizing extreme propensities, so the confused reader can view this as a work deriving simple asymptotics for trimmed AIPW regardless of their preferred meaning of "trim."

Tamer (2010) show that very weak overlap yields an irregularly identified parameter, an infinite semiparametric efficiency bound, a limitation to slow estimation rates for the traditional average causal effects, and no clear notion of best estimator. An important theoretical literature has proposed novel point and confidence interval estimators with desirable properties under weak overlap (Rothe, 2017; Armstrong and Kolesár, 2017, 2021; Sasaki and Ura, 2022; Ma et al., 2023; Chaudhuri and Hill, 2024), but to my knowledge there has been little take-up by practitioners. Ma and Wang (2020) and Khan and Ugander (2022) show that sufficiently trimmed AIPW and IPW can remain asymptotically normal, but at the cost of introducing first-order bias for the standard causal effects that often calls for a nonstandard debiasing strategy. Crump et al. (2009), Yang and Ding (2018), Li et al. (2018), and Goldsmith-Pinkham et al. (2024) propose targeting estimators that are easier to estimate under weak overlap; when the econometrician prefers to target the traditional average causal effect that I consider here, then these proposals introduce a discontinuous estimand based on whether or not the econometrician detects meaningful overlap weakness. Ma and Wang (2020) and Heiler and Kazak (2021) propose using self-normalized subsampling methods that enable valid statistical inference for standard estimands without clipping or trimming, but empirical practice has favored simple t-tests. Heiler and Kazak also find that estimated untrimmed AIPW is first-order equivalent to an oracle estimator with an asymptotic alpha-stable distribution if the product of nuisance estimation rates is of a lower order than the oracle standard deviation; I find this result does not extend to the clipped AIPW estimator that I show is asymptotically normal. Ma et al. (2024) and Lei et al. (2021) propose statistical tests under a null of sufficient or strict overlap, respectively, presumably in the hopes of avoiding these complications. Relative to these procedures, I analyze a standard procedure of estimating the standard outcome and propensity nuisance functions, clipping or trimming extreme propensity scores, and then building Wald confidence intervals. The only difference from common practice under strict overlap is the simple clipping or trimming step, and the only difference from common practice under weak overlap is the use of a clipping threshold that goes to zero and the requirement of an AIPW estimator that debiases error in the clipped region.

The plan of the paper is as follows. Section 2 presents the setting and main theoretical results. Section 3 interprets these results by considering special cases, proving limitations of these results, and proposing rules of thumb for empirical use. Section 4 presents numerical results for simulations and the empirical application to right-heart catheterization. Section 5 concludes.

Notation. I follow Heiler and Kazak (2021) and use "strict overlap" to refer to the case in which there is some $\epsilon > 0$ such that $e(X) \geq \epsilon$ almost surely; I use "weak overlap" to refer to the case in which the infimum of the support of $e(X)$ is zero, which is sometimes called "limited overlap" (Khan and Tamer, 2010; Chaudhuri and Hill, 2024). I focus my attention on distributions with weak overlap that may possess subexponential tails. I use "very weak overlap" to refer to case in which the associated heavy tails

can fail to generate inverse propensity second moments, a class which is sometimes called "heavy tailed" (Chaudhuri and Hill, 2024). I use "somewhat weak overlap" to refer to the case in which I allow only subexponential tails that do not yield very weak overlap, a class which is sometimes said to satisfy "strict overlap" or "strong overlap" (Heiler and Kazak, 2021). I write $\tilde{\psi}^{AIPW}_{(Orcl)}(b_n) = \frac{1}{n}\sum \phi(Z \mid b_n, \eta)$ for the oracle AIPW estimate with clipping threshold b_n and $\sigma_n = n^{-1/2}\sqrt{\frac{1}{n}\sum \phi(Z \mid b_n, \eta)^2 - \tilde{\psi}^{AIPW}_{(Orcl)}(b_n)^2}$ and $\hat{\sigma}_n = n^{-1/2}\sqrt{\frac{1}{n}\sum \phi(Z \mid b_n, \hat{\eta})^2 - \left(\frac{1}{n}\sum \phi(Z \mid b_n, \hat{\eta})\right)^2}$ for the associated oracle and estimated sample standard deviation, respectively. I refer to regions of the covariate space in which the propensity can be arbitrarily close to zero as singularities. I use the notation $E_P[\cdot]$ and $E[\cdot]$ to refer to the expectation under the maintained distribution P, and I use the notation $\psi(P)$ to refer to $E_P[E_P[Y \mid X, D = 1]]$ where the right-hand side is well-defined under P. I abuse notation and write $\sup_{P \in A} B$ to refer to the supremum of B over distributions P in A under any maintained restrictions on the distribution and nuisance functions. I write that a set of nuisance functions are cross-fit if the data is partitioned into K folds and the nuisance functions in fold k are independent of the data in fold k. I write $A_n \leq_P B_n$ to refer to the case that for all $\epsilon > 0$, $P(A_n > B_n + \epsilon) \to 0$. I write $P(E_n)$ for the probability of event E_n occurring under the distribution P, with the number of draws n sometimes left implicit. I use the notation $c_n \ll d_n$ for nonnegative sequences c_n, d_n to indicate that $d_n > 0$ for all n large enough and $c_n/d_n \to 0$. I use the notation $c_n \precsim d_n$ and $d_n \succsim c_n$ to indicate that there is some $\delta > 0$ such that $d_n \geq \delta c_n$ for all n large enough. I write $c_n = o_P(d_n)$ for sequence of $d_n > 0$ to indicate that for all $\delta > 0$, $P(|c_n|/d_n > \delta) \to 0$; if there is only one distribution in a statement, $c_n = o(d_n)$ should be understood to mean $c_n = o_P(d_n)$. I use log to refer to the natural logarithm and $a \vee b$ to indicate $\max\{a, b\}$. I define Hölder smoothness using a multivariate version of the notation of Tsybakov (2009): a function f is in the Hölder smoothness class $\Sigma(\beta, L)$ if the $\lfloor\beta\rfloor$-order multivariate derivatives $D^\alpha f$ satisfy $\|D^\alpha f(x) - D^\alpha f(x')\| \leq L\|x - x'\|^{\beta - \lfloor\beta\rfloor}$. For simplicity, I use Nadaraya-Watson regression to refer to specifically regression with uniform bandwidth: I write $\hat{\mu}^{(NW)}(x \mid h) = \frac{\sum D1\{\|X-x\| \leq h\}Y}{\sum D1\{\|X-x\| \leq h\}}$ when feasible and $\hat{\mu}^{(NW)}(x \mid h) = 0$ when no nearby treated observations are available.

2 Setting, Consistency, and Asymptotic Normality

This section presents the core theoretical results for asymptotic normality.

2.1 Setting

I derive uniform convergence rates under lower bounds on overlap weakness. I follow Ma and Wang (2020) and parameterize overlap weakness through a tail parameter γ_0. Unlike their analysis, the results will be uniform over a model family $\mathscr{P}$ satisfying certain restrictions. The first is some basic regularity conditions.

Assumption 1. Let $\mathscr{P}$ be a nonempty family of distributions that satisfy the following conditions for some $M, q, \sigma_{\min}, \pi_{\min}, C, \gamma_0$:

(a) *Conditional moments.* $\mathbb{E}[|Y - E[Y \mid X, D]|^q \mid X, D] \leq M^q < \infty$ almost surely for some $q > 3$.

(b) *Unconditional moments.* $Var(E[Y \mid X, D = d]) \leq M$.

(c) *Residuals.* $\text{Var}(Y \mid X, D) \geq \sigma^2_{\min} > 0$ almost surely.

(d) *Treated fraction.* $1 - \pi_{\min} \geq P(D = 1) \geq \pi_{\min} > 0$.

(e) *Propensity tail.* $P(e(X) \leq \pi) \leq C\pi^{\gamma_0 - 1}$ for all $\pi \in [0, 1]$ and some $\gamma_0 > 1$.

For a distribution $P \in \mathscr{P}$, I abuse notation by writing $\psi = E_P[E_P[Y \mid X, D = 1]]$.

Definition 1 generalizes Ma and Wang (2020)'s slowly varying tails assumption. Assumptions 1(a) through 1(d) are regularity conditions that rule out cases like perfectly predictable outcomes. Assumption 1(e) provides the substantial restriction on $\mathscr{P}$: overlap may be weak in the sense that γ_0 is finite, but there is some minimal γ_0 and C that provides a lower bound on the propensity's tail behavior. Under strict overlap, Assumption 1(e) holds for any finite $\gamma_0 > 1$, and most results here will hold by replacing γ_0 with infinity. Under weak overlap, Assumption 1(e) may only hold for some values of γ_0, in which case the inverse propensity distribution may be heavy-tailed. As Ma and Wang note, the case $\gamma_0 = 2$ roughly corresponds to a uniform distribution of propensities near the origin, and is the knife-edge case at which unclipped IPW and AIPW estimators become non-Gaussian. I refer to the case $\gamma_0 > 2$ as the "somewhat weak overlap" case and refer to the case of $\gamma_0 < 2$ as the "very weak overlap" case. As γ_0 shrinks below 2, overlap is permitted to be increasingly weak. $\gamma_0 \leq 1$ corresponds to no bound on the propensity distribution.

I will require certain rates on the nuisance functions $e(X)$ and $\mu(X)$. I write the worst-case rates as $r_{e,n}$ and $r_{\mu,n}$.

Assumption 2 (Cross-fitting)**.** The nuisances $\hat{\mu}$ and $\hat{e}$ are estimated with cross-fitting with a fixed number of folds K. If n_k is the number of observations per fold, then $\inf_k n_k / \sup_k n_k \to 1$. Further, for all $k \in 1, \ldots, K$ and all $P \in \mathscr{P}$, the cross-fit nuisances satisfy the uniform consistency rates $\mathbb{E}_P[\|\hat{\mu}_n^{(-k)} - \mu\|_\infty] \leq r_{\mu,n}$ and $\mathbb{E}_P[\|\hat{e}_n^{(-k)} - e\|_\infty] \leq r_{e,n}$ where $r_{\mu,n}, r_{e,n}$ are uniformly bounded above.

Cross-fitting is a common strategy for simplifying the analysis of Neyman-orthogonal estimators like AIPW (Chernozhukov et al., 2018). In practice, nuisances satisfying Assumption 2 may only be achieved with arbitrarily high probability. The uniformity condition is needed to handle regions of x with singularities, but can be bypassed with L_2 error conditions in other regions. Such uniformity assumptions are standard in studying semiparametric estimators under irregular identification (Semenova, 2024).

2.2 Estimator and Consistency

My formal analysis considers the clipped AIPW estimator with cross-fit nuisance function estimates. I begin by providing sufficient conditions for consistency.

Recall that the clipped AIPW estimator of ψ is:

$$\hat{\psi}_{clip}^{AIPW}(b_n) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{F}^k}\phi\left(Z_i \mid b_n, \hat{\eta}^{(-k)}\right),$$

where $\phi(Z \mid b, \hat{\eta}) = \hat{\mu}(X) + \frac{D(Y-\hat{\mu}(X))}{\max\{\hat{e}(X),b\}}$. In that equation, $\mathcal{F}^k$ is the set of observations i randomly partitioned in fold k, $\hat{\eta}^{(-k)}$ is the nuisance function estimates constructed only on observations in folds other than k, and ϕ is defined in Equation (1). The unclipped AIPW estimator is the special case of $b_n = 0$. I analyze the clipped AIPW estimator because results for the trimmed AIPW estimator follow somewhat more easily.

A standard result for the unclipped AIPW estimator is double robustness: when $e(X)$ is bounded away from zero, unclipped AIPW is consistent for ψ if either $r_{e,n}$ or $r_{\mu,n}$ tends to zero. The existence of weak overlap introduces a subtlety to double robustness.

Proposition 1 (Consistency)**.** *Suppose b_n satisfies $n^{-1/2} \ll b_n \ll 1$, the conditions of Assumption 2 hold, and either (i) $r_{e,n} b_n^{\min\{\gamma_0-2,0\}} \to 0$ or (ii) $r_{\mu,n}\frac{r_{e,n}+b_n}{b_n} \to 0$. Then for all $\epsilon > 0$,*

$$\sup_{P\in\mathscr{P}} P\left(\left|\hat{\psi}_{clip}^{AIPW}(b_n) - \psi(P)\right| > \epsilon\right) \to 0.$$

Condition (i) is a stronger condition than the classic strict overlap condition that $r_{e,n}$ or $r_{\mu,n}$ tends to zero. It requires that $r_{e,n}$ go to zero faster than $b_n^{2-\gamma_0}$, so that as overlap is allowed to be weaker, the propensity consistency rate may need to be as fast as b_n itself. Condition (ii) is also a stronger condition than the classic $r_{\mu,n} \to 0$ condition. The condition allows for a meaningful fraction of the data may be clipped even asymptotically, in which case the outcome regression error rate must offset the positive probability of assigning an inverse propensity weight of b_n^{-1}.

These theoretical results provide sufficient conditions for black-box nuisance estimators and clipping thresholds to yield valid Wald confidence for the usual average treatment effect. In the next section, I attempt to interpret these results, provide some key limitations that may not be obvious on first appearance, and use these theoretical results to propose some rules of thumb for empirical use.

2.3 Statistical Inference

This subsection presents the main theoretical claims of the paper. It shows that under suitable rate restrictions, the clipped AIPW estimator is first-order equivalent to an oracle clipped AIPW estimator, both estimators are consistent and asymptotically normal, and simple Wald confidence intervals are well-calibrated.

A common strategy for deriving confidence intervals for unclipped AIPW under strict overlap is Neyman orthogonality. In those classic settings, the difference between the feasible AIPW estimator with estimated nuisance functions and the hypothetical oracle AIPW estimator with known nuisance functions is

$$\frac{1}{n}\sum \phi(Z \mid 0, \hat{\eta}) - \phi(Z \mid 0, \eta) = \frac{1}{n}\sum (\hat{\mu} - \mu)\left(\frac{D}{\hat{e}} - 1\right) + (Y - \mu)\left(\frac{D}{\hat{e}} - \frac{D}{e}\right).$$

Intuitively, the regression errors $\hat{\mu} - \mu$ are debiased by the inverse propensity $\frac{D}{\hat{e}}$ estimates of the number one. As a result, in classical settings, slowly consistent nuisance estimates can yield quickly consistent causal estimates. When all nuisances are consistent at $o(n^{-1/4})$ rates and $e(X)$ is bounded away from zero, estimation error in inverse propensities is of the same order as estimation error in the propensities themselves, classical AIPW estimates are first-order equivalent to oracle estimates with known nuisances, and simple Wald confidence intervals cover the true causal effect by appeal to the oracle AIPW estimator.

Under very weak overlap, the clipped AIPW estimator does not obtain the standard AIPW orthogonality benefit. The analogous decomposition for clipped AIPW is

$$\frac{1}{n}\sum \phi(Z \mid b_n, \hat{\eta}) - \phi(Z \mid b_n, \eta) = \frac{1}{n}\sum (\hat{\mu} - \mu)\left(\frac{D}{\max\{\hat{e}, b_n\}} - 1\right) + (Y - \mu)\left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}}\right).$$

Above the clipping threshold b_n, the clipped AIPW estimator's nuisance estimation error enjoys a product-of-errors character that is similar to classical settings, albeit without the usual ability to interchange between propensity and inverse propensity error. Below the clipping threshold, the regression errors are not orthogonalized by a consistent propensity estimate. No wonder the previous literature for IPW under weak overlap has targeted nonstandard estimands or appealed to nonstandard procedures.

I exploit that the clipped AIPW estimator enjoys a subtly different form of orthogonality. For observations below the clipping threshold, the clipped AIPW estimator fails to orthogonalize regression error using the propensity estimates. However, as the clipping threshold tends to zero, increasingly little mass is clipped so that regression error in the clipped region has an increasingly small effect on bias.

The orthogonality-by-reducing-b_n character of clipped AIPW introduces a tradeoff. Smaller clipping thresholds b_n introduce less bias from clipping. Larger clipping thresholds b_n make the product-of-errors

condition above b_n more manageable. My formal contribution is to show that there can be a Goldilocks range where b_n tends to zero neither too quickly nor too slowly, so that the clipped AIPW estimator is first-order equivalent to the oracle AIPW estimator and is asymptotically normal.

My results for asymptotic normality and statistical inference will proceed under the following rate requirements.

Assumption 3 (Minimal rates)**.** Assumption 2 holds, with the following rates on the regression error $r_{\mu,n}$ and the propensity error $r_{e,n}$:

(a) *Consistency.* $r_{\mu,n}, r_{e,n} \to 0$.

(b) *Product of errors.* For some $\eta > 0$, $r_{\mu,n} r_{e,n} \left(1 + b_n^{(\gamma_0-2)/2} n^{\eta}\right) \ll n^{-1/2}$.

(c) *Regression error near singularities.* $r_{\mu,n} b_n^{\gamma_0/2} \ll n^{-1/2}$.

(d) *Asymptotically known thresholding.* $r_{e,n} \ll b_n$.

I discuss these rates further in Section 3.1: for example, when $\gamma_0 = 1.5$ and $r_{\mu,n} = n^{-1/5}$, then $r_{e,n} \ll n^{-0.4}$ will suffice for these conditions, provided $r_{e,n} \ll b_n \ll n^{-0.4}$. Shared regression rates of $n^{-1/3}$ will always suffice for these conditions, provided the clipping threshold b_n goes to zero at a rate sufficiently close to $n^{-1/3}$. Under very weak overlap ($\gamma_0 \in (1, 2)$), the product-of-errors condition (b) is stronger than the standard product-of-errors condition $r_{\mu,n} r_{e,n} \ll n^{-1/2}$. I only analyze cases in which $r_{e,n} \ll b_n$, so that the expected misclassification rate between oracle and estimated clipping tends to zero. These rates are always weaker than joint $n^{-1/2}$ rates achievable with parametric assumptions. Propensity rates under weak overlap are no more difficult than common practice. Outcome regression rates can be more difficult in the regions of singularities with small propensity values, both because of reduced treated observations and the possibility of irregular designs.

Stronger rates will be needed to handle unusual distributions. I therefore provide two alternative further assumptions: a distributional smoothness assumption and a faster-rate assumption.

Assumption 4 (Nongeneracy or faster rates)**.** One of the following two conditions hold:

(i) *Nondegenerate overlap.* There exists some $\rho > 0$ such that for all $P \in \mathscr{P}$ and $\pi \in [0, 1]$, $P(e(X) \leq \pi/2) \leq (1 - \rho) P(e(X) \leq \pi)$.

(ii) *Faster rates.* $r_{\mu,n} b_n^{(\gamma_0-1)2/\gamma_0} \ll n^{-1/2}$.

Assumption 4(i) is a uniform version of the requirement that $P(e(X) \leq x) = c(x) x^{\gamma_0 - 1}$ for $c(x)$ tending to a constant at zero. The definition formalizes the notation that a distribution may place some weight near the origin, but it may not place weight at arbitrarily adversarial points near the origin. When $\gamma_0 < 2$,

Assumption 4(ii) is stronger than Assumption 3(c). As γ_0 tends to one, the condition approaches the parametric requirement $r_{\mu,n} = O(n^{-1/2})$.

I now provide the main theoretical result.

Theorem 1 ((Slow) Asymptotic Normality)**.** *Suppose b_n satisfies $n^{-1/2} \ll b_n \ll 1$, and Assumptions 1, 2, 3, and 4 hold. Then the clipped AIPW estimator is oracle-equivalent:*

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}} \sigma_n^{-2} E_P \left[\left(\hat{\psi}^{AIPW}_{clip}(b_n) - \tilde{\psi}^{AIPW}_{(Orcl)}(b_n) \right)^2 \right] = 0.$$

Further, clipped AIPW is asymptotically normal:

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}} \sup_{t\in\mathbb{R}} \left| P\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n) - \psi(P)}{\hat{\sigma}_n} \leq t \right) - \Phi(t) \right| = 0.$$

Theorem 1 is the core theoretical claim of this paper. The first result shows that the clipped AIPW estimator is first-order equivalent to an oracle estimator with known nuisances: the effect of nuisance estimation error on the treatment effect estimate tends to zero faster than the standard deviation of the oracle estimator. The second result leverages this first-order equivalence to characterize the asymptotic distribution of the clipped AIPW estimates and t-statistics: the estimator is asymptotically normal, and estimated t-statistics are asymptotically standard normal.

Both results are standard for AIPW under strict overlap, but substantial care is required to handle unbounded inverse propensities under weak overlap. The argument for normality builds on Ma and Wang (2020)'s proof that aggressively-trimmed IPW with known propensities achieves asymptotic normality with first-order bias. I extend their argument to a uniform family of distributions using the Berry-Esseen Theorem. Because AIPW is a debiasing estimator, this first extension shows that trimmed or clipped AIPW with known nuisance functions achieves asymptotic normality with zero bias. The task of Theorem 1 is to show that replacing the true nuisances with estimated nuisances has a second-order effect on clipped AIPW estimates under appropriate conditions. This is nontrivial, because under weak overlap, there is an asymptotically unbounded number of observations with arbitrarily large inverse propensities with even known nuisance functions. Nevertheless, by taking appropriate care and leveraging that clipping introduces bias by increasing propensities and reducing inverse propensities, I am able to show that the effect of nuisance estimation is second-order even under such weak overlap that unadjusted AIPW fails to be asymptotically normal and no regular root-n estimators exist.

A careful analysis Assumption 3 suggests that smaller values of b_n are preferable because they admit weaker rate requirements. However, the following result shows that such robustness does not come for free.

I first characterize the consistency rate of a generic estimate as follows.

Proposition 2 (Black-box consistency rate)**.** *Suppose the assumptions of Proposition 1 hold. Then there exist positive constants* $c_{\min}$ *and* $c_{\max}$ *such that* $c_{\min} n^{-1} \mathbb{E}_P \left[\frac{D}{\max\{e(X), b_n\}^2} \right] \leq \sigma_n^2 \leq c_{\max} n^{-1} \mathbb{E}_P \left[\frac{D}{\max\{e(X), b_n\}^2} \right]$ *for all* $P \in \mathscr{P}$*, where* $\sigma_n^2 = n^{-1} \left(\frac{1}{n} \sum \phi(Z \mid b_n, \eta)^2 - \tilde{\psi}^{AIPW}_{(Orcl)}(b_n)^2 \right)$ *is the oracle sample variance.*

Weaker overlap corresponds to larger values of $\mathbb{E}_P[D/\max\{e(X), b_n\}^2]$ and slower consistency rates. Conditional on P, larger values of b_n correspond to a smaller value of $\mathbb{E}_P[D/\max\{e(X), b_n\}^2]$, faster oracle consistency rate, and greater asymptotic power.

Proposition 2 implies a worst-case consistency rate over distributions in $\mathscr{P}$.

Corollary 1 (Worst-case consistency rate)**.** *Suppose* $\gamma_0 < 2$ *and let* b_n *be a fixed sequence of* b_n *satisfying* $1 \gg b_n \gg n^{-1/2}$*. There exists a* $C' > 0$ *such that for* $\mathscr{P}$ *satisfying Assumption 1,* $C' n^{-1} b_n^{\gamma_0 - 2} \geq \sup_{P \in \mathscr{P}} \sigma_n^2$ *for all* n *large enough. Further, there exists a (single-element) family* $\mathscr{P}$ *satisfying Assumption 1 and a* $C'' \in (0, C')$ *such that* $\sup_{P \in \mathscr{P}} \sigma_n^2 \geq C'' n^{-1} b_n^{\gamma_0 - 2}$ *for all* n *large enough.*

The combination of Corollary 1 and Theorem 1 yields a trade-off: smaller values of b_n yield laxer requirements on regression estimation near singularities, but lead to larger variance and slower consistency. The rate $n^{-1} b_n^{\gamma_0 - 2}$ is a worst-case consistency rate in b_n: every distribution in $\mathscr{P}$ achieves a consistency at least as fast as $n^{-1} b_n^{\gamma_0 - 2}$, and it is possible to find a distribution for which the consistency rate is no faster.

Finally, I show that Theorem 1 yields the natural result for inference: simple t-tests based on Wald confidence intervals are well-calibrated.

Corollary 2 (T-tests are well-calibrated)**.** *Suppose the conditions of Theorem 1 hold. Consider the Wald confidence interval* $\hat{\mathcal{C}}_n = \left[\hat{\psi}^{AIPW}_{clip}(b_n) + z_{\alpha/2} \hat{\sigma}_n, \hat{\psi}^{AIPW}_{clip}(b_n) + z_{1-\alpha/2} \hat{\sigma}_n \right]$*. Then*

$$\limsup_{n \to \infty} \sup_{P \in \mathscr{P}} \left| P(\psi(P) \in \hat{\mathcal{C}}_n) - (1 - \alpha) \right| = 0.$$

Taken together, this sub-section yields a remarkable result for practice. The distribution P may place so much propensity mass near the origin that the semiparametric efficiency bound is infinite, the lower bound on the density of propensity mass near the origin can be so weak that identification nearly fails, and the nuisance estimator may be so poorly designed that it pushes all observations' estimated propensities towards the origin at a slower-than-parametric rate. Nevertheless, Neyman orthogonality is sufficiently powerful to ensure the validity of the simple t-test.

The next sub-section interprets the rate requirements I impose in a few special cases, and then derives sufficient conditions to achieve these nuisance rates.

3 Interpretation and Lessons for Empirical Practice

This section analyzes the regression rates needed to achieve Theorem 1 under Assumption 4(i). First, I use some special cases to provide intuition for the nuisance error rates I request: for example, a shared consistency rate of $n^{-1/3}$ will always suffice for valid Wald confidence intervals to exist. I then provide some key limitations of these results: the required product-of-errors rate may be more stringent than the usual $n^{-1/2}$ requirement, and a given outcome regression rate may be more difficult to achieve. In the third subsection, I prove some rules of thumb for choosing a clipping threshold for practitioners that do not feel comfortable taking a strong stance on regression rates. Finally, I provide some intuition and AIPW and IPW with parametric estimator functions that may be correctly or incorrectly specified.

3.1 Rate Condition Special Cases

This section interprets the rate requirements of Theorem 1 under various special cases. The main requirement is that $r_{\mu,n} r_{e,n}^{\gamma_0/2}$ goes to zero faster than $n^{-1/2}$. As a result, outcome regression rates are more valuable than nominally equivalent propensity rates under very weak overlap.

I consider the Assumption 4(i) rate requirements in a few special cases. I omit an analysis of the stronger rate requirement in Assumption 4(ii) that would be needed to handle degenerate distributions.

Assumption 5. Assumptions 1, 2, and 4(i) hold, and $r_{e,n}, r_{\mu,n} \to 0$.

This assumption rules out degenerate forms of weak overlap.

I now provide sufficient conditions for Wald confidence interval validity under various special cases of overlap weakness.

Example 1 (Strict overlap)**.** Suppose Assumption 5 holds and either (i) there is strict overlap and $r_{\mu,n} r_{e,n} \ll n^{-1/2}$, or (ii) there is somewhat weak overlap $r_{e,n} = o(n^{-\eta})$ for some fixed $\eta > 0$, and $r_{\mu,n} r_{e,n} \ll n^{-1/2}$. Then there exists a $b_n \to 0$ such that clipped AIPW t-statistics are asymptotically well-calibrated.

Example 2 (Second moments barely fail to exist)**.** Suppose Assumption 5 holds for $\gamma_0 = 2$ and there is some $\eta > 0$ such that $r_{\mu,n}, r_{e,n} \ll n^{-1/4-\eta}$. Then there exists a $b_n \to 0$ such that clipped AIPW t-statistics are asymptotically well-calibrated.

Example 3 (Shared rates, very weak overlap)**.** Suppose Assumption 5 holds for some $\gamma_0 > 1$ and $r_{\mu,n}, r_{e,n} \ll n^{-1/3}$. Then there exists a $b_n \to 0$ such that clipped AIPW t-statistics are asymptotically well-calibrated.

Example 4 (Parametric rates)**.** Suppose Assumption 5 holds for some $\gamma_0 > 1$ and there is some fixed $\eta > 0$ such that either (i) $r_{\mu,n} = O(n^{-1/2})$ and $r_{e,n} = o(n^{-\eta})$ or (ii) $r_{e,n} = O(n^{-1/2})$ and $r_{\mu,n} = o(n^{(\gamma_0-2)/4-\eta})$. Then there exists a $b_n \to 0$ such that clipped AIPW t-statistics are asymptotically well-calibrated.

These examples help characterize the product-of-rates condition I ask for under weak overlap. I next provide two important limitations of these characterizations: the product of errors condition is always more stringent than the usual product requirement, and any given outcome regression rate may also be more difficult to achieve.

3.2 Limitations

So far, this work has focused on the theoretical advantages of clipped or trimmed AIPW for analysis: under suitable outcome and propensity regression rates and threshold rates, the standard Wald confidence intervals exhibit asymptotically exact coverage of the standard causal estimands. In practice, there are important limitations to these results. I formally prove that under weak overlap, more the required product of errors of nuisance errors and any given rate for outcome errors both become more difficult to achieve.

The usual product-of-errors condition under strict overlap often takes a form like $r_{\mu,n} r_{e,n} \ll \sigma_n$, where σ_n is the standard deviation of the Oracle estimator. For example, Heiler and Kazak (2021) argue that this condition is sufficient for estimated unclipped AIPW to be first-order equivalent to an oracle estimator. An alternative characterization of the usual product-of-errors condition that is more stringent under weak overlap is a requirement $r_{\mu,n} r_{e,n} \ll n^{-1/2}$. I now show that even this requirement is insufficient for clipped AIPW Wald confidence intervals to be valid.

Corollary 3 (Clipping makes product-of-errors more stringent)**.** *Fix $\gamma_0 \in (1, 2)$, some target coverage level $\alpha \in (0, 1)$, some sequence of clipping thresholds b_n such that $n^{-1/2} \ll b_n \ll 1$, and some sequence of $r_{e,n}$ and $r_{\mu,n}$ such that Assumptions 3(a) and 3(d) hold, but Assumption 3(c) does not hold. Then there is a family $\mathscr{P}$ and nuisance estimators $\hat{\mu}$ and $\hat{e}$ such that Assumption 1, Assumption 2, and Assumption 4(i) hold for these rates and this γ_0, but Wald confidence intervals have the zero-coverage property that for all $P \in \mathscr{P}$, $P(\psi(P) \in \hat{\mathcal{C}}_n) \to 0$.*

The intuition is the heuristic from Section 3.1: a sufficient condition for Wald confidence interval validity is $r_{\mu,n} r_{e,n}^{\min\{\gamma_0,2\}/2} \ll n^{-1/2}$. When $\gamma_0 < 2$, there is a range of nuisance estimates such that $r_{\mu,n} r_{e,n} \ll n^{-1/2} \ll r_{\mu,n} r_{e,n}^{\min\{\gamma_0,2\}/2}$ and estimation bias can be of a higher order than the oracle variance.

A given product of rates can also be more difficult to achieve under weak overlap. The added challenge comes from outcome regression. Clipped AIPW needs to estimate $E[Y \mid X, D = 1]$ accurately in precisely the regions in which $P(D = 1 \mid X)$ is smallest. As a result, an outcome regression rate $r_{\mu,n}$ can become more

difficult to achieve under weak overlap: there are fewer treated observations in some regions, and the density of treated observations can be chosen to introduce a shape for which outcome regression more difficult. I characterize the harm in the case in which the harm is solely the reduction in treated observations, and leave a characterization of outcome regression with potential degenerate designs to future work.

I characterize optimal outcome regression rates for Nadaraya-Watson regression. Nadaraya-Watson regression estimates $E[Y \mid X, D = 1]$ using a kernel-weighted average of observed treated outcomes. Under strict overlap and with Hölder continuity of order $\beta_\mu \in (0, 1]$, the Nadaraya-Watson estimator achieves the optimal pointwise consistency rate of $n^{-\beta_\mu/(2\beta_\mu+d)}$, and can be used to achieve the optimal global consistency rate with an associated polylog penalty (Stone, 1982). Note that I focus on kernel regression estimates of $E[Y \mid X, D = 1]$. Faster rates can be obtained by thoughtfully leveraging smoothness assumptions in $E[Y \mid e(X), D = 1]$ (Ma and Wang, 2020; Sasaki and Ura, 2022; Ma et al., 2023). It may also be possible to obtain faster rates for $E[Y \mid X, D = 1]$ under stronger outcome smoothness assumptions, but I focus on Nadaraya Watson to avoid subtle design degeneracy issues under weak overlap.

I characterize the optimal Nadaraya-Watson regression rate under Hölder continuity. For convenience, I fix the domain of the covariates to be a specific hypercube in $\mathbb{R}^d$.

Assumption 6 (Hölder smoothness and fixed domain)**.** Assumption 4(i) holds for some fixed $\rho > 0$ and for all $P \in \mathscr{P}$, X is continuously distributed over $[-1, 1]^d$ with a uniform lower density bound and $\mu(X) = E_P[Y \mid X, D = 1]$ is in the Hölder smoothness class $\Sigma(\beta_\mu, L)$ for some $\beta_\mu, L > 0$. .

Hölder continuity is the standard assumption to motivate Nadaraya-Watson regression. If X is distributed uniformly on $[-1, 1]^d$, then a simple worst-case propensity score is some function proportional to $\|X\|^{d/(\gamma_0-1)}$. Such a propensity score has a propensity tail of the form $P(e(X) \leq \pi) \sim \pi^{\gamma_0-1}$ for small enough π, and ensures the fewest possible treated observations near the point $X = 0$.

For convenience, I write Θ for the space of parameters defining the restrictions in Assumption 6, and abuse notation and write $\mathcal{P}(\theta)$ for the set of families $\mathscr{P}$ that satisfy this assumption for a set of parameters θ. I focus on the Nadaraya-Watson case and pointwise rates for simplicity in this work. Higher-order smoothness assumptions allow for local polynomial regression with potentially subtle dynamics of associated eigenvalues (Hall et al., 1997; Gaïffas, 2005, 2009), and it unclear whether the usual polylog penalty for global rates is necessary under weak overlap.

In the worst case, weak overlap of order $\gamma_0 - 1$ plays a role equivalent to increasing the number of covariates by $d/(\gamma_0 - 1)$.

Proposition 3 (Minimax Nadaraya Watson regression rates)**.** *The optimal Nadaraya-Watson pointwise regression rate is $r_n = n^{-\beta_\mu/(2\beta_\mu+d+d/(\gamma_0-1))}$: there is a function $C' : \Theta \to \mathbb{R}$ and a sequence of functions*

$h_n : \Theta \to \mathbb{R}$ *such that:*

(i) If $\mathscr{P} \in \mathcal{P}(\theta)$ *and* $\delta_n \to \infty$, *then* $\limsup_{n\to\infty} \sup_{P\in\mathscr{P},x_0} P\left(\left|\hat{\mu}^{(NW)}(x_0 \mid h_n(\theta)) - \mu(x_0)\right| \geq \delta_n r_n\right) \to 0$.

(ii) For every $\gamma_0 > 1$, *there is an associated* $\theta \in \Theta$, $\mathscr{P} \in \mathcal{P}(\theta)$, *and* $x_0 \in [-1,1]^d$ *such that*

$$\liminf_{n\to\infty} \inf_{P\in\mathscr{P},h_n>0} P\left(\left|\hat{\mu}^{(NW)}(x_0 \mid h_n) - \mu(x_0)\right| \geq C'(\theta) r_n\right) > 0.$$

Under strict overlap, the known optimal pointwise rate $n^{-\beta_\mu/(2\beta_\mu+d)}$ corresponds to the case that γ_0 infinite. When γ_0 is finite, the rate in Proposition 3 is slower, with a loss equivalent to adding $d/(\gamma_0 - 1)$ dimensions under strict overlap. As overlap is allowed to become increasingly weak and other parameters are held constant, γ_0 is reduced, there can be regions of the covariate space with increasingly few treated observations, and the optimal Nadaraya-Watson rate is worse. In the limit in which γ_0 tends to one, the rate in Proposition 3 can become arbitrarily poor.

Proposition 3 yields minimal smoothness assumptions for Wald confidence interval validity. Recall that the optimal propensity estimation rate under a Hölder smoothness restriction of order β_e is $n^{-\beta_e/(2\beta_e+d)}$. Also recall that the heuristic sufficient condition for Wald confidence intervals to be valid for some clipping threshold under Assumption 4(i) is $r_{\mu,n} r_{e,n}^{\gamma_0/2} \ll n^{-1/2}$, or equivalently, $\log(r_{\mu,n}) + (\gamma_0/2)\log(r_{e,n}) \ll \log(n)^{-1/2}$. Taken together, these conditions require:

$$\frac{2\beta_\mu}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)} + \frac{\gamma_0\beta_e}{2\beta_e + d} > 1. \tag{2}$$

In the Lipschitz-continuity case $\beta_\mu = 1$, Equation (2) further reduces to $\beta_e > \frac{d^2}{2(\gamma_0-1)-d(2-\gamma_0)}$, or equivalently, $\gamma_0 > \frac{2(d+1)+d^2/\beta_e}{d+2}$. When $\beta_e > d^2/2$, clipped AIPW can achieve asymptotically valid Wald confidence intervals for some $\gamma_0 < 2$. When $d > 1$, the econometrician must assume stronger smoothness restrictions than Lipschitz continuity in order to achieve the necessary nuisance rate guarantees.

Equation (2) provides sufficient smoothness assumptions for some sequence of valid clipping thresholds to exist. Before proceeding to applying clipped AIPW, I propose some rules of thumb for choosing a plausibly valid clipping threshold.

3.3 Choice of Clipping Threshold

The theoretical analysis above shows that for some sequence of outcome and propensity error rates, the clipped AIPW estimator can be asymptotically normal and centered around the true causal estimand for some clipping rate. However, these results do not provide guidance for how to choose the clipping rate.

I now propose some rules of thumb for the choice of a clipping or trimming threshold. I favor thresholds that impose the weakest possible rate requirements necessary to achieve asymptotic normality for the average treatment effect. that requirement corresponds to a smaller threshold with fewer clipped observations. Where the econometrician is confident that they can achieve faster nuisance rate estimation, then a larger threshold with more clipped observations will be able to achieve more precise estimates and better power than the rules of thumb proposed here. I will always implicitly require that b_n be no larger than $n^{-1/2}\log(n)$ to ensure that at least the oracle clipped AIPW estimator will be asymptotically normal. I describe these rules of thumb as being rules for clipped AIPW, but the logic also applies to trimmed AIPW.

The first two rules of thumb are based on scenarios in which the econometrician is confident of nuisance estimation rates. I begin with the case in which the econometrician has implied a minimal rate of propensity convergence. Suppose $\hat{e}(x)$ is estimated through a version of local polynomial regression with ℓ_e derivatives. The econometrician has implicitly assumed that $e(x)$ has $\beta_e > \ell_e$ degrees of Hölder smoothness, in which case one can achieve a global consistency rate of $(n/\log(n))^{-\beta_e/(2\beta_e+d)}$ (Stone, 1982). In such a case, the propensity rule of thumb would choose a clipping threshold on the order of $b_n = n^{-\ell_e/(2\ell_e+d)}$. Any slower rate would impose needless restrictions on the outcome regression rate, and any faster rate would only be valid if the econometrician is willing to assume a smoothness level of some specific $\beta_e > \ell_e$. An interesting avenue for future work is whether there is a convenient way to choose the constant in this process.

Under weak overlap, it is often easier to achieve faster convergence for propensity estimation than outcome regression. A second rule of thumb is therefore based on minimal outcome regression rate $\underline{r}_{\mu,n}$. I propose choosing b_n to set $g_n(b_n) = 0$, where

$$g_n(b) = \frac{1}{n}\sum \frac{\underline{r}_{\mu,n}\frac{1}{n}\sum 1\{\hat{e} \leq b\}}{\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}} + r_{\mu,n}\underline{r}_{\mu}\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}} - n^{-1/2}. \tag{3}$$

Informally, this rule of thumb finds a clipping thresold b_n such that if $r_{\mu,n} \ll \underline{r}_{\mu,n}$ and $r_{e,n} \ll b_n$, then clipped AIPW will achieve asymptotic normality around the target causal estimand. Any slower rate would achieve needless restrictions on the propensity estimation rate, and any faster rate would only be valid if the econometrician is willing to assume a faster outcome regression rate than $\underline{r}_{\mu,n}$. This function is a plug-in version of a combination of technical conditions presented in Assumption 3'. This rule of thumb has the drawback that theoretical guarantees for $r_{\mu,n}$ are not yet well-developed. I find in simulations that this rule of thumb performs well when outcome regression rate guarantees are available.

A third rule of thumb is theoretically attractive, and prevents the need for the empirical researcher to

specify any nuisance consistency rates. This proposal is to choose b_n to solve $f_n(b_n) = 0$, where:

$$f_n(b) = \frac{b\frac{1}{n}\sum 1\{\hat{e} \leq b\}}{\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}} + b^2\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}} - n^{-1/2}. \tag{4}$$

This proposal is a plug-in version of the rule for g_n. Informally, it corresponds to finding a sequence of b_n such that if $r_{\mu,n}, r_{e,n} \ll b_n$, then the clipped AIPW estimator will achieve asymptotic normality. This rule of thumb is always feasible.

Lemma 1 (Well-defined rule of thumb)**.** *Suppose* $\hat{e} \in (0,1]$ *and* $\sum D/\hat{e} > 0$*. Then there is exactly one* b_n *such that* $\limsup_{b\to b_n^-} f_n(b) \leq 0 \leq \liminf_{b\to b_n^+} f_n(b_n)$.

In smooth applications, the rule of thumb in Equation (4) produces a clipping threshold on the order of $n^{-1/(2+\gamma_0)}$. If $r_{\mu,n}, r_{e,n}$ both go to zero more quickly than this clipping threshold, then clipped AIPW estimation with b_n chosen to solve Equation (4) will produce valid confidence intervals for the standard causal estimand. If $r_{\mu,n}, r_{e,n}$ both go to zero more slowly than this clipping threshold, then no clipping threshold will allow Wald confidence intervals to cover the standard causal estimand. If the econometrician is to do better than this rule of thumb, then they must have application-specific knowledge of one of the two nuisance rates that calls for a more specific approach than a generic rule of thumb. However, in practice, a given outcome regression rate is often more difficult to achieve than a given propensity rate under weak overlap, so I generally recommend using one of the first two rules of thumb where feasible.

3.4 Parametric Estimators and Misspecification

When both nuisance functions are estimated nonparametrically, then consistency is achievable and AIPW is generally preferable under strict overlap. When both nuisance functions are estimated parametrically, then it is possible for one or both nuisance function to be inconsistent and the choice of estimator may be ambiguous. I now provide some intuition on the two estimators when nuisance functions are estimated parametrically and through cross-fitting. I will consider IPW and AIPW with the same sequence of thresholds b_n satisfying $1 \gg b_n \gg n^{-1/2}$.

In this subsection, I will assume that parametric nuisance estimators $\hat{\eta}$ achieve an L_∞ error relative to a limiting nuisance function $\bar{\eta}$ that is the order of $n^{-1/2}$. For example, consider logit estimation of a propensity model of the form $\bar{e}(X) = \frac{exp(X'\beta)}{1+exp(X'\beta)}$ for a pseudo-true parameter β. If the support of X is bounded, then $n^{-1/2}$-consistent estimate of β is sufficient to achieve $n^{-1/2}$-consistent estimation of $\bar{e}(X)$ everywhere. However, weak overlap may emerge from unbounded tails, in which case the L_∞ rate may not go to zero. Unbounded covariates are an important case in general. For example, Ma and Wang (2020) motivate

weak overlap tails through the distribution of covariates under a logistic propensity model. Nevertheless, a careful treatment of parametric estimation of nuisances with unbounded covariates is outside the scope of this work. I write that a nuisance estimate $\hat{\eta}$ is consistent if it tends to the correct limit $\bar{\eta} = \eta$, and I write that $\hat{\eta}$ is inconsistent otherwise.

The analysis above is easiest to extend when either both or neither nuisance function is consistent. If both the propensity and outcome regression estimates are inconsistent, then both the IPW and AIPW estimators fail to be consistent, and as in the case of inconsistent nuisance functions with strict overlap, there is no general reason to prefer one or the other. If both nuisance estimates are consistent, then the AIPW and IPW estimators will be consistent and will have variance on the same order, but the IPW estimator may have higher-order bias than the AIPW estimator. This higher-order bias follows because IPW can be viewed as a particular case of AIPW with an inconsistent outcome regression estimator.

When the outcome regression estimate is inconsistent, there is no general reason to prefer IPW or AIPW, but both estimators may have bias that is of a higher-order than the estimator's standard deviation. When $\hat{\mu}$ is inconsistent, both IPW and AIPW can be viewed as instances of AIPW with an inconsistent outcome regression estimate. Suppose P is a distribution from the second half of Corollary 1, which has $P(e(X) \leq \pi) \sim \pi^{\gamma_0 - 1}$ for all π small enough. The bias in the clipped (or trimmed) region with an inconsistent outcome regression estimate is generally on the order of $P(e(X) \leq b_n) \sim b_n^{\gamma_0 - 1}$. However, by Corollary 1, the oracle AIPW (and oracle IPW) standard deviation is on the order of $n^{-1/2} b_n^{\gamma_0/2 - 1} \ll b_n^{\gamma_0 - 1}$. This heuristic analysis suggests that in many cases, IPW or AIPW-with-inconsistent-outcome-regression will have bias that is of a higher order than the estimator's standard error. That intuition is similar to Ma et al. (2023)'s analysis of trimmed AIPW with a tailored debiasing procedure.

The case of a consistent outcome regression estimate with inconsistent propensity estimates is more interesting. In this case, AIPW should have lower-order bias than IPW. The Berry-Esseen argument for AIPW asymptotic normality with known nuisance functions only requires cross-fitting and b_n to go to zero slower than $n^{-1/2}$, so that clipped or trimmed AIPW should also be asymptotically normal under appropriate error product conditions. In this case, it is possible for there to be a blessing of weak overlap: if there are enough clipped observations to induce a slower-than-$n^{-1/2}$ standard error, then it is not clear that with a consistent parametric outcome regression, nuisance estimation error has a first-order effect on the causal estimate. This robustness intuition is useful, because I apply parametric nuisance estimators in the application to right heart catheterization. Careful treatment of the parametric case is left for future work.

4 Applications

In this section, I present simulated results for the clipped AIPW estimator as well as empirical results from an application to right heart catheterization. I find that clipped AIPW performs very well under weak overlap, producing nearly normal t-statistics and almost perfectly calibrated p-values with only 1,000 observations. When studying the right heart catheterization data, I find that the rule of thumb approach increases the estimated harm of the procedure by 0.37 standard errors relative to the usual 10% trimming rule, while reducing the estimated standard error by 2.3%.

4.1 Simulation Evidence

I now study the performance of the clipped AIPW estimator in simulations.

My simulation design is based on the design in Ma and Wang (2020). As in their work, I simulate data with $P(e(X) \leq \pi) = \pi^{\gamma-1}$ and $DY = \kappa D(1-e(X)) + D(\varepsilon - 4)/\sqrt{8}$, where $\varepsilon \mid X, D \sim \xi_4^2$ is scaled to achieve zero mean and unit variance. However, I increase γ from 1.5 to 1.8 to ensure feasible outcome regression rates, set $\kappa = 2$ rather than $\kappa = 1$ to avoid a coincidental offset of IPW lower- and upper-tail bias in small samples, and reduce DY by $\kappa E[D(1-e(X))]$ so that the true average potential outcome is zero. I achieve this propensity distribution by taking $X \sim Unif([0,1])$ i.i.d. and setting $e(X) = X^{1/(\gamma_0-1)}$. I present results for 5,000 simulations of increasingly large samples.

I estimate both the propensity and outcome regressions with five-fold cross-fitting. I use shrinkage cubic splines and REML estimation, as implemented by the `mgcv` package in `R`. In this setting, Proposition 3 establishes that outcome regression can achieve a pointwise rate of $n^{-1/(3+1/(\gamma_0-1))}$. I therefore choose the clipping threshold b_n based on Equation (3), assuming $r_{\mu,n} \ll n^{-1/5}$, which is implied by $\gamma_0 > 1.5$.

I begin by summarizing point estimates in Figure 1. The unclipped estimators are approximately median-unbiased, but possess sufficiently heavy inverse propensity tails that the mean performance gets worse with increasing sample sizes. The clipped estimators perform much better, but the clipped IPW estimator exhibits its known first-order bias. The clipped AIPW estimator exhibits less bias than the clipped IPW estimator, and has slightly better performance in terms of mean squared error.

I find in Figure 2 that the clipped AIPW estimator's t-statistics are reasonably well-calibrated. The plot presents t-statistics on the true average potential outcome $1-(\gamma_0-1)/\gamma_0$. The t-statistics of unclipped IPW and AIPW estimators are visibly non-Gaussian, and often exhibit a multimodal distribution. This poor performance is unsurprising: the unclipped IPW and AIPW estimators are known to fail to be asymptotically normal in this setting. Both the clipped IPW and AIPW estimators are known to be asymptotically normal

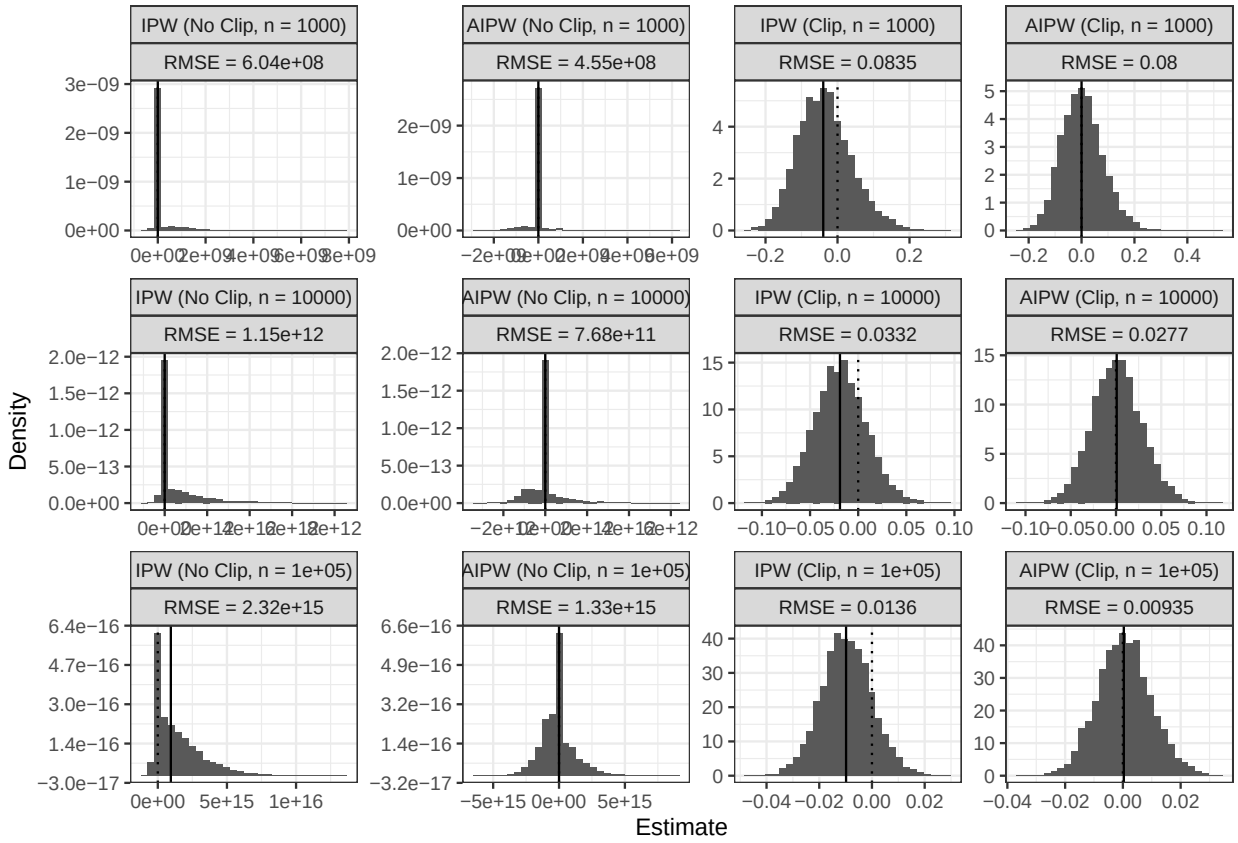

Figure 1: Histograms of point estimates in simulations for the various methods considered in the simulations. Vertical dotted and solid lines indicate true causal effect and median estimate, respectively. Clipped estimators achieve much better performance than unclipped estimators, and clipped AIPW's debiasing property is also apparent.

in this setting, and both the asymptotic normality and the clipped IPW estimator's first-order bias are visible to the naked eye, although the clipped IPW estimator also exhibits visible skew in small samples. I test for t-statistic normality using a Shapiro-Wilk test. The test rejects normality for both clipped estimates. Still, the clipped AIPW estimator's violations are less severe by this criterion.

I find in Figure 3 that the clipped AIPW estimator's p-values are well-calibrated in sufficiently large samples. I use Wald confidence intervals to calculate two-sided p-values on the null of the true average potential outcome. If Wald confidence intervals are well-calibrated, then the simulated p-values on the true average potential outcome will be exactly uniformly distributed. The unclipped IPW and AIPW estimators exhibit known poor performance. The clipped IPW estimator exhibits over-rejection even with large samples, as it provides well-calibrated inference for a first-order-biased estimand. The clipped AIPW estimator also over-rejects in small samples, but the bias is less severe: with 1,000 observations, clipped IPW rejects the true null in 12.0% of simulations, while clipped AIPW rejects in 8.8% of simulations. As the sample size increases,

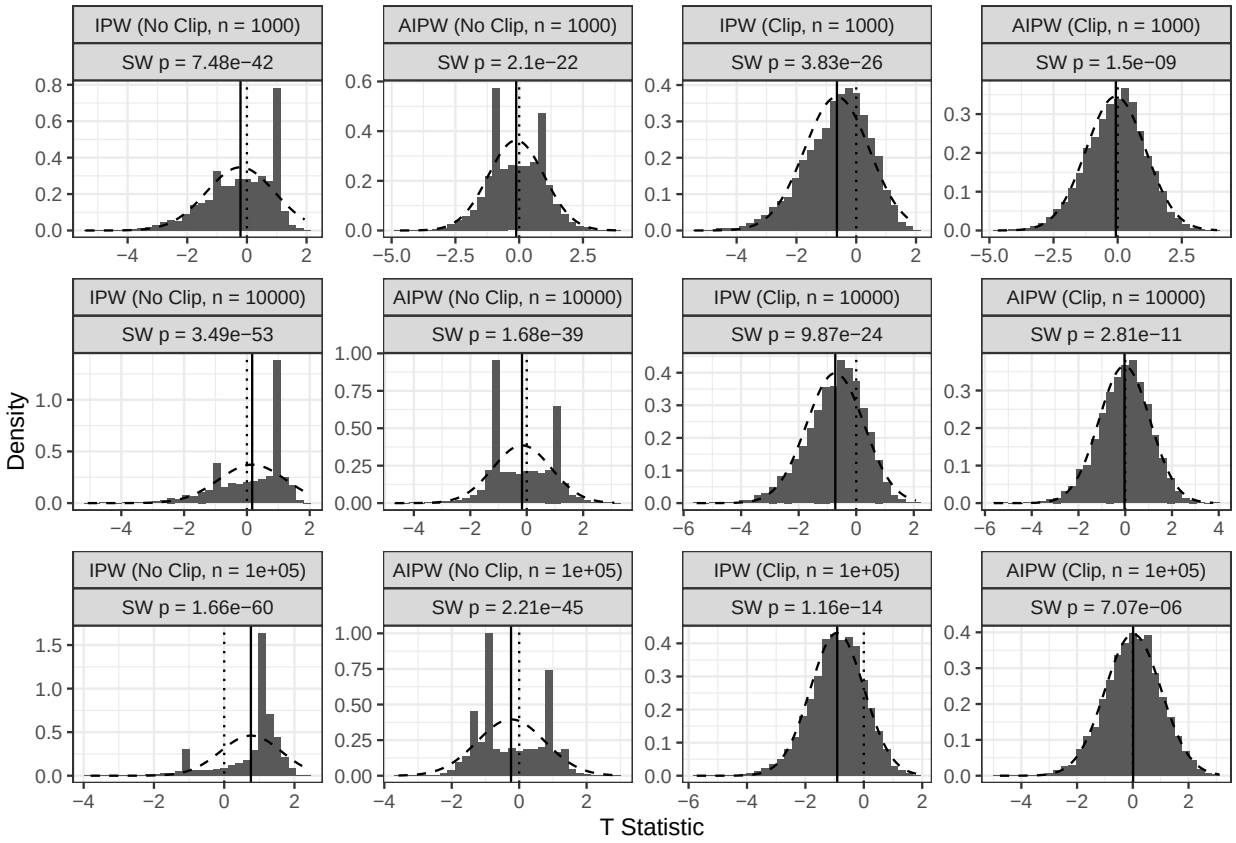

Figure 2: Histograms of simulation t-statistics for various sample sizes. Vertical solid and dotted lines indicate mean t-statistic and target mean t-statistic of zero, respectively. Dashed line corresponds to the calibrated Gaussian density targeted in the Shaprio-Wilk test for normality.

the asymptotic normality of Theorem 1 becomes apparent. With 100,000 observations, clipped IPW rejects the true null hypothesis in 12.8% of simulations, while clipped AIPW rejects in 5.3% of simulations. The p-value on exact calibration of the two-sided test statistics for clipped AIPW with 100,000 observations is 0.692. This is a remarkable result: despite the known extreme difficulty of statistical inference in this setting, 5,000 simulated draws are insufficient to detect a meaningful failure of Wald confidence intervals based on the clipped AIPW estimator.

In moderate samples, clipped AIPW can undercover due to the challenge of outcome regression under weak overlap. In Appendix A (Figures 8 through 10), I conduct the same experiments, but with the estimated outcome regression function replaced by the oracle true function. The root-mean-squared error and failures of normality are comparable, suggesting these non-inferential patterns are driven by propensity estimation and clipping. However, the two-sided p-values exhibit better performance in small samples, and if anything slightly under-reject with 100,000 observations. In fact, outcome regression is difficult in small samples. Figure 4 presents an example with 1,000 observations. It is rare to have treated observations with small

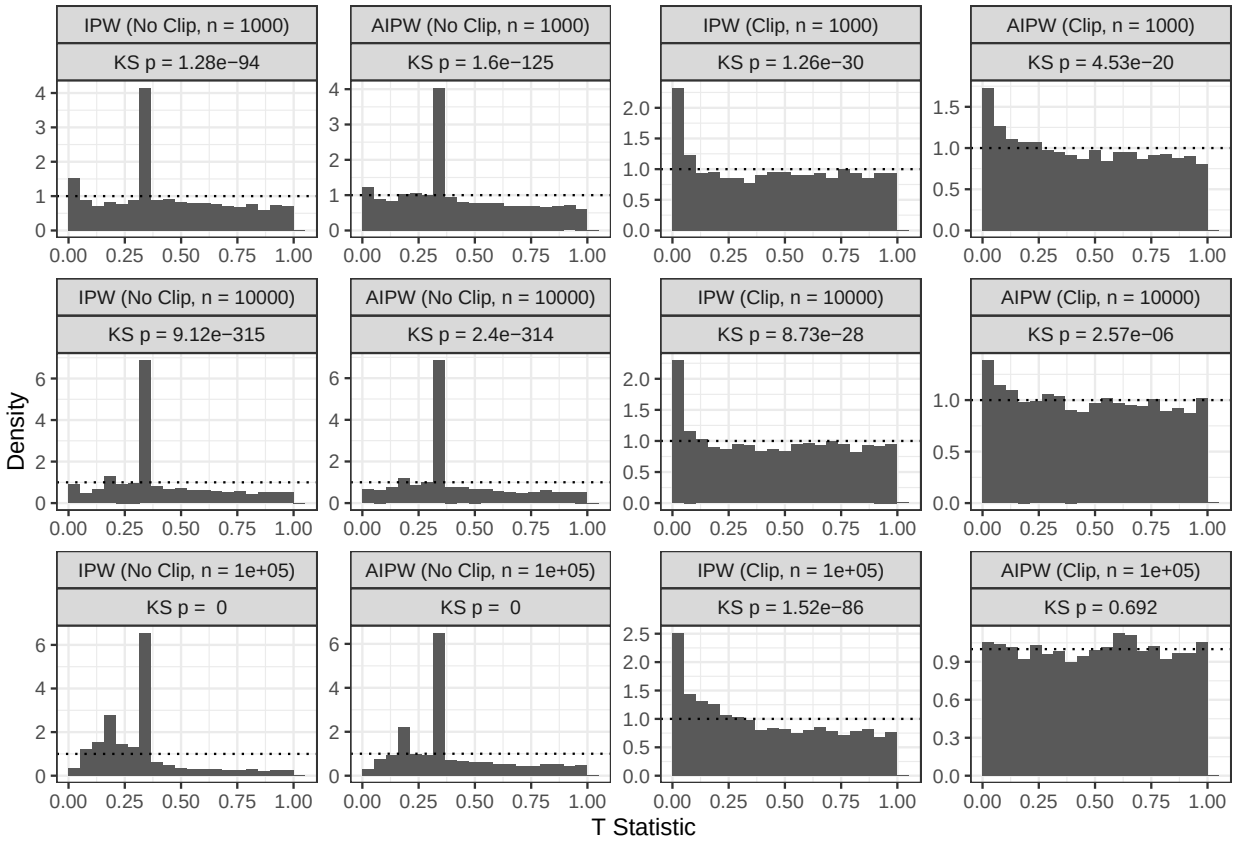

Figure 3: Histograms of simulation p-values on null hypothesis of true APO for various sample sizes. Dotted lines correspond to the target Uniform(0, 1) density. P-values in labels correspond to Kolmogorov-Smirnov tests for the Uniform(0, 1) distribution.

values of $e(X)$. As a result, when such observations are treated, a small number of observations can receive substantial leverage in outcome regression, and the predictions of $E[Y \mid X = 0, D = 1]$ can be driven by a small number of outcome residuals. An important avenue for future work is exploring better methods for outcome regression estimation strategies in small samples under weak overlap.

In Appendix A (Figures 11 through 13), I show that these conclusions largely carry through if clipping were replaced by trimming. The notable differences are that trimmed AIPW exhibits slightly better estimation performance in small samples, while if anything trimmed IPW is slightly worse; trimmed t-statistics exhibit less severe violations of normality; and p-values based on trimmed propensities exhibit more severe undercoverage for both IPW and AIPW.

4.2 Application to Right Heart Catheterization

I apply the clipped AIPW estimator to study the effect of right-heart cathterization (RHC) on survival. This dataset was first analyzed by Connors et al. (1996), and is a common benchmark in the weak overlap

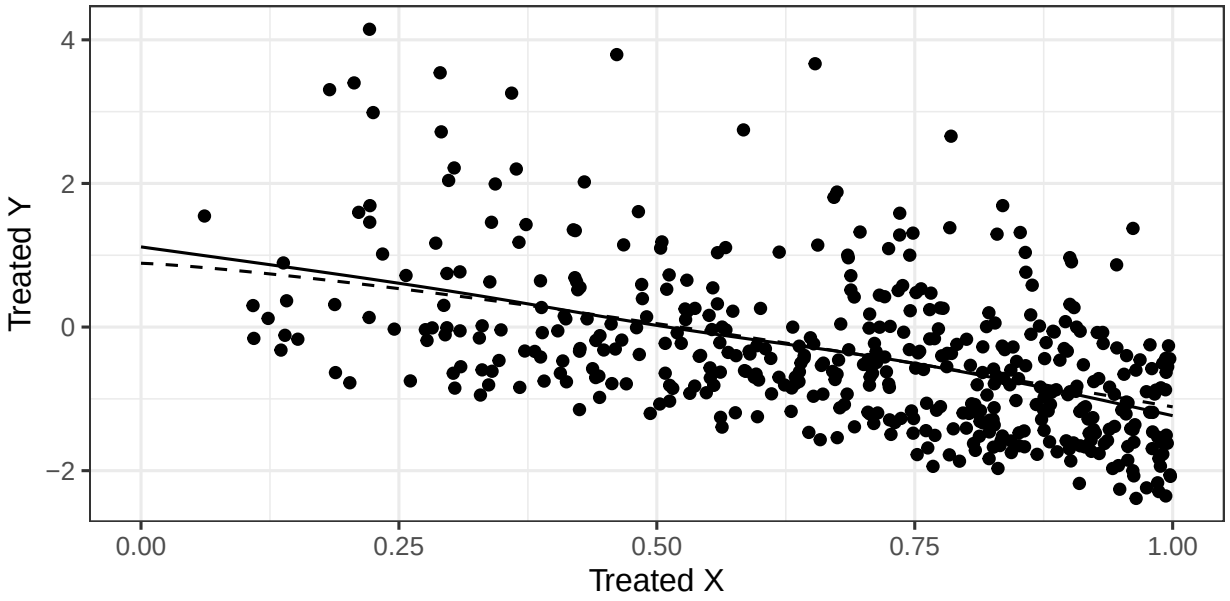

Figure 4: Simulated treated observations for one simulation of 1,000 observations. It is rare to see treated observations with small X, which corresponds to small values of $e(X) = X^{1/(\gamma_0-1)}$. As a result, such observations can have high leverage when predicting $E[Y \mid X = 0, D = 1]$, and can yield to important errors between the true (dashed) and predicted (solid) regression lines.

literature (Crump et al., 2009; Armstrong and Kolesár, 2017).

I analyze a version of the dataset from Armstrong and Kolesár (2017). The dataset is comprised of 5,735 adult patients, and the treatment D corresponds to receiving RHC within 24 hours of admission. The target causal effect is the average treatment effect of RHC on 30-day survival. The data includes 52 covariates X (72 covariates if counting factor levels separately). I estimate the nuisance functions $e(X)$ and $\mu(X)$ using five-fold cross-fitting. I estimate nuisance functions with logistic regression to align with Crump et al. (2009)'s empirical application. I estimate standard errors by bootstrapping the procedure. I keep fold assignment fixed in bootstraps to minimize the risk of over-fitting.

Crump et al. propose a weak overlap rule of thumb that estimates the treatment effect for the subpopulation with propensity scores between 10% and 90%. This rule-of-thumb trimming rule is chosen to approximately minimize asymptotic variance. This strategy ensures asymptotic normality, but changes the target estimand even asymptotically. By comparison, the clipped and trimmed AIPW estimators I analyze have thresholds b_n that tend to zero asymptotically. As a result, the estimators proposed here are able to target full population average treatment effect, potentially at the cost of increased variance even asymptotically. I compare these procedures, as well as other potential fixed trimming rules, using the same nuisance estimates.

I present the distribution of estimated propensity scores for treated and control units in Figure 5. The

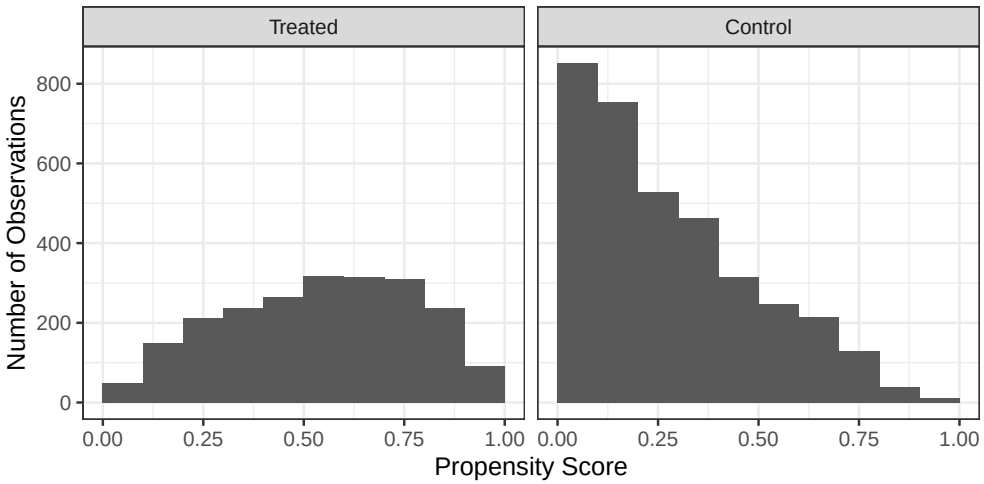

Figure 5: Histogram of estimated propensity scores for treated (left) and control (right) observations in right heart catheterization data. The plot is designed to parallel Figure 1 in Crump et al. (2009). Slight differences reflect the use of cross-fitting.

figure is an analog of Crump et al. (2009)'s Figure 1. There is a meaningful density of units with estimated propensities near zero, suggesting weak overlap. This pattern is similar to the findings of Crump et al., although there are slight differences, presumably due to my use of cross-fitting.

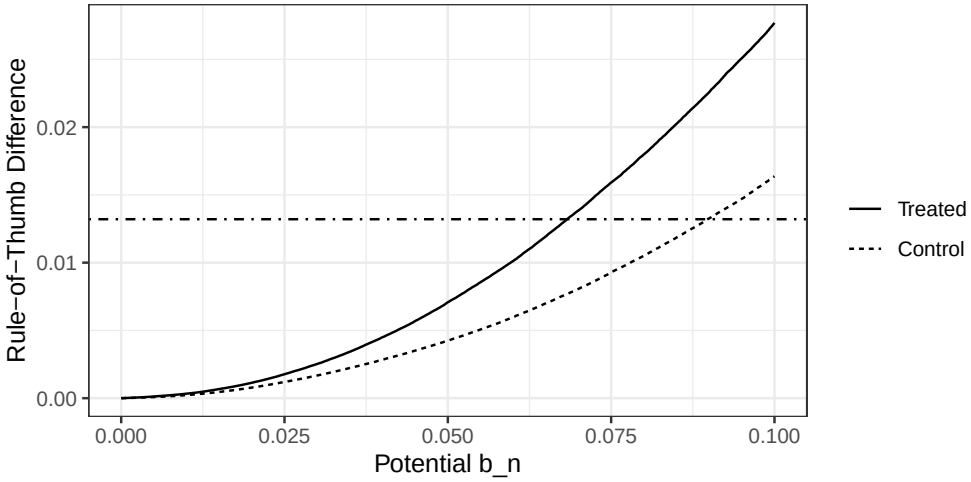

Figure 6: The value of $f_n(b) + n^{-1/2}$ from Equation (4), where the equal-rates rule of thumb chooses b_n to set $f_n(b)$ equal to zero. $n^{-1/2}$ is indicated by horizontal dashed line. The more favorable distribution of estimated treatment propensities allows for a more aggressive clipping threshold.

I compare AIPW estimators for various trimmed subsamples to the clipped AIPW estimator. I choose the clipping threshold b_n through the equal-rates Equation (4) because I estimate both nuisance functions parametrically. This strategy chooses treated and control clipping thresholds to set a data-dependent function equal to $n^{-1/2}$, so I plot the data-dependent functions in Figure 6. The estimated lower clipping threshold

is 0.068 and affects 10.5% of observations. The Crump et al. 10% rule of thumb would exclude 16.3% of observations below. The estimated upper clipping threshold is 0.09 below one: there are few observations with large estimated propensities, so the rule of thumb concludes there is no need to trim observations with large estimated propensities. This upper threshold affects 1.4% of observations, comparable to the 1.8% of observations excluded above by the 10% rule of thumb.

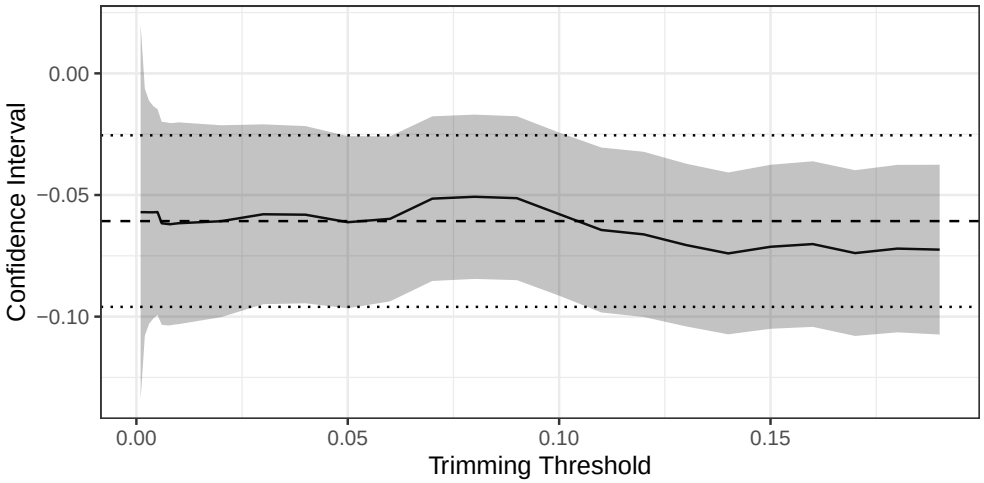

Figure 7: Estimated effects (solid line) and 95% confidence interval (shaded region) for AIPW applied to various trimmed subsamples. Estimate and confidence for clipped AIPW is represented by the dashed and dotted horizontal lines, respectively. Clipped AIPW produces similar estimates and standard errors as the clipping procedure while targeting a more interpretable estimand. A threshold of zero is omitted from the graph because the resulting confidence interval of [-568.8, 581.3] would make the graph difficult to read.

I present estimated effects and confidence intervals for various potential fixed trimming rules in Figure 7. The 10% trimming rule yields an estimated reduction in survival rates of -5.79 percentage points among the trimmed sample, with an estimated 95% Wald confidence interval of [-9.14, -2.43]. Other trimming rules would yield larger confidence intervals, as expected because the 10% rule is chosen to roughly minimize asymptotic variance over target populations.

I compare the trimmed-sample AIPW estimates to a clipped AIPW estimator that targets the full population. The estimated harm increases to -6.07 percentage points, a change of 0.168 standard errors under the 10% rule of thumb estimator. The clipped AIPW confidence interval of [-9.6, -2.55], has a 5.14% larger width than the 10% trimmed sample interval. The clipped AIPW point estimates are similar to the point estimates under a 1% or 5% trimming rule, but the associated confidence interval is slightly narrower. Part of the clipped AIPW standard error is driven by using clipping in the modified population with $\hat{e}(X)$ below b_n: if I used a trimmed, rather than clipped, AIPW estimator, the estimated effect would move by 0.256 standard errors, and the standard error would only increase by 0.54%. However, the simulation results of Section 4.1 suggest the trimmed AIPW estimator may slightly under-cover.

Taken together, these results illustrate that under weak overlap, targeting the causal effect within the full population need not come at a large precision cost. In this application, clipped AIPW with a rule-of-thumb clipping rate yields similar estimates to estimators that target a fixed trimmed sample, while targeting a population that is often more relevant than the fixed-trimming sample and adding only a small precision cost. These results suggest that clipped AIPW is a viable alternative to fixed trimming. At a minimum, practitioners can easily report results under both strategies. When, as here, the fixed-trimming and sequence-of-clipping responses to weak overlap yield similar causal conclusions, then there is strong evidence that these conclusions are not driven by the treatment of observations with small estimated propensities.

5 Conclusion

This work shows that standard Wald confidence intervals for clipped AIPW can achieve target coverage for standard causal effects under plausible conditions. I provide sufficient conditions on nuisance regression rates for clipped (or trimmed) AIPW to be uniformly valid over distributions with even very weak overlap. I use these theoretical results to derive new rules of thumb for choosing a threshold. I find that Wald confidence intervals perform very well in simulations, and can achieve comparable precision to a fixed 10% trimming rule in practice.

This work can be applied in many interesting directions. This work exploits Neyman orthogonality to achieve standard statistical inference in the presence of a small region of irregular identification. Semenova (2024) applies similar techniques to intersection bounds, where at a high level a margin condition plays the role of the minimal overlap bound here. Perhaps similar ideas could apply to other forms of irregular identification. Sasaki and Ura (2022) and Ma et al. (2023) propose estimators for ratio estimands beyond IPW; the arguments here are likely to extend to their more general framework. Issues of weak overlap hold for inverse propensity and other importance sampling estimators in settings like difference-in-difference estimation (Callaway and Sant'Anna, 2021) or statistical inference for parameters that are identified at infinity (Andrews and Schafgans, 1998; Khan and Nekipelov, 2024); the results and rules of thumb here can likely be adapted to those settings.

References

Andrews, D. W. K. and Schafgans, M. M. A. (1998). Semiparametric estimation of the intercept of a sample selection model. *The Review of Economic Studies*, 65(3):497–517.

Armstrong, T. B. and Kolesár, M. (2017). A simple adjustment for bandwidth snooping. *The Review of Economic Studies*, 85(2):732–765.

Armstrong, T. B. and Kolesár, M. (2021). Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177.

Bailey, M. J. and Goodman-Bacon, A. (2015). The war on poverty's experiment in public medicine: Community health centers and the mortality of older americans. *American Economic Review*, 105(3):1067–1104.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230. Themed Issue: Treatment Effect 1.

Chaudhuri, S. and Hill, J. B. (2024). Heavy tail robust estimation and inference for average treatment effects. *Econometric Reviews*.

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

Connors, Alfred F., J., Speroff, T., Dawson, N. V., Thomas, C., Harrell, Frank E., J., Wagner, D., Desbiens, N., Goldman, L., Wu, A. W., Califf, R. M., Fulkerson, William J., J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897.

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

Currie, J. and Walker, R. (2011). Traffic congestion and infant health: Evidence from E-ZPass. *American Economic Journal: Applied Economics*, 3(1):65–90.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Galiani, S., Gertler, P., and Schargrodsky, E. (2005). Water for life: The impact of the privatization of water services on child mortality. *Journal of Political Economy*, 113(1):83–120.

Gaïffas, S. (2005). Convergence rates for pointwise curve estimation with a degenerate design. *Mathematical Methods of Statistics*, 14(1).

Gaïffas, S. (2009). Uniform estimation of a signal based on inhomogeneous data. *Statistica Sinica*, 19(2):427–447.

Goldsmith-Pinkham, P., Hull, P., and Kolesár, M. (2024). Contamination bias in linear regressions. *American Economic Review*, 114(12):4015–51.

Hall, P., Marron, J. S., Neumann, M. H., and Titterington, D. M. (1997). Curve estimation when the design density is low. *The Annals of Statistics*, 25(2):756 – 770.

Heiler, P. and Kazak, E. (2021). Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics*, 222(2):1083–1108.

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

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311.

Khan, S. and Nekipelov, D. (2024). On uniform inference in nonlinear models with endogeneity. *Journal of Econometrics*, 240(2):105261.

Khan, S. and Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042.

Khan, S. and Ugander, J. (2022). Doubly-robust and heteroscedasticity-aware sample trimming for causal inference.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLoS ONE*, 6(3).

Lei, L., D'Amour, A., Ding, P., Feller, A., and Sekhon, J. (2021). Distribution-free assessment of population overlap in observational studies.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Ma, X., Sasaki, Y., and Wang, Y. (2024). Testing limited overlap. *Econometric Theory*.

Ma, X. and Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860.

Ma, Y., Sant'Anna, P. H. C., Sasaki, Y., and Ura, T. (2023). Doubly robust estimators with weak overlap.

Rothe, C. (2017). Robust confidence intervals for average treatment effects under limited overlap. *Econometrica*, 85(2):645–660.

Sasaki, Y. and Ura, T. (2022). Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory*, 38(1):66–112.

Semenova, V. (2024). Aggregated intersection bounds and aggregated minimax values.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric estimators. *The Annals of Statistics*, 10(4):1040–1053.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493.

A Other Simulation Evidence

In this section, I presented simulated evidence for trimmed estimators.

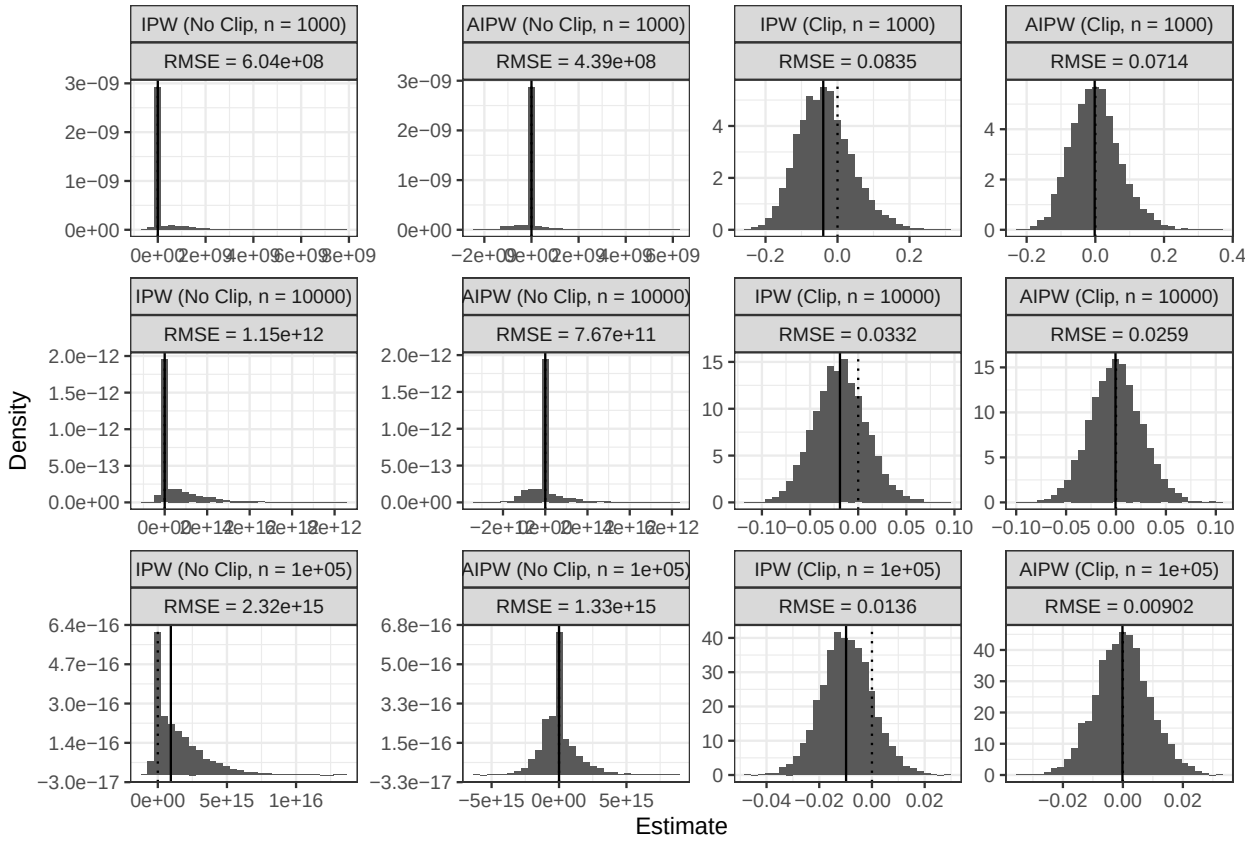

Figure 8: Histograms of point estimates in simulations for the various methods considered in the simulations, but using the oracle μ regression function instead of the estimated $\hat{\mu}$ regression function.

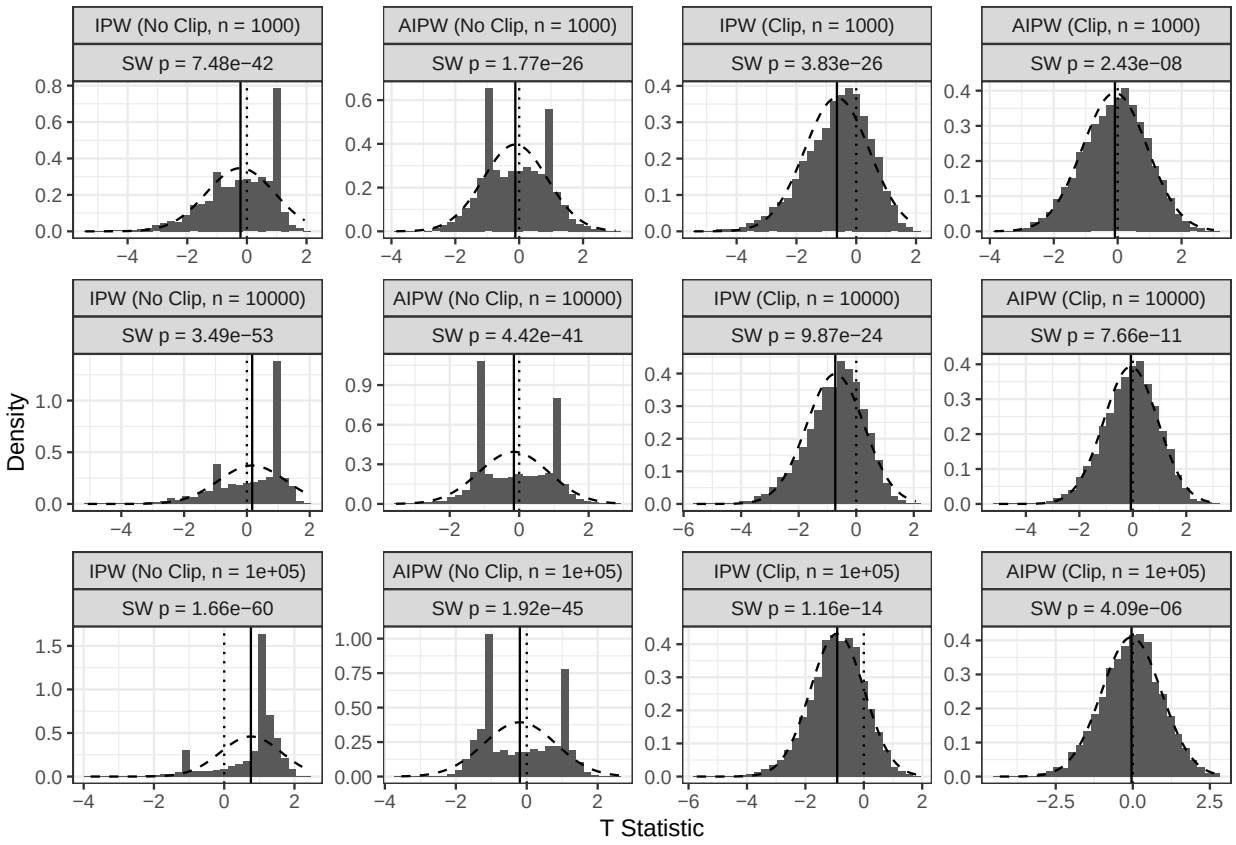

Figure 9: Histograms of simulation t-statistics for various sample sizes, but using the oracle μ regression function instead of the estimated $\hat{\mu}$ regression function.

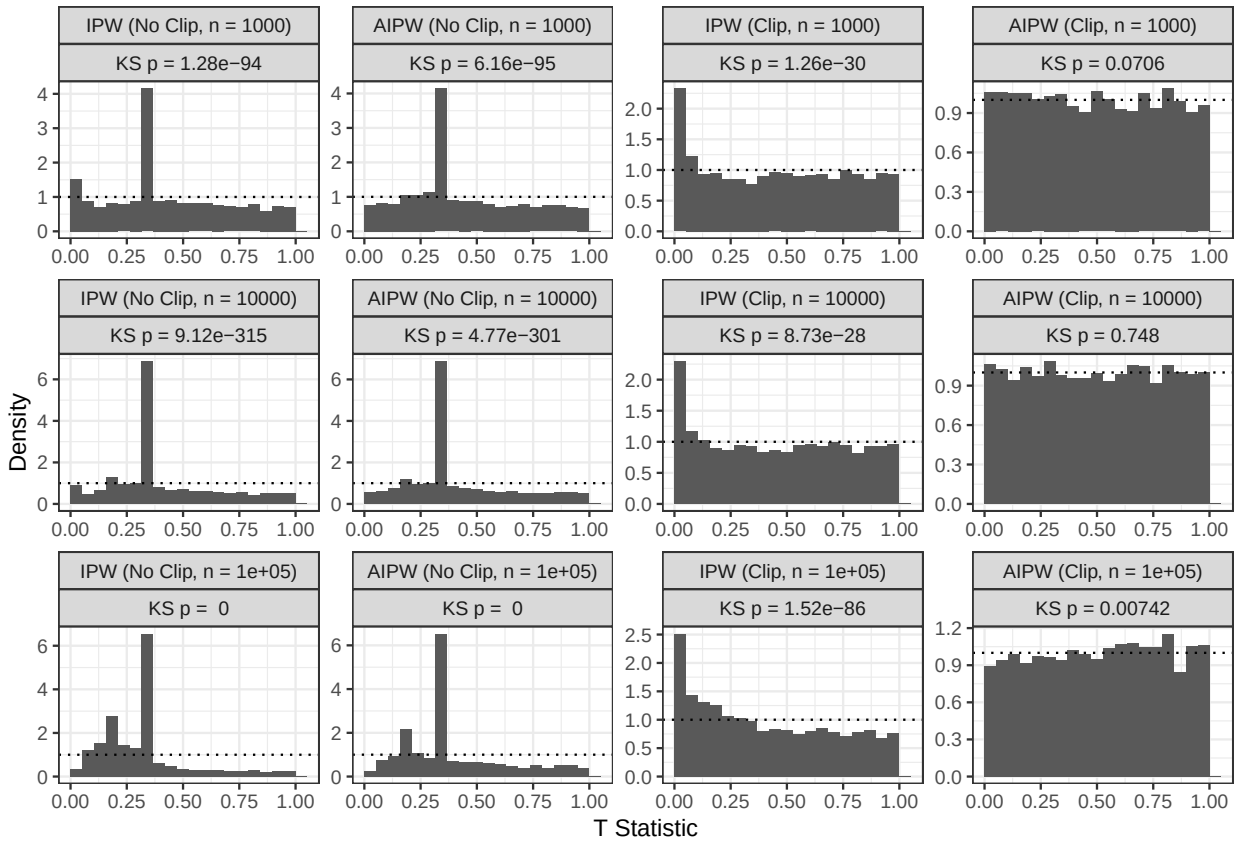

Figure 10: Histograms of simulation p-values on null hypothesis of true APO for various sample sizes, but using the oracle μ regression function instead of the estimated $\hat{\mu}$ regression function.

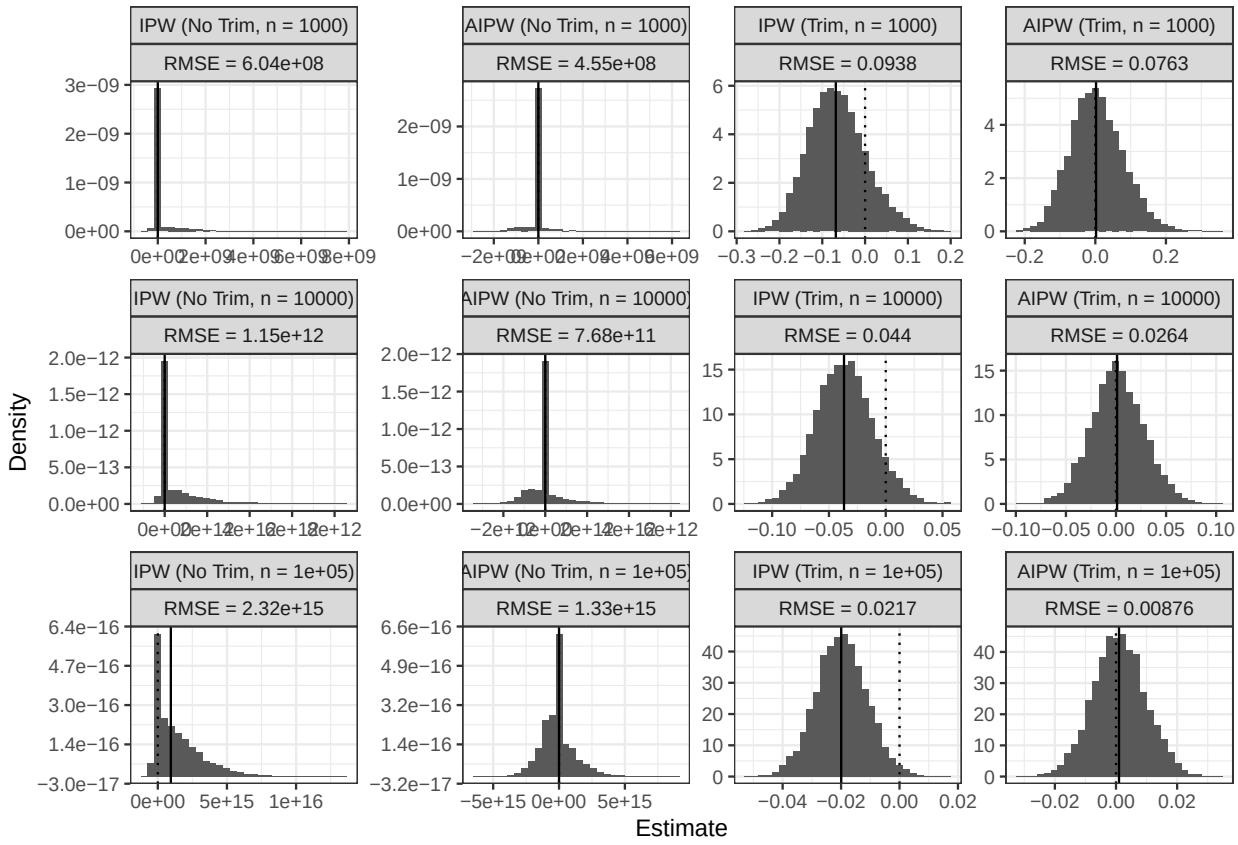

Figure 11: Histograms of point estimates in simulations for the various methods considered in the simulations, but with trimming instead of clipping.

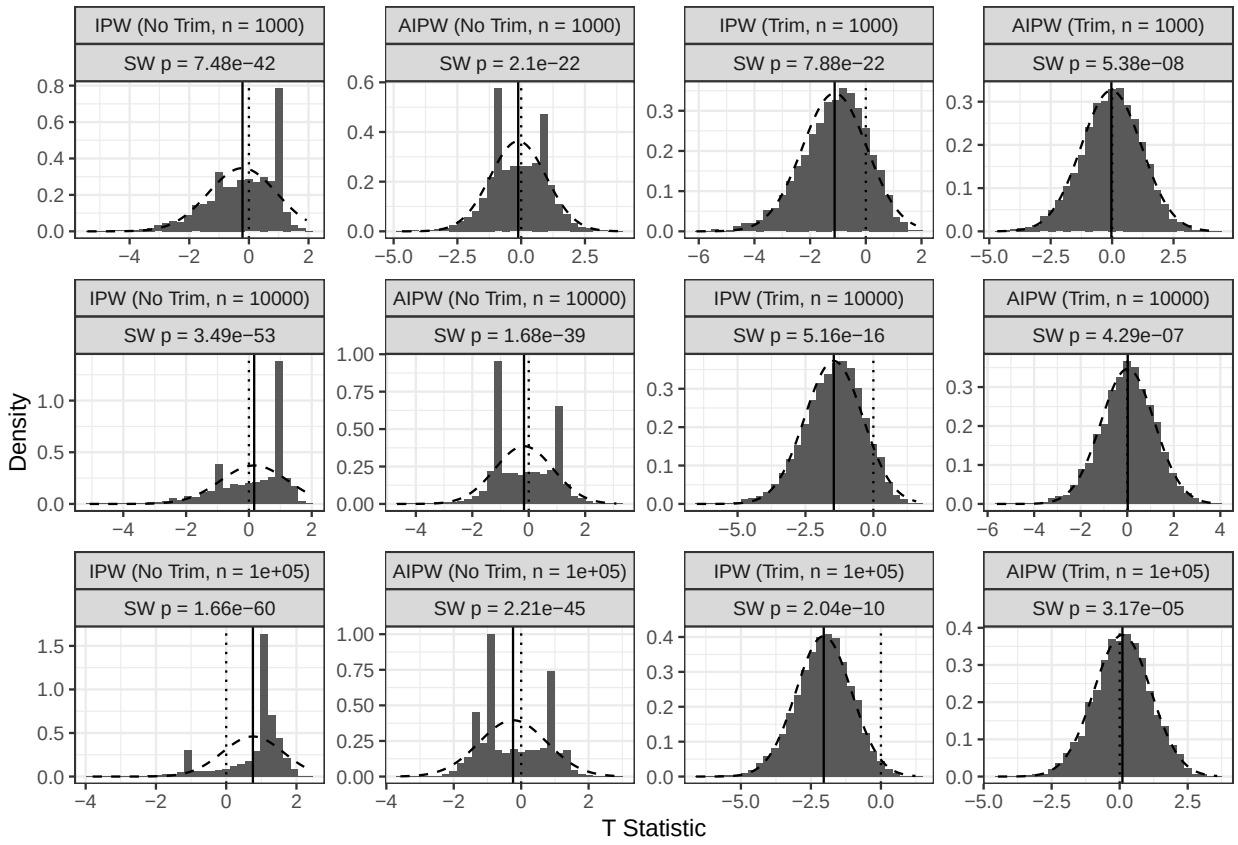

Figure 12: Histograms of simulation t-statistics for various sample sizes, but with trimming instead of clipping.

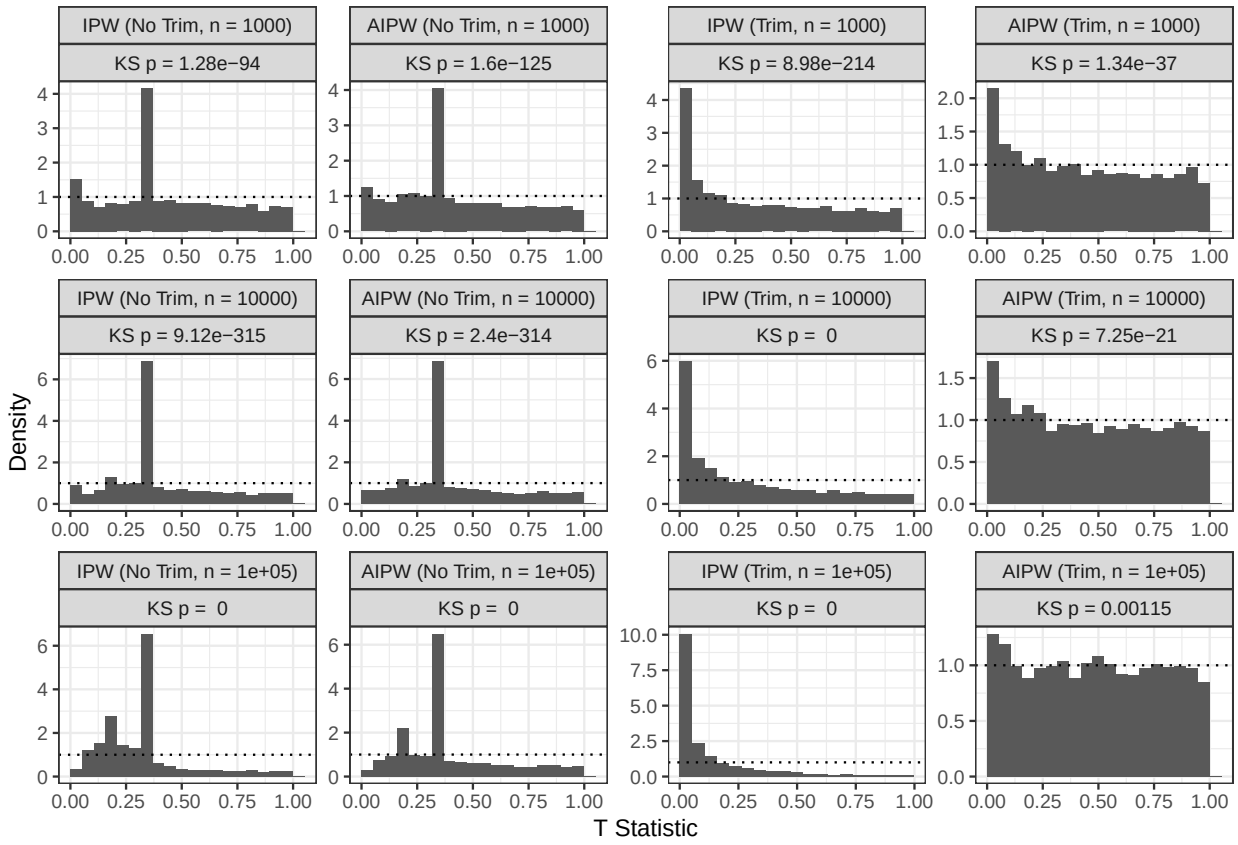

Figure 13: Histograms of simulation p-values on null hypothesis of true APO for various sample sizes, but with trimming instead of clipping.

B Proofs

The plan of the proofs is as follows. Appendix B.1 presents key technical claims in the proof of Theorem 1: Assumption 3' presents the technical rate requirements in terms of AIPW quantities; Proposition 4 shows that the oracle clipped AIPW estimator's t-statistics are well-calibrated, and then the interior Theorem 1' shows that the estimated t-statistics are well-calibrated. These claims rely on several technical claims, which are organized into sections showing asymptotic properties of oracle clipped AIPW (Appendix B.2), consistency of estimated clipped AIPW (Appendix B.3), and asymptotic properties of estimated clipped AIPW (Appendix B.4). Finally, Appendix B.5, Appendix B.6, and Appendix B.7 prove claims about the rate interpretations, proposed rules of thumb, and practical limitations, respectively.

Notation. In these proofs, I use $\mathbb{P}(n)$ to refer to an arbitrary sequence of distributions for the purposes of computing suprema; for such sequences, I use $\psi_n = \psi(\mathbb{P}(n))$ to denote the sequence of average potential outcomes. I use $P_n[c_n]$ to refer to the average of c_n over n draws from P (sometimes abusing notation and including nuisance functions in c_n), and I use $P[c_n]$ to refer to the expectation of c_n over P. This can occasionally lead to unfortunate notation like $\mathbb{P}(n)_n(E_n)$ for a sequence of event probabilities under a sequence of distributions. I write $\lim_{x\to z^+} f(x)$ and $\lim_{x\to z^-}$ for the right- and left-hand limits of f at z. I write $c_n = o_{\mathbb{P}(n)}(1)$ if for all $\delta > 0$, $\mathbb{P}(n)(|c_n|/d_n > \delta) \to 0$, and if no $\mathbb{P}(n)$ is defined, I use $c_n = o_{\mathbb{P}(n)}(d_n)$ to mean that for any sequence of $\mathbb{P}(n) \subset \mathscr{P}$, $c_n = o_{\mathbb{P}(n)}(d_n)$. I write $c_n = O_{\mathbb{P}(n)}(1)$ if for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\mathbb{P}(n)(|c_n|/d_n > \delta) < \epsilon$. If there is a sequence of distributions to be considered, then I use $o(d_n)$ and $O(d_n)$ to implicitly refer to $o_{\mathbb{P}(n)}(d_n)$ and $O_{\mathbb{P}(n)}(d_n)$. I write that $c_n \overset{\mathbb{P}(n)}{\rightsquigarrow} N(0,1)$ if $\sup_{t\in\mathbb{R}} |\mathbb{P}(n)(c_n \le t) - \Phi(t)| \to 0$, where Φ is the standard normal cumulative distribution function; I write that $c_n \to_{\mathbb{P}(n)} c$ if $c_n - c = o_{\mathbb{P}(n)}(1)$; and I write that $c_n \xrightarrow{\mathscr{P}} c$ if for all sequences of $\mathbb{P}(n) \in \mathscr{P}$, $c_n \to_{\mathbb{P}(n)} c$. I write $c_n = \Theta(d_n)$ if there exists a $k_1, k_2 > 0$ such that $\mathbb{P}(n)[c_n \in [k_1 d_n, k_2 d_n]] \to 1$.

B.1 Key Technical Claims

I make use of the following assumptions.

Assumption 3'. Assumption 2 holds, with the following rates on the regression error $r_{\mu,n}$ and the propensity error $r_{e,n}$ for any sequence of $\mathbb{P}(n) \in \mathscr{P}$:

(a) *Consistency.* $r_{\mu,n}, r_{e,n} \to 0$.

(b) *Product of errors.* $r_{\mu,n} r_{e,n} \sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X), b_n\}^2}\right]} \ll n^{-1/2}$.

(c) *Regression error near singularities.* $r_{\mu,n} \frac{\mathbb{P}(n)(e(X) \le b_n)}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e, b_n\}^2}\right]}} \ll n^{-1/2}$.

(d) *Asymptotically known thresholding.* $r_{e,n} \ll b_n$.

These conditions adapt to the distributions in the sequence $\mathbb{P}(n)$, and are weaker than the more interpretable conditions in the main text.

I will proceed under these rate assumptions, which are implied by the assumptions above.

Corollary 4 (Sufficiency of Assumption 3'). *Suppose Assumptions 3, 4, and Assumption 4(i) hold and let $\rho > 0$ be given. Then for any sequence of $\mathbb{P}(n) \in \mathscr{P}$, Assumption 3' holds.*

I will show that the feasible clipped estimator $\hat{\psi}^{AIPW}_{clip}(b_n)$ is first-order equivalent to the oracle clipped estimator $\tilde{\psi}^{AIPW}_{(Orcl)}(b_n)$. The oracle clipped AIPW estimator is asymptotically normal by the trimmed IPW arguments in Ma and Wang (2020). By construction, the oracle clipped AIPW estimator is finite-sample unbiased. The following asymptotic normality follows as a result.

Proposition 4 (Oracle asymptotic normality). *Suppose b_n satisfies $n^{-1/2} \ll b_n \ll 1$. Then the oracle clipped AIPW estimator has uniform convergence to a normal distribution in the sense that*

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}} \sup_{t\in\mathbb{R}} \left| P\left(\frac{\tilde{\psi}^{AIPW}_{(Orcl)}(b_n) - \psi(P)}{\sigma_n} \leq t \right) - \Phi(t) \right| = 0.$$

Proposition 4 will be an extension of the following claim. In addition to this modified theorem, Theorem 1 replaces the oracle standard deviation σ_n with the estimated standard deviation $\hat{\sigma}_n$ when constructing t-statistics.

Theorem 1' ((Slow) Asymptotic Normality). Suppose the conditions of Theorem 1 hold, and $\mathbb{P}(n)$ is a sequence of distributions $P \in \mathscr{P}$. Then $\sigma_n^{-1}\left(\hat{\psi}^{AIPW}_{clip}(b_n) - \psi_n\right) \overset{\mathbb{P}(n)}{\rightsquigarrow} N(0,1)$, where σ_n is the oracle standard deviation defined in Proposition 4.

B.2 Oracle Normality

Lemma 2. *Assume $b_n \to 0$. Then for all large n, the following inequalities hold throughout $\mathscr{P}$:*

(i) $P(e(X) > \pi_{\min}/2) \geq \pi_{\min}/2$

(ii) $\mathbb{E}[e(X)/\{e(X) \vee b_n\}^2] \geq \pi_{\min}/2$

(iii) $\mathbb{E}[|\phi_n - \mathbb{E}_{\mathbb{P}(n)}[\phi_n]|^q] \leq (4M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2}$

(iv) $\mathbb{E}[|\phi_n|^q] \leq (8M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2}$

Proof of Lemma 2. I take these proofs one at a time.

(i) Start from the following inequalities:

$$\pi_{\min} \leq \mathbb{E}[e(X)]$$

$$
\begin{aligned}
&= \mathbb{E}[e(X)\mathbf{1}\{e(X) \leq \pi_{\min}/2\}] + \mathbb{E}[e(X)\mathbf{1}\{e(X) > \pi_{\min}/2\}] \\
&\leq (\pi_{\min}/2)[1 - P(e(X) > \pi_{\min}/2)] + P(e(X) > \pi_{\min}/2) \\
&< \pi_{\min}/2 + P(e(X) > \pi_{\min}/2)
\end{aligned}
$$

Subtracting $\pi_{\min}/2$ from the far left- and right-hand sides of this inequality gives the desired conclusion.

(ii) If $b_n \leq \pi_{\min}/2$ (which happens for all large n), then:

$$
\mathbb{E}[e(X)/\{e(X) \vee b_n\}^2] \geq \mathbb{E}[1/e(X)\mathbf{1}\{e(X) \geq b_n\}] \geq P(e(X) \geq b_n) \geq P(e(X) \geq \pi_{\min}/2) \geq \pi_{\min}/2.
$$

(iii) By Jensen's inequality:

$$
\begin{aligned}
\mathbb{E}[|\phi_n - \mathbb{E}_{\mathbb{P}(n)}[\phi_n]|^q] &\leq 2^{q-1}(\mathbb{E}[|\mu(X) - \mathbb{E}_{\mathbb{P}(n)}[\mu(x)]|^q] + \mathbb{E}[|Y - \mu(X)|^q D/\{e(X) \vee b_n\}^q]) \\
&\leq 2^{q-1}(2^q \mathbb{E}[|\mu(X)|^q] + \mathbb{E}[\mathbb{E}[|Y - \mu(X)|^q \mid X, D = 1]e(X)/\{e(X) \vee b_n\}^q]) \\
&\leq 2^{q-1}(2^q M^q + 2^q \mathbb{E}[\mathbb{E}[|Y|^q \mid X, D = 1]e(X)/\{e(X) \vee b_n\}^q]) \\
&\leq 2^{q-1}(2^q M^q + 2^q M^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2 \times 1/\{e(X) \vee b_n\}^{q-2}]) \\
&\leq 2^{q-1}(2^q M^q + 2^q M^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2})
\end{aligned}
$$

Since $\mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2} \geq \pi_{\min}/2b_n^{q-2} \to \infty$ by Item (ii), I may further bound the above quantity by $2^{2q} M^q \mathbb{E}[e(X)/\{e(X) \vee b_n^2\}]/b_n^{q-2}$ once n is large enough.

(iv) By Jensen's inequality:

$$
\begin{aligned}
\mathbb{E}[|\phi_n|^q] &= \mathbb{E}[|\phi_n - \mathbb{E}_{\mathbb{P}(n)}[\phi_n] + \mathbb{E}_{\mathbb{P}(n)}[\phi_n]|^q] \\
&\leq 2^{q-1}(\mathbb{E}[|\phi_n - \mathbb{E}_{\mathbb{P}(n)}[\phi_n]|^q] + |\mathbb{E}_{\mathbb{P}(n)}[\phi_n]|^q]) \\
&\leq 2^{q-1}(4M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2} + 2^{q-1}|\mathbb{E}_{\mathbb{P}(n)}[\mu(x)]|^q && \text{(Item (iii))} \\
&\leq 2^{q-1}(4M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2/b_n^{q-2}] + 2^{q-1}\mathbb{E}[\mathbb{E}[|Y|^q \mid X, D = 1]] && \text{(Jensen)} \\
&\leq 2^{q-1}(4M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2} + 2^{q-1} M^q
\end{aligned}
$$

As before, since $\mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2} \to \infty$, the first term in the upper bound is eventually larger than the second and I may bound the whole expression by $(8M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2}$ once n is large enough.

□

Lemma 3. *Let $c(\gamma) = \frac{\gamma-1}{\gamma} C^{-1/(\gamma-1)} > 0$. Then for any $P \in \mathscr{P}$, I have:*

$$\mathbb{E}_P[e(X)\mathbf{1}\{e(X) \le b_n\}] \ge c(\gamma) P(e(X) \le b_n)^{\gamma/(\gamma-1)}. \tag{5}$$

This lower bound is attained when $P(e(X) \le t) = t^{\gamma-1}$.

Proof of Lemma 3. Let $p = P(e(X) \le b_n)$. If $p = 0$, then the bound holds trivially so I will assume throughout that $p > 0$. Then I may write:

$$\begin{aligned}
\mathbb{E}_P[e(X)\mathbf{1}\{e(X) \le b_n\}] &= \int_0^\infty P(e(X)\mathbf{1}\{e(X) \le b_n\} > t) dt \\
&= \int_0^{b_n} P(t < e(X) \le b_n) dt \\
&= \int_0^{b_n} p - P(e(X) \le t) dt \\
&\ge b_n p - \int_0^{b_n} \min\{p, Ct^{\gamma-1}\} dt \\
&= b_n p - (C/\gamma)(p/C)^{\gamma/(\gamma-1)} - b_n p + p(p/C)^{1/(\gamma-1)} \\
&= c(\gamma) p^{\gamma/(\gamma-1)}.
\end{aligned}$$

This proves the lower bound. When $P(e(X) \le t) = t^{\gamma-1}$, a direct calculation gives $\mathbb{E}_P[e(X)\mathbf{1}\{e(X) \le b_n\}] = [(\gamma-1)/\gamma] b_n^\gamma = [(\gamma-1)/\gamma] P(e(X) \le b_n)^{\gamma/(\gamma-1)}$. Therefore, the lower bound is also sharp. □

Lemma 4. *For any $P \in \mathscr{P}$,*

$$\begin{aligned}
Var_P(\phi(Z \mid b_n, \eta)) &\ge \sigma^2_{\min} \mathbb{E}_P\left[e(X)/\max\{e(X), b_n\}^2\right] \\
&\ge \sigma^2_{\min}[c(\gamma) P(e(X) \le b_n)^{\gamma/(\gamma-1)}/b_n^2 + \pi_{\min}/2] \\
&\ge \sigma^2_{\min} \pi_{\min}/2 > 0.
\end{aligned}$$

Proof of Lemma 4. For the first line:

$$\begin{aligned}
\mathrm{Var}_P(\phi_n) &= \mathbb{E}[\mathrm{Var}(\phi_n \mid X)] + \mathrm{Var}(\mathbb{E}[\phi_n \mid X]) \\
&= \mathbb{E}[\mathrm{Var}(\phi_n \mid X)] \\
&= \mathbb{E}[\mathbb{E}[|Y - \mu(X)|^2 \mid X, D = 1] e(X)/\{e(X) \vee b_n\}^2] \\
&\ge \sigma^2_{\min} \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2].
\end{aligned}$$

Since Definition 1 implies $e(X) > 0$, $\mathrm{Var}_P(\phi_n) > 0$.

For the second line, I assume n is so large that $b_n \leq \pi_{\min}/2$. Then:

$$\begin{aligned}
\mathbb{E}[e(X)/\max\{e(X), b_n\}^2] &= \mathbb{E}[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + \mathbb{E}[1/e(X)\mathbf{1}\{e(X) > b_n\}] \\
&\geq \mathbb{E}[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + P(e(X) > b_n) \\
&\geq \mathbb{E}[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + P(e(X) > \pi_{\min}) \\
&\geq \mathbb{E}[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + \pi_{\min}/2 && \text{(Lemma 2.(i))} \\
&\geq c(\gamma)(1/b_n)^2 P(e(X) \leq b_n)^{\gamma/(\gamma-1)} + \pi_{\min}/2. && \text{(Lemma 3)}
\end{aligned}$$

The final line is immediate. □

Lemma 5. $\frac{1}{\sqrt{Var(\phi(Z|b_n,\eta))}} \leq \frac{1}{\sqrt{\sigma_{\min}^2 \mathbb{E}_{\mathbb{P}(n)}[D/\max\{e(X),b_n\}^2]}}$.

Proof of Lemma 5. By Lemma 4, I have:

$$\sigma_n^{-1} \leq n^{1/2}/\sqrt{\sigma_{\min}^2 \mathbb{E}_{\mathbb{P}(n)}\left[D/\max\{e(X), b_n\}^2\right]},$$

where $\sigma_n^{-1} = n^{-1/2}/\sqrt{Var_{\mathbb{P}(n)}(\phi_n)}$. □

Lemma 6. *Define* $\tilde{\phi}(Z \mid b, P) \equiv \phi(Z \mid b, \eta(\mathbb{P})) - \mathbb{E}_P[\mu(X)]$ *for* $P \in \mathscr{P}$. *Further define* $\rho(b, P) \equiv \mathbb{E}_P[|\tilde{\phi}(Z \mid b, P)|^3]$ *and* $\sigma(b, P) \equiv \sqrt{Var_P(\tilde{\phi}(Z \mid b, P))}$.

Then the following hold:

1. $\mathbb{E}_P[\tilde{\phi}(Z \mid b, P)] = 0$
2. $\sigma(b, P) > 0$
3. $\rho(b, P) < \infty$ *(though it may be arbitrarily large)*
4. *If* b_n *be a sequence of positive real numbers such that* $n^{-1/2} \ll b_n$ *and* $\mathbb{P}(n)$ *be a sequence of distributions in* $\mathscr{P}$, *then* $\frac{\rho(b_n, \mathbb{P}(n))}{\sigma(b_n, \mathbb{P}(n))^3 \sqrt{n}} = o(1)$.

Proof of Lemma 6. $\mathbb{E}_P[\tilde{\phi}(Z \mid b_n, P)] = 0$ is immediate.

$Var_P[\tilde{\phi}(Z \mid b, P)] > 0$ follows by Lemma 4.

For the third moment being finite:

$$\begin{aligned}
\rho(b, P) = \mathbb{E}_P[|\tilde{\phi}(Z \mid b, P)|^3] &\leq 8\mathbb{E}_P\left[|\mu(X) - \mathbb{E}_P[\mu(X)]|^3 + b^{-3}|Y - \mu(X)|^3\right] \\
&\leq O(M^q) + 16b^{-3}\mathbb{E}_P\left[|Y|^3\right].
\end{aligned}$$

This is finite (and $O(b^{-3})O(M^q)$) by assumption.

Finally, I have the claim for sequences. Recall that by Lemmas 4 and 2, $\frac{1}{\sigma(b_n,\mathbb{P}(n))^3\sqrt{n}} = o(1)$ and $\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] \geq \sigma^2_{\min}/2$. As a result:

$$\begin{aligned}
\frac{\rho(b_n,\mathbb{P}(n))}{\sigma(b_n,\mathbb{P}(n))^3\sqrt{n}} &\leq 8\frac{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D|Y-\mu(X)|^3}{\max\{e(X),b_n\}^3} + |\mu(X)-\mathbb{E}_{\mathbb{P}(n)}[\mu(X)]|^3\right]}{\sigma(b_n,\mathbb{P}(n))^3\sqrt{n}} \\
&\leq O(M^q)\frac{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right]}{b_n\sigma(b_n,\mathbb{P}(n))^3\sqrt{n}} + \frac{O(M^q)}{\sigma(b_n,\mathbb{P}(n))^3\sqrt{n}} \\
&= O(M^q,\sigma^2_{\min})\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right]^{-1/2}(b_n^2 n)^{-1/2} + o(1) \\
&= o(1).
\end{aligned}$$

□

Proof of Proposition 4. Let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$. By Lemma 6 and the Berry Esseen Theorem, the difference between the CDF of oracle clipped AIPW t-statistic $\frac{\tilde{\psi}^{AIPW}_{clip}-\psi_n}{\sigma_n} = \frac{\sum\tilde{\phi}(Z|b_n,\mathbb{P}(n))}{\sqrt{Var(\phi(Z|b_n,\eta))}\sqrt{n}}$ and the standard normal CDF Φ is uniformly bounded above by $\frac{3\rho(b_n,\mathbb{P}(n))}{\sigma(b_n,\mathbb{P}(n))^3\sqrt{n}}$. By Lemma 6.4, this difference tends to zero. Therefore:

$$\limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\sup_{t\in\mathbb{R}}\left|P\left(\frac{\tilde{\psi}^{AIPW}_{(Orcl)}(b_n)-\psi(P)}{\sigma_n}\leq t\right)-\Phi(t)\right| = \limsup_{n\to\infty} o(1) = 0.$$

□

B.3 Consistency

Lemma 7. *Under cross-fitting, I have the bias bound:*

$$\left|\mathbb{E}\left[\hat{\psi}^{AIPW}_{clip}(b_n)-\psi_n\right] \mid \hat{\mu},\hat{e}\right| \leq r_{\mu,n}\mathbb{P}(n)\left(e(X)\leq b_n+r_{e,n}\right) + r_{\mu,n}r_{e,n}\mathbb{E}\left[\frac{\mathbf{1}\{e>b_n+r_{e,n}\}}{e-r_{e,n}}\right].$$

Proof. Fix one fold and take $\hat{\mu}^{(-k)}$ and $\hat{e}^{(-k)}$ as given. I write the bias relative to oracle clipped AIPW as:

$$\mathbb{E}\left[(\hat{\mu}-\mu)\left(1-\frac{D}{\max\{\hat{e},b_n\}}\right)+(\mu-Y)\left(\frac{D}{\max\{e,b_n\}}-\frac{D}{\max\{\hat{e},b_n\}}\right)\right] = \mathbb{E}\left[(\hat{\mu}-\mu)\left(1-\frac{D}{\max\{\hat{e},b_n\}}\right)\right],$$

with the equality following by cross-fitting.

For $p = 1, 2$, let $c_{n,p}$ solve:

$$\frac{|\max\{c_{n,p}, b_n\}^p - \max\{c_{n,p} - r_{e,n}, b_n\}^p|}{\max\{c_{n,p} - r_{e,n}, b_n\}^p} = \frac{|\max\{c_{n,p}, b_n\}^p - \max\{c_{n,p} + r_{e,n}, b_n\}^p|}{\max\{c_{n,p} + r_{e,n}, b_n\}^p}. \tag{6}$$

$c_{n,p}$ is useful because it is the changeover point between whether the worst-case $\hat{e}$ is above or below e. Note that $c_{n,p} \in (b_n, b_n + r_{e,n})$ by the intermediate value theorem: when $c_{n,p} = b_n$, the left-hand side is zero, while when $c_{n,p} = b_n + r_{e,n}$, the left hand side has the smaller denominator but equal numerator.

$c_{n,1}$ is useful, because when $e(X) = c_{n,1}$, $\hat{e}(X) = e(X) - r_{e,n}$ and $\hat{e}(X) = e(X) + r_{e,n}$ produce equal levels of observation-wise bias from clipped inverse propensities relative to unity.

Then I have the bound:

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{\psi}^{AIPW}_{clip}(b_n) - \psi_n\right]\right| &\leq \left|\mathbb{E}\left[(\hat{\mu} - \mu)\frac{\max\{\hat{e}, b_n\} - e}{\max\{\hat{e}, b_n\}}\mathbf{1}\{e \leq b_n - r_{e,n}\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\quad + \left|\mathbb{E}\left[(\hat{\mu} - \mu)\frac{\max\{\hat{e}, b_n\} - e}{\max\{\hat{e}, b_n\}}\mathbf{1}\{e \in (b_n - r_{e,n}, c_{n,1}]\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\quad + \left|\mathbb{E}\left[(\hat{\mu} - \mu)\frac{\max\{\hat{e}, b_n\} - e}{\max\{\hat{e}, b_n\}}\mathbf{1}\{e \in (c_{n,1}, b_n + r_{e,n}]\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\quad + \left|\mathbb{E}\left[(\hat{\mu} - \mu)\frac{\max\{\hat{e}, b_n\} - e}{\max\{\hat{e}, b_n\}}\mathbf{1}\{e > b_n + r_{e,n}\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\leq \left|\mathbb{E}\left[r_{\mu,n}\frac{b_n - e}{b_n}\mathbf{1}\{e \leq b_n - r_{e,n}\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\quad + \left|\mathbb{E}\left[r_{\mu,n}\frac{r_{e,n}}{e + r_{e,n}}\mathbf{1}\{e \in (b_n - r_{e,n}, c_{n,1}]\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\quad + \left|\mathbb{E}\left[r_{\mu,n}\frac{e - b_n}{b_n}\mathbf{1}\{e \in (c_{n,1}, b_n + r_{e,n}]\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\quad + \left|\mathbb{E}\left[(\hat{\mu} - \mu)\frac{r_{e,n}}{e - r_{e,n}}\mathbf{1}\{e > b_n + r_{e,n}\}\right] \mid \hat{e}, \hat{\mu}\right| \\
&\leq r_{\mu,n}\mathbb{P}(n)(e(X) \leq b_n + r_{e,n}) + r_{\mu,n}r_{e,n}\mathbb{E}\left[\frac{1}{e - r_{e,n}}\mathbf{1}\{e > b_n + r_{e,n}\}\right],
\end{aligned}$$

where the final line follows because $r_{\mu,n}$ is always multiplied by a term that is bounded above by one for all $e(X) \leq b_n + r_{e,n}$. □

Proof of Proposition 1. Let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$, and fix some $k \in 1, ..., K$.

Write $\hat{\psi}^{AIPW}_{clip}(b_n) = \frac{1}{K}\sum_k \hat{\psi}^{AIPW}_{clip,k}(b_n)$, where $\hat{\psi}^{AIPW}_{clip,k}(b_n)$ is the fold$-k$ APO estimate. I will show that $\hat{\psi}^{AIPW}_{clip,k}(b_n) - \psi_n = o_{\mathbb{P}(n)}(1)$.

First I show that $\mathbb{E}\left[\hat{\psi}^{AIPW}_{clip,k}(b_n) - \psi_n \mid \hat{\mu}^{(-k)}, \hat{e}^{(-k)}\right] = o_{\mathbb{P}(n)}(1)$. This holds by the assumptions of Proposition 1 applied to the bias bound from Lemma 7:

$$\left|E\left[\hat{\psi}^{AIPW}_{clip,k}(b_n) - \psi_n \mid \hat{\mu}^{(-k)}, \hat{e}^{(-k)}\right]\right| \leq Cr_{\mu,n}(b_n + r_{e,n})^{\gamma_0 - 1} + r_{\mu,n}r_{e,n}E\left[\frac{1\{e(X) > b_n + r_{e,n}}{e(X) - r_{e,n}}\right].$$

If $r_{\mu,n}\frac{r_{e,n}+b_n}{b_n} \to_{\mathbb{P}(n)} 0$, then this term is $o_{\mathbb{P}(n)}(1)$ because $r_{\mu,n}$ and $r_{e,n}+b_n$ are bounded above by Assumption 2, and at least one of the two terms must tend to zero because $b_n \to_{\mathbb{P}(n)} 0$ and $r_{\mu,n}\frac{r_{e,n}+b_n}{b_n} \to_{\mathbb{P}(n)} 0$. If $r_{e,n}b_n^{\min\{\gamma_0-2,0\}} \to_{\mathbb{P}(n)} 0$, then there is a $\delta_n \to \infty$ such that $r_{e,n}b_n^{\min\{\gamma_0-2,0\}}\delta_n \to 0$. For the first term, $r_{\mu,n}(b_n+r_{e,n})^{\gamma_0-1} \to_{\mathbb{P}(n)} 0$ because $r_{\mu,n}$ is bounded above and $\gamma_0 > 1$. For the final term, suppose that $\gamma_0 < 2$, so that the claim is not immediate. If $r_{e,n} \leq b_n\delta_n^{1/(1-\gamma_0)}$, then $r_{e,n}\left[\frac{1\{e(X)>b_n+r_{e,n}\}}{e(X)-r_{e,n}}\right] \leq r_{e,n}/b_n \leq 1/\delta_n \to 0$. If $r_{e,n} \geq b_n\delta_n^{1/(1-\gamma_0)}$:

$$\begin{aligned}
E_{P(n)}\left[\frac{1\{e(X)>b_n+r_{e,n}\}}{e(X)-r_{e,n}}\right] &= C(b_n+r_{e,n})^{\gamma_0-1}b_n^{-1} + C(\gamma_0-1)\int_{b_n+r_{e,n}}^{C^{-1/(\gamma_0-1)}} (t-r_{e,n})^{-1}t^{\gamma_0-2}dt \\
&\leq C(b_n+r_{e,n})^{\gamma_0-1}b_n^{-1} + C(\gamma_0-1)\int_{b_n+r_{e,n}}^{C^{-1/(\gamma_0-1)}} (t-r_{e,n})^{\gamma_0-3}dt \\
&\leq C(b_n+r_{e,n})^{\gamma_0-1}b_n^{-1} + C(\gamma_0-1)\int_{b_n}^{1} x^{\gamma_0-3}dx \\
&= C(b_n+r_{e,n})^{\gamma_0-1}b_n^{-1} + \frac{C(\gamma_0-1)}{2-\gamma_0}\left(b_n^{\gamma_0-2}-1\right) \\
&\leq \left(2^{\gamma_0-1}C + \frac{\gamma_0-1}{2-\gamma_0}C\right)b_n^{\gamma_0-2} + 2Cr_{e,n}^{\gamma_0-1}b_n^{-1} \\
&\leq O(1)b_n^{\gamma_0-2}\delta_n.
\end{aligned}$$

Note that this bound may be lax. Whether the propensity rate requirement could be weakened is an open question for future work. Regardless, in this remaining case under the propensity rate requirement (i), $r_{\mu,n}r_{e,n}b_n^{\gamma_0-2}\delta_n \to_{\mathbb{P}(n)} 0$ by construction of δ_n, so that the final term of the bias bound tends to zero.

Next, I show that $V(\hat{\psi}_{clip,k}^{AIPW}(b_n) \mid \hat{\mu}^{(-k)}, \hat{e}^{(-k)}) = o_{\mathbb{P}(n)}(1)$. I have:

$$\begin{aligned}
V(\hat{\psi}_{clip,k}^{AIPW}(b_n) \mid \hat{\mu}^{(-k)}, \hat{e}^{(-k)}) &\leq n^{-1}\mathbb{E}\left[\left(\hat{\mu}+\frac{D(Y-\hat{\mu})}{\max\{\hat{e},b_n\}}\right)^2 \mid \hat{\mu},\hat{e}\right] \\
&\leq 8n^{-1}b_n^{-2} \qquad \text{(Lemma 2.(iv))} \\
&= o(1).
\end{aligned}$$

Therefore $\mathbb{E}\left[\left(\hat{\psi}_{clip,k}^{AIPW}(b_n)-\psi_n\right)^2 \mid \hat{\mu}^{(-k)}, \hat{e}^{(-k)}\right] = o_{\mathbb{P}(n)}(1)$. Therefore, for all $\epsilon > 0$, every $\delta > 0$, and every sequence of $\mathbb{P}(n)$, there is an n large enough such that $\mathbb{P}(n)\left(\left(\hat{\psi}_{clip,k}^{AIPW}(b_n)-\psi_n\right)^2 > \epsilon^2\right) \leq \delta$. Thus, consistency holds. □

Proof of Proposition 2. Define $\sigma^2_{\max} = \sup_{\mathbb{P}\in\mathscr{P}}\sup_{X,D} Var(Y \mid X, D)$. By the presence of $q > 2$ moments, $\sigma^2_{\max}$ is finite.

By Lemma 4, $\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\sigma^2_{\min}}{\max\{e(X),b_n\}^2}\right] \not\to 0$.

By iid sampling and the oracle AIPW conditional mean being equal to $\mu(X)$, I obtain:

$$\begin{aligned}
Var_{\mathbb{P}(n)}\left(\tilde{\psi}^{AIPW}_{(Orcl)}(b_n)\right) - n^{-1}Var_{\mathbb{P}(n)}\left(\mu(X)\right) &= \mathbb{E}_{\mathbb{P}(n)}\left[Var_{\mathbb{P}(n)}\left(\tilde{\psi}^{AIPW}_{(Orcl)}(b_n) \mid \{X\}\right)\right] \\
&= n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[Var_{\mathbb{P}(n)}\left(\frac{D(Y-\mu(X))}{\max\{e(X),b_n\}} \mid X\right)\right] \\
&= n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[e(X)Var_{\mathbb{P}(n)}\left(\frac{(Y-\mu(X))}{\max\{e(X),b_n\}} \mid X, D=1\right)\right] \\
&= n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[e(X)\frac{Var(Y \mid X, D=1)}{\max\{e(X),b_n\}^2}\right] \\
&= \Theta\left(n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right]\right).
\end{aligned}$$

In addition, $n^{-1}Var_{\mathbb{P}(n)}\left(\mu(X)\right) \leq n^{-1}M = O\left(n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right]\right)$, proving the claim. □

Proof of Corollary 1. Let n be large enough that $b_n \leq 1$ and $b_n^{\gamma_0-2} > 2$. Recall the definition of $\sigma^2_{\max}$ from the proof of Proposition 2.

For the upper bound, let $\mathscr{P}$ be arbitrary:

$$\begin{aligned}
\sigma_n^2 - n^{-1}Var(\mu) &= n^{-1}Var_P\left(\mu(X) + \frac{D(Y-\mu(X))}{\max\{e(X),b_n\}}\right) - n^{-1}Var_P(\mu(X)) \\
&= n^{-1}E_P\left[\frac{D(Y-\mu(X))^2}{\max\{e(X),b_n\}^2}\right] \\
&\leq n^{-1}E_P\left[\frac{D\sigma^2_{\max}}{\max\{e(X),b_n\}^2}\right] \\
&= \sigma^2_{\max}E_P\left[\frac{e(X)}{\max\{e(X),b_n\}^2}\right] \\
&= n^{-1}\sigma^2_{\max}\int_0^\infty P\left(\frac{e(X)}{\max\{e(X),b_n\}^2} \geq t\right)dt \\
&= n^{-1}\sigma^2_{\max}\int_0^\infty P\left(e(X) \leq b_n, e(X) \geq tb_n^2\right)dt \\
&+ n^{-1}\sigma^2_{\max}\int_0^\infty P\left(e(X) > b_n, e(X) \leq 1/t\right)dt \\
&= n^{-1}\sigma^2_{\max}\int_0^{b_n^{-1}} P\left(e(X) \in [tb_n^2, b_n]\right)dt + n^{-1}\sigma^2_{\max}\int_0^{b_n^{-1}} P\left(e(X) \in [b_n, 1/t]\right)dt \\
&= n^{-1}\sigma^2_{\max}\int_0^{b_n^{-1}} P\left(e(X) \in [tb_n^2, 1/t]\right)dt \\
&\leq n^{-1}\sigma^2_{\max}\int_0^{b_n^{-1}} P\left(e(X) \leq 1/t\right)dt \\
&\leq Cn^{-1}\sigma^2_{\max}\int_0^{b_n^{-1}} t^{1-\gamma_0}dt \\
&= \underbrace{\frac{C\sigma^2_{\max}}{\gamma_0-2}}_{C'} n^{-1}b_n^{\gamma_0-2}.
\end{aligned}$$

For the remaining term, $n^{-1}Var(\mu) = O(n^{-1}) = o\left(C'n^{-1}b_n^{\gamma_0-2}\right)$.

For the lower bound, define $\mathscr{P} = \{P\}$, where P is the distribution which draws $e(X)$ from the CDF $P(e(X) \leq \pi) = (1-\pi_{\min})\min\{C\pi^{\gamma_0-1}, 1\} + \pi_{\min}1\{\pi \geq 1\}$ and $Y \mid X, D \sim \mathcal{N}(0, \sigma^2_{\min})$. This distribution has valid conditional moments and residual variance by the minimal value of M and the choice of $Var(Y \mid X < D)$. The treated fraction is at least $\pi_{\min} > 0$. For all $\pi < C^{-1/(\gamma_0-1)}$, $P(e(X) \leq \pi) = (1-\pi_{\min})C\pi^{\gamma_0-1} \leq C\pi^{\gamma_0-1}$; for all $\pi > C^{-1/(\gamma_0-1)}$, $P(e(X) \leq \pi) = (1-\pi_{\min}) + \pi_{\min}1\{\pi = 1\}$, which must be below $C\pi^{\gamma_0-1}$ for all such π in order for $\mathscr{P}$ to be non-empty. Finally, note that:

$$\begin{aligned}
\sigma_n^2 - n^{-1}Var_P(\mu) &= n^{-1}E_P\left[\frac{D(Y-\mu(X))^2}{\max\{e, b_n\}^2}\right] \\
&= n^{-1}\sigma^2_{\min}\left(\begin{array}{lll} & \int_0^{b_n} \frac{t}{b_n^2}(1-\pi_{\min})(\gamma_0-1)Ct^{\gamma_0-2}dt \\ + & \int_{b_n}^1 \frac{1}{t}(1-\pi_{\min})(\gamma_0-1)Ct^{\gamma_0-2}dt \\ + & \pi_{\min} \end{array}\right) \\
&= n^{-1}\sigma^2_{\min}\left(\begin{array}{lll} & b_n^{-2}(1-\pi_{\min})(\gamma_0-1)C\int_0^{b_n} t^{\gamma_0-1}dt \\ + & (1-\pi_{\min})(\gamma_0-1)C\int_{b_n}^1 t^{\gamma_0-3}dt \\ + & \pi_{\min} \end{array}\right) \\
&= n^{-1}\sigma^2_{\min}\left(\begin{array}{lll} & b_n^{\gamma_0-2}(1-\pi_{\min})C\left(\frac{\gamma_0-1}{\gamma_0} + \frac{\gamma_0-1}{2-\gamma_0}\right) \\ + & \pi_{\min} - (1-\pi_{\min})C\frac{\gamma_0-1}{2-\gamma_0} \end{array}\right) \\
&\geq \underbrace{\sigma^2_{\min}C(1-\pi_{\min})\left(\frac{\gamma_0-1}{\gamma_0} + \frac{\gamma_0-1}{2(2-\gamma_0)}\right)}_{C''} n^{-1}b_n^{\gamma_0-2}.
\end{aligned}$$

Note also that $C'' > 0$. Thus, $C''n^{-1}b_n^{\gamma_0-2} \leq \sup_{P\in\mathscr{P}}\sigma_n^2 - \mathrm{Var}_P(\mu(X)) \leq C'n^{-1}b_n^{\gamma_0-2}$. Analogously to before, $n^{-1}Var(\mu) = o\left(C''n^{-1}b_n^{\gamma_0-2}\right)$, completing the proof. □

B.4 Asymptotic Normality and Rates

Lemma 8. *Suppose the conditions of Proposition 4 hold and let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$. Then for all $\alpha > 0$, $\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] = O\left(1 + n^{\alpha}b_n^{\gamma_0-2}\right)$.*

Proof of Lemma 8. Let $\alpha > 0$ be given and let m be an integer above $1/(2\alpha)$. Let $M_0 \equiv \mathbf{1}\{e(X) \leq b_n\}$ and for $j = 1, ..., m$, write $M_j \equiv \mathbf{1}\{b_n n^{(j-1)\alpha} < e(X) \leq b_n n^{j\alpha}\}$. For n large enough, $\sum_{j=0}^m M_j = 1$ with

probability one, because by assumption, $b_n n^{m\alpha} \gg 1$. For all n large enough:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] &= \sum_{j=0}^{m} \mathbb{E}_{\mathbb{P}(n)}\left[M_j \frac{D}{\max\{e(X),b_n\}^2}\right] \\
&= \mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e(X)\le b_n\}}{\max\{e(X),b_n\}^2}\right] + \sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{b_n n^{(j-1)\alpha} \le e(X) \le b_n n^{j\alpha}\}}{\max\{e(X),b_n\}^2}\right] \\
&\le \mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e(X)\le b_n\}}{b_n^2}\right] + \sum_{j=1}^{m} \mathbb{E}_{\mathbb{P}(n)}\left[\frac{\mathbf{1}\{b_n n^{(j-1)\alpha} e(X) \le b_n n^{j\alpha}\}}{e(X)}\right] \\
&\le b_n^{-1} b_n^{\gamma_0 - 1} + \sum_{j=1}^{m} \left(b_n n^{j\alpha}\right)^{\gamma_0 - 1} b_n^{-1} n^{(1-j)\alpha} \\
&= b_n^{\gamma_0 - 2}\left(1 + \sum_{j=1}^{m} n^{\alpha(1+j(\gamma_0-2))}\right) \le m b_n^{\gamma_0-2}\left(1 + n^{\alpha(\gamma_0-1)}\right) \le m n^{\alpha} b_n^{\gamma_0-2} + O(1).
\end{aligned}$$

□

Lemma 9. *Suppose the conditions of Proposition 4 hold and $\mathbb{P}(n)$ is a sequence of distributions in $\mathscr{P}$. Then*

$$\frac{1}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}} = O\left(\frac{1}{\sqrt{1 + b_n^{-2}\mathbb{P}(n)(e(X)\le b_n)^{\gamma_0/(\gamma_0-1)}}}\right).$$

Proof of Lemma 9. For any $m \ge 0$, define $\mathscr{P}_{m,n} = \{P \in \mathscr{P} \mid P(e(X) \le b_n) \le m)$. For each n, define $m_n = \mathbb{P}(n)(e(X) \le b_n)^{\gamma_0/(\gamma_0-1)}$.

I have:

$$\begin{aligned}
\sup_{P\in\mathscr{P}_{m_n,n}} \mathbb{E}_P\left[\frac{D}{\max\{e,b_n\}^2}\right] &= \sup_{P\in\mathscr{P}_{m_n,n}} \mathbb{E}_P\left[\frac{D\mathbf{1}\{e(X) > b_n\}}{\max\{e,b_n\}^2}\right] + \mathbb{E}_P\left[\frac{D\mathbf{1}\{e(X)\le b_n\}}{b_n^2}\right] \\
&\ge 1 + \sup_{P\in\mathscr{P}_{m_n,n}} b_n^{-2}\mathbb{E}_P[e(X)\mathbf{1}\{e(X)\le b_n\}] \\
&\ge 1 + c(\gamma_0) \sup_{P\in\mathscr{P}_{m_n,n}} P(e(X)\le b_n)^{\gamma_0/(\gamma_0-1)} \qquad \text{(Lemma 3)} \\
&= 1 + c(\gamma_0) b_n^{-2} m_n^{\gamma_0/(\gamma_0-1)}.
\end{aligned}$$

Therefore $\left(1 + b_n^{-2}\mathbb{P}(n)(e(X)\le b_n)^{\gamma_0/(\gamma_0-1)}\right)^2 = O\left(\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right)$. Taking the square root and inverting both sides completes the proof. □

Proof of Corollary 4. The two changes are the product-of-errors term being stated as 3'(b) and the regression error near singularities term being stated as 3'(c).

I begin with the product-of-errors term (b). Suppose Assumption 3(b) holds. I wish to show that

if $r_{\mu,n}r_{e,n}\left(1+b_n^{(\gamma_0-2)/2}n^{\alpha}\right)=o(n^{-1/2})$, then $r_{\mu,n}r_{e,n}\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n^2\}}\right]}=o_{\mathbb{P}(n)}(1)$. Let $\alpha>0$ from Assumption 3 be given. By Lemma 8, for all $\alpha>0$,

$$r_{\mu,n}r_{e,n}\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e(X),b_n^2\}}\right]}=O\left(r_{\mu,n}r_{e,n}\left(1+n^{\alpha/2}b_n^{(\gamma_0-2)/2}\right)\right).$$

By Assumption 3(b), there is an $\alpha>0$ such that this term is $o(1)$, implying Assumption 3'(b) holds.

Next I consider the regression error near the singularities term (c). I first verify (c) for a sequence of $\mathbb{P}(n)\in\mathscr{P}$ under Assumption 4(i). Let a sequence of $\mathbb{P}(n)\in\mathscr{P}$ and b_n be given, and consider an arbitrary sub-sequence. If there is a further sub-sub-sequence with $\mathbb{P}(n)(e(X)\leq b_n)=0$, take this sub-sub-sequence and the claim holds. If not, take a sub-sub-sequence with $\mathbb{P}(n)(e(X)\leq b_n)>0$. On this sub-sub-sequence, I have the bound:

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right] &\geq \mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e\in(b_n/2,b_n]\}}{\max\{e,b_n\}^2}\right] \\
&\geq \frac{1}{2b_n}\mathbb{P}(n)(e\in(b_n/2,b_n]) \\
&= \frac{1}{2b_n}\left(\mathbb{P}(n)(e\leq b_n)-\mathbb{P}(n)(e\leq b_n/2)\right) \\
&\geq \frac{\rho}{2b_n}\mathbb{P}(n)(e\leq b_n). \qquad \text{(Assumption 4(i))}
\end{aligned}$$

Then, applying Assumption 1(e),

$$r_{\mu,n}\frac{\mathbb{P}(n)(e(X)\leq b_n)}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}}\leq r_{\mu,n}\left(\frac{2b_n\mathbb{P}(n)(e\leq b_n)}{\rho}\right)^{1/2}\leq\sqrt{\frac{2C}{\rho}}r_{\mu,n}b_n^{\gamma_0/2}=o(n^{-1/2}).$$

Finally, I verify (c) assuming Assumption 4(ii) holds. I wish to show that if there is a sequence of $\mathbb{P}(n)\in\mathscr{P}$ and associated constants such that $r_{\mu,n}b_n^{\gamma_0/2}=o(n^{-1/2})$, then $r_{\mu,n}\frac{\mathbb{P}(n)(e(X)\leq b_n)}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}}=o(n^{-1/2})$. Under somewhat weak overlap ($\gamma_0>2$), then

$$r_{\mu,n}\frac{\mathbb{P}(n)(e(X)\leq b_n)}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}}=O\left(r_{\mu,n}b_n^{\gamma_0-1}\right)=O\left(r_{\mu,n}b_n^{2(\gamma_0-1)/\gamma_0+(\gamma_0^2+2-3\gamma_0)/\gamma_0}\right)=o(n^{-1/2}).$$

I therefore proceed for $\gamma_0\leq 2$. I will use the bound from Lemma 9:

$$r_{\mu,n}\frac{\mathbb{P}(n)(e(X)\leq b_n)}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}}\leq r_{\mu,n}\frac{\mathbb{P}(n)(e(X)\leq b_n)}{\sqrt{1+b_n^{-2}\mathbb{P}(n)(e(X)\leq b_n)^{\gamma_0/(\gamma_0-1)}}}=\text{“}RHS.\text{”}$$

Consider a sub-sequence of $\mathbb{P}(n) \in \mathscr{P}$ and constants. I will show that there is a further sub-sub-sequence for which this right-hand side RHS is $o(n^{-1/2})$. Suppose there is a sub-sub-sequence such that $b_n^{-2}\mathbb{P}(n)(e(X) \leq b_n)^{\gamma_0/(\gamma_0-1)} \to 0$. Then for that sub-sub-sequence, I have:

$$RHS \leq r_{\mu,n}\mathbb{P}(n)(e(X) \leq b_n) = r_{\mu,n} o\left(b_n^{2(\gamma_0-1)/\gamma_0}\right) = o(n^{-1/2}).$$

If not, then $b_n^2 \precsim \mathbb{P}(n)(e(X) \leq b_n)^{\gamma_0/(\gamma_0-1)}$ and I have the bound:

$$\begin{aligned}
RHS &\leq O\left(r_{\mu,n} b_n \mathbb{P}(n)(e(X) \leq b_n)^{1-\gamma_0/(2(\gamma_0-1))}\right) \\
&= O\left(r_{\mu,n} b_n \left(\mathbb{P}(n)(e(X) \leq b_n)^{(1/2-1/\gamma_0)*\gamma_0/(\gamma_0-1)}\right)\right) \\
&= O\left(r_{\mu,n} b_n \left(\mathbb{P}(n)(e(X) \leq b_n)^{\gamma_0/(\gamma_0-1)}\right)^{(\gamma_0-2)/(2\gamma_0)}\right) \\
&= O\left(r_{\mu,n} b_n \left(b_n^2\right)^{(\gamma_0-2)/(2\gamma_0)}\right) \\
&= O\left(r_{\mu,n} b_n^{2(\gamma_0-1)/\gamma_0}\right) = o(n^{-1/2}).
\end{aligned}$$

Therefore Assumption 3'(c) holds. □

Lemma 10. *Suppose the requirements of Theorem 1' hold. Then by implication,* $r_{\mu,n}\mathbb{P}(n)(e(X) \leq b_n) \ll n^{-1/2}\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}$.

Proof of Lemma 10.

$$\begin{aligned}
r_{\mu,n}\mathbb{P}(n)(e(X) \leq b_n) &= \left(r_{\mu,n}\frac{\mathbb{P}(n)(e(X) \leq b_n)}{\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}}\right)\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]} \\
&\ll n^{-1/2}\sqrt{\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}. \qquad \text{(A3'(c))}
\end{aligned}$$

□

Lemma 11 (Oracle consistency)**.** *If* $n^{-1/2} \ll b_n \ll 1$*, then* $|P_n[\phi_n] - \mathbb{E}_{\mathbb{P}(n)}[\mu(x)]| \xrightarrow{\mathscr{P}} 0$.

Proof of Lemma 11. Let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$. For any $t > 0$, I have:

$$\begin{aligned}
\mathbb{P}(n)\left(|\mathbb{P}(n)_n[\phi_n] - \mathbb{E}_{\mathbb{P}(n)}[\mu(x)]| > t\right) &\leq \frac{\mathbb{E}[|\phi_n - \mathbb{E}_{\mathbb{P}(n)}[\mu(x)]|^2]}{nt^2} && \text{(Chebyshev's inequality)} \\
&\leq \frac{\mathbb{E}[|\phi_n - \mathbb{E}_{\mathbb{P}(n)}[\phi_n]|^q]^{2/q}}{nt^2} && \text{(Jensen's inequality)} \\
&\leq \frac{[(4M)^q\mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]]^{2/q}}{t^2 n b_n^{2(q-2)/q}} && \text{(Lemma 2.(iii))}
\end{aligned}$$

$$\leq \frac{(4M)^2}{t^2}\frac{1}{nb_n^2}.$$

This upper bound tends to zero and holds simultaneously for all $P \in \mathscr{P}$. Hence, $|\mathbb{P}(n)_n[\phi_n] - \mathbb{E}_{\mathbb{P}(n)}[\mu(x)]| = o_{\mathbb{P}(n)}(1)$. □

Lemma 12 (Oracle variance consistency). *Let $\sigma_n^2 = n^{-1}(P_n[\phi_n^2] - P_n[\phi_n]^2)$ be the oracle sample variance. If $n^{-1/2} \ll b_n \ll 1$, then $n\sigma_n^2/\mathrm{Var}_{\mathbb{P}(n)}(\phi_n) \xrightarrow{\mathscr{P}} 1$.*

Proof of Lemma 12. Let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$.

First, I argue that for any $q > 2$:

$$\begin{aligned}
\mathbb{P}_P\left(\left|\frac{\mathbb{P}(n)_n[\phi_n^2] - P[\phi_n^2]}{\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)}\right| > t\right) &\leq \frac{\mathbb{E}\{|\mathbb{P}(n)_n[\phi_n^2] - P[\phi_n^2]|^{q/2}\}}{t^{q/2} Var_{\mathbb{P}(n)}(\phi_n)^{q/2}} && \text{(Markov inequality)}\\
&\leq \frac{2}{t^{q/2} n^{q/2-1}} \frac{\mathbb{E}\{|\phi_n^2 - P[\phi_n^2]|^{q/2}\}}{Var_{\mathbb{P}(n)}(\phi_n)^{q/2}} && \text{(von Bahr-Esseen inequality)}\\
&\leq \frac{2^{q/2+1}}{t^{q/2} n^{q/2-1}} \frac{\mathbb{E}[|\phi_n|^q]}{Var_{\mathbb{P}(n)}(\phi_n)^{q/2}} && \text{(Jensen's inequality)}\\
&\leq \frac{2^{q/2+1}}{t^{q/2} n^{q/2-1}} \frac{(8M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]}{b_n^{q-2}(\mathrm{Var}_{\mathbb{P}(n)}(\phi_n))^{q/2}} && \text{(Lemma 2.(iv))}\\
&\leq \frac{2^{q/2+1}}{t^{q/2} n^{q/2-1}} \frac{(8M)^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]}{b_n^{q-2}\sigma_{\min}^q \mathbb{E}[e(X)/\{e(X) \vee b_n\}^2]^{q/2}} && \text{(Lemma 4)}\\
&\leq \frac{(8M)^q 2^{q/2+1}}{t^{q/2}\sigma_{\min}^q(\pi_{\min}/2)^{q/2-1}} \frac{1}{n^{q/2-1} b_n^{q-2}}. && \text{(Lemma 4)}
\end{aligned}$$

Since $b_n \gg n^{-1/2}$, $n^{q/2-1} b_n^{q-2} \to \infty$, so that $\left|\frac{\mathbb{P}(n)_n[\phi_n^2] - P[\phi_n^2]}{\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)}\right| = o_{\mathbb{P}(n)}(1)$.

Then, by the triangle inequality:

$$\begin{aligned}
|n\sigma_n^2/\mathrm{Var}_{\mathbb{P}(n)}(\phi_n) - 1| &\leq \left|\frac{\mathbb{P}(n)_n[\phi_n]^2 - P[\phi_n]^2}{\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)}\right| + \left|\frac{\mathbb{P}(n)_n[\phi_n^2] - P[\phi_n^2]}{\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)}\right|\\
&\leq |\mathbb{P}(n)_n[\phi_n] + P[\phi_n]| \times |\mathbb{P}(n)_n[\phi_n] - P[\phi_n]|/(\sigma_{\min}^2 \pi_{\min}/2) + o_{\mathbb{P}(n)}(1)\\
& && \text{(Lemma 4 + above)}\\
&\leq (2M + o_{\mathbb{P}(n)}(1)) o_{\mathbb{P}(n)}(1) O_{\mathbb{P}(n)}(1) + o_{\mathbb{P}(n)}(1) && (P[\phi_n] \leq M + \text{Lemma 11})\\
&= o_{\mathbb{P}(n)}(1),
\end{aligned}$$

where σ_n^2 is the oracle sample variance. Therefore, this upper bound tends to zero uniformly over $\mathscr{P}$. □

Lemma 13 (Orthogonalized inverse propensities). *Suppose the conditions of Proposition 1 hold, $r_{e,n} \ll b_n$,*

and $\mathbb{P}(n)$ is a sequence of distributions in $\mathscr{P}$. Then

$$\mathbb{E}_{\mathbb{P}(n)}\left[\left(\frac{D}{\max\{\hat{e},b_n\}}-\frac{D}{\max\{e,b_n\}}\right)^2\right]=o_{\mathbb{P}(n)}\left(\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right).$$

Proof of Lemma 13. Write $(I)=\mathbb{E}_{\mathbb{P}(n)}\left[\left(\frac{D}{\max\{\hat{e},b_n\}}-\frac{D}{\max\{e,b_n\}}\right)^2\right]$. Since $r_{e,n}\ll b_n$, let n be large enough that $b_n\geq 2r_{e,n}$, so that $e\geq b_n+r_{e,n}$ implies $e-r_{e,n}\geq e/2$. Then the squared error has the following decomposition:

$$\begin{aligned}
(I)&=\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{\hat{e},b_n\}^2}-\frac{D}{\max\{e,b_n\}^2}\right]-2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}}\left(\frac{D}{\max\{\hat{e},b_n\}}-\frac{D}{\max\{e,b_n\}}\right)\right]\\
&=\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\frac{\max\{e,b_n\}^2-\max\{\hat{e},b_n\}^2}{\max\{\hat{e},b_n\}^2}\right]-2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{\max\{e,b_n\}-\max\{\hat{e},b_n\}}{\max\{\hat{e},b_n\}}\right)\right]\\
&=\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\frac{\max\{e,b_n\}^2-\max\{\hat{e},b_n\}^2}{\max\{\hat{e},b_n\}^2}\right]+2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{\max\{\hat{e},b_n\}-\max\{e,b_n\}}{\max\{\hat{e},b_n\}}\right)\right]\\
&\leq\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{\hat{e}\leq e\}}{\max\{e,b_n\}^2}\frac{\max\{e,b_n\}^2-\max\{\hat{e},b_n\}^2}{\max\{\hat{e},b_n\}^2}\right]+2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{r_{e,n}}{\max\{e+r_{e,n},b_n\}}\right)\right]\\
&\leq\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\frac{\max\{e,b_n\}^2-\max\{e-r_{e,n},b_n\}^2}{\max\{e-r_{e,n},b_n\}^2}\right]+2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{r_{e,n}}{b_n}\right)\right]\\
&\leq\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e\in[b_n,b_n+r_{e,n})\}}{\max\{e,b_n\}^2}\frac{(b_n+r_{e,n})^2-b_n^2}{b_n^2}\right]+\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e\in[b_n+r_{e,n},1]\}}{\max\{e,b_n\}^2}\frac{e^2-(e-r_{e,n})^2}{(e-r_{e,n})^2}\right]\\
&\quad+2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{r_{e,n}}{b_n}\right)\right]\\
&\leq\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e\in[b_n,b_n+r_{e,n})\}}{\max\{e,b_n\}^2}\frac{2b_nr_{e,n}+r_{e,n}^2}{b_n^2}\right]+\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e\in[b_n+r_{e,n},1]\}}{\max\{e,b_n\}^2}\frac{2er_{e,n}}{(e-r_{e,n})^2}\right]\\
&\quad+2\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{r_{e,n}}{b_n}\right)\right]\\
&\leq\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{4b_nr_{e,n}+r_{e,n}^2}{b_n^2}\right)\right]+4\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D\mathbf{1}\{e\in[b_n+r_{e,n},1]\}}{\max\{e,b_n\}^2}\frac{2r_{e,n}}{e}\right]\\
&\leq\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\left(\frac{12b_nr_{e,n}+r_{e,n}^2}{b_n^2}\right)\right].
\end{aligned}$$

Since $r_{e,n}=o(b_n)$ by assumption, this upper bound is $o\left(\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right)$. □

Lemma 14. *For all $P\in\mathscr{P}$, $\mathbb{E}_P[D/\max\{e(X),b_n\}^2]\geq 1-\left(b_n^2\frac{\gamma_0-1}{c(\gamma_0)}\right)^{\gamma_0-1}\gamma_0^{-\gamma_0}$.*

Proof of Lemma 14.

$$\begin{aligned}
\mathbb{E}_P[D/\max\{e(X),b_n\}^2]&=\mathbb{E}_P[D\mathbf{1}\{e(X)\leq b_n\}/\max\{e(X),b_n\}^2]+\mathbb{E}_P[D\mathbf{1}\{e(X)>b_n\}/\max\{e(X),b_n\}^2]\\
&\geq b_n^{-2}\mathbb{E}_P[D\mathbf{1}\{e(X)\leq b_n\}]+(1-P(e(X)\leq b_n))\\
&\geq b_n^{-2}c(\gamma_0)P(e(X)\leq b_n)^{\gamma_0/(\gamma_0-1)}+(1-P(e(X)\leq b_n)) \qquad \text{(Lemma 3)}
\end{aligned}$$

$$= 1 + P(e(X) \leq b_n) \left(c(\gamma_0) b_n^{-2} P(e(X) \leq b_n)^{1/(\gamma_0 - 1)} - 1 \right).$$

This term is minimized over $P(e(X) \leq b_n)$ at $P(e(X) \leq b_n) = \left(\frac{\gamma_0 - 1}{c(\gamma_0) b_n^{-2} \gamma_0} \right)^{\gamma_0 - 1}$, which produces

$$\mathbb{E}_P[D / \max\{e(X), b_n\}^2] = 1 - \frac{P(e(X) \leq b_n)}{\gamma_0} = 1 - \left(b_n^2 \frac{\gamma_0 - 1}{c(\gamma_0)} \right)^{\gamma_0 - 1} \gamma_0^{-\gamma_0}.$$

□

Lemma 15. *Suppose the conditions of Lemma 13 hold and $r_{\mu,n} \to 0$. Let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$. Recall the definitions of ϕ_n as the oracle clipped influence function and $\hat{\phi}_n$ the estimated influence function. Then $\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2] = Var_{\mathbb{P}(n)}(\phi_n) + O(1)$ and $\mathbb{P}(n)_n \left[\hat{\phi}_n^2 - \phi_n^2 \right] = o_{\mathbb{P}(n)} \left(Var_{\mathbb{P}(n)}(\phi_n) \right)$.*

Proof of Lemma 15. First, note that:

$$\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2] = Var_{\mathbb{P}(n)}(\phi_n) + \mathbb{E}_{\mathbb{P}(n)}[\phi_n]^2 = Var_{\mathbb{P}(n)}(\phi_n) + O(1)^2. \qquad \text{(Assumption 1(a))}$$

Next, I show that $\mathbb{P}(n)_n \left[\hat{\phi}_n^2 - \phi_n^2 \right] = o_{\mathbb{P}(n)} \left(\mathbb{E}_{\mathbb{P}(n)} \left[D / \max\{e(X), b_n\}^2 \right] \right)$. I have:

$$\begin{aligned}
\left| \mathbb{P}(n)_n \left[\hat{\phi}_n^2 - \phi_n^2 \right] \right| &= \mathbb{P}(n)_n [(\hat{\phi}_n - \phi_n)^2] + \left| \mathbb{P}(n)_n \left[2(\hat{\phi}_n - \phi_n)\phi_n \right] \right| \\
&\leq \mathbb{P}(n)_n [(\hat{\phi}_n - \phi_n)^2] + 2 \sqrt{\mathbb{P}(n)_n \left[(\hat{\phi}_n - \phi_n)^2 \right] \mathbb{P}(n)_n [\phi_n^2]} && \text{(Cauchy-Schwarz)} \\
&= \mathbb{P}(n)_n [(\hat{\phi}_n - \phi_n)^2] + \sqrt{\mathbb{P}(n)_n \left[(\hat{\phi}_n - \phi_n)^2 \right] O_{\mathbb{P}(n)} \left(\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2] \right)} && \text{(Lemma 12)} \\
&= \mathbb{E}_{\mathbb{P}(n)} [(\hat{\phi}_n - \phi_n)^2] + \sqrt{\mathbb{E}_{\mathbb{P}(n)} \left[(\hat{\phi}_n - \phi_n)^2 \right] O_{\mathbb{P}(n)} \left(\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2] \right)} \\
&\quad + o_{\mathbb{P}(n)} \left(\sqrt{\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2]} \right). && \text{(SLLN)}
\end{aligned}$$

I decompose the nuisance error as:

$$\begin{aligned}
(\hat{\phi}_n - \phi_n)^2 &\leq \left(|\hat{\mu} - \mu| \left| 1 - \frac{D}{\max\{e, b_n\}} \right| + |Y - \hat{\mu}| \left| \frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}} \right| \right)^2 && \text{(Tri. Ineq.)} \\
&\leq 4 \left(r_{\mu,n}^2 \left(1 + \frac{D}{\max\{e, b_n\}^2} \right) + \left((Y - \mu)^2 + o_{\mathbb{P}(n)}(r_{\mu,n}) \right) \left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}} \right)^2 \right) \\
&\precsim r_{\mu,n}^2 \frac{D}{\max\{e, b_n\}^2} + \left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}} \right)^2 + o_{\mathbb{P}(n)}(1) \\
&= o_{\mathbb{P}(n)}(1) \frac{D}{\max\{e, b_n\}^2} \mathbb{E}_{\mathbb{P}(n)} \left[\frac{D}{\max\{e, b_n\}^2} \right] + \left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}} \right)^2 + o_{\mathbb{P}(n)}(1) \\
&&& (r_{\mu,n} \to 0, \text{ Lemma 14})
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}(n)}\left[(\hat{\phi}_n-\phi_n)^2\right] &= o_{\mathbb{P}(n)}\left(\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right)\\
&\quad+\mathbb{E}_{\mathbb{P}(n)}\left[\left(\frac{D}{\max\{\hat{e},b_n\}}-\frac{D}{\max\{e,b_n\}}\right)^2\right]+o_{\mathbb{P}(n)}(1)\\
&= o_{\mathbb{P}(n)}\left(\mathbb{E}_{\mathbb{P}(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right)+o_{\mathbb{P}(n)}(1). && \text{(Lemma 13)}\\
&= o_{\mathbb{P}(n)}\left(Var_{\mathbb{P}(n)}(\phi_n)\right)+o_{\mathbb{P}(n)}(1). && \text{(Lemma 4)}
\end{aligned}$$

As a result:

$$\begin{aligned}
\left|\mathbb{P}(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right| &= \sqrt{o_{\mathbb{P}(n)}\left(Var_{\mathbb{P}(n)}(\phi_n)\right)O_{\mathbb{P}(n)}\left(\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2]\right)}+o_{\mathbb{P}(n)}\left(Var_{\mathbb{P}(n)}(\phi_n)\right)+o_{\mathbb{P}(n)}\left(\sqrt{\mathbb{E}_{\mathbb{P}(n)}[\phi_n^2]}\right)\\
&= o_{\mathbb{P}(n)}\left(Var_{\mathbb{P}(n)}(\phi_n)+\sqrt{Var_{\mathbb{P}(n)}(\phi_n)}\right)=o_{\mathbb{P}(n)}\left(Var_{\mathbb{P}(n)}(\phi_n)\right). && \text{(Lemma 14)}
\end{aligned}$$

□

Lemma 16 (Estimated variance consistency)**.** *Suppose the assumptions of Proposition 2 and Lemma 15 hold. Let $\mathbb{P}(n)$ be a sequence of distributions in $\mathscr{P}$. Then $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{\mathscr{P}} 1$.*

Proof of Lemma 16. Recall the definition $\bar{\sigma}_n^2 = n^{-1}\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)$.

Let $\sigma_n^2 = n^{-1}(\mathbb{P}(n)_n[\phi_n^2]-\mathbb{P}(n)_n[\phi_n]^2)$ be the oracle sample variance. By Lemma 12, $\sigma_n^2/\bar{\sigma}_n^2 \xrightarrow{\mathscr{P}} 1$. Therefore it suffices to show that $(\hat{\sigma}_n^2-\sigma_n^2)/\bar{\sigma}_n^2 = \frac{\mathbb{P}(n)_n[\hat{\phi}_n^2]-\mathbb{P}(n)_n[\phi_n^2]-\hat{\psi}_{clip}^{AIPW}(b_n)^2+\tilde{\psi}_{(Orcl)}^{AIPW}(b_n)^2}{Var_{\mathbb{P}(n)}(\phi_n)} \xrightarrow{\mathscr{P}} 0$.

Note that the assumptions of Proposition 1 hold, because $r_{\mu,n}\frac{r_{e,n}+b_n}{b_n} = r_{\mu,n}O(1) = o(1)$.

Note that by Proposition 2, $\mathbb{E}_{\mathbb{P}(n)}\left[D/\max\{e(X),b_n\}^2\right] = \Theta(n\bar{\sigma}_n^2) = \Theta_{\mathbb{P}(n)}\left(\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)\right)$.

By the triangle inequality:

$$\begin{aligned}
\left|\frac{\hat{\sigma}_n^2-\sigma_n^2}{\bar{\sigma}_n^2}\right| &\le \left|\frac{\mathbb{P}(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]}{Var_{\mathbb{P}(n)}(\phi_n)}\right|+\left|\frac{\hat{\psi}_{clip}^{AIPW}(b_n)^2-\tilde{\psi}_{(Orcl)}^{AIPW}(b_n)^2}{Var_{\mathbb{P}(n)}(\phi_n)}\right|\\
&\precsim \left|\mathbb{P}(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right|O\left(\frac{1}{\mathbb{E}_{\mathbb{P}(n)}\left[D/\max\{e(X),b_n\}^2\right]}\right) && \text{(Lemma 5)}\\
&\quad+O_{\mathbb{P}(n)}\left(\left|\hat{\psi}_{clip}^{AIPW}(b_n)-\tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right|\right) && \text{(Lemma 4)}\\
&= \left|\mathbb{P}(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right|O_{\mathbb{P}(n)}\left(\frac{1}{\mathbb{E}_{\mathbb{P}(n)}\left[D/\max\{e(X),b_n\}^2\right]}\right)+o_{\mathbb{P}(n)}(1) && \text{(Proposition 1)}\\
&= \left|\mathbb{P}(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right|O_{\mathbb{P}(n)}\left(1/Var_{\mathbb{P}(n)}(\phi_n)\right)+o_{\mathbb{P}(n)}(1) && \text{(Proposition 2)}\\
&= o_{\mathbb{P}(n)}(Var_{\mathbb{P}(n)}(\phi_n))O_{\mathbb{P}(n)}\left(1/Var_{\mathbb{P}(n)}(\phi_n)\right)+o_{\mathbb{P}(n)}(1) && \text{(Lemma 15)}\\
&= o_{\mathbb{P}(n)}(1).
\end{aligned}$$

Therefore $(\hat{\sigma}_n^2 - \sigma_n^2)/\bar{\sigma}_n^2 \to_{\mathbb{P}(n)} 0$. By Lemma 12, $\sigma_n^2/\bar{\sigma}_n^2 \to_{\mathbb{P}(n)} 1$. As a result, $(\hat{\sigma}_n^2 - \bar{\sigma}_n^2)/\bar{\sigma}_n^2 \to_{\mathbb{P}(n)} 0$ and $\hat{\sigma}_n^2/\bar{\sigma}_n^2 \xrightarrow{\mathscr{P}} 1$. □

Lemma 17. *Suppose the conditions of Theorem 1' hold. Then* $\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right) = o_{\mathbb{P}(n)}(1)$.

Proof of Lemma 17. I write $k(i)$ for observation i's fold and n_k for the number of observations in fold k. Then the oracle and clipped AIPW estimators are:

$$\tilde{\psi}_{(Orcl)}^{AIPW}(b_n) = \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i \mid b_n, \eta) = \sum_k \frac{n_k}{n}\overbrace{\frac{1}{n_k}\sum_{i:k(i)=k}\phi(Z_i \mid b_n, \eta)}^{\text{“}\tilde{\psi}_{clip}^{AIPW,(k)}(b_n)\text{”}}$$

$$\hat{\psi}_{clip}^{AIPW}(b_n) = \frac{1}{n}\sum_{i=1}^{n}\phi(Z_i \mid b_n, \hat{\eta}^{(-k)}) = \sum_k \frac{n_k}{n}\underbrace{\frac{1}{n_k}\sum_{i:k(i)=k}\phi(Z_i \mid b_n, \hat{\eta}^{(-k)})}_{\text{“}\hat{\psi}_{clip}^{AIPW,(k)}(b_n)\text{”}}.$$

I write $\hat{r}_k \equiv \sigma_n^{-1}\left(\tilde{\psi}_{clip}^{AIPW,(k)}(b_n) - \hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right)$. I wish to show that $\sum_k \frac{n_k}{n}\hat{r}_k = o_{\mathbb{P}(n)}(1)$. I consider an arbitrary k and quantify the bias and variance of $\hat{r}_k$ given the data and nuisance estimates from the other folds $-k$.

I write the standard decomposition:

$$\hat{r}_k = \sigma_n^{-1}P_n\left[(\hat{\mu} - \mu)\left(1 - \frac{D}{\max\{\hat{e}, b_n\}}\right) + (\mu - Y)\left(\frac{D}{\max\{e, b_n\}} - \frac{D}{\max\{\hat{e}, b_n\}}\right)\right].$$

By cross-fitting, the bias satisfies $\mathbb{E}[\hat{r}_k \mid \hat{\eta}^{(-k)}] = \sigma_n^{-1}\mathbb{E}\left[(\hat{\mu} - \mu)\frac{\max\{\hat{e}, b_n\} - e}{\max\{\hat{e}, b_n\}}\right]$. I now bound this term.

$$\begin{aligned}
\left|\mathbb{E}[\hat{r}_k \mid \hat{e}^{(-k)}, \hat{\mu}^{(-k)}]\right| &\le \sigma_n^{-1}r_{\mu,n}\mathbb{P}(n)(e(X) \le b_n + r_{e,n}) \\
&\quad + \sigma_n^{-1}r_{\mu,n}r_{e,n}\mathbb{E}_{\mathbb{P}(n)}\left[\frac{1}{e - r_{e,n}}\mathbf{1}\{e > b_n + r_{e,n}\}\right] && \text{(Lemma 7)} \\
&\le o_{\mathbb{P}(n)(1)} + n^{1/2}r_{\mu,n}r_{e,n}\frac{\mathbb{E}\left[D/\max\{e, b_n\}^2\right]}{\sqrt{\mathbb{E}\left[D/\max\{e, b_n\}^2\right]}} && \text{(Lemma 10 + Proposition 2)} \\
&= o_{\mathbb{P}(n)}(1). && \text{(Assumption 3'(b))}
\end{aligned}$$

Next, I show that $Var_{\mathbb{P}(n)}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right) = o_{\mathbb{P}(n)}(\bar{\sigma}_n^2)$, where $\bar{\sigma}_n^2 = n^{-1}\mathrm{Var}_{\mathbb{P}(n)}(\phi_n)$. . Consider the estimates in one fold k, with the nuisances from other folds fixed. For the estimates in that fold:

$$Var_{\mathbb{P}(n)}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right) = Var\left(\mathbb{P}(n)_n\hat{\phi} - \phi\right)$$

$$
\begin{aligned}
&= n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[(\hat{\phi}-\phi)^2 - \mathbb{E}_{\mathbb{P}(n)}E[\hat{\phi}-\phi]^2\right] \\
&= n^{-1}\mathbb{E}_{\mathbb{P}(n)}\left[(\hat{\phi}-\phi)^2 - o_{\mathbb{P}(n)}(1)\right] && \text{(Bias)} \\
&= n^{-1}o_{\mathbb{P}(n)}\left(Var_{\mathbb{P}(n)}(\phi_n)\right) + o_{\mathbb{P}(n)}(1) && \text{(Lemma 15)} \\
&= n^{-1}o_{\mathbb{P}(n)}\left(\mathbb{E}_{\mathbb{P}(n)}[D/\max\{e(X), b_n\}^2]^2\right) + o_{\mathbb{P}(n)}(1) && \text{(Lemma 2.(iii))} \\
&= o_{\mathbb{P}(n)}(\bar{\sigma}_n^2 + 1) && \text{(Lemma 5)} \\
&= o_{\mathbb{P}(n)}\left(\frac{E_{\mathbb{P}(n)}[D/\max\{e(X), b_n\}^2]}{n} + 1\right) && \text{(Proposition 2)} \\
&= o_{\mathbb{P}(n)}(\bar{\sigma}_n^2). && (b_n \gg n^{-1/2})
\end{aligned}
$$

As a result:

$$
E\left[\hat{r}_k^2 \mid \hat{\eta}^{(-k)}\right] = E\left[\hat{r}_k \mid \hat{\eta}^{(-k)}\right]^2 + Var\left(\hat{r}_k \mid \hat{\eta}^{(-k)}\right) = o_{\mathbb{P}(n)}(1)
$$

$$
\hat{r}_k = o_{\mathbb{P}(n)}(1)
$$

$$
\left|\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right) = o_{\mathbb{P}(n)}(1)\right| \leq \sum_k \frac{n_k}{n}|\hat{r}_k| = \sum_k \frac{n_k}{n}o_{\mathbb{P}(n)}(1) = o_{\mathbb{P}(n)}(1).
$$

Finally, I am ready to prove the central claims of this work. □

Proof of Theorem 1'. By Lemma 17, $\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right) = o_{\mathbb{P}(n)}(1)$. Therefore, by Proposition 4, $\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n\right) = \sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Orcl)}^{AIPW}(b_n)\right) + \sigma_n^{-1}\left(\tilde{\psi}_{(Orcl)}^{AIPW}(b_n) - \psi_n\right) \overset{\mathbb{P}(n)}{\rightsquigarrow} N(0,1)$. □

Proof of Theorem 1. For either claim, let $\mathbb{P}(n)$ be a sequence of distributions P in the relevant set. Note that in either case, the assumptions of Theorem 1' hold by Corollary 4. Therefore, by Theorem 1',

$$
\sup_{t\in\mathbb{R}}\left|\mathbb{P}(n)_n\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n}{\sigma_n} \leq t\right) - \Phi(t)\right| \to 0,
$$

where $\mathbb{P}(n)_n$ denotes the empirical average under distribution $\mathbb{P}(n)$ and σ_n is defined in Proposition 4.

Now I expand the empirical t-statistics for any fixed t:

$$
\begin{aligned}
\mathbb{P}(n)_n\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n}{\hat{\sigma}_n} \leq t\right) &= \mathbb{P}(n)_n\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n}{\sigma_n}\left(\frac{\sigma_n}{\hat{\sigma}_n}\right) \leq t\right) \\
&= \mathbb{P}(n)_n\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n}{\sigma_n} \leq t\frac{\hat{\sigma}_n}{\sigma_n}\right) \\
&= \mathbb{P}(n)_n\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n}{\sigma_n} - t\frac{\hat{\sigma}_n}{\sigma_n} \leq 0\right)
\end{aligned}
$$

$$\to \Phi(t),$$

with the final result holding by Slutsky's theorem.

Therefore,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}(n)_n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi_n}{\hat{\sigma}_n}\leq t\right)-\Phi(t)\right|\to 0,$$

by properties of a cumulative distribution function. □

Proof of Corollary 2. For simplicity of exposition, I prove the result for the class $\mathscr{P}$ under Assumption 4(ii):

$$\begin{aligned}
\limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\left|P(\psi(P)\in\hat{\mathcal{C}}_n)-(1-\alpha)\right| &= \limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\left|P^n\left(\frac{\psi(P)-\hat{\psi}^{AIPW}_{clip}(b_n)}{\hat{\sigma}_n}\in\left[z_{\alpha/2},z_{1-\alpha/2}\right]\right)-(1-\alpha)\right| \\
&= \limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\left|\begin{array}{c}\left(P^n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)}{\hat{\sigma}_n}>z_{1-\alpha/2}\right)-\alpha/2\right) \\ -\left(P^n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)}{\hat{\sigma}_n}>z_{\alpha/2}\right)-(1-\alpha/2)\right)\end{array}\right| \\
&= \limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\left|\begin{array}{c}\left(P^n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)}{\hat{\sigma}_n}<z_{1-\alpha/2}\right)-(1-\alpha/2)\right) \\ -\left(P^n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)}{\hat{\sigma}_n}<z_{\alpha/2}\right)-\alpha/2\right)\end{array}\right| \\
&\leq \limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\left|P^n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)}{\hat{\sigma}_n}<z_{1-\alpha/2}\right)-\Phi(z_{1-\alpha/2})\right| \\
&\quad+\limsup_{n\to\infty}\sup_{P\in\mathscr{P}}\left|P^n\left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)}{\hat{\sigma}_n}<z_{\alpha/2}\right)-\Phi(z_{\alpha/2})\right| \\
&= 0. \qquad \text{(Theorem 1)}
\end{aligned}$$

□

B.5 Rates and Rate Requirements

Proof of Example 1. For simplicity, I proceed assuming $\liminf_{n\to\infty} n^{1/2}r_{e,n}>0$.

By Theorem 1, it remains to show that there is a $b_n\to 0$ such that:

3(b) For some $\eta'>0$, $r_{\mu,n}r_{e,n}\left(1+b_n^{(\gamma_0-2)/2}n^{\eta'}\right)\ll n^{-1/2}$.

3(c) $r_{\mu,n}b_n^{\gamma_0/2}\ll n^{-1/2}$.

3(d) $r_{e,n}\ll b_n$.

(i) Without loss of generality, assume $\eta < 1/2$ is small enough. Represent the unclipped estimator as $b_n = n^{-\eta/2}$, take η' arbitrary, and let γ_0 be arbitrarily large. The remaining claims hold by inspection.

(ii) Because $\gamma_0 > 2$ and $r_{\mu,n} r_{e,n} \ll n^{-1/2}$, there is some $\eta'' > 0$ such that $r_{\mu,n}(r_{e,n} n^{2\eta''})^{\gamma_0/2} \ll n^{-1/2}$ as well. Choose some b_n such that $r_{e,n} \ll b_n \ll r_{e,n} n^{\eta''} \ll r_{e,n} n^{2\eta''}$. Choose some η' such that $n^{\eta'} \ll b_n^{(2-\gamma_0)/2}$. Then the remaining claims hold by inspection. □

Proof of Example 2. Take $b_n = n^{-1/4-\eta/2}$ and $\eta' = \eta$. The conditions listed in the proof of Example 1 hold by inspection. □

Proof of Example 3. If $\gamma_0 \geq 2$, the claim holds by Example 1 and Example 2.

Now suppose that $\gamma_0 < 2$. By construction, there is an $\eta'' \in \left(0, \frac{2-\gamma_0}{12\gamma_0}\right)$ such that $n^{-2/(3(2-\gamma_0))+4\eta''/(2-\gamma_0)} \ll n^{-1/2}$. Take this η'', and take $b_n = n^{-1/(3(2-\gamma_0))+4\eta''/(2-\gamma_0)}$. By construction, $r_{e,n} \ll b_n \ll 1$. Take $\eta' = \eta$ For such a b_n:

$$r_{\mu,n} r_{e,n} b_n^{(\gamma_0-2)/2} n^{\eta'} \ll n^{-2/3} n^{1/6} n^{-2\eta''} n^{\eta'} \ll n^{-1/2},$$

and $r_{\mu,n} b_n^{\gamma_0/2} \ll n^{\frac{-1}{3}+\frac{-\gamma_0}{3(2-\gamma_0)}+\eta\frac{2\gamma_0}{2-\gamma_0}} = n^{\frac{-2}{3(2-\gamma_0)}+\eta\frac{2\gamma_0}{2-\gamma_0}} \ll n^{-2/3+\eta\frac{2\gamma_0}{2-\gamma_0}} \ll n^{-2/3+1/6} = n^{-1/2}$. The conditions listed in the proof of Example 1 hold by inspection. □

Proof of Example 4. Without loss of generality, suppose $\eta < 1/2$. I verify the conditions listed in the proof of Example 1 in both cases as follows.

(i) $r_{\mu,n} = O(n^{-1/2})$, take $b_n = n^{-\eta/2}$. Note that $1 \gg b_n \gg r_{e,n}$. Then, taking $\eta' = \eta/2$, I obtain

$$r_{\mu,n} r_{e,n} b_n^{(\gamma_0-2)/2} n^{\eta/2} \ll r_{\mu,n} n^{-\eta} n^{\eta/4} n^{\eta/2} \ll n^{-1/2},$$

and $r_{\mu,n} b_n^{\gamma_0/2} = r_{\mu,n} o(1) = o(n^{-1/2})$.

(ii) Take $b_n = n^{-1/2} \log(n)$. Note that $1 \gg b_n \gg r_{e,n}$. Then, taking $\eta' = \eta$, I obtain:

$$r_{\mu,n} r_{e,n} b_n^{(\gamma_0-2)/2} n^{\eta} = \log(n)^{(\gamma_0-2)/2} r_{\mu,n} r_{e,n} n^{(2-\gamma_0)/4} n^{\eta} \ll r_{e,n} = O(n^{-1/2}),$$

and $r_{\mu,n} b_n^{\gamma_0/2} \ll n^{(\gamma_0-2)/4} n^{-\gamma_0/4} = n^{-1/2}$. □

B.6 Rules of Thumb

Proof of Lemma 1. First, I show that there is at least one such solution.

Recall the equation:

$$f_n(b) = \frac{b\frac{1}{n}\sum 1\{\hat{e}(X) \leq b\}}{\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}} + b^2\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}} - n^{-1/2}.$$

When $b = 0$, $f_n(b)$ is well-defined: $\sum D/\bar{e}$ is finite, so $\sup D/\bar{e}^2$ is finite. Because the first two terms of $f_n(b)$ include multiplication by b, $f_n(0) = 0$.

When $b = 1$:

$$\begin{aligned} f_n(1) &= \left(\sqrt{\frac{1}{n}\sum D}\right)^{-1} + \sqrt{\frac{1}{n}\sum D} - n^{-1/2} \\ &> \left(\sqrt{\frac{1}{n}\sum D}\right)^{-1} - 1 \geq 0. \end{aligned}$$

The final line holds because $\frac{1}{n}\sum D \in (0, 1]$ by assumption.

Define $b_n^- = \sup b \leq 1 \mid f_n(b) \leq 0$. Define $b_n^+ = \inf b \geq b_n^- \mid f_n(b) \geq 0$. Because $f_n(0) \leq 0 \leq f_n(1)$, both of these values are well-defined. Therefore, for every b satisfying $b_n^- < b < b_n^+$, it is the case that $f_n(b)$ is a well-defined real number that satisfies both $f_n(b) > 0$ and $f_n(b) < 0$. No such number exists, so it must be that $b_n^- = b_n^+$. Define b_n to be that value.

Next, I show that there is a unique solution. In particular, I show that $\hat{g}_n(b) \equiv \frac{b\frac{1}{n}\sum 1\{\hat{e}(X)\leq b\}}{\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}} + b^2\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}$ is a strictly increasing function of b for $b \geq \min_i \hat{e}_i$. As b increases, the first term's numerator strictly increases and the denominator weakly decreases. As a result, the first term strictly increases in that range. For $b < \min_i \hat{e}_i$, the first term is zero and as a result is weakly increasing. The second term can be rewritten as

$$\sqrt{\frac{1}{n}\sum D\min\{\hat{e}^{-2}b^4, b^2\}},$$

which is a strictly increasing function. As a result, $f_n(b)$ is a strictly increasing function in the desired range, so that there can be at most one solution. □

B.7 Limitations

Proof of Corollary 3. Take P to be the distribution that draws $e(X)$ from the CDF $P(e(X) \leq \pi) = \pi^{\gamma_0 - 1}$ and draws $Y \mid X, D \sim \mathcal{N}(0, 1)$. let $\mathscr{P} = \{P\}$. Such a family confirms to the requirements of Assumption 1 by construction.

Because $r_{e,n} \ll b_n$, let n be large enough that $b_n > 2r_{e,n}$ and $(b_n - r_{e,n})^{\gamma_0 - 2} > 2$. Recall that $\gamma_0 < 2$ by

assumption, so that I can divide by $2-\gamma_0$.

Take the nuisance estimate for the sequence $P(n)=P$ as $\hat{\mu}(X)=\mu(X)-r_{\mu,n}$ and $\hat{e}(X)=e(X)+r_{e,n}$. The bias of the clipped estimator $\hat{\psi}^{AIPW}_{clip}(b_n)$ with n observations is:

$$\begin{aligned}
E_P\left[\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)\right] &= E\left[(\hat{\mu}(X)-\mu(X))\left(\frac{D}{\max\{\hat{e}(X),b_n\}}-1\right)\right] \\
&= r_{\mu,n}E\left[\left(1-\frac{D}{\max\{e(X)+r_{e,n},b_n\}}\right)\right] \\
&= r_{\mu,n}E\left[\mathbf{1}\{e(X)\leq b_n-r_{e,n}\}\frac{b_n-e(X)}{b_n}+\mathbf{1}\{e(X)>b_n-r_{e,n}\}\frac{r_{e,n}}{e(X)+r_{e,n}}\right] \\
&\geq r_{\mu,n}b_n^{-1}E\left[\mathbf{1}\{e(X)\leq b_n-r_{e,n}\}(b_n-e(X))\right] \\
&\geq r_{\mu,n}b_n^{-1}E\left[\mathbf{1}\{e(X)\leq b_n/2\}(b_n-e(X))\right] \\
&\geq \frac{r_{\mu,n}}{2}E\left[\mathbf{1}\{e(X)\leq b_n/2\}\right] \\
&= r_{\mu,n}b_n^{\gamma_0-1}2^{\gamma_0-2}.
\end{aligned}$$

It is convenient to write $B_n=\frac{E_P\left[\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)\right]}{\sigma_n}$ for this proof.

By the proof of Corollary 1, there is a $C^{-1}>0$ such that $\sigma_n\geq Cn^{-1/2}b_n^{\gamma_0/2-1}$ for all n large enough. For such n:

$$\begin{aligned}
B_n &= \frac{E_P\left[\hat{\psi}^{AIPW}_{clip}(b_n)-\psi(P)\right]}{\sigma_n} \\
&\geq C2^{\gamma_0-2}n^{1/2}r_{\mu,n}b_n^{\gamma_0-1}b_n^{1-\gamma_0/2} \\
&= C2^{\gamma_0-2}n^{1/2}r_{\mu,n}b_n^{\gamma_0/2} \\
&\geq C2^{\gamma_0-2}n^{1/2}r_{\mu,n}r_{e,n}^{\gamma_0/2}\to\infty.
\end{aligned}$$

Note also that:

$$\begin{aligned}
Var\left(\hat{\psi}^{AIPW}_{clip}(b_n)\right) &= \frac{1}{n}Var\left(\mu+\frac{D}{\max\{\hat{e},b_n\}}(Y-\mu)+r_{\mu,n}\mathbf{1}\{e\geq b_n,X\in\mathcal{X}^Q\}\}\left(1-\frac{D}{e+r_{e,n}}\right)\right) \\
&= \frac{1}{n}Var\left(\mu+\frac{D}{\max\{\hat{e},b_n\}}(Y-\mu)\right) \\
&\quad + r_{\mu,n}\frac{1}{n}Var\left(\mathbf{1}\{e\geq b_n,X\in\mathcal{X}^Q\}\}\left(1-\frac{D}{e+r_{e,n}}\right)\right) \\
&\leq \sigma_n^2+\frac{\epsilon r_{\mu,n}}{n}\left(Var_Q\left(\frac{r_{e,n}\mathbf{1}\{e\geq b_n\}}{e+r_{e,n}}\right)+E_Q\left[\frac{e(1-e)\mathbf{1}\{e\geq b_n\}}{(e+r_{e,n})^2}\right]\right) \\
&\leq \sigma_n^2+\frac{\epsilon r_{\mu,n}}{n}\left(E_Q\left[\frac{\mathbf{1}\{e\geq b_n\}\left(r_{e,n}^2+e\right)}{e^2}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sigma_n^2 + \frac{2\epsilon r_{\mu,n}}{n} E_Q\left[\mathbf{1}\{e \geq b_n\} e^{-1}\right] \\
&= \sigma_n^2 + \frac{2C\epsilon r_{\mu,n}(\gamma_0 - 1)}{n} \int_{b_n}^1 t^{\gamma_0 - 3} dt \\
&= \sigma_n^2 + \frac{2C\epsilon r_{\mu,n}(\gamma_0 - 1)}{n(2-\gamma_0)} \int_{b_n}^1 (b_n^{\gamma_0 - 2} - 1) \\
&\leq \sigma_n^2 + r_{\mu,n} \frac{2C\epsilon(\gamma_0 - 1)}{2 - \gamma_0} \frac{b_n^{\gamma_0 - 2}}{n} \\
&= \sigma_n^2 + o(\sigma_n^2). \qquad \text{(Proof of Corollary 1)}
\end{aligned}
$$

Next, I show that the conditions of Lemma 16 hold. The requirements are that the conditions of Proposition 1, $r_{e,n} \ll b_n$, and $r_{\mu,n} \to 0$. By assumption, $n^{-1/2} \ll r_{e,n} \ll b_n \ll 1$ and $r_{\mu,n} \to 0$. Therefore $r_{\mu,n} \frac{r_{e,n} + b_n}{b_n} \to 0$ and the conditions of Proposition 1 hold as well. Therefore, Lemma 16 applies, $\hat{\sigma}_n^2 / Var(\hat{\psi}_{clip}^{AIPW}(b_n)) \to^P 1$ and $E\left[\hat{\psi}_{clip}^{AIPW}(b_n) - \psi(P)\right] \to^P \infty$. As a result, $\frac{E_P[\hat{\psi}_{clip}^{AIPW}(b_n) - \psi(P)]}{\hat{\sigma}_n} \to^P \infty$ and for any fixed $\alpha > 0$,

$$
\begin{aligned}
P(\psi(P) \in \hat{\mathcal{C}}_n) &= P\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - \psi(P)}{\hat{\sigma}_n} \in [z_{\alpha/2}, z_{1-\alpha/2}]\right) \\
&= P\left(\frac{\hat{\psi}_{clip}^{AIPW}(b_n) - E\left[\hat{\psi}_{clip}^{AIPW}(b_n)\right]}{\sigma_n + o_P(\sigma_n)} \in \left[B_n + z_{\alpha/2} + o_P(1), B_n + z_{1-\alpha/2} + o_P(1)\right]\right) \\
&= P\left(\frac{O_P(\sigma_n)}{\sigma_n + o_P(\sigma_n)} \in \left[B_n + z_{\alpha/2} + o_P(1), B_n + z_{1-\alpha/2} + o_P(1)\right]\right) \to^P 0,
\end{aligned}
$$

with the limit holding because B_n tends to infinity. □

Proof of Proposition 3. First, I show that this rate of convergence is achievable for Nadaraya-Regression for any distribution $P \in P$ and an arbitrary $x_0 \in [-1,1]^d$. Let the lower density bound be $\underline{f}$. For simplicity, assume that $e(X)$ has a continuous density. Fix some such distribution $P \in \mathscr{P}$ and take a bandwidth $h_n = n^{-1/(2\beta_\mu + d + d/(\gamma_0 - 1))}$. Note that:

$$
\begin{aligned}
P(D = 1, \|X - x_0\| \leq h_n) &\geq P\left(D = 1, e(X) \leq F_{e(X)}^{-1}\left(P(\|X - x_0\|_\infty \leq h_n)\right)\right) \\
&\geq P\left(D = 1, e(X) \leq F_{e(X)}^{-1}\left(h_n^d\right)\right) \\
&\geq \int_0^{(C/\underline{f})^{-1/(\gamma_0 - 1)} h_n^{d/(\gamma_0 - 1)}} C(\gamma_0 - 1) t t^{\gamma_0 - 2} dt \\
&= \underline{f}^{\gamma_0/(\gamma_0 - 1)} C \frac{\gamma_0 - 1}{\gamma_0 - 2} \left(C^{-1/(\gamma_0 - 1)} h_n^{d/(\gamma_0 - 1)}\right)^{\gamma_0} \\
&= \underline{f}^{\gamma_0/(\gamma_0 - 1)} C^{-1/(\gamma_0 - 1)} \frac{\gamma_0 - 1}{\gamma_0 - 2} h_n^{d + d/(\gamma_0 - 1)}.
\end{aligned}
$$

Since $Var(Y \mid X, D) \leq M$, by standard arguments, the Nadaraya Watson variance conditional on the $\{X, D\}$ data is with high probability upper bounded by a term on the order of $n^{-1}h_n^{d+d/(\gamma_0-1)} = n^{-2\beta_\mu/(2\beta_\mu+d+d/(\gamma_0-1))}$. Further, with high probability, the conditional prediction bias is bounded by $Lh_n^{\beta_\mu} = Ln^{-\beta_\mu/(2\beta_\mu+d+d/(\gamma_0-1))}$ by standard Hölder smoothness arguments. Therefore, for every $x_0 \in [-1,1]^d$, the conditional mean squared error of $\hat{\mu}^{(NW)}(x_0 \mid h_n(\theta))$ is $O_P(r_n^2)$, with a constant that only depends on θ. (i) then follows by Markov's inequality.

Next, I show that there is a family for which this rate is the optimal rate of convergence for Nadaraya-Watson regression. For every $\gamma_0 > 1$, let $\mathcal{P}$ be the family of distributions for which $X \sim Unif([-1,1]^d)$, $D \mid X \sim Bern(\|X\|_\infty^{d/(\gamma_0-1)})$, $\mu(X)$ is in $\Sigma(\beta_\mu, L)$, and $Y \mid X, D \sim \mathcal{N}(\mu(X), 1)$. By analysis of the Irwin-Hall distribution, $P(e(X) \leq \pi) = P(\|X\|_\infty \leq \pi^{(\gamma_0-1)/d}) = C\pi^{\gamma_0-1}$ for some $C > 0$ and all π small enough. As a result, this family satisfies the constraints of Assumption 1 and Proposition 3 for some fixed $\theta^* \in \Theta$.

Now consider the variance and bias of Nadaraya-Watson regression for predicting $E[Y \mid X, D = 1]$ over a sequence of $P(n) \in \mathscr{P}$. Take an arbitrary sequence of positive bandwidths $h_n \to 0$ so that there is a hope for consistency. For simplicity, consider the uniform kernel with L_∞ distance. Let $M_n(h_n)$ be the number of treated observations with $\|X - 0\|_\infty \leq h_n$ for a bandwidth h_n. Fix some $h_n < \min\{1, \eta/d\}$. Because $\hat{\mu}(x_0 \mid h_n)$ is the sum of independent normal random variables, there is a closed-form:

$$\hat{\mu}(x_0 \mid h_n) - \mu(x_0) \mid \{X, D\} \sim \mathcal{N}\left(\frac{\sum D1\{\|X - x_0\|_\infty \leq h_n\}(\mu(X) - \mu(x_0))}{\sum D1\{\|X - x_0\|_\infty \leq h_n\}}, 1/M_n(h_n)\right).$$

By standard arguments, there is a $P \in \mathscr{P}$, $\alpha > 0$, and a $c(\theta)$ such that for every sequence of positive $h_n \to 0$,

$$\liminf_{n\to\infty} P\left(\frac{\sum D1\{\|X - x_0\|_\infty \leq h_n\}(\mu(X) - \mu(x_0))}{\sum D1\{\|X - x_0\|_\infty \leq h_n\}} \geq c(\theta)h_n^{\beta_\mu}\right) > \alpha$$
$$\liminf_{n\to\infty} P\left(1/M_n(h_n) \geq c(\theta)^2 n^{-1} h_n^{-d-d/(\gamma_0-1)}\right) > \alpha.$$

Therefore, for every sequence of $h_n \to 0$ and $P(n) \in \mathscr{P}$,

$$\begin{aligned} E_{P(n)}[(\hat{\mu}^{(NW)}(x_0 \mid h_n) - \mu(x_0))^2 \mid \{X, D\}] &\gtrsim_{P(n)} n^{-1}h_n^{-d-d/(\gamma_0-1)} + n^{-1/(2\beta_\mu+d+d/(\gamma_0-1))} \\ &\geq \min_{h_n>0} n^{-1}h_n^{-d-d/(\gamma_0-1)} + n^{-1/(2\beta_\mu+d+d/(\gamma_0-1))} \\ &\gtrsim n^{-\beta_\mu/(2\beta_\mu+d+d/(\gamma_0-1))}, \end{aligned}$$

with associated constants in the $\gtrsim$ terms that only depend on θ. Therefore there is a $C'(\theta)$ and $C''(\theta)$ such

that

$$P(n)\left(P(n)\left((\hat{\mu}^{(NW)}(x_0 \mid h_n) - \mu(x_0))^2 \mid \{X, D\} \geq C'(\theta) r_n\right)\right) > C''(\theta) > 0$$

for all n large enough. Thus, (ii) holds. □