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

Jacob Dorn*

July 9, 2025

Abstract

Propensity scores near zero or one, a condition known as weak overlap, distort the coverage of standard confidence intervals. This paper shows that a remedy is available by applying the common practice of thresholding to doubly robust estimators. The key is to threshold at a rate shrinking slowly enough to restore asymptotic normality, but quickly enough to only introduce second-order bias. I characterize the added burden in terms of nuisance function accuracy and smoothness conditions, and show these conditions are achievable without specifying the degree of overlap weakness. The theoretical results motivate a data-adaptive procedure for threshold selection, which exhibits near-exact coverage in large simulated samples, and yields comparable precision in an empirical application to a heuristic 10% fixed-trimming approach that changes the target parameter.

1 Introduction

This paper studies statistical inference on treatment effects when there are covariate regions in which the distributions among treated and control units only overlap weakly. Under weak overlap, it is difficult to find enough units in one population to predict counterfactual outcomes in the other population, weighting approaches may fail to be asymptotically normal, Wald "estimate ± 1.96 standard error" confidence intervals may exclude the treatment effect asymptotically, and there may be no semiparametric estimator capable of achieving the usual consistency rate (Khan and Tamer, 2010). A burgeoning literature has proposed strategies for adjusting confidence intervals for the possibility of weak overlap, but empirical practice has favored fixed-thresholding strategies that allow Wald confidence intervals to cover some treatment effect, at the cost of targeting a new and less interpretable population.

*I am grateful for suggestions from Xiaohong Chen, Rebecca Dorn, Kevin Guo, Edward Kennedy, Samir Khan, Michal Kolesár, Lihua Lei, Xinwei Ma, Ulrich Müller, Yuya Sasaki, Yulong Wang, and Larry Wasserman, and participants at seminars at the University of Pennsylvania. Artificial intelligence was used to suggest changes and identify potential errors. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656 and by grant T32 HS026116 from the Agency for Healthcare Research and Quality. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation or the Agency for Healthcare Research and Quality.

This paper shows that a simple extension of the standard practice of thresholding can restore nominal coverage of Wald confidence intervals. The intuition is as follows. Thresholded inverse propensity weighting (IPW) estimators in the presence of weak overlap are known to be biased but asymptotically normal, provided the threshold tends to zero slowly enough (Ma and Wang, 2020). The usual double robust estimator, Augmented IPW (AIPW), can be interpreted as leveraging an outcome regression estimate to reduce IPW's bias under strict overlap. If AIPW is thresholded at a rate tending to zero *quickly* enough, then thresholding bias will be of a lower order than sampling uncertainty. The main task of this paper is to establish when there exists a Goldilocks region of thresholds that are neither too small nor too large, so that Wald confidence intervals constructed assuming an appropriate central limit theorem will be well-calibrated.

I derive sufficient conditions for coverage under two frameworks. First, I follow the literature on semiparametric estimation (Chernozhukov et al., 2018) and derive sufficient convergence rates for the nuisance functions used in AIPW to ensure validity of "estimate $\pm$ 1.96 standard error" confidence intervals. These results are "black-box" in the sense that they are agnostic towards the details of the nuisance estimators and specific distribution beyond assuming a lower bound on the degree of overlap weakness. Second, I establish sufficient conditions for those black-box rates to be attained under possible weak overlap. In establishing the feasibility of achieving the black-box rates, I follow the nonparametric statistics literature in deriving sufficient conditions under Hölder smoothness assumptions (Stone, 1982), and propose a data-adaptive estimator that can achieve optimal regression rates without knowledge of the degree of overlap weakness.

I follow Ma and Wang (2020) and measure overlap weakness with a tail parameter γ_0. Values of γ_0 above two guarantee that the density of propensity scores tends to zero at zero, a case which I call *somewhat weak overlap*. Under somewhat weak overlap, it is known that AIPW can achieve the best possible consistency rate available under semiparametric assumptions (Newey, 1994; Hahn, 1998). My analysis shows that semiparametric efficiency continues to hold when the true nuisance functions are replaced by estimated nuisance functions, so long as the nuisance estimates achieve an appropriate product rate and the threshold is chosen appropriately. However, the required product rate can be stronger than the usual sufficient condition that the product of nuisance error rates goes to zero faster than $n^{-1/2}$; when only L^2-norm nuisance guarantees are available, the analogous requirement is that the product goes to zero faster than $n^{-\gamma_0/(2(\gamma_0-1))}$. This is because convergence rates on the estimated propensity score do not imply the same convergence rates on the estimated inverse propensity score.

Values of γ_0 below two allow the density of propensity scores to be unbounded at zero, a case which I call *very* weak overlap. It is the very weak overlap case in which semiparametric estimators fail to achieve their usual consistency rate and unthresholded IPW estimators fail to be asymptotically normal. This paper establishes that semiparametric guarantees are sufficient to establish validity of Wald confidence intervals

under very weak overlap. However, I show that $n^{-1/2}$ error products for even L^∞-norm errors are insufficient to establish the validity of Wald confidence intervals, and the associated estimates converge at a slower rate than is available under somewhat weak overlap.

Both regimes require stronger smoothness guarantees to ensure a given black-box error rate is feasible. The reason is that under any degree of weak overlap, there are covariate regions with few samples available for use in outcome regression. I quantify this added difficulty for the case of regression within a Hölder smoothness class of order $\beta_\mu > 0$. I show that weak overlap degrades the effective outcome smoothness by a factor of $1 - 1/\gamma_0$, exhibiting a cost of weak overlap even in the somewhat weak overlap regime in which the other asymptotic results carry through nearly unchanged. I show that this rate can be attained through a data-adaptive estimator that does not use knowledge of γ_0, and leverage a novel partitioning argument to show that the optimal pointwise rate can be achieved uniformly without the usual polylogarithmic factor.

Taken together, these results provide a precise answer to the question posed by this work's title: doubly robust t-statistics can handle weak overlap of tail bound γ_0, provided the outcome and propensity nuisance functions are in Hölder smoothness classes of order β_μ and β_e and

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

In this case, the threshold $b_n = n^{-\beta_e/(2\beta_e+d)} \log(n)^{(3\beta_e+d)/(2\beta_e+d)}$ suffices, regardless of the weak overlap parameter γ_0. When the outcome and propensity smoothness orders are the same $\beta > 0$, then thresholded AIPW can handle weak overlap of order γ_0, so long as:

$$\gamma_0 > \max\left\{\frac{2\beta^2 + 2\beta d + d^2}{\beta(2\beta + d)}, \frac{4\beta^2}{4\beta^2 - d^2}\right\}. \tag{2}$$

Under Lipschitz continuity of both nuisance functions in one dimension, doubly robust t-statistics can handle weak overlap of order $\gamma_0 > \frac{5}{3}$. In higher dimensions, there is always some sufficient smoothness order that yields valid t-statistics for any fixed tail bound.

The conditions yield new rules of thumb for thresholding in applied work. In my favored regime, the econometrician is willing to posit a minimal consistency rate for one of the two nuisance function estimates. Given such a minimal rate, a simple plug-in procedure predicts the threshold with the laxest restriction on the other nuisance function needed to achieve well-calibrated Wald confidence intervals. In the absence of any such information, a third rule of thumb derives a threshold that imposes the laxest equal minimal consistency rate on both nuisance estimates. None of these rules of thumb depend directly on knowledge of the tail bound parameter γ_0.

In simulations, I find that thresholded AIPW achieves the promised properties asymptotically. I consider a setting of very weak overlap with nonparametric outcome regression and propensity estimates, and focus on clipping (Winsorizing) extreme propensity scores. Unthresholded IPW and AIPW estimators perform poorly, with large errors and nonnormal asymptotic distributions. In this setting, clipped IPW displays its known first-order bias, and clipped AIPW displays the second-order bias justified by the theoretical analysis. 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 causal 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 proposal of a 10% fixed-trimming rule of thumb. I compare the clipped AIPW estimator that targets the full-population effect to estimators that apply AIPW to a sample trimmed based on a fixed rule. 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 relative to the 10% fixed-trimming rule, while increasing the estimated standard error by only 5.1%. These results show that targeting the full-population treatment effect does not need to introduce a major efficiency loss, and show that thresholded AIPW can easily be added as a robustness test when practitioners apply a fixed-trimming rule.

Weak overlap is a common and serious problem for statistical inference (Khan and Tamer, 2010; D'Amour et al., 2021). Existing approaches for statistical inference involve either targeting nonstandard estimands or leveraging a nonstandard estimator. The standard inverse propensity estimator in the presence of weak overlap involves trimming: dropping samples with small propensity estimates in order to estimate average effects within a better-behaved population (Imbens, 2004; 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), or clipping strategies that Winsorize weights (Ionides, 2008; Lee et al., 2011).[1] These and other proposals that reweight towards higher-precision populations (Yang and Ding, 2018; Li et al., 2018; Goldsmith-Pinkham et al., 2024) allow standard Wald confidence intervals based on standard estimators to cover some treatment effect, but come at the cost of discontinuously targeting a nonstandard and often less interpretable estimand. An important theoretical literature has proposed novel point and confidence interval estimators with desirable properties for statistical inference on the standard average treatment effects under weak overlap (Rothe, 2017;

[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."

Armstrong and Kolesár, 2017, 2021; Ma and Wang, 2020; Hirshberg and Wager, 2021; Heiler and Kazak, 2021; Sasaki and Ura, 2022; Ma et al., 2023; Chaudhuri and Hill, 2024). However, these proposals introduce new and less familiar estimators in the presence of weak overlap, and there has been little take-up by practitioners.

This paper instead shows that a standard estimator (thresholed AIPW) can provide valid confidence intervals on the standard average treatment effects. As a result, the analysis here builds most directly on work studying the behavior of inverse propensity estimators under weak overlap: Ma and Wang (2020) and Khan and Ugander (2022) show that thresholded IPW and AIPW can remain asymptotically normal, but at the cost of introducing first-order bias. My work is also related to work showing unthresholed AIPW can be oracle-equivalent with either parametric propensity estimates (Chen et al., 2008; Heiler and Kazak, 2021) or direct estimates of inverse propensities (Hirshberg and Wager, 2021). My analysis of nonparametric regression rates is also relevant to the literature on regression with degenerate designs; to my knowledge, the possibility of a uniform regression rate with no polylogarithmic penalty is new (Stone, 1982; Hall et al., 1997; Gaïffas, 2005; Mou et al., 2023).

The plan of the paper is as follows. Section 2 presents the setting and main theoretical results. Section 3 interprets these results in terms of minimal black-box consistency or smoothness rates, and presents a table summarizing some key results. Section 4 considers implications for parametric estimators and derives some rules of thumb for empirical use. Section 5 presents numerical results for simulations and the empirical application to right-heart catheterization. Section 6 concludes.

Notation. I use the notation $E_P[\cdot]$ and $E[\cdot]$ to refer to the expectation under the maintained distribution P over data Z. I abuse notation and write ψ for the target causal estimand under P, which in the technical analysis is the average potential outcome, and use $\sup_{P\in A} B$ to refer to the supremum of B over distributions P in A under any maintained restrictions on the distribution and nuisance functions. I write $A_n \leq_P B_n$ 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 $D^\alpha f = \frac{\partial^{\|\alpha\|_1} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots}$ satisfy $\|D^\alpha f(x) - D^\alpha f(x')\| \leq L\|x - x'\|^{\beta - \lfloor\beta\rfloor}$, where I write $D^\alpha f(x)$ for $D^\alpha f$ evaluated at x. For simplicity, I use local polynomial regression to refer to specifically kernel

regression with uniform bandwidth: $\hat{\mu}^{(NW)}(x \mid h) = \frac{\sum D\mathbf{1}\{\|X-x\|\leq h\}Y}{\sum D\mathbf{1}\{\|X-x\|\leq h\}}$ when feasible and $\hat{\mu}^{(NW)}(x \mid h) = 0$ when no nearby treated observations are available. I write the L^p norm as $\|f\|_{L^p(P)} = E_P[|f|^p]^{1/p}$ for finite p and $\|f\|_{L^\infty(P)} = \lim_{p\to\infty} \|f\|_{L^p(P)}$.

2 Setting, Consistency, and Asymptotic Normality

This section presents asymptotic results under black-box nuisance conditions.

2.1 Setting

I follow the standard semiparametric setup with a binary treatment. I assume the econometrician has access to n samples Z of data (X, D, Y), where $X \in \mathbb{R}^d$ are covariates, $D \in \{0, 1\}$ is a binary treatment, and $Y \in \mathbb{R}$ is an observed outcome. Often, the econometrician is interested in the average treatment effect $E_P\left[E_P[Y \mid X, D = 1] - E_P[Y \mid X, D = 0]\right]$.

For simplicity, I focus the theoretical analysis on estimating the average potential outcome $\psi(P) = E_P[E[Y \mid X, D = 1]]$. The average treatment effect follows as a corollary. I assume that the propensity score $e(X) = \mathbb{P}(D = 1 \mid X)$ is in $(0, 1)$ almost surely, so that the average potential outcome can be identified as $\psi(P) = E_P[DY/e(X)]$. I refer to regions of the covariate space in which the propensity can be arbitrarily close to zero as singularities. Strict overlap rules out singularities.

I derive uniform convergence rates under lower bounds on overlap weakness. I follow Ma and Wang (2020), who provide important building blocks in my analysis, and parameterize overlap weakness through a tail parameter γ_0. I extend their results to a model family $\mathscr{P}$ over distributions whose overlap is at least as strong as that tail parameter, in addition to some regularity conditions.

Assumption 1 (Distribution moments and tail bound)**.** Let $\mathscr{P}$ be a nonempty family of distributions, and write $e(X) = P(D = 1 \mid X)$ and $\mu(X) = E_P[Y \mid X, D = 1]$. Then every $P \in \mathscr{P}$ is a distribution over $(X, D, Y) \in \mathbb{R}^d \times \{0, 1\} \times \mathbb{R}$ satisfying the following conditions for some $q > 3, M, \sigma_{\min}, C > 0$, and $\gamma_0 > 1$:

(a) *Conditional moments.* $E[|Y - \mu(X)|^q \mid X, D = 1] \leq M^q < \infty$ almost surely.

(b) *Unconditional moments.* $Var(\mu(X)) \leq M$.

(c) *Residuals.* $\mathrm{Var}(Y \mid X, D) \geq \sigma^2_{\min}$.

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

Definition 1 generalizes Ma and Wang (2020)'s slowly varying tails assumption. Assumptions 1(a) through 1(c) are regularity conditions that rule out cases like perfectly predictable outcomes. Assumption 1(d) provides the substantial restriction on $\mathscr{P}$: overlap may be weak in the sense that γ_0 is finite, but there

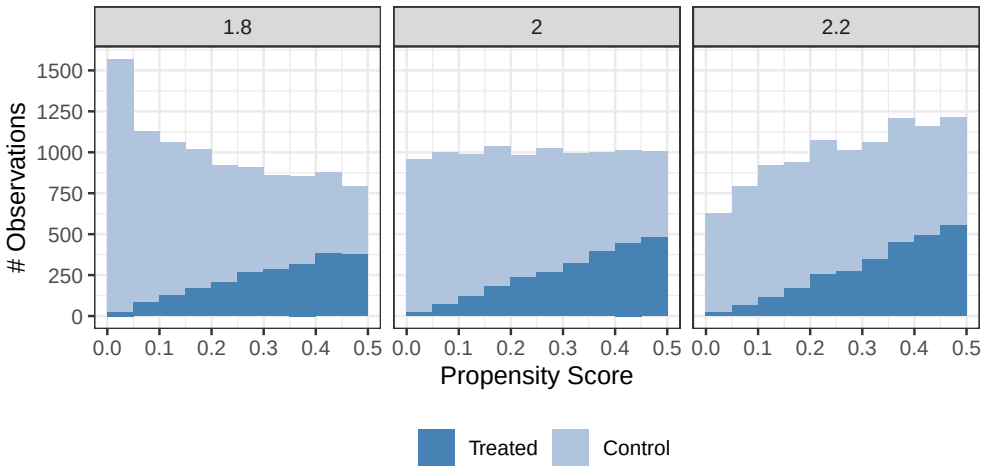

Figure 1: Simulations of 10,000 observations of $e(X)$ with $P(e(X) \leq \pi) = \pi^{\gamma_0 - 1}$ for increasing values of γ_0.

is some minimal γ_0 and C that provides a lower bound on the propensity's tail behavior. 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 distinguish between three cases of the degree of overlap weakness. I follow Heiler and Kazak (2021) and use *strict overlap* to refer to the case in which the propensity score is bounded away from zero almost surely, in which case Assumption 1(d) holds for any finite $\gamma_0 > 1$, and most results here hold after replacing γ_0 with infinity. I use "weak overlap" to refer to the case in which the infimum of the support of the propensity score is zero, which is sometimes called "limited overlap" (Khan and Tamer, 2010; Chaudhuri and Hill, 2024). In particular, finite values of γ_0 allow the inverse propensity distribution may be heavy-tailed. Within weak overlap, I distinguish between the case of *somewhat weak* overlap ($\gamma_0 > 2$) and *very* weak overlap ($\gamma_0 < 2$).[2]

Figure 1 illustrates behavior for simulated data with various values of γ_0. When the propensity score $e(X)$ has a well-defined density, $\gamma_0 = 2$ corresponds to a roughly uniform distribution of propensity scores (Ma and Wang, 2020). When γ_0 is above two, the density of propensity scores tends to zero at zero; when γ_0 is below two, the density of propensity scores can tend to infinity at zero. Heuristically, there are never too many treated observations with very small propensity scores, but γ_0 governs the degree to which there can be many untreated observations with very small propensity scores.

A phase transition occurs when γ_0 crosses two. Above two, the semiparametric efficiency bound is finite and $\sqrt{n}$-consistent estimation is feasible without parametric knowledge. Below two, the semiparametric efficiency bound is infinite and IPW with known propensity scores fails to be asymptotically normal (Khan

[2]Distributions satisfying somewhat weak overlap are sometimes said to satisfy "strict overlap" or "overlap" (Heiler and Kazak, 2021; Bruns-Smith et al., 2024), and distributions that exhibit very weak overlap are sometimes called "heavy tailed" (Chaudhuri and Hill, 2024).

and Tamer, 2010; Ma and Wang, 2020).

I will require certain rates on the estimates of 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: the observations Z_i are partitioned into K folds at random such that the distribution of the fold-k nuisance function estimates $\{\hat{\mu}_i, \hat{e}_i\}_{k_i=k}$ for each k is degenerate conditional on the other-fold data $\{Z_i\}_{k_i \neq k}$. If n_k is the number of observations per fold, then $\inf_k n_k / \sup_k n_k \to 1$. Further, there exist $p_\mu, p_e \geq 2$ and sequences $r_{\mu,n} = O(1)$ and $r_{e,n} = O(1)$ such that

$$\sup_{P \in \mathscr{P}} P\left(\|\hat{\mu}_n^{(-k)} - \mu\|_{L^{p_\mu}(P)} > r_{\mu,n} \text{ or } \|\hat{e}_n^{(-k)} - e\|_{L^{p_e}(P)}] > r_{e,n} \text{ for any } k \right) = o(1).$$

Cross-fitting is a common strategy for simplifying the analysis of Neyman-orthogonal estimators like AIPW (Chernozhukov et al., 2018). I assume that rates hold uniformly rather than in high probability to simplify the guarantees for uniform performance across a family of distributions in $\mathscr{P}$. Under strict overlap, it is standard to assume that $\frac{1}{p_\mu} + \frac{1}{p_e} < 1$. The version here is slightly stronger, and is much weaker than the sup-norm assumption ($p_e = p_\mu = \infty$) that is standard when 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. Guarantees for the other standard thresholding procedure, trimming, generally follow by the same arguments. I begin by providing sufficient conditions for consistency.

The clipped AIPW estimator of the average potential outcome ψ is:

$$\hat{\psi}_{clip}^{AIPW}(b_n) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{F}^k} \phi\left(Z_i \mid b_n, \hat{\eta}^{(-k)}\right), \text{ where } \phi(Z \mid b, \bar{\eta}) = \bar{\mu}(X) + \frac{D(Y - \bar{\mu}(X))}{\max\{\bar{e}(X), b\}}. \tag{3}$$

In that equation, $\mathcal{F}^k$ is the set of observations i randomly partitioned in fold k, $\hat{\eta}^{(-k)} = \left(\hat{\mu}^{(-k)}, \hat{\mu}^{(-e)}\right)$ is the nuisance function estimates constructed only on observations in folds other than k. The unthresholded AIPW estimator is the special case of $b_n = 0$.

A standard result for the unthresholded AIPW estimator is double robustness: when $e(X)$ is bounded away from zero, unthresholded 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) $r_{e,n} \ll b_n^{\frac{\gamma_0 - 1 + p_e}{(\gamma_0 - 1)(p_\mu p_e + 1 - p_\mu - p_e) + p_e}}$ or (ii) $r_{\mu,n} \ll b_n r_{e,n}^{-\frac{(\gamma_0 - 1)(p_\mu p_e + 1 - p_\mu - p_e) + p_e}{\gamma_0 - 1 + p_e}}$. Then for all $\epsilon > 0$,*

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

When at least one of γ_0, p_μ, or p_μ is finite, these conditions are stronger than the classic double robustness result that $r_{e,n}$ or $r_{\mu,n}$ tending to zero implies estimator consistency under strict overlap. For example, with an inconsistent propensity score, a meaningful fraction of the data may be thresholded asymptotically even under strict overlap.

2.3 Sufficient Rates for Confidence Interval Coverage

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 justifying Wald confidence intervals for unthresholded AIPW under strict overlap leverages Neyman orthogonality. It is useful to write the oracle AIPW estimator with known nuisance functions as $\tilde{\psi}_{(Oracle)}^{AIPW}(b_n) = \frac{1}{n}\sum \phi(Z \mid b, \eta)$. Under strict overlap, the difference between the feasible AIPW estimator with estimated nuisance functions and the hypothetical oracle AIPW estimator with known nuisance functions is

$$\hat{\psi}_{clip}^{AIPW}(0) - \tilde{\psi}_{(Oracle)}^{AIPW}(0) = \frac{1}{n}\sum (\hat{\mu} - \mu)\left(\frac{D}{\hat{e}} - 1\right) + (Y - \mu)\left(\frac{D}{\hat{e}} - \frac{D}{e}\right). \tag{4}$$

The first term debiases the inverse propensity estimate $\frac{D}{\hat{e}}$ of the number one with the regression error $\hat{\mu} - \mu$; the second term is zero-mean, and second-order under minor consistency conditions. As a result, slowly consistent nuisance estimates can yield a combined causal estimate whose difference from the oracle estimator is small. When all nuisances are consistent at $o(n^{-1/4})$ rates and $\hat{e}(X)$ is bounded away from zero, inverse propensity errors are of the same order as propensity errors, classical AIPW estimates are first-order equivalent to oracle estimates with known nuisances, simple Wald confidence intervals cover the true causal effect by appeal to the asymptotically normal oracle AIPW estimator.

Under very weak overlap, thresholded AIPW does not obtain the standard debiasing benefit. The anal-

ogous decomposition to Equation (4) for clipped AIPW is

$$\hat{\psi}^{AIPW}_{clip}(b_n) - \tilde{\psi}^{AIPW}_{(Oracle)}(b_n) = \frac{1}{n}\sum (\hat{\mu} - \mu)\left(\frac{D}{\max\{\hat{e}, b_n\}} - 1\right) + (Y - \mu)\left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}}\right). \tag{5}$$

Intuitively, clipping can be viewed as introducing a systematic error in propensity estimates in the direction of stability. Above the clipping threshold b_n, thresholded AIPW's nuisance estimation error enjoys a product-of-errors character that is similar to classical settings, albeit with multiplication by weights as large as b_n^{-1}. Below the clipping threshold, the regression errors are not debiased by any inverse propensity estimate. Instead, there is a subtly different form of debiasing: as the threshold tends to zero, increasingly little mass is thresholded. If the threshold b_n tends to zero quickly enough, the bias from outcome regression error in the thresholded region can be adequately debiased by the threshold itself.

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

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

(a) *Outcome consistency.* $r_{\mu,n} \ll b_n^{\frac{\min\{\gamma_0 - 1, 1\}}{p_\mu}}$ or $\exists\, \delta_n \to 0$ such that $\sup_{P \in \mathscr{P}} P\left(\max_k \|\hat{\mu}^{(-k)} - \mu\|^2_{L^\infty(P)} \geq \delta_n\right) = o(1)$.

(b) *Asymptotically known thresholding.* $r_{e,n} \ll b_n^{1 + \frac{\min\{\gamma_0 - 1, 1\}}{p_e}}$.

(c) *Second-order bias near singularities.* Either $b_n^{\gamma_0 - 1} \ll n^{-1/2}$ or $r_{\mu,n} b_n^{(\gamma_0 - 1)\frac{p_\mu - 1}{p_\mu} + \frac{\max\{2 - \gamma_0, 0\}}{2}} \ll n^{-1/2}$.

(d) *Second-order bias away from singularities.* $r_{\mu,n} r_{e,n}\left(b_n^{\zeta - \frac{\min\{\gamma_0, 2\}}{2}} + \log(1/b_n)^{1\{\zeta = 1\}}\right) \ll n^{-1/2}$, where $\zeta = (\gamma_0 - 1)\frac{p_\mu p_e - p_\mu - p_e}{p_e p_\mu}$.

Under strict overlap, the usual product rate condition is $r_{\mu,n} r_{e,n} \ll n^{-1/2}$. The conditions here are more stringent if either p_μ or p_e is finite, or if both are infinite (so that rates are sup-norm) but γ_0 is below two (so that there is very weak overlap). For example, when $\gamma_0 \geq 1.5$, $p_\mu = p_e = \infty$, and $r_{\mu,n} = n^{-1/4}$, then $r_{e,n} \ll n^{-1/3}$ will suffice, provided $r_{e,n} \ll b_n \ll n^{-1/3}$. However, these conditions never require parametric $n^{-1/2}$ consistency rates: shared L^2 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 $n^{-1/3}$. Condition Under somewhat weak overlap, (c) allows for thresholding bias to be second-order from the combination of $\hat{\mu}$ being bounded and b_n tending to zero quickly enough. Condition (d) is weaker than the naive sufficient condition $r_{\mu,n} r_{e,n} b_n^{-1} \ll n^{-1/2}$ that allows every observation to be thresholded, and reduces to the usual product rate condition provided at least one of p_μ or p_e is above two and γ_0 is large enough.

Certain technical possibilities call for one of two alternative further assumptions: a distributional smoothness assumption or a stronger rate assumption.

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

(i) *Tail lower bound.* There exists some $C' > 0$ such that $P(e(X) \leq \pi) \geq C'\pi^{\gamma_0 - 1}$ for all $\pi \in [0, 1]$.

(ii) *Faster rates.* $r_{\mu,n} \ll b_n^{\frac{1}{p_\mu}}$ or $\exists \delta_n \to 0$ such that $\sup_P P(\max_k \|\hat{\mu}^{(-k)} - \mu\|_{L^\infty(P)} \geq \delta_n) = o(1)$, $r_{e,n} \ll b_n^{\frac{p_e+1}{p_e}}$, and $r_{\mu,n} b_n^{\min\{\gamma_0 - 1, 1\}\frac{p_\mu - 1}{p_\mu}} \ll n^{-1/2}$.

Assumption 4(i) is a uniform version of Ma and Wang (2020)'s regularly varying tails assumption that $\lim_{t \to 0^+} \frac{P(e(X) \leq t\pi)}{P(e(X) \leq \pi)} = \pi^{\gamma_0 - 1}$ for all fixed $\pi > 0$. Most acutely, 4(i) rules out distributions that place a large point mass of propensities at points tending to zero slowly. When $\gamma_0 < 2$, Assumption 4(ii) is stronger than Assumption 3(c).

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 in the sense that with high probability:*

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

where $\sigma_n = n^{-1/2} \sqrt{\frac{1}{n} \sum \phi(Z \mid b_n, \eta)^2 - \tilde{\psi}^{AIPW}_{(Oracle)}(b_n)^2}$ is the oracle sample standard deviation. Further, clipped AIPW is asymptotically normal:

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

where $\hat{\sigma}_n = n^{-1/2} \sqrt{\frac{1}{n} \sum \phi(Z \mid b_n, \hat{\eta}_i)^2 - \left(\frac{1}{n} \sum \phi(Z \mid b_n, \hat{\eta}) \right)^2}$ is the feasible sample standard deviation. As a result, the Wald confidence interval $\hat{\mathcal{C}}_n(\alpha) = \left[\hat{\psi}^{AIPW}_{clip}(b_n) + z_{\alpha/2}\hat{\sigma}_n, \hat{\psi}^{AIPW}_{clip}(b_n) + z_{1-\alpha/2}\hat{\sigma}_n \right]$ for any fixed $\alpha \in (0, 1/2)$ has asymptotic coverage equal to α.

Theorem 1 is the core theoretical claim of this paper. The first result shows that thresholded AIPW 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, estimated t-statistics are asymptotically standard normal, and Wald confidence intervals constructed using these t-statistics are well-calibrated.

These 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 oracle 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 and note that oracle AIPW must have zero finite-sample bias.

The main 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 even with sup-norm convergence ($p = \infty$), because under weak overlap, there is an asymptotically unbounded number of observations with arbitrarily large inverse propensities with even known nuisance functions. The task is even more delicate with finite-norm convergence, which allows the estimated propensity distribution to exhibit weak overlap under strict overlap, and allows unbounded outcome regression bias near singularities under weak overlap. Nevertheless, by taking appropriate care and leveraging that clipping introduces bias by reducing inverse propensities, I am able to show that the effect of nuisance estimation is second-order even under the very weak overlap case in which unthresholded AIPW fails to be asymptotically normal and no regular root-n estimators exist.

Under somewhat weak overlap, unthresholded AIPW is semiparametric efficient. I now show that under moderate conditions, thresholding is also unnecessary in this case.

Corollary 1 (Semiparametric efficiency under somewhat weak overlap)**.** *Suppose Assumption 2 holds for some $\gamma_0 > 2$, $r_{e,n} \to 0$, $r_{\mu,n} \ll n^{\frac{-1}{2p_\mu}}$, and $r_{\mu,n} r_{e,n}^{\frac{p_e(p_\mu-1)}{(p_e+1)p_\mu}} \ll n^{-1/2}$. Let AV_n be semiparametric asymptotic variance bound (Hahn, 1998). Then there is a sequence of $b_n \to 0$ such that the thresholded AIPW estimator's variance is asymptotically optimal in the sense that for any $\delta > 0$, $\sup_{P \in \mathscr{P}} P\left(\left|nVar\left(\hat{\psi}_{clip}^{AIPW}(b_n)\right)/AV_n - 1\right| > \delta\right) = o(1)$. If in addition $r_{e,n} \ll n^{\frac{-(p_e+\gamma_0-1)}{p_e\gamma_0}}$, then the unthresholded AIPW estimator $\hat{\psi}_{clip}^{AIPW}(0)$ also has this property.*

Corollary 1 shows that it is possible for an unthresholded AIPW estimator to be semiparametrically efficient under weak overlap with unbounded propensity errors. The logic involves first verifying that there is a sequence of $b_n \to 0$ satisfying Assumption 4(ii), and then showing that the probability of there being a treated observation with a small estimated propensity score is asymptotically negligible. Further, as the nuisance norms p_e and p_μ become large, these conditions approach the standard $r_{\mu,n} r_{e,n} \ll n^{-1/2}$ condition, coupled with a requirement that $r_{e,n} \ll n^{-1/\gamma_0}$.

Taken together, this subsection 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, the and the

Table 1: Summary of degradation of asymptotic behavior and requirements as overlap is permitted to be increasingly weak under either L^2-norm or L^∞-norm consistency guarantees. Black box requirements are conditions for a sequence of valid b_n to exist.

Overlap Phase	**Strict** ($\inf e(x) > 0$)	**Somewhat Weak** ($\gamma_0 > 2$)	**Very Weak** ($\gamma_0 < 2$)
Double Robustness			
Consistency Conditions (L^2)	$\frac{r_{e,n}}{b_n} \to 0$ or $r_{\mu,n}\frac{r_{e,n}}{b_n} \to 0$	$\frac{r_{e,n}}{b_n} \to 0$ or $r_{\mu,n}\frac{r_{e,n}}{b_n} \to 0$	$\frac{r_{e,n}}{b_n} \to 0$ or $r_{\mu,n}\frac{r_{e,n}}{b_n} \to 0$
Consistency Conditions (L^∞)	$r_{e,n} \to 0$ or $r_{\mu,n}\frac{r_{e,n}}{b_n} \to 0$	$r_{e,n} \to 0$ or $r_{\mu,n}\frac{r_{e,n}}{b_n} \to 0$	$r_{e,n} \to 0$ or $r_{\mu,n}\frac{r_{e,n}}{b_n} \to 0$
Oracle AIPW Asymptotics			
Unthresholded Distribution	Normal	Normal	Nonnormal
Thresholded Convergence Rate	$n^{-1/2}$	$n^{-1/2}$	$n^{-1/2}b_n^{(\gamma_0-2)/2}$
Semiparametric efficient?	Yes	Yes	No
Nuisance Requirements			
Black Box (L^2)	$n^{1/2}r_{\mu,n}r_{e,n} \to 0$	$n^{\frac{\gamma_0}{2(\gamma_0-1)}}r_{\mu,n}r_{e,n} \to 0$ AND $n^{\frac{1}{3(\gamma_0-1)}}r_{e,n} \to 0$ OR $n^{1/2}r_{\mu,n}r_{e,n}^{1/3} \to 0$	$n^{1/2}r_{\mu,n}r_{e,n}^{1/(\gamma_0+1)} \to 0$
Black Box (L^∞)	$n^{1/2}r_{\mu,n}r_{e,n} \to 0$	$n^{1/2}r_{\mu,n}r_{e,n} \to 0$	$n^{1/2}r_{\mu,n}r_{e,n}^{\gamma_0/2} \to 0$
Smoothness:	$\frac{\beta_\mu}{2\beta_\mu+d} + \frac{\beta_e}{2\beta_e+d} > \frac{1}{2}$	$\frac{\beta_\mu}{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}} + \frac{\beta_e}{2\beta_e+d} > \frac{1}{2}$	$\frac{\beta_\mu}{2\beta_\mu+d\frac{\gamma_0}{\gamma_0-1}} + \frac{\frac{\gamma_0}{2}\beta_e}{2\beta_e+d} > \frac{1}{2}$
Regression Rates			
Pointwise optimum:	$n^{\frac{-\beta_\mu}{2\beta_\mu+d}}$	$n^{\frac{-\beta_\mu(1-1/\gamma_0)}{2\beta_\mu(1-1/\gamma_0)+d}}$	$n^{\frac{-\beta_\mu(1-1/\gamma_0)}{2\beta_\mu(1-1/\gamma_0)+d}}$
Uniform optimum:	$(n/\log(n))^{\frac{-\beta_\mu}{2\beta_\mu+d}}$	$n^{\frac{-\beta_\mu(1-1/\gamma_0)}{2\beta_\mu(1-1/\gamma_0)+d}}$	$n^{\frac{-\beta_\mu(1-1/\gamma_0)}{2\beta_\mu(1-1/\gamma_0)+d}}$

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 section interprets the rate requirements of Assumption 3.

3 Interpretation of Nuisance Requirements

This section interprets the rate requirements for Wald confidence intervals to cover asymptotically. I summarize the results in Table 1. Under somewhat weak overlap, thresholded and unthresholded AIPW with known nuisance functions remain semiparametric efficient and $\sqrt{n}$-consistent, and the traditional product of errors nuisance condition is sufficient with a modification to an L^∞ on errors. Under very weak overlap, clipped AIPW achieves a slower consistency rate and the required black-box nuisance rates are more stringent even with L^∞-norm guarantees. Both cases make outcome regression more difficult, but never so difficult as to require parametric assumptions. As a byproduct of this analysis, I show that the optimal pointwise and uniform regression rates under weak overlap are the same, without the usual polylogarithmic factor in the uniform rate.

3.1 Degradation of Consistency Rate

The previous analysis suggests that smaller values of b_n are preferable because they admit weaker black-box requirements. However, under very weak overlap, larger values of b_n correspond to faster AIPW rates.

I characterize the consistency rate of any oracle-equivalent estimator as follows.

Proposition 2 (Consistency rate)**.** *There exist positive constants* $c_{\min}$ *and* $c_{\max}$ *such that* $c_{\min} n^{-1} E_P\left[\frac{D}{\max\{e(X), b_n\}^2}\right] \leq \sigma_n^2 \leq c_{\max} n^{-1} E_P\left[\frac{D}{\max\{e(X), b_n\}^2}\right]$ *for all* $P \in \mathscr{P}$*, where* $\sigma_n^2 = n^{-1}\left(\frac{1}{n}\sum \phi(Z \mid b_n, \eta)^2 - \tilde{\psi}^{AIPW}_{(Oracle)}(b_n)^2\right)$ *is the oracle sample variance.*

If the estimator were trimmed instead of clipped, $E_P\left[\frac{D}{\max\{e(X), b_n\}^2}\right]$ would be replaced by $E_P\left[\frac{D\mathbf{1}\{e(X) \geq b_n\}}{e(X)^2}\right]$. Weaker overlap corresponds to larger values of $E_P[D/\max\{e(X), b_n\}^2]$ and slower consistency rates. Conditional on P, larger values of b_n correspond to a smaller value of $E_P[D/\max\{e(X), b_n\}^2]$, faster oracle consistency, and greater asymptotic power.

Proposition 2 implies a worst-case consistency rate over distributions in $\mathscr{P}$. I focus on the case of very weak overlap, because Corollary 1 shows that under somewhat weak overlap, clipped and traditional AIPW achieve the usual $\sqrt{n}$ consistency rate.

Corollary 2 (Worst-case consistency rate)**.** *Write* $v_n(b) = n^{-1}\left(b^{\gamma_0 - 2} + \log(1/b)^{1\{\gamma_0 = 2\}}\right)$*. Then for any* $P \in \mathscr{P}$*,* $\sup_{P \in \mathscr{P}} \sigma_n^2 = O\left(v_n(b_n)\right)$*. If in addition Assumption 4(i) holds, then* $\inf_{P \in \mathscr{P}} \sigma_n^2 \gtrsim v_n(b_n)$*.*

The rate v_n 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. The combination of Corollary 2 and Theorem 1 yields a trade-off under very weak overlap: smaller values of b_n yield laxer requirements on $\hat{\mu}$, but lead to larger variance and slower consistency.

3.2 Degradation of Black-Box Nuisance Requirements

Under very weak overlap, the black-box rates of Assumption 3 are more stringent than the usual $r_{\mu,n} r_{e,n} \ll n^{-1/2}$ condition. The main requirement is that $r_{\mu,n} r_{e,n}^{\min\{\gamma_0/2, 1\}}$ 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.

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 unthresholded AIPW with parametric nuisance function estimates to be first-order equivalent to an oracle estimator. An alternative characterization of the usual product-of-errors condition that $r_{\mu,n} r_{e,n} \ll n^{-1/2}$; this requirement is equivalent under somewhat weak overlap, but is more stringent

under very weak overlap. Under very weak overlap, even this stronger product-of-errors requirement is insufficient for Wald confidence intervals to be valid.

Corollary 3 (Under weak overlap, faster rates can be necessary)**.** *Take some $\gamma_0 > 1, p_\mu, p_e \geq 2$ such that $(\gamma_0 - 1)\frac{p_e - 1}{p_e} < 1$. Then for any sequence b_n satisfying $n^{-1/2} \ll b_n \ll 1$, there is a family $\mathscr{P}$ and cross-fit nuisance estimators $\hat{\mu}$ and $\hat{e}$ such that:*

1. Overlap is not too weak. *$\mathscr{P}$ satisfies Assumption 1 for this γ_0.*
2. Nuisances satisfy cross-fitting and separate rates. *Assumptions 2 and 3(a),(b) hold.*
3. The usual product rate holds. $n^{1/2} r_{\mu,n} r_{e,n} \to 0$.
4. Wald inference fails. *For any fixed target coverage level $\alpha \in (0, 1)$, the Wald confidence interval $\hat{\mathcal{C}}_n(\alpha))$ based on AIPW with clipping at b_n has the zero-coverage property* $\sup_{P \in \mathscr{P}} P(\psi(P) \in \hat{\mathcal{C}}_n(\alpha)) \to 0$.

The condition $(\gamma_0 - 1)\frac{p_e - 1}{p_e} < 1$ ensures that $\zeta < 1$ even under sup-norm rate guarantees; otherwise, the conditions for Wald confidence interval failure are more subtle.

A sufficient condition for the existence of a sequence of b_n such that Wald confidence intervals are valid under sup-norm rates 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. Under L^2 rates, the associated condition is $r_{\mu,n} r_{e,n}^{1/\min\{\gamma_0 + 1, 3\}} \ll n^{-1/2}$ for $r_{e,n} \gtrsim n^{-1/2}$. There is a blessing of very weak overlap: with L^2 nuisance guarantees (or otherwise if $p_\mu + p_e > p_\mu p_e / 2$), smaller values of γ_0 yield a more lax product of errors condition, and therefore a slightly more copacetic black-box requirement to achieve valid Wald confidence intervals. Nevertheless, as I show later, this blessing of very weak overlap is likely outweighed by the added difficulty of outcome regression.

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

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

I characterize the following special cases.

Example 1 (Somewhat weak overlap)**.** Suppose Assumption 5 holds, $\gamma_0 > 2$, 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 $r_{\mu,n} r_{e,n} \log(1/r_{e,n}) \ll n^{-1/2}$. Then there exists a $b_n \to 0$ such that clipped AIPW t-statistics are asymptotically well-calibrated.

Example 3 (Very weak overlap, shared rates)**.** 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 either (i) $r_{\mu,n} = O(n^{-1/2})$ and $r_{e,n} = o(1)$ or (ii) $r_{e,n} = O(n^{-1/2})$ and $r_{\mu,n} = o(n^{(\gamma_0-2)/4})$. Then there exists a $b_n \to 0$ such that clipped AIPW t-statistics are asymptotically well-calibrated.

I now unpack the black box and quantify the degree to which weak overlap makes a given outcome regression rate more difficult to achieve.

3.3 Necessary Smoothness Conditions

Weak overlap makes outcome regression more difficult: there may be few treated observations in precisely the regions in which thresholded AIPW depends most acutely on outcome regression. In this section, I show that for pointwise rates, even somewhat weak overlap degrades the effective outcome smoothness for optimal local polynomial estimators. There is a slight blessing of weak overlap: the optimal uniform rate is equal to the optimal pointwise rate, without the usual polylogarithmic penalty.

I characterize optimal nonparametric regression rates under Hölder continuity. For convenience, I fix the covariates to be uniform over a specific hypercube in $\mathbb{R}^d$ with constant variance.

Assumption 6 (Hölder smoothness and fixed domain)**.** $\mathscr{P}$ satisfies Assumptions 1 and 4(i), and for all $P \in \mathscr{P}$, $X \sim Unif([-1,1]^d)$ and $Y \mid X, D \sim \mathcal{N}(D\mu_P(X) + (1-D)\mu'_P(X), \sigma^2_{\min})$, with μ, μ' in the Hölder smoothness class $\Sigma(\beta_\mu, L)$ for some fixed $\beta_\mu, L > 0$.

Most of these assumptions are standard assumptions for studying local polynomial regression under strict overlap (Stone, 1982). I assume normal outcomes in order to simplify the characterization of the optimal rate. It is known that in this case but under strict overlap, the optimal pointwise rate is $n^{\frac{-\beta_\mu}{2\beta_\mu+d}}$, and the optimal uniform rate $(n/\log(n))^{\frac{-\beta_\mu}{2\beta_\mu+d}}$ has a polylogarithmic penalty, in the sense that the optimal uniform rate is worse by some polynomial factor of $\log(n)$.

I require that treated observations cannot concentrate in small regions.

Assumption 7 (Non-trivial concentration)**.** There are parameters $\rho, \nu > 0$ such that for all $h > 0$ small enough and all $P \in \mathscr{P}$ and $x_0 \in [-1,1]^d$, $P(e(X) \geq \rho \sup_{\|x-x_0\| \leq h} e(x) \mid D = 1, \|X - x_0\| \leq h) > \nu$.

I show in Appendix Proposition 4 that this condition holds if the propensity function is sufficiently smooth. When the propensity function is nonsmooth, it is possible for nature to choose a distribution with better local overlap properties that induces local polynomial degeneracy issues. It is likely that the feasibility

proof could bypass Assumption 7 through use of some hypothetical degeneracy-robust estimator, but that would be outside the scope of this work.

In the worst case, weak overlap of order $\gamma_0 > 1$ plays a role equivalent to scaling the effective outcome smoothness downward by $(1 - 1/\gamma_0)$.

Theorem 2 (Weak overlap reduces effective outcome smoothness)**.** *Define* $\psi_n = n^{\frac{-\beta^*}{2\beta^*+d}}$*, where* $\beta^* = \beta_\mu (1 - 1/\gamma_0)$*. Then*

(i) ψ_n is a pointwise (and uniform) rate upper bound. *There exists a* $c > 0$ *and a* $\mathscr{P}$ *satisfying Assumptions 6 and 7 such that*

$$\liminf_{n\to\infty} \inf_{\hat{\mu}} \sup_{P\in\mathscr{P}} P\left(|\hat{\mu}(0) - \mu(0)| > c\psi_n\right) > 0.$$

(ii) ψ_n is an achievable uniform (and pointwise) rate. *Suppose* $\mathscr{P}$ *satisfies Assumptions 6 and 7. Then there exists an estimator* $\hat{\mu}(x)$ *such that for all* $\epsilon > 0$*, there is a finite* $c(\epsilon)$ *such that*

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}} P\left(\|\hat{\mu} - \mu\|_{L^\infty(P)} > c(\epsilon)\psi_n\right) \leq \epsilon.$$

Further, the estimator can be computed without knowledge of the overlap bound γ_0*.*

Recall that under strict overlap, the optimal pointwise convergence rate for a Hölder-smooth regression function is $n^{\frac{-\beta_\mu}{2\beta_\mu+d}}$. Theorem 2 shows that weak overlap has the effect of degrading the effective smoothness rate from β_μ to $\beta_\mu(1 - 1/\gamma_0)$. As overlap is allowed to become increasingly weak and other parameters are held constant, there can be regions of the covariate space with increasingly few treated observations so that the optimal pointwise regression rate is slower. This penalty occurs even under somewhat weak overlap. In the limit in which γ_0 tends to one, the rate in Theorem 2 can become arbitrarily poor. For example, the difficulty of estimating a twice continuously differentiable function when $\gamma_0 = 2$ is comparable to the difficulty of estimating a Lipschitz-continuous function under strict overlap.

At a high level, the construction follows a standard framework by constructing a grid over $[-1, 1]^d$, choosing bandwidths for each gridpoint to balance the local polynomial bias (increasing in bandwidth) and variance (decreasing in bandwidth), estimating $\hat{\mu}(x)$ at each gridpoint via local polynomial regression, and then interpolating between gridpoints. However, the construction here avoids knowledge of γ_0 by exploiting that for any fixed x with the worst possible overlap under $\mathscr{P}$, taking h to solve $\sum_i 1\{\|X_i - x\| \leq h, D_i = 1\} = h^{-2\beta_\mu}$ yields a bandwidth on the order of $n^{\frac{-1}{2\beta_\mu + d\gamma_0/(\gamma_0-1)}}$. I construct a data-adaptive grid by starting with a hypothetical grid that would be appropriate under strict overlap, but otherwise is too dense. I then use the dense grid's implied weak overlap bound to shrink the gridpoint density until I can ensure that every

gridpoint has sufficient treated observations within some hypothetical largest gridpoint. I then adapt each gridpoint's bandwidth downward to solve $\sum_i 1\{\|X_i - x\| \leq h, D_i = 1\} = h^{-2\beta_\mu}$, which allows the estimator to exploit better overlap properties away from singularities.

This construction turns out to avoid the usual polylogarithmic penalty. Intuitively, this is because no distribution can have too many separated points with the weakest possible overlap; the implied global overlap bound would be even weaker. Formally, I partition the gridpoints into increasingly large groups with increasingly small bandwidths (and therefore increasingly strong overlap guarantees). I show in Appendix Proposition 5 that when the groups are constructed appropriately and the first group is *large* enough, the expected largest regression error in each subsequent group shrinks geometrically. As a result, the expected largest regression error across groups is controlled by the sum of the expected largest group errors, which in turn is controlled by the large, but bounded, expected largest error in the smallest group.

Theorem 2 yields minimal smoothness assumptions for Wald confidence interval validity.

Corollary 4 (Minimal smoothness conditions)**.** *Suppose Assumption 6 holds and there is a $\beta_e > 0$ such that $e(X) \in \Sigma(\beta_e, L)$ and*

$$\frac{\beta_\mu}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)} + \frac{\min\{\gamma_0/2, 1\}\beta_e}{2\beta_e + d} > 1/2. \tag{6}$$

Then there is a sequence of nuisance estimators and thresholds that are independent of γ_0 such that for all $\gamma_0 > 1$, the associated Wald confidence interval $\hat{\mathcal{C}}_n(\alpha)$ constructed using $\hat{\psi}^{AIPW}_{clip}(b_n)$ satisfies

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

Further, this estimator can be computed without knowledge of the overlap bound γ_0.

In one dimension, thresholded AIPW with Lipschitz-continuous conditional outcome mean and propensity function can handle weak overlap of order $\gamma_0 > \frac{4+1/\beta_e}{3}$. In multiple dimensions, the econometrician must assume stronger smoothness restrictions than Lipschitz continuity in order to achieve the necessary nuisance rate guarantees under even strict overlap. When the propensity function is infinitely-differentiable, thresholded AIPW with a Lipschitz-continuous conditional outcome mean can handle weak overlap of order $\gamma_0 > \frac{2(d+1)}{d+2}$. More generally, under very weak overlap, if $\beta_\mu, \beta_e > \frac{d\left(\sqrt{\gamma_0^2+4\gamma_0-4}+2-\gamma_0\right)}{4(\gamma_0-1)}$, then it is feasible to achieve standard inference with thresholded AIPW.

This concludes the substantive theoretical analysis. In the next section, I use these results to infer lessons for empirical practice.

4 Lessons for Empirical Practice

This section leverages the theoretical analysis to consider misspecified parametric estimators and some rules of thumb for empirical use.

4.1 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 consider IPW and AIPW with the same sequence of thresholds b_n satisfying $1 \gg b_n \gg n^{-1/2}$. I write that a nuisance estimate $\hat{\eta}$ is consistent if it tends to the correct limit η, and I write that $\hat{\eta}$ is inconsistent otherwise.

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.

The analysis above is easiest to extend when either both or neither nuisance function is consistent. 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. 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.

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 2, which has $P(e(X) \leq \pi) \sim \pi^{\gamma_0 - 1}$ for all π small enough. The bias in the thresholded 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 2, 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, because IPW will be inconsistent. 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 thresholded AIPW should also be asymptotically normal under appropriate error product conditions. However, it is unclear how the bias compares to sampling error. In any event, 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.

Before proceeding to apply clipped AIPW, I derive some rules of thumb for choosing a threshold.

4.2 Choice of Threshold

The theoretical analysis above provides conditions under which there is some sequence of thresholds for which AIPW is asymptotically normal and centered around the true causal estimand. I now provide guidance for how to choose the threshold.

I propose different rules of thumb based on whether the econometrician is willing to provide an upper bound on the rate of convergence for the propensity estimate, the outcome regression estimate, or both. The combined proposal is presented in Algorithm 1.

In practice, it is often relatively easy to identify an upper bound on the propensity rate of convergence. For instance, if the propensity score is estimated with local polynomial regression of order p_e, then the econometrician is implicitly asserting a Hölder smoothness of some order $\beta_e > \ell_e$, and a feasible uniform consistency rate of $(n/\log(n))^{-\beta_e/(2\beta_e + d)}$ (Stone, 1982). In this case, the econometrician can safely conjecture that if their estimator is well-founded, then it will achieve a uniform consistency rate $r_{e,n} \ll n^{-\ell_e/(2\ell_e + d)}$, and a threshold of $b_n = n^{-\ell_e/(2\ell_e + d)} = \text{“}\bar{r}_{e,n}\text{”}$. This rule of thumb is practical and the safest rule with respect to outcome regression that does not impose stronger requirements on the propensity estimate, but also often corresponds to a relatively slow consistency rate with relatively little power.

Three alternative rules of thumb target faster consistency rates and laxer propensity requirements.

Input: Rate upper bounds $\bar{r}_{\mu,n}$ and $\bar{r}_{e,n}$ (maybe null), data $\{D_i, \hat{e}_i\}_{i=1}^n$

Output: Rule-of-thumb threshold b_n

Function `calculateRuleofThumb`$(\bar{r}_{\mu,n}, \bar{r}_{e,n}, \{D_i, \hat{e}_i\}_{i=1}^n)$**:**
  if *is.null*$(\bar{r}_{\mu,n})$ and *!is.null*$(\bar{r}_{e,n})$ **then**
    $b_n \leftarrow \bar{r}_{e,n}$
  else
    $A_\mu \leftarrow$*!is.null*$(\bar{r}_{\mu,n})$
    $A_e \leftarrow$*!is.null*$(\bar{r}_{e,n})$
    $b_n \leftarrow \sup b : 0 \geq$`errorBoundDiff`$(b, \bar{r}_{\mu,n}, \bar{r}_{e,n}, \{D_i, \hat{e}_i\}_{i=1}^n, A_\mu, A_e)$
  end
  return b_n
Function `errorBoundDiff`$(b, \bar{r}_{\mu,n}, \bar{r}_{e,n}, \{D_i, \hat{e}_i\}_{i=1}^n, A_\mu, A_e)$**:**
  $\tilde{r}_{\mu,n} \leftarrow A_\mu \bar{r}_{\mu,n} + (1 - A_\mu)b$
  $\tilde{r}_{e,n} \leftarrow A_e \bar{r}_{e,n} + (1 - A_e)b$
  `second_moment` $\leftarrow \frac{1}{n}\sum_{i=1}^n \frac{D_i}{\max\{\hat{e}_i, b\}^2}$
  `error_bound` $\leftarrow \tilde{r}_{\mu,n}\frac{\frac{1}{n}\sum \mathbf{1}\{\hat{e}_i \leq b\}}{\sqrt{\texttt{second\_moment}}} + \tilde{r}_{\mu,n}\tilde{r}_{e,n}\sqrt{\texttt{second\_moment}}$
  return `error_bound` $- n^{-1/2}$

Algorithm 1: Rule-of-thumb for choice of threshold b_n given (possibly null) rate upper bounds $\bar{r}_{\mu,n}$ and $\bar{r}_{\mu,n}$, to ensure that L^p-norm guarantees of the form $r_{\mu,n} \ll \bar{r}_{\mu,n}$ (or $r_{\mu,n} \ll b_n$ if no bound is given) and $r_{e,n} \ll \bar{r}_{e,n}$ (or $r_{e,n} \ll b_n$ if no bound is given) is sufficient to achieve well-calibrated Wald confidence intervals.

While the main text focuses on black-box requirements on consistency rates in terms γ_0, the proof goes through weaker and less interpretable conditions in Assumption 3'. However, these conditions can be translated into data-adaptive sufficient conditions on b_n by leveraging the inequality $b_n^{\gamma_0 - 1}\log(1/b_n)^{1\{\gamma_0 = 2\}} \precsim b_n E_P[D/\max\{e(X), b_n\}^2]$. For example, when $p_\mu = p_e = \infty$, I obtain sufficient conditions that b_n tends to zero slower than $n^{-1/2}$, $r_{e,n} \ll b_n$, and

$$r_{\mu,n}\frac{P(e(X) \leq b_n)}{\sqrt{E_P\left[\frac{D}{\max\{e(X), b_n\}^2}\right]}} + r_{\mu,n}r_{e,n}\sqrt{E_P\left[\frac{D}{\max\{e(X), b_n\}^2}\right]} \ll n^{-1/2}.$$

In these alternative cases, I propose replacing b_n in the right-hand side with the putative threshold b and $r_{\mu,n}$ and $r_{e,n}$ in the left-hand side with predicted upper bounds $\bar{r}_{\mu,n}$ and $\bar{r}_{e,n}$ where feasible and with the putative threshold b otherwise. This yields an empirical function `error_bound`(b). I then solve for the putative threshold b where `error_bound`(b) crosses $n^{-1/2}$. The result is a threshold b_n aimed to target maximal efficiency and minimal propensity consistency requirements, calculated without use of the outcome data or direct knowledge of γ_0, and with the property that if the nuisance errors go to zero faster than the specified upper bound (or estimated threshold), then the resulting Wald confidence intervals will be well-calibrated. Alternative rules of thumb can be constructed using Appendix Assumption 3' when only finite-order consistency rate guarantees are available. An interesting avenue for future work is whether there

is a convenient choice of context-dependent constant multiples in the `error_bound` function.

This rule of thumb is also always feasible, for example in the case of nonspecified upper bounds.

Lemma 1 (Well-defined rule of thumb)**.** *Suppose* $\hat{e} \in (0, 1]$ *and* $\sum D/\hat{e} > 0$*. Let* $f_n(b)$ *be the* `error_bound` *function with no upper bound nuisance rates given. Then there is exactly one* b_n *such that* $\limsup_{b \to b_n^-} f_n(b) \leq 0 \leq \liminf_{b \to b_n^+} f_n(b_n)$.

A rule of thumb with nonspecified rates seems particularly attractive, since it is an entirely data-driven way to choose the AIPW threshold. However, in practice, a given outcome regression rate is more difficult to achieve than a given propensity rate under weak overlap, so that rules of thumb with specified nuisance rates is more appropriate for nonparametric outcome regression estimates.

5 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 well asymptotically, producing near-perfect calibration of p-values with 100,000 observations, but exhibits some undercoverage in small samples. When studying the right heart catheterization data, I find that the rule of thumb approach increases the estimated harm of the procedure by 0.17 standard errors relative to the usual 10% trimming rule, while increasing the estimated standard error by 5.1%.

5.1 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 $P(e(X) \leq \pi) = \pi^{\gamma_0 - 1}$ 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_0 from 1.5 to 1.8 to ensure feasible outcome regression rates, set $\kappa = 2$ rather than $\kappa = 1$ to avoid coincidental offsetting bias of IPW lower and upper tails 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, Theorem 2 establishes that kernel regression can achieve a pointwise rate of $n^{-1/(3+1/(\gamma_0 - 1))}$. I conjecture that $r_{\mu,n} \ll n^{-1/5}$, which is feasible if $\gamma_0 > 1.5$, and choose the clipping threshold b_n based on Algorithm 1.

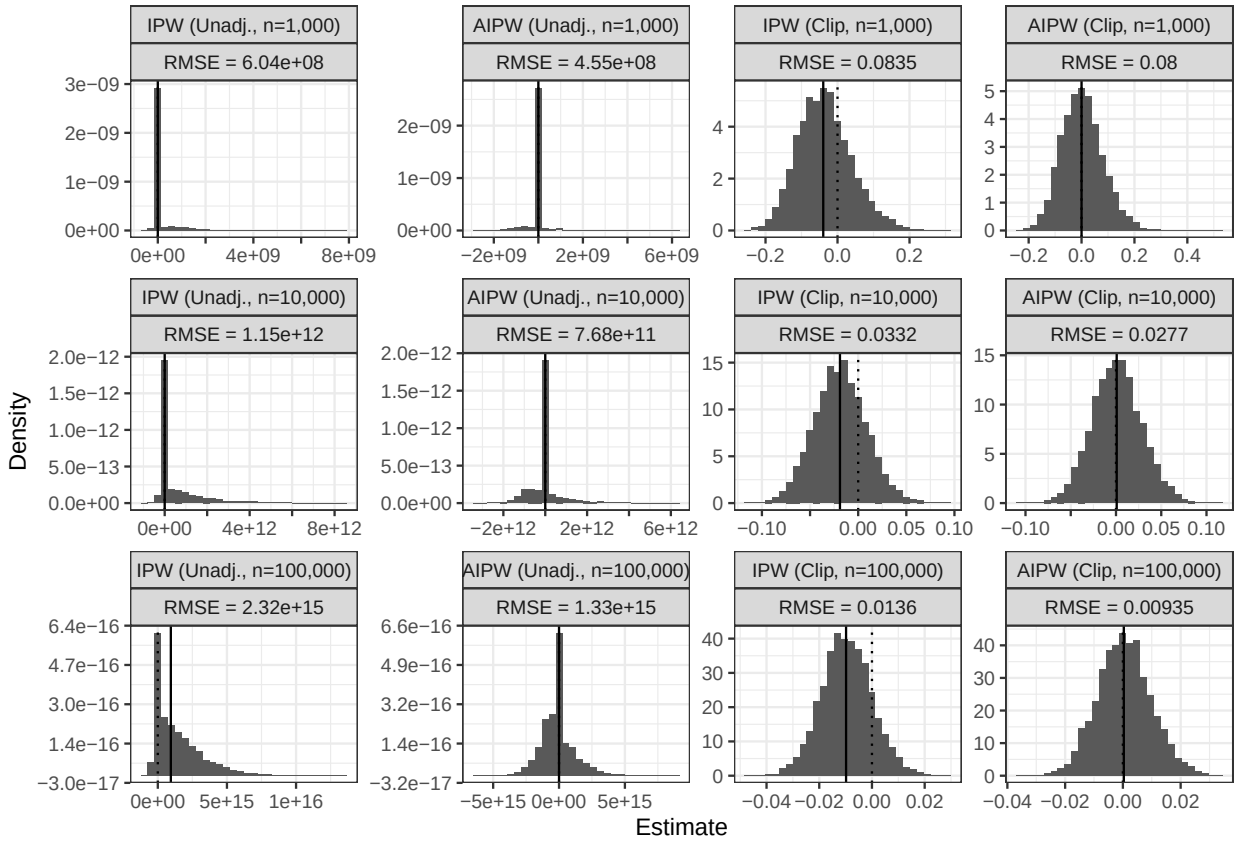

Figure 2: Histograms of point estimates in simulations for the various methods considered in the simulations. Vertical dotted and solid lines indicate true causal effect and median estimate, respectively. Clipped estimators achieve much better performance than unthresholded estimators, and clipped AIPW's debiasing property is also apparent.

I begin by summarizing point estimates in Figure 2. The unthresholded estimators are approximately median-unbiased, but possess sufficiently heavy inverse propensity tails that the mean performance degrades with increasing sample size. 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 3 that the clipped AIPW estimator's t-statistics are reasonably well-calibrated. The plot presents t-statistics on the true average potential outcome. The t-statistics of unthresholded IPW and AIPW estimators are visibly non-Gaussian, and often exhibit a multimodal distribution. This poor performance is unsurprising: unthresholded estimators are known to fail to be asymptotically normal in this setting. Both thresholded estimators are known to be asymptotically normal in this setting when the propensity score is known, 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

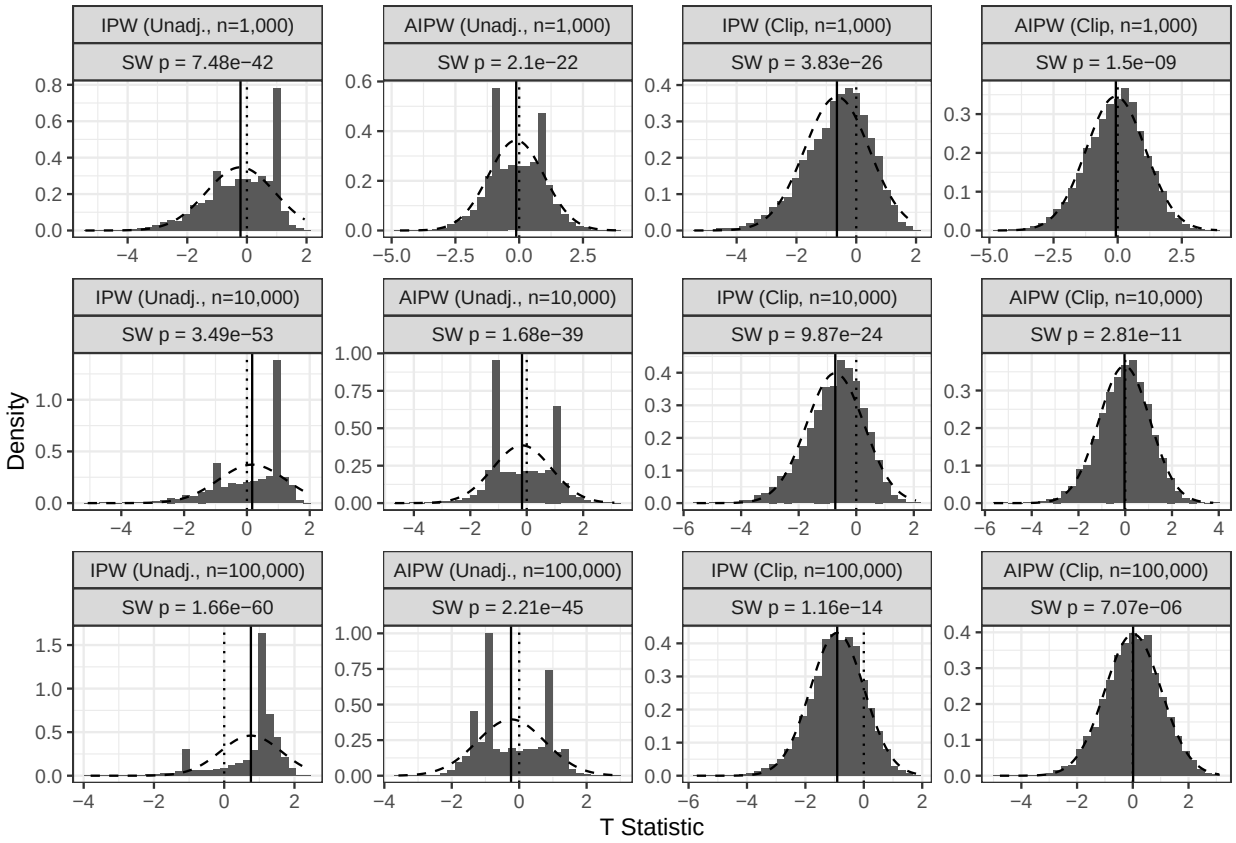

Figure 3: Histograms of simulated t-statistics on the true null hypothesis for various sample sizes. Vertical solid and dotted lines indicate mean t-statistic and target mean t-statistic of zero, respectively. Dashed line corresponds to the calibrated Gaussian density targeted in the Shaprio-Wilk test for normality.

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 4 that the clipped AIPW estimator's p-values are well-calibrated in 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 unthresholded IPW and AIPW estimators exhibit known poor performance. The clipped IPW estimator exhibits overrejection even with large samples, as even oracle clipped IPW would provide well-calibrated inference for a biased estimand. The clipped AIPW estimator also overrejects 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, the asymptotic calibration 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 Kolmogorov-Smirnov p-value on exact calibration of the two-sided test statistics for

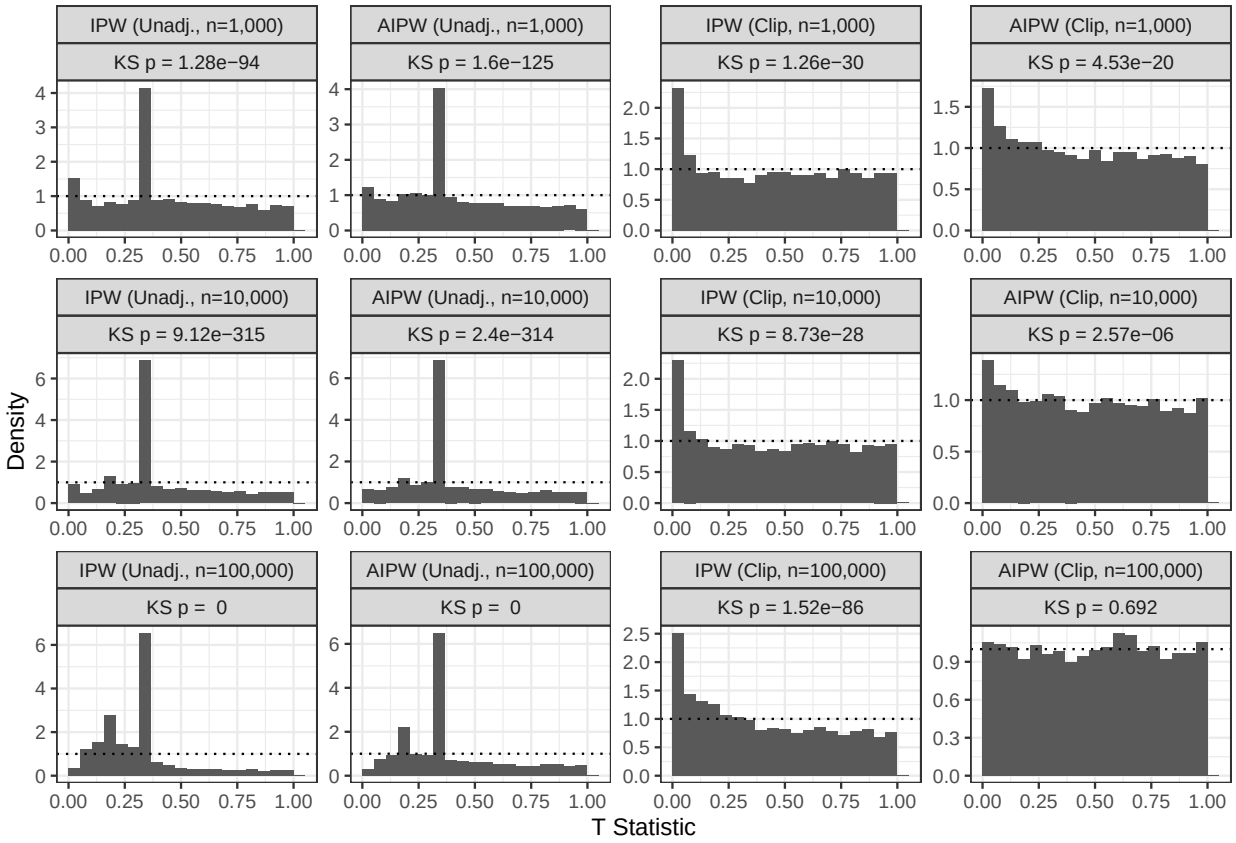

Figure 4: Histograms of simulation p-values on null hypothesis of true average potential outcome for various sample sizes. Dotted lines correspond to the target Uniform(0, 1) density. P-values in labels correspond to Kolmogorov-Smirnov tests for the Uniform(0, 1) distribution.

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 difficulty of outcome regression in this setting. Figure 5 presents an example with 1,000 observations. It is rare to have treated observations with small 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 observations. In Appendix B (Figures 9 through 11), I conduct the same experiments, but with the estimated outcome regression function replaced by the true outcome regression 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 underreject with 100,000 observations.

In Appendix B (Figures 12 through 14), I show that these conclusions would largely carry through

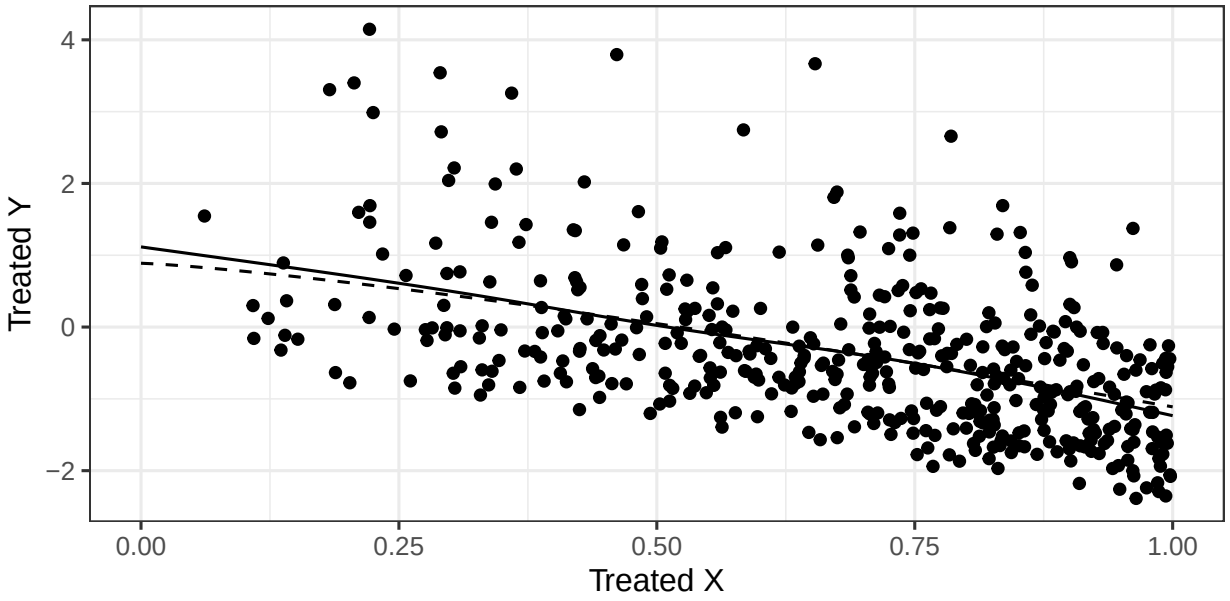

Figure 5: Simulated treated observations for one simulation of 1,000 observations. It is rare to see treated observations with small X, which corresponds to small values of $e(X) = X^{1/(\gamma_0-1)}$. As a result, such observations can have high leverage when predicting $E[Y \mid X = 0, D = 1]$, and can yield to important errors between the true (dashed) and predicted (solid) regression lines.

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.

5.2 Application to Right Heart Catheterization

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

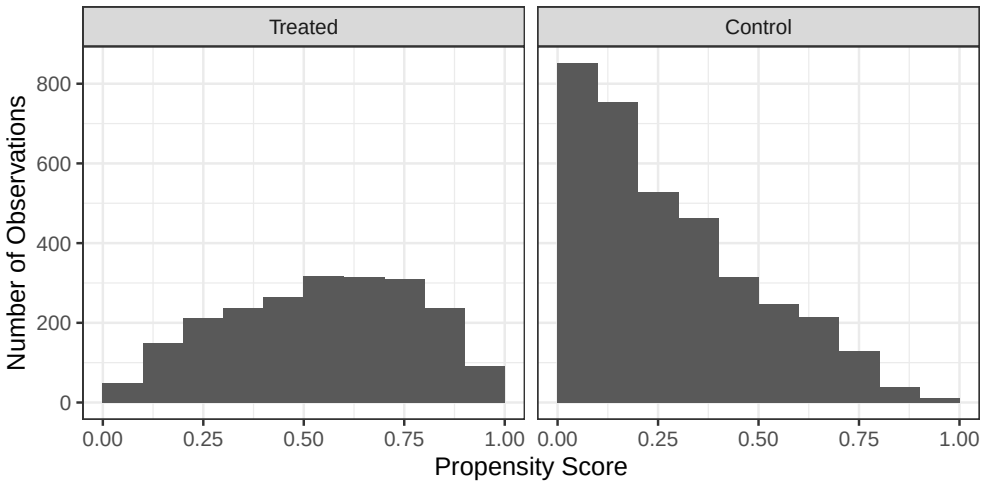

Figure 6: Histogram of estimated propensity scores for treated (left) and control (right) observations in right heart catheterization data. The plot is designed to parallel Figure 1 in Crump et al. (2009). Slight differences reflect the use of cross-fitting.

population 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. I compare these procedures to the 10% rule and other fixed trimming rules using the same nuisance estimates.

I present the distribution of estimated propensity scores for treated and control units in Figure 6. The 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.

I compare AIPW estimators for various trimmed subsamples to the clipped AIPW estimator. I choose the clipping threshold b_n through the no-specified-upper-bound version of Algorithm 1 because I estimate both nuisance functions parametrically. I plot the functions used in choosing b_n in Figure 7. 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.

I present estimated effects and confidence intervals for various potential fixed trimming rules in Figure 8. The 10% trimming rule yields an estimated reduction in survival rates of 5.79 percentage points among the

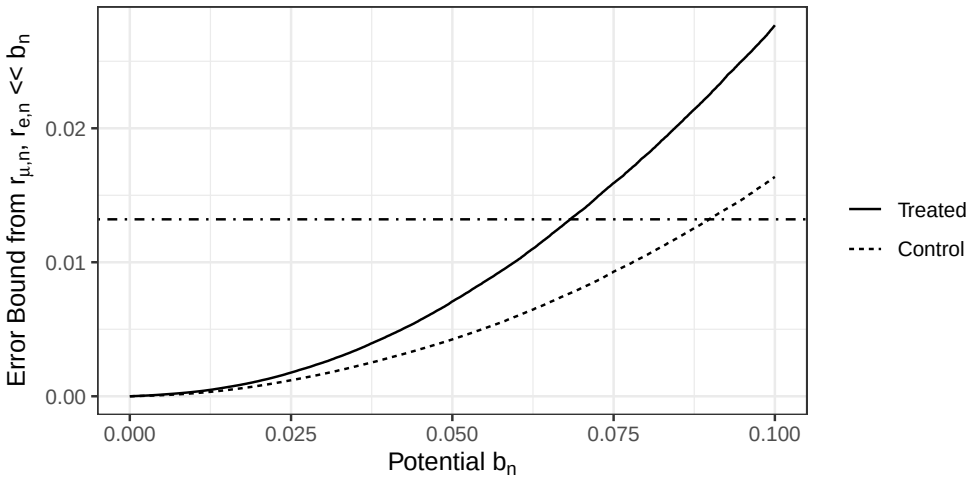

Figure 7: The value of `error_bound`(b) from Algorithm 1 for $e(x)$ and $1-e(x)$ thresholding with no specified nuisance upper bounds, which corresponds to an error bound implied by $r_{\mu,n}, r_{e,n} \ll b_n$. The procedure chooses b_n to set this function equal to $n^{-1/2}$, which corresponds to the horizontal line. $n^{-1/2}$ is indicated by horizontal dashed line. The more favorable distribution of estimated treatment propensities allows for a more aggressive clipping threshold.

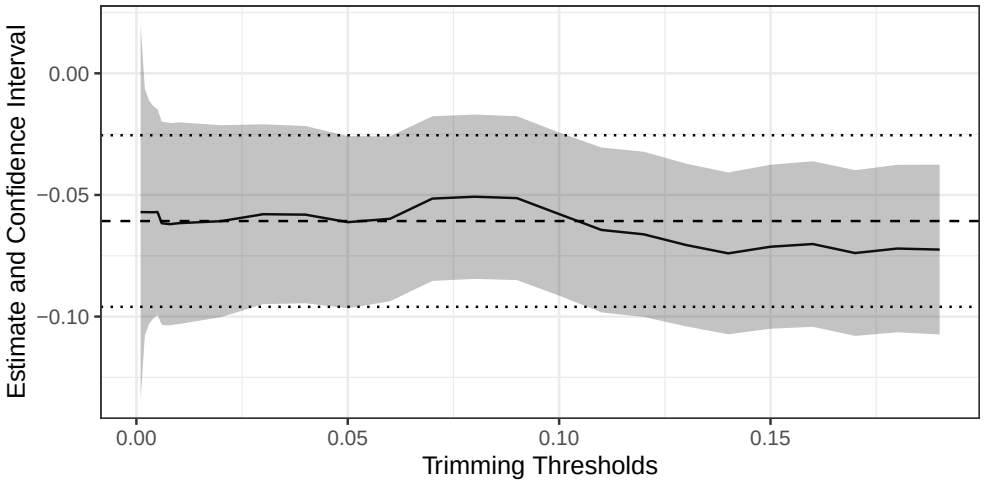

Figure 8: Estimated effects (solid line) and 95% confidence interval (shaded region) for AIPW applied to various trimmed subsamples. Estimate and confidence interval for clipped AIPW with rule-of-thumb bandwidth are represented by the dashed and dotted horizontal lines, respectively. Clipped AIPW produces similar estimates and standard errors as the fixed-trimming 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.

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 fixed-trimmed-sample AIPW estimates to a clipped AIPW estimator that targets the full population treatment effect. 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 narrower under the full-population estimator. Part of the added width is driven by inverse propensities among clipped observations: 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 5.1 suggest that trimmed AIPW may slightly undercover.

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 and adding only a small precision cost.

6 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 well in simulations, especially in large samples, and can achieve comparable precision to a fixed 10% trimming rule in practice.

These results can be extended in many interesting directions. This work exploits Neyman orthogonality to achieve standard statistical inference in the presence of a small region 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. Semenova (2024) applies thresholding strategies 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.

The results here suggest that thresholded AIPW is a viable alternative to fixed-trimming rules. I provide rules of thumb that enable practitioners to easily report results that target the population average effect. When, as in my empirical application, the fixed-trimming and sequence-of-thresholds approaches yield similar

causal conclusions, then there is strong evidence that causal conclusions are driven by causal effects, and not how the researcher treats observations with extreme propensity scores.

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.

Bruns-Smith, D., Dukes, O., Feller, A., and Ogburn, E. L. (2024). Augmented balancing weights as linear regression.

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*.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in gmm models of nonclassical measurement errors, missing data and treatment effects. Technical Report 1644, Cowles Foundation.

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).

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

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

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.

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

Horn, R. A. and Johnson, C. R. (2013). *Matrix Analysis: Second Edition*. Cambridge University Press.

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).

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. 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.

Mou, W., Ding, P., Wainwright, M. J., and Bartlett, P. L. (2023). Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, 62(6):1349–1382.

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 Key Technical Assumptions and 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 $P(n) \in \mathscr{P}$:

(a) *Outcome Consistency.* $r_{\mu,n} \ll \left(b_n E_{P(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right)^{\frac{1}{p_\mu}}$ or $\|\hat{\mu} - \mu\|^2_{L^\infty(P(n))} = o(1)$.

(b) *Asymptotically known thresholding.* $r_{e,n} \ll b_n \left(b_n E_{P(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]\right)^{\frac{1}{p_e}}$.

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

(d) *Product of errors.* $r_{\mu,n} r_{e,n} \frac{b_n^{(\gamma_0 - 1)\frac{p_\mu p_e - p_\mu - p_e}{p_e p_\mu} - 1} + \log\left(\frac{1}{b_n}\right)^{1\left\{(\gamma_0-1)\frac{p_\mu p_e - p_\mu - p_e}{p_e p_\mu} = 1\right\}}}{\sqrt{E_{P(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right]}} \ll n^{-1/2}$.

These conditions adapt to the distributions in the sequence $P(n)$. Note that condition (d) can often be relaxed for specific $P(n)$, but the form here suffices to bound the product-of-errors term using a relatively simple optimization problem.

The conditions of Assumption 3' are weaker and less interpretable than the conditions in the main text.

Corollary 5 (Sufficiency of Assumption 3'). *Suppose Assumption 2 holds, and either Assumption 4(ii) holds or both Assumption 3 and Assumption 4(i) hold. Then 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 $\tilde{\psi}^{AIPW}_{(Oracle)}(b_n)$. 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 3 (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*

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

Proposition 3 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 $P(n)$ is a sequence of distributions $P \in \mathscr{P}$. Then $\sigma_n^{-1}\left(\hat{\psi}^{AIPW}_{clip}(b_n) - \psi_n\right) \overset{P(n)}{\rightsquigarrow} N(0,1)$, where σ_n is the oracle standard deviation defined in Proposition 3.

Next, I describe the key new results for nonparametric regression. This result shows that in nonparametric regression, if the propensity function is sufficiently smooth, then nature cannot severely concentrate treated observations within a given bandwidth of any point. The non-concentration ensures that the eigenvalues of the local polynomial regression matrix are nondegenerate.

Proposition 4 (Sufficient conditions for non-trivial concentration)**.** *Suppose Assumption 6 holds and there is an* $L > 0, \beta_e > \frac{d}{\gamma_0 - 1}$ *such that* $P(D = 1 \mid X) \in \Sigma(\beta_e, L)$ *for all* $P \in \mathscr{P}$*. Then Assumption 7 holds.*

Note that $\frac{d}{\gamma_0 - 1}$ is also a key parameter in Mou et al. (2023). A broader connection is outside the scope of this work.

The next result presents the main construction involved in avoiding a polylogarithmic penalty under weak overlap. At a high-level, the idea is to eventually choose gridpoints of a grid covering $[-1, 1]^d$, and then split these gridpoints into increasingly large groups of up to $m_n^{(k)}$ gridpoints with at least $\left(h_n^{(1)} \frac{t_n^{(1))}}{t_n^{(k)}}\right)^{-2\beta_\mu}$ treated observations nearby, where for $m_n^{(k)}$ and $t_n^{(k)}$ increase in k, $h_n^{(1)}$ is an initial bandwidth $h_n^{(1)}$ I specify later, and "nearby" depends on the gridpoint's associated bandwidth. I show that under certain conditions, $m_n^{(k)}$ increases in k and increases slowly enough to ensure that the conditional expected largest local polynomial regression error in class k, which is on the order of $\log\left(m_n^{(k)}\right)\left(h_n^{(1)} \frac{t_n^{(k)}}{t_n^{(1)}}\right)^{2\beta_\mu}$, shrinks geometrically (in particular by factors of one-half). The key requirement turns out to be that the number of gridpoints in the smallest class must be *large* enough to achieve sustainable growth of $m_n^{(k)}$ to infinity.

Proposition 5 (Inductive grouping)**.** *Fix some* $\beta_\mu > 0$*,* $\gamma_0 > 1$*, and* $d \geq 1$*. Then there is a* $\underline{\delta} > 0$ *and a* $\pi \in (0, 2^{-1/\beta_\mu}]$ *such that if* $\delta \geq \underline{\delta}$*, then the sequence defined by* $m_n^{(0)} = \delta$*,* $m_n^{(1)} = exp(\delta/2)$*, and* $m_n^{(k+1)} = exp\left(\delta 2^{-(k+1)}\left(m_n^{(k)}/m_n^{(0)}\right)^{2\beta_\mu(\gamma_0-1)/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}\right)$ *diverges to infinity, and the sequence of* $t_n^{(k)} = \left(m_n^{(k-1)}/m_n^{(0)}\right)^{(\gamma_0-1)/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}$ *satisfies* $t_n^{(k+1)} \geq t_n^{(k)}/\pi$ *for* $k = 1, 2, \ldots$.

B Other Simulation Evidence

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

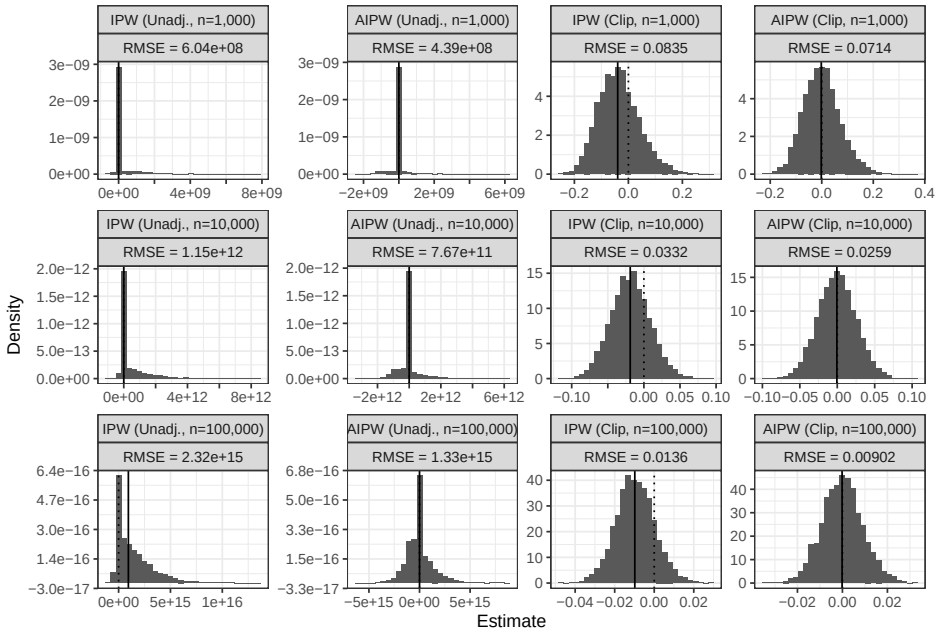

Figure 9: Histograms of point estimates in simulations for the various methods considered in the simulations, but using the oracle μ regression function instead of the estimated $\hat{\mu}$ regression function.

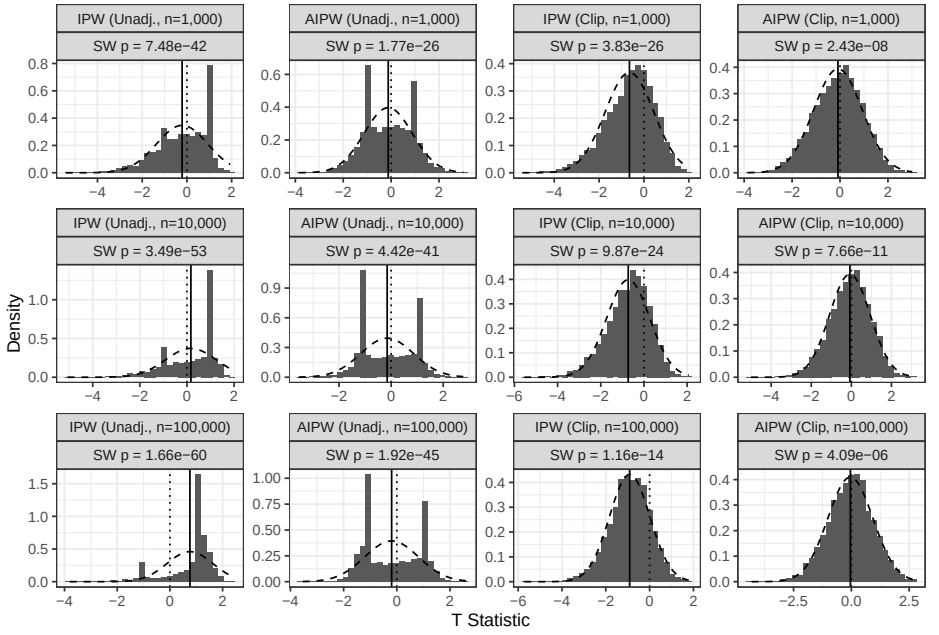

Figure 10: Histograms of simulation t-statistics for various sample sizes, but using the oracle μ regression function instead of the estimated $\hat{\mu}$ regression function.

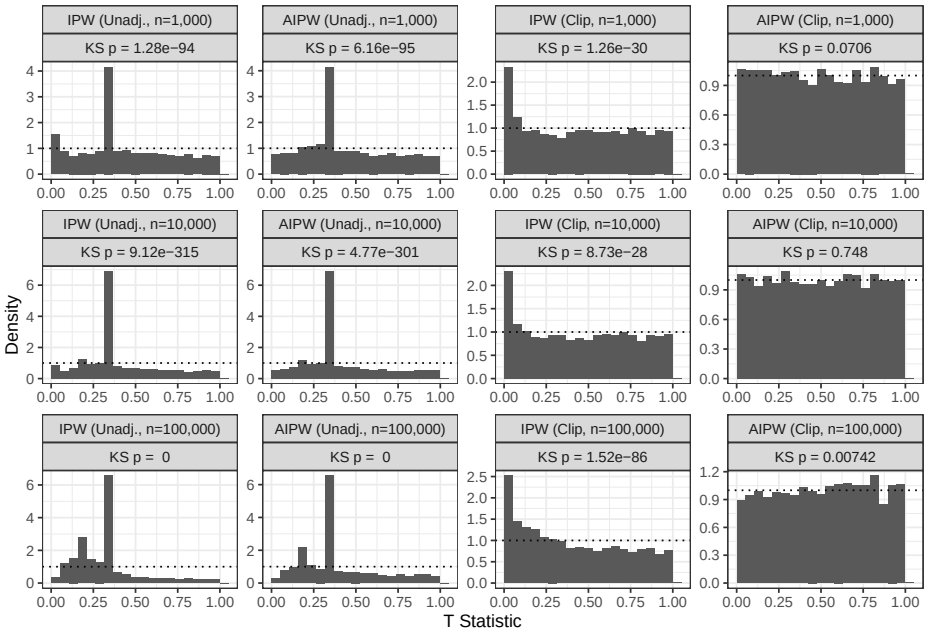

Figure 11: Histograms of simulation p-values on null hypothesis of true average potential outcome for various sample sizes, but using the oracle μ regression function instead of the estimated $\hat{\mu}$ regression function.

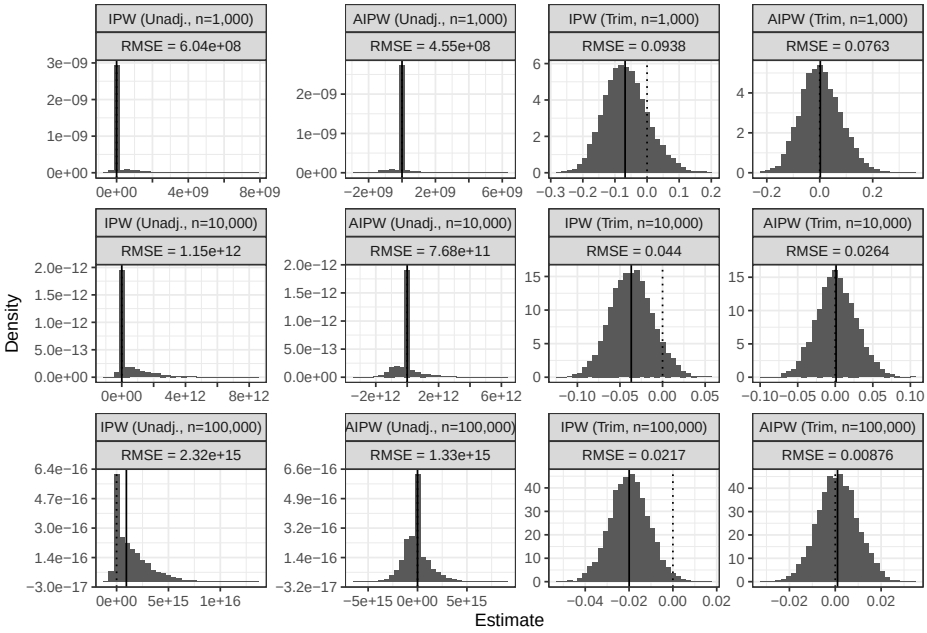

Figure 12: Histograms of point estimates in simulations for the various methods considered in the simulations, but with trimming instead of clipping.

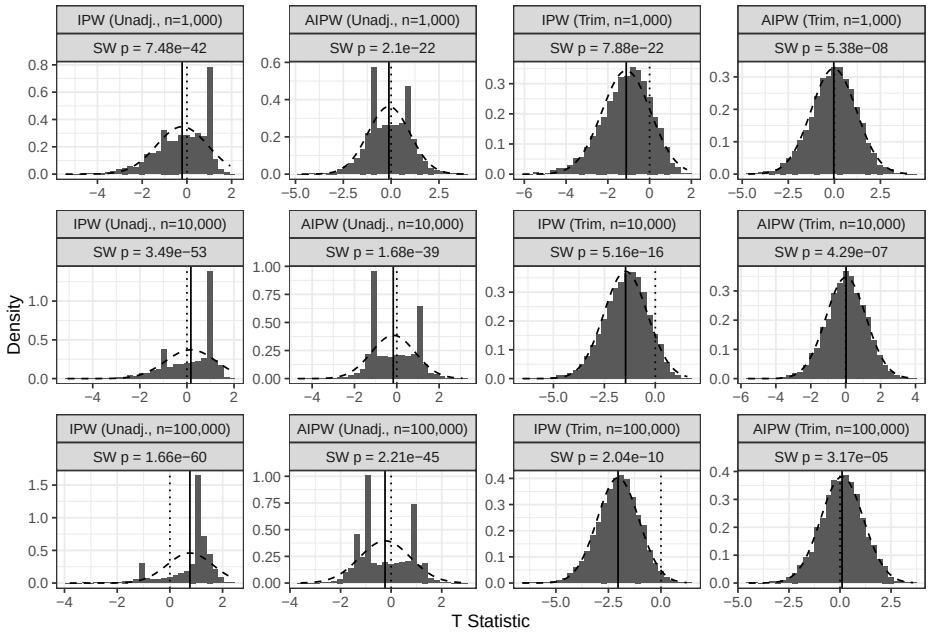

Figure 13: Histograms of simulation t-statistics for various sample sizes, but with trimming instead of clipping.

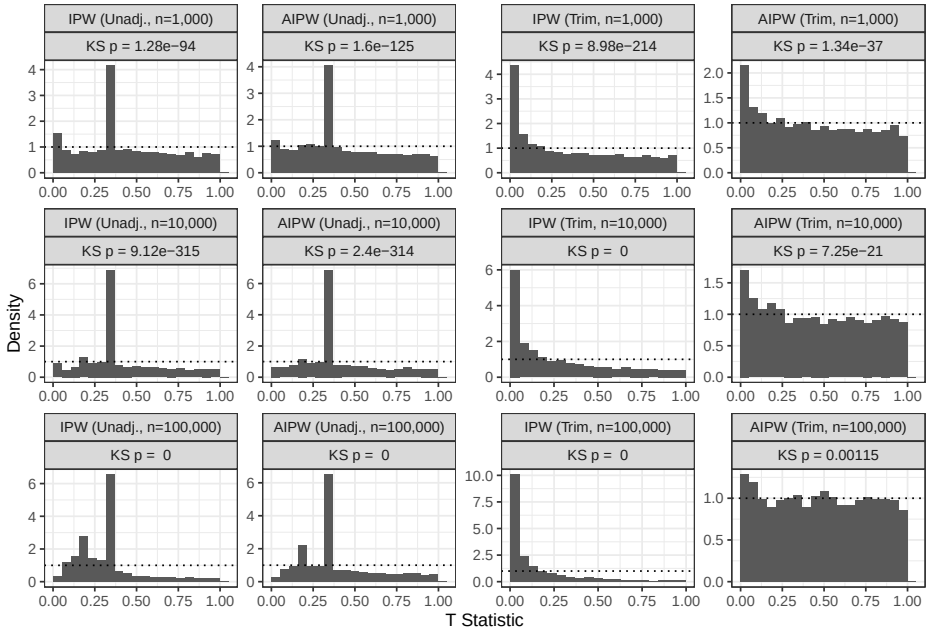

Figure 14: Histograms of simulation p-values on null hypothesis of true average potential outcome for various sample sizes, but with trimming instead of clipping.

C Proofs

The proofs, including proofs of the claims in Appendix A, are split into sections showing asymptotic properties of oracle clipped AIPW (Appendix C.1), consistency of estimated clipped AIPW (Appendix C.2), and black-box consistency rates (Appendix C.3). Note that Appendix C.3 is out of order from the perspective of the main text and corresponds to claims in Section 3.1, but Proposition 2 is used to show claims that appear earlier in the main text. Appendix C.4 presents proofs for the main claims: the black-box asymptotic properties of estimated clipped AIPW. Finally, Appendix C.5, Appendix C.6, and Appendix C.7 prove claims about black-box nuisance rates, outcome regression rates, and rules of thumb, respectively.

Additional Notation. In these proofs, I use $P(n)$ to refer to an arbitrary sequence of distributions for the purposes of computing suprema; for such sequences, I use $\psi_n = \psi(P(n))$ 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 $P(n)_n(E_n)$ 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 $c_n = o_{P(n)}(1)$ if for all $\delta > 0$, $P(n)(|c_n|/d_n > \delta) \to 0$, and if no $P(n)$ is defined, I use $c_n = o_{P(n)}(d_n)$ to mean that for any sequence of $P(n) \subset \mathscr{P}$, $c_n = o_{P(n)}(d_n)$. I write $c_n = O_{P(n)}(1)$ if for all $\epsilon > 0$, there exists a $\delta > 0$ such that $P(n)(|c_n|/d_n > \delta) < \epsilon$. If there is a sequence of distributions to be considered, then I use $o(d_n)$ and $O(d_n)$ to implicitly refer to $o_{P(n)}(d_n)$ and $O_{P(n)}(d_n)$. I write that $c_n \overset{P(n)}{\rightsquigarrow} N(0,1)$ if $\sup_{t\in\mathbb{R}} |P(n)(c_n \le t) - \Phi(t)| \to 0$, where Φ is the standard normal cumulative distribution function; I write that $c_n \to_{P(n)} c$ if $c_n - c = o_{P(n)}(1)$; and I write that $c_n \xrightarrow{\mathscr{P}} c$ if for all sequences of $P(n) \in \mathscr{P}$, $c_n \to_{P(n)} c$. I write $c_n = \Theta(d_n)$ if there exists a $k_1, k_2 > 0$ such that $P(n)[c_n \in [k_1 d_n, k_2 d_n]] \to 1$, and I write $c_n = \Omega(d_n)$ if there exists a $k_1 > 0$ such that $P(n)[c_n \ge k_1 d_n] \to 0$.

C.1 Oracle Normality

Lemma 2. *Define* $\pi_{\min} = 2^{\frac{-\gamma_0}{\gamma_0-1}} C^{\frac{-1}{\gamma_0-1}}$. *Then* $P(e(X) \ge 2\pi_{\min}) \ge 1/2$ *and* $\inf_{P\in\mathscr{P}} P(D=1) \ge \pi_{\min} > 0$.

Proof of Lemma 2. For any $P \in \mathscr{P}$, $P(e(X) \le 2\pi_{\min}) \le C\left((2C)^{\frac{-1}{\gamma_0-1}}\right)^{\gamma_0-1} = 1/2$, so that $P(D=1) \ge P(e(X) \ge 2\pi_{\min}, D=1) \ge (1/2)(2\pi_{\min}) = \pi_{\min}$. □

Lemma 3. *Assume* $b_n \to 0$. *Then for all large* n, *the following inequalities hold throughout* $\mathscr{P}$:

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

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

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

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

Proof of Lemma 3. I take these proofs one at a time.

(i) Start from the following inequalities:

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

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:

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

(iii) By Jensen's inequality:

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

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

(iv) By Jensen's inequality:

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

$$\leq 2^{q-1}(4M)^q E[e(X)/\{e(X) \vee b_n\}^2]/b_n^{q-2} + 2^{q-1}M^q.$$

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

□

Lemma 4. *Let $c(\gamma) = \frac{\gamma-1}{\gamma} C^{-1/(\gamma-1)} > 0$. Then for any $P \in \mathscr{P}$,*

$$E_P[e(X)\mathbf{1}\{e(X) \leq b_n\}] \geq c(\gamma) P(e(X) \leq b_n)^{\gamma/(\gamma-1)}. \tag{7}$$

This lower bound is attained when $P(e(X) \leq t) = Ct^{\gamma_0 - 1}$.

Proof of Lemma 4. Let $p = P(e(X) \leq b_n)$. If $p = 0$, then the bound holds trivially so I will assume throughout that $p > 0$. Then I may write:

$$\begin{aligned}
E_P[e(X)\mathbf{1}\{e(X) \leq b_n\}] &= \int_0^\infty P(e(X)\mathbf{1}\{e(X) \leq b_n\} > t)dt = \int_0^{b_n} P(t < e(X) \leq b_n)dt = \int_0^{b_n} p - P(e(X) \leq t)dt \\
&= b_n p - \int_0^{b_n} P(e(X) \leq t)dt \geq b_n p - \int_0^{b_n} \min\{p, Ct^{\gamma_0 - 1}\}dt \\
&= b_n p - \int_0^{\left(\frac{p}{c}\right)^{\frac{1}{\gamma_0 - 1}}} Ct^{\gamma_0 - 1}dt - \int_{\left(\frac{p}{c}\right)^{\frac{1}{\gamma_0 - 1}}}^{b_n} pdt = b_n p - \frac{p^{\frac{\gamma_0}{\gamma_0 - 1}} C^{\frac{-1}{\gamma_0 - 1}}}{\gamma_0} - b_n p + p^{\frac{\gamma_0}{\gamma_0 - 1}} C^{\frac{-1}{\gamma_0 - 1}} = c(\gamma_0) p^{\frac{\gamma_0}{\gamma_0 - 1}}.
\end{aligned}$$

This proves the lower bound. When $P(e(X) \leq t) = Ct^{\gamma_0 - 1}$, a direct calculation gives $E_P[e(X)\mathbf{1}\{e(X) \leq b_n\}] = [(\gamma_0 - 1)/\gamma_0] b_n^{\gamma_0} = [(\gamma_0 - 1)/\gamma_0] P(e(X) \leq b_n)^{\gamma_0/(\gamma_0 - 1)}$. Therefore, the lower bound is also sharp. □

Lemma 5. *For any $P \in \mathscr{P}$,*

$$\begin{aligned}
\mathrm{Var}_P(\phi(Z \mid b_n, \eta)) &\geq \sigma_{\min}^2 E_P\left[e(X)/\max\{e(X), b_n\}^2\right] \\
&\geq \sigma_{\min}^2 [c(\gamma_0) P(e(X) \leq b_n)^{\gamma_0/(\gamma_0 - 1)}/b_n^2 + \pi_{\min}/2] \\
&\geq \sigma_{\min}^2 \pi_{\min}/2 > 0.
\end{aligned}$$

Proof of Lemma 5. For the first line:

$$\begin{aligned}
\mathrm{Var}_P(\phi_n) &= E[\mathrm{Var}(\phi_n \mid X)] + \mathrm{Var}(E[\phi_n \mid X]) = E[\mathrm{Var}(\phi_n \mid X)] \\
&= E\left[E[|Y - \mu(X)|^2 \mid X, D = 1] \frac{e(X)}{\{e(X) \vee b_n\}^2}\right] \geq \sigma_{\min}^2 E\left[\frac{e(X)}{\{e(X) \vee b_n\}^2}\right].
\end{aligned}$$

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

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

$$\begin{aligned}
E[e(X)/\max\{e(X), b_n\}^2] &= E[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + E[1/e(X)\mathbf{1}\{e(X) > b_n\}] \\
&\geq E[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + P(e(X) > b_n) \\
&\geq E[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + P(e(X) > \pi_{\min}) \\
&\geq E[e(X)/b_n^2 \mathbf{1}\{e(X) \leq b_n\}] + \pi_{\min}/2 && \text{(Lemma 3.(i))} \\
&\geq c(\gamma_0)(1/b_n)^2 P(e(X) \leq b_n)^{\gamma_0/(\gamma_0 - 1)} + \pi_{\min}/2. && \text{(Lemma 4)}
\end{aligned}$$

The final line is immediate. □

Lemma 6. $\frac{1}{\sqrt{Var(\phi(Z|b_n,\eta))}} \leq \frac{1}{\sqrt{\sigma_{\min}^2 E_{P(n)}[D/\max\{e(X),b_n\}^2]}}$.

Proof of Lemma 6. By Lemma 5, I have:

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

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

Lemma 7. *Define* $\tilde{\phi}(Z \mid b, P) \equiv \phi(Z \mid b, \eta(P)) - E_P[\mu(X)]$ *for* $P \in \mathscr{P}$. *Further define* $\rho(b, P) \equiv E_P[|\tilde{\phi}(Z \mid b, P)|^3]$ *and* $\sigma(b, P) \equiv \sqrt{Var_P(\tilde{\phi}(Z \mid b, P))}$.

Then the following hold:

1. $E_P[\tilde{\phi}(Z \mid b, P)] = 0$
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* $P(n)$ *be a sequence of distributions in* $\mathscr{P}$, *then* $\frac{\rho(b_n, P(n))}{\sigma(b_n, P(n))^3 \sqrt{n}} = o(1)$.

Proof of Lemma 7. $E_P[\tilde{\phi}(Z \mid b_n, P)] = 0$ is immediate.

$Var_P[\tilde{\phi}(Z \mid b, P)] > 0$ follows by Lemma 5.

For the third moment being finite:

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

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

Finally, I have the claim for sequences. Recall that by Lemmas 5 and 3, $\frac{1}{\sigma(b_n,P(n))^3\sqrt{n}} = o(1)$ and $E_{P(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] \geq \sigma^2_{\min}/2$. As a result:

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

□

Proof of Proposition 3. Let $P(n)$ be a sequence of distributions in $\mathscr{P}$. By Lemma 7 and the Berry Esseen Theorem, the difference between the CDF of oracle clipped AIPW t-statistic $\frac{\tilde{\psi}^{AIPW}_{clip}-\psi_n}{\sigma_n} = \frac{\sum\tilde{\phi}(Z|b_n,P(n))}{\sqrt{Var(\phi(Z|b_n,\eta))}\sqrt{n}}$ and the standard normal CDF Φ is uniformly bounded above by $\frac{3\rho(b_n,P(n))}{\sigma(b_n,P(n))^3\sqrt{n}}$. By Lemma 7.4, this difference tends to zero. Therefore:

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

□

C.2 Consistency

Lemma 8 (Worst-case distribution simplification)**.** *Suppose p is finite. Then:*

$$\sup_{P\in\mathscr{P}}\sup_{E_P[|\hat{e}-e|^{p_e}]\leq r^{p_e}_{e,n}} E_P\left[\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}}-1\right|^p\right] = \sup_{P\in\mathscr{P},d_e\in[0,e-b_n]} E_P\left[\left(\frac{\min\{d_e,e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right] \quad s.t. \quad \begin{matrix} e\geq b_n \ a.s., \\ E_P\left[d_e^{p_e}\right]\leq r^{p_e}_{e,n}\end{matrix}.$$

Proof of Lemma 8. First, suppose $P(e(X)\leq b_n)>0$ for some $P\in\mathscr{P}$. Then consider P' constructed by drawing $(X,e,\hat{e},y)\sim P$ and returning $(X,e+\max\{\hat{e}-b_n,0\},\hat{e}+\max\{\hat{e}-b_n,0\},y)$. It is clear that P' is in the constraint set because $P'(e\leq\pi)\leq P(e\leq\pi)$ and $E_{P'}[|\hat{e}-e|^{p_e}] = E_P[|\hat{e}-e|^{p_e}]$. But also not that $\left|\frac{\max\{e+\max\{\hat{e}-b_n,0\},b_n\}}{\max\{\hat{e}+\max\{\hat{e}-b_n,0\},b_n\}}-1\right|$ is equal to $\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}}-1\right|$ if $e\geq b_n$ or $\hat{e}+\max\{\hat{e}-b_n,0\},b_n\}\leq b_n$, and is equal to $\left|\frac{\max\{e,b_n\}}{\hat{e}+\max\{\hat{e}-b_n,0\},b_n\}}-1\right| = 1-\frac{b_n}{\hat{e}+\max\{\hat{e}-b_n,0\},b_n\}}\geq 1-\frac{b_n}{\hat{e}}$ otherwise. Therefore $\sup_{P\in\mathscr{P}}\sup_{E_P[|\hat{e}-e|^{p_e}]\leq r^{p_e}_{e.n}} E_P\left[\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}}-1\right|^p\right]$ $\sup_{P\in\mathscr{P}}\sup_{E_P[|\hat{e}-e|^{p_e}]\leq r^{p_e}_{e.n}} E_P\left[\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}}-1\right|^p\right]$ s.t. $e\geq b_n$ a.s..

Now consider P with $e\geq b_n$ almost surely. Then consider P' constructed by drawing $(X,e,\hat{e},y)\sim P$ and returning $(X,\hat{e},e,y)$. It is once again clear that P' is in the constraint set because $P'(e\leq\pi)\leq P(e\leq\pi)$ and

$E_{P'}[|\hat{e}-e|^{p_e}] = E_P[|\hat{e}-e|^{p_e}]$. But notice also that P' weakly increases $\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}} - 1\right| = \frac{|\max\{e,b_n\}-\max\{\hat{e},b_n\}|}{\max\{\hat{e},b_n\}}$. Therefore $\sup_{P\in\mathscr{P}} \sup_{E_P[|\hat{e}-e|^{p_e}]\leq r_{e.n}^{p_e}} E_P\left[\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}} - 1\right|^p\right] = \sup_{P\in\mathscr{P}} \sup_{E_P[|\hat{e}-e|^{p_e}]\leq r_{e.n}^{p_e}} E_P\left[\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}} - 1\right|^p\right]$ s.t. $e \geq b_n$ a.s., $\hat{e} \leq e$ a.s..

But then in this new problem, adding a constraint $d_e = e - \hat{e} \geq e - b_n$ yields the desired optimization problem and has no effect on the objective. □

Lemma 9 (Expectation constraint for scaling propensity error)**.** *Suppose $b_n \to 0$ and $\pi^* \geq b_n$ solves $(\pi^* - b_n)^{p_e} C(\pi^*)^{\gamma_0 - 1} = r_{e,n}^{p_e}$. Then $\pi^* - b_n = O\left(r_{e,n}\left(b_n^{\frac{-(\gamma_0-1)}{p_e}} + r_{e,n}^{\frac{-(\gamma_0-1)}{p_e+\gamma_0-1}}\right)\right)$ and for every finite p,*

$$\sup_{P\in\mathscr{P}, d_e\geq 0} E_{P^*}\left[\left(\frac{d_e}{e/2}\right)^p\right] \text{ s.t. } d_e \leq \pi^* - b_n, E_{P^*}\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e} = O\left(r_{e,n}^p b_n^{-p}\left(\frac{r_{e,n}}{\pi^* - b_n}\right)^{p_e - p} + r_{e,n}^p \log\left(\frac{1}{b_n}\right)^{1\left\{(\gamma_0-1)\frac{p_e-p}{p_e} = p\right\}}\right).$$

Proof of Lemma 9. First, I bound the order of $\pi^* - b_n$. Notice that if $\pi^* \leq 2b_n$, then $r_{e,n}^{p_e} \geq = C 2^{\gamma_0 - 1}(\pi^* - b_n)^{p_e} b_n^{\gamma_0 - 1}$, so that $\pi^* - b_n = O\left(r_{e,n} b_n^{-(\gamma_0-1)/p_e}\right)$. On the other hand, if $\pi^* \geq 2b_n$, then $r_{e,n}^{p_e} \geq C 2^{-p_e}(\pi^*)^{p_e+\gamma_0-1}$, so that $\pi^* = O\left(r_{e,n}^{1-\frac{\gamma_0-1}{p_e+\gamma_0-1}}\right)$. As a result, $\pi^* - b_n = O\left(r_{e,n}\left(b_n^{\frac{-(\gamma_0-1)}{p_e}} + r_{e,n}^{\frac{-(\gamma_0-1)}{p_e+\gamma_0-1}}\right)\right)$.

Now, write the worst-case objective as B_9. The relevant Lagrangian is $d_e^p \pi^{-p} - \frac{p}{p_e}\lambda^{p-p_e}\left(d_e^{p_e} - r_{e,n}^{p_e}\right)$, with first-order condition $d_e^{p_e - p} = \lambda^{p_e - p}\pi^{-p}$. Thus, for the worst case d_e, $d_e = \lambda\pi^{\frac{-p}{p_e-p}}$ and $d_e^p \pi^{-p} = \lambda^p\left(\pi^{\frac{-p-(p_e-p)}{p_e-p}}\right)^p = \lambda^p\left(\pi^{\frac{-p}{p_e-p}}\right)^{p_e} = \lambda^{p-p_e} d_e^{p_e}$.

Therefore a worst-case objective is $\int_{\pi^*}^1 d_e^p \pi^{-p}\pi^{\gamma_0-2}d\pi = \int_{\pi^*}^1 \lambda^{p-p_e} d_e^{p_e}\pi^{\gamma_0-2}d\pi = \lambda^{p-p_e} r_{e,n}^{p_e}$, where to ensure $\int d_e^{p_e}\pi^{\gamma_0-2}d\pi = r_{e,n}^{p_e}$, the Lagrangian multiplier is $\lambda = r_{e,n}\left(\int_{\pi^*}^1 \pi^{\gamma_0-2-p\frac{p_e}{p_e-p}}d\pi\right)^{-1/p_e}$. Therefore: $B_9 = O\left(r_{e,n}^p\left(\int_{\pi^*}^1 \pi^{\gamma_0-2-p\frac{p_e}{p_e-p}}d\pi\right)^{(p_e-p)/p_e}\right)$.

If $(\gamma_0 - 1)\frac{p_e-p}{p_e} = p$, then $B_9 = O\left(r_{e,n}^p\left(\int_{\pi^*}^1 \pi^{\gamma_0-2-p\frac{p_e}{p_e-p}}d\pi\right)^{(p_e-p)/p_e}\right) = O\left(r_{e,n}^p \log(1/\pi^*)^{(p_e-p)/p_e}\right) = O\left(r_{e,n}^p \log(1/b_n)\right)$ because $b_n \leq \pi^*$ and $(p_e - p)/p_e \in [0,1]$. Similarly, if $(\gamma_0 - 1)\frac{p_e-p}{p_e} > p$ so that $\gamma_0 - 1 - p\frac{p_e}{p_e-p} > 0$, then $B_9 = O\left(r_{e,n}^p\right) = O\left(r_{e,n}^p \log(1/b_n)\right)$. Finally, if $(\gamma_0 - 1)\frac{p_e-p}{p_e} < 1$, then:

$$B_9 = O\left(r_{e,n}^p(\pi^*)^{(\gamma_0-1)\frac{p_e-p}{p_e}-p}\right) = O\left(r_{e,n}^p b_n^{-p}(\pi^*)^{\left((\gamma_0-1)\frac{p_e-p}{p_e}\right)}\right) = O\left(r_{e,n}^p b_n^{-p}\left(\frac{r_{e,n}}{\pi^* - b_n}\right)^{p_e-p}\right).$$

□

Lemma 10 (Scaled propensity error bound)**.** *Fix some $p \in (1, p_e]$ and write $\tilde{\zeta} = (\gamma_0 - 1)\left(1 - \frac{p}{p_e}\right)$. Then if $b_n \to 0$:*

$$\sup_{P\in\mathscr{P}} \sup_{\|\hat{e}-e\|_{L^{p_e}(P)}\leq r_{e,n}} \left\|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}} - 1\right\|_{L^p(P)} = O\left(r_{e,n}\left(b_n^{\frac{\tilde{\zeta}}{p}-1} + \log(1/b_n)^{1\{\tilde{\zeta}=p\}/p} + r_{e,n}^{\frac{\tilde{\zeta}}{p}\frac{p_e}{p_e+\gamma_0-1}} b_n^{-p}\right)\right).$$

In the special case of $p = p_e$, this bound is $O\left((r_{e,n}/b_n)\right)$.

Proof of Lemma 10. First, suppose p is infinite, so that p_e is infinite and $\tilde{\zeta} = 0$. Then the largest feasible value of $\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}} - 1\right|$ is $e = b_n + r_{e,n}$ and $\hat{e} = b_n$, yielding $\left|\frac{\max\{e,b_n\}}{\max\{\hat{e},b_n\}} - 1\right| = r_{e,n}/b_n$ as claimed. Now proceed assuming p is finite.

By Lemma 8, it suffices to bound

$$\text{“}B_{10}\text{”} = \sup_{P\in\mathscr{P},d_e} E_P\left[\left(\frac{\min\{d_e, e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right] \text{ s.t. } e \geq b_n \text{ a.s.}, d_e \in [0, e-b_n] \text{ a.s.}, E_P\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e}.$$

It is clear that the claim holds by $B = O\left(r_{e,n}^p b_n^{-p}\right)$ if $p_e = p > \tilde{\zeta}$, so I proceed assuming $p_e > p$.

Now let π^* be the largest propensity at which P can place a point mass with $\hat{e} - b_n$ and $E_P[(\hat{e}-b_n)^{p_e}] \leq r_{e,n}^{p_e}$, i.e. , π^* is as in Lemma 9. Recall that $\pi^* - b_n = O\left(r_{e,n}\left(b_n^{\frac{-(\gamma_0-1)}{p_e}} + r_{e,n}^{\frac{-(\gamma_0-1)}{p_e+\gamma_0-1}}\right)\right)$.

Let P^* be a distribution in $\mathscr{P}$ with $P^*(e(X) \leq \pi) = \min\{1, C\pi^{\gamma_0-1}1\{\pi \geq \pi^*\}\}$, i.e. which has the smallest distribution of propensities subject to $e(X) \geq \pi^*$ almost surely. I then split the optimization problem between the events (i) with d_e either above $\pi^* - b_n$ or equal to zero, and events with $d_e \in [0, \pi^* - b_n]$ and either (ii) $e \leq 2\pi^*$ or (iii) $e \geq 2\pi^*$. These events are mutually exclusive. Then (with almost sure caveats implicit for space):

$$\begin{aligned}
B_{10} \leq & \sup_{P\in\mathscr{P},d_e} E_P\left[\left(\frac{\min\{d_e, e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right] \text{ s.t. } e \geq b_n, d_e \in [0, e-b_n], E_P\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e} \\
\leq & \sup_{P\in\mathscr{P},d_e} E_P\left[\left(\frac{\min\{d_e, e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right] \text{ s.t. } e \geq b_n, d_e \notin (0, \pi^* - b_n), E_P\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e} \qquad \text{(i)} \\
& + \sup_{P\in\mathscr{P},d_e} E_P\left[\left(\frac{\min\{d_e, e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right] \text{ s.t. } e \geq b_n, d_e \in [0, (\pi^* - b_n)1\{e \leq 2\pi^*\}\}], E_P\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e} \\
& \qquad \text{(ii)} \\
& + \sup_{P\in\mathscr{P},d_e} E_P\left[\left(\frac{\min\{d_e, e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right] \text{ s.t. } e \geq b_n, d_e \in [0, (\pi^* - b_n)1\{e \geq 2\pi^*\}], E_P\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e}. \\
& \qquad \text{(iii)}
\end{aligned}$$

The term (i) is bounded above by constraining $P = P^*$: on the event $e(X) < \pi^*$, the objective of (i) can be increased by increasing $e(X)$ to π^* and not changing d_e, and if $P(e(X) \leq \pi) < C\pi^{\gamma_0-1}$ for some $\pi \in [\pi^*, C^{1/(1-\gamma_0)}]$, then the objective of (i) can be increased by reducing those values of $e(X)$ without changing the value of d_e. Then the remaining problem $\sup_{d_e \geq} E_{P^*}\left[\left(\frac{\min\{d_e,e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right]$ s.t. $d_e \notin (0, \pi^* - b_n), E_{P^*}\left[d_e^{p_e}\right] \leq r_{e,n}^{p_e}$ is solved by setting $d_e = (\pi^* - b_n)1\{e = \pi^*\}$: this distribution has $E_{P^*}\left[d_e^{p_e}\right] = r_{e,n}^{p_e}$ by construction, and any other assignment of d_e would place d_e less adversarially and have a smaller value of $E_{P^*}\left[\left(\frac{\min\{d_e,e-b_n\}}{e-\min\{d_e,e-b_n\}}\right)^p\right]$. But then $(i) = P^*(e(X) = \pi^*)\left(\frac{\pi^*-b_n}{\pi^*-(\pi^*-b_n)}\right)^p = C(\pi^*)^{\gamma_0-1}\left(\frac{\pi^*-b_n}{b_n}\right)^p$. The term

(ii) is bounded by the same order via a simpler argument:

$$(ii) \leq \sup_{P\in\mathscr{P},d_e} E_P\left[1\{e\leq 2\pi^*\}\left(\frac{\pi^*-b_n}{b_n}\right)^p\right] = C(2\pi^*)^{\gamma_0-1}\left(\frac{\pi^*-b_n}{b_n}\right)^p = O\left((\pi^*)^{\gamma_0-1}\left(\frac{\pi^*-b_n}{b_n}\right)^p\right).$$

For the term (iii), note that $e\geq 2\pi^*$, $d_e\leq \pi^*-b_n$ implies $e-\min\{d_e, e-b_n\}\geq e-(\pi^*-b_n)\geq e-\pi^*\geq e/2$. Also note that one can replace the first two constraints with $e\geq 2\pi^*, d_e\in[0,\pi^*-b_n]$ without changing the objective, and then replace $e\geq 2\pi^*$ with $e\geq\pi^*$ while weakly increasing the objective. As a result,

$$\begin{aligned} B_{10} &\leq O\left((\pi^*)^{\gamma_0-1}\left(\frac{\pi^*-b_n}{b_n}\right)^p\right) + \sup_{P\in\mathscr{P},d_e\geq 0} E_{P^*}\left[\left(\frac{d_e}{e/2}\right)^p\right] \text{ s.t. } d_e\leq\pi^*-b_n, E_{P^*}\left[d_e^{p_e}\right]\leq r_{e,n}^{p_e} \\ &= O\left(r_{e,n}^p b_n^{-p}\left(\frac{r_{e,n}}{\pi^*-b_n}\right)^{p_e-p} + r_{e,n}^p\log\left(\frac{1}{b_n}\right)^{1\{\tilde{\zeta}=p\}}\right). \qquad \text{(Lemma 9)} \end{aligned}$$

Notice that

$$\begin{aligned} (\pi^*)^{\gamma_0-1}\left(\frac{\pi^*-b_n}{b_n}\right)^p &= (\pi^*)^{\gamma_0-1}(\pi^*-b_n)^{p_e}(\pi^*-b_n)^{p-p_e}b_n^{-p} = C^{-1}r_{e,n}^{p_e}b_n^{-p}(\pi^*-b_n)^{p-p_e} \\ &= O\left(r_{e,n}^p b_n^{-p}\left(\frac{r_{e,n}}{\pi^*-b_n}\right)^{p_e-p}\right) = O\left(r_{e,n}^p b_n^{-p}\left(b_n + r_{e,n}^{\frac{p_e}{p_e+\gamma_0-1}} + b_n^{-p}\log\left(\frac{1}{b_n}\right)^{1\{\tilde{\zeta}=p\}}\right).\right). \end{aligned}$$

As a result, $B_{10} = O\left(r_{e,n}^p b_n^{-p}\left(\frac{r_{e,n}}{\pi^*-b_n}\right)^{p_e-p} + r_{e,n}^p\log\left(\frac{1}{b_n}\right)^{1\{\tilde{\zeta}=p\}}\right)$. But then because $b_n, r_{e,n} = O(1)$,

$$B_{10} = O\left(r_{e,n}^p b_n^{-p}\left(\frac{r_{e,n}}{\pi^*-b_n}\right)^{p_e-p} + r_{e,n}^p\log\left(\frac{1}{b_n}\right)^{1\{\tilde{\zeta}=p\}}\right) = O\left(r_{e,n}^p b_n^{-p}\left(b_n^{\tilde{\zeta}} + r_{e,n}^{\frac{\tilde{\zeta}p_e}{p_e+\gamma_0-1}} + b_n^p\log\left(\frac{1}{b_n}\right)^{1\{\tilde{\zeta}=p\}}\right)\right).$$

Taking this bound to the power of $1/p$, which is feasible because I am proceeding under the assumption that p is finite, completes the claim. □

Lemma 11. *Under cross-fitting, for any sequence of $P(n)\in\mathscr{P}$, there is the bias bound in terms of $\zeta = (\gamma_0-1)\frac{p_\mu p_e - p_\mu - p_e}{p_e p_\mu}$ with probability tending to one:*

$$\left|E_{P(n)}\left[\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n \mid \hat{\mu},\hat{e}\right]\right| = O\left(\tilde{r}_{\mu,n}P(n)(e(X)\leq b_n)^{\frac{p_\mu-1}{p_\mu}} + r_{\mu,n}r_{e,n}\left(b_n^{\zeta-1} + \log\left(\frac{1}{b_n}\right)^{1\{\zeta=1\}} + r_{e,n}^{\zeta\frac{p_e}{p_e+\gamma_0-1}}b_n^{-1}\right)\right),$$

where $\tilde{r}_{\mu,n} = \min\left\{r_{\mu,n}, P(e(X)\leq b_n)^{1/p_\mu}\right\}$ and the constant is uniform over $P\in\mathscr{P}$.

Proof of Lemma 11. Let a sequence of $P(n)$ be given and left implicit. Proceed under the high probability event under Assumption 2.

Recall that by the assumption $p_\mu, p_e \geq 2$, so that $\frac{p_\mu}{p_\mu - 1} \leq p_e$, and:

$$\begin{aligned}
\left| E\left[\hat{\psi}_{clip}^{AIPW}(b_n) - \psi_n \mid \hat{\mu}, \hat{e}\right]\right| &= \left| E\left[\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{clip}^{AIPW,(k)}(b_n) \mid \hat{\mu}, \hat{e}\right]\right| \\
&= \left| E\left[(\hat{\mu} - \mu)\frac{\max\{e, b_n\} - D}{\max\{e, b_n\}} + (Y - \hat{\mu})\left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}}\right) \mid \hat{\mu}, \hat{e}\right]\right| \\
&\leq E\left[|\hat{\mu} - \mu|\left(1\{e \leq b_n\} + |\max\{e, b_n\}/\max\{\hat{e}, b_n\} - 1|\right) \mid \{\hat{\mu}, \hat{e}\}\right] \\
&\leq r_{\mu,n}\left(P(e(X) \leq b_n)^{\frac{p_\mu - 1}{p_\mu}} + \|\max\{e, b_n\}/\max\{\hat{e}, b_n\} - 1\|_{L^{\frac{p_\mu}{p_\mu - 1}}(P)}\right) \\
&= O\left(r_{\mu,n}P(e(X) \leq b_n)^{\frac{p_\mu - 1}{p_\mu}} + r_{\mu,n}r_{e,n}\left(b_n^{\zeta - 1} + \log\left(\frac{1}{b_n}\right)^{1\{\zeta = 1\}} + r_{e,n}^{\zeta\frac{p_e}{p_e + \gamma_0 - 1}}b_n^{-1}\right)\right),
\end{aligned}$$

with the final line holding by applying Lemma 10 to $p = \frac{p_\mu}{p_\mu - 1} \leq p_e$, where $\frac{\tilde{\zeta}}{p} = (\gamma_0 - 1)\left(\frac{1}{p} - \frac{1}{p_e}\right) = \zeta$. More generally, the outcome boundedness assumption implies $E\left[|\hat{\mu} - \mu|1\{e \leq b_n\}\right] = O\left(P(e \leq b_n)\right)$, so I may substitute $r_{\mu,n}$ with $\tilde{r}_{\mu,n}$ in the final line. □

Proof of Proposition 1. Let $P(n)$ be a sequence of distributions in $\mathscr{P}$, and fix some $k \in 1, ..., K$. If $r_{\mu,n} \not\to 0$, then without loss of generality take $p_\mu = \infty$ because of the uniform bound Assumption 2. Take the high probability event under Assumption 2.

I first show the bias tends to zero. Recall that by Lemma 11,

$$E_{P(n)}\left[\hat{\psi}_{clip,k}^{AIPW}(b_n) - \tilde{\psi}_{clip}^{AIPW,(k)}(b_n)\right] = O\left(\tilde{r}_{\mu,n}P(n)(e(X) \leq b_n)^{\frac{p_\mu - 1}{p_\mu}} + r_{\mu,n}r_{e,n}\left(b_n^{\zeta - 1} + \log\left(\frac{1}{b_n}\right)^{1\{\zeta = 1\}} + r_{e,n}^{\zeta\frac{p_e}{p_e + \gamma_0 - 1}}b_n^{-1}\right)\right)$$

with $\zeta = (\gamma_0 - 1)\frac{p_\mu p_e - p_\mu - p_e}{p_e p_\mu}$ and $\tilde{r}_{\mu,n} = \min\{r_{\mu,n}, P(e(X) \leq b_n)^{1/b_n}\}$. By construction, $\tilde{r}_{\mu,n}$ is bounded and $b_n \to 0$, so that the first term is $o(1)$. Also when $\zeta = 1$ and either condition (i) or (ii) hold, then there is an $\eta > 0$ such that $r_{\mu,n}r_{e,n} \ll b_n^\eta = o\left(\log(1/b_n)\right))$, so that I obtain the bias bound:

$$E_{P(n)}\left[\hat{\psi}_{clip,k}^{AIPW}(b_n) - \tilde{\psi}_{clip}^{AIPW,(k)}(b_n)\right] = O\left(r_{\mu,n}r_{e,n}b_n^{\zeta - 1} + r_{\mu,n}r_{e,n}^{1 + \zeta\frac{p_e}{p_e + \gamma_0 - 1}}b_n^{-1}\right) + o(1) = \text{“}B\text{”} + o(1).$$

Suppose condition (i) holds, so that $r_{e,n}^{1 + \zeta\frac{p_e}{p_e + \gamma_0 - 1}} \ll b_n$. Then there is a $\bar{r}_{e,n} \gtrsim r_{e,n}$ such that $\max\left\{r_{e,n}, b_n^{\frac{\gamma_0 - 1 + p_e}{p_e}}\right\} \ll \bar{r}_{e,n} \ll b_n^{\frac{1}{1 + \zeta\frac{p_e}{p_e + \gamma_0 - 1}}}$, and $B = O\left(r_{\mu,n}\bar{r}_{e,n}^{1 + \zeta\frac{p_e}{p_e + \gamma_0 - 1}}b_n^{-1}\right) = o(1)$.

Now suppose condition (ii) holds. For any subsequence of n with $r_{e,n} \leq b_n^{\frac{\gamma_0 - 1 + p_e}{p_e}}$, then $B = O\left(r_{\mu,n}r_{e,n}b_n^{\zeta - 1}\right) = O\left(r_{\mu,n}b_n^{1 + \frac{\gamma_0 - 1}{p_e}}b_n^{\zeta - 1}\right) = o(1)$ on that subsequence. For any subsequence of n with $r_{e,n} > b_n^{\frac{\gamma_0 - 1 + p_e}{p_e}}$, then $B = O\left(r_{\mu,n}r_{e,n}^{1 + \zeta\frac{p_e}{p_e + \gamma_0 - 1}}b_n^{-1}\right) = o(1)$ by condition (ii). Therefore $E_{P(n)}\left[\hat{\psi}_{clip,k}^{AIPW}(b_n) - \tilde{\psi}_{clip}^{AIPW,(k)}(b_n)\right] = o(1)$.

Now, write $\hat{\psi}_{clip}^{AIPW}(b_n) = \frac{1}{K}\sum_k \hat{\psi}_{clip,k}^{AIPW}(b_n)$, where $\hat{\psi}_{clip,k}^{AIPW}(b_n)$ is the fold$-k$ average potential outcome

estimate. I will show that $\hat{\psi}^{AIPW}_{clip,k}(b_n) - \psi_n = o_{P(n)}(1)$. Fix the nuisance estimates $\hat{\mu}$, $\hat{e}$ from other folds implicitly. Then:

$$\begin{aligned}
E_{P(n)}\left[\left(\hat{\psi}^{AIPW}_{clip,k}(b_n) - \psi_n\right)^2\right] &\leq 3E_{P(n)}\left[\left(\hat{\psi}^{AIPW}_{clip,k}(b_n) - \tilde{\psi}^{AIPW,(k)}_{clip}(b_n)\right)^2 + \left(\tilde{\psi}^{AIPW,(k)}_{clip}(b_n) - \psi_n\right)^2\right] \\
&= O\left(E_{P(n)}\left[\left(\hat{\psi}^{AIPW}_{clip,k}(b_n) - \tilde{\psi}^{AIPW,(k)}_{clip}(b_n)\right)^2\right] + n^{-1}E\left[R\right]\right) \qquad \text{(Lemma 5)} \\
&= O\left(E_{P(n)}\left[\hat{\psi}^{AIPW}_{clip,k}(b_n) - \tilde{\psi}^{AIPW,(k)}_{clip}(b_n)\right]^2 + n^{-1}E_{P(n)}\left[(\hat{\phi}_n - \phi_n)^2\right] + n^{-1}E\left[R\right]\right) \\
&= o(1) + O\left(n^{-1}b_n^{-2}\right) + o(1) = o(1).
\end{aligned}$$

Therefore:

$$E_{P(n)}\left[\left(\hat{\psi}^{AIPW}_{clip}(b_n) - \psi_n\right)^2\right] = E_{P(n)}\left[\left(\frac{1}{K}\sum_k \hat{\psi}^{AIPW}_{clip,k}(b_n) - \psi_n\right)^2\right] \leq K^2 E_{P(n)}\left[\max_k\left(\hat{\psi}^{AIPW}_{clip}(b_n) - \psi_n\right)^2\right] = o(1).$$

□

C.3 Degradation of Consistency Rate

Lemma 12. *Suppose the conditions of Proposition 3 hold and let $P(n)$ be a sequence of distributions in $\mathscr{P}$. Then, $E_{P(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] = O\left(1 + b_n^{\gamma_0-2}\log(1/b_n)^{\mathbf{1}\{\gamma_0-2\}}\right)$, with a constant that only depends on C and γ_0.*

Proof of Lemma 12. Let $P(n)$ be given. Then:

$$\begin{aligned}
E_{P(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] = E_{P(n)}\left[\frac{e(X)}{\max\{e(X),b_n\}^2}\right] &\leq E_{P(n)}\left[\frac{1}{\max\{e(X),b_n\}}\right] = \int_0^\infty P(n)\left(\frac{1}{\max\{e(X),b_n\}} > t\right)dt \\
&= \int_0^\infty P(n)\left(\max\{e(X),b_n\} < 1/t\right)dt = 1 + \int_1^\infty P(n)\left(\max\{e(X),b_n\} < 1/t\right)dt \\
&= 1 + \int_1^{1/b_n} P(n)\left(\max\{e(X),b_n\} < 1/t\right)dt \leq 1 + C\int_1^{b_n^{-1}} t^{1-\gamma_0}dt.
\end{aligned}$$

First, suppose $\gamma_0 = 2$. Then $E_{P(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] \leq 1 + C\left(\log(1/b_n) - 1\right)$. Alternatively, suppose $\gamma_0 \neq 2$. Then $E_{P(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right] \leq 1 + \frac{C}{\gamma_0-2}\left(1 - b_n^{\gamma_0-2}\right)$. □

Lemma 13. *Suppose the conditions of Proposition 3 hold and $P(n)$ is a sequence of distributions in $\mathscr{P}$. Then*

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

Proof of Lemma 13. For any $m \geq 0$, define $\mathscr{P}_{m,n} = \{P \in \mathscr{P} \mid P(e(X) \leq b_n) \leq m)$. For each n, define $m_n = P(n)(e(X) \leq b_n)^{\gamma_0/(\gamma_0-1)}$.

I have:

$$\begin{aligned}
\sup_{P \in \mathscr{P}_{m_n,n}} E_P\left[\frac{D}{\max\{e,b_n\}^2}\right] &= \sup_{P \in \mathscr{P}_{m_n,n}} E_P\left[\frac{D\mathbf{1}\{e(X) > b_n\}}{\max\{e,b_n\}^2}\right] + E_P\left[\frac{D\mathbf{1}\{e(X) \leq b_n\}}{b_n^2}\right] \\
&\geq 1 + \sup_{P \in \mathscr{P}_{m_n,n}} b_n^{-2} E_P[e(X)\mathbf{1}\{e(X) \leq b_n\}] \\
&\geq 1 + c(\gamma_0) \sup_{P \in \mathscr{P}_{m_n,n}} P(e(X) \leq b_n)^{\gamma_0/(\gamma_0-1)} \qquad \text{(Lemma 4)} \\
&= 1 + c(\gamma_0) b_n^{-2} m_n^{\gamma_0/(\gamma_0-1)}.
\end{aligned}$$

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

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

By Lemma 5, $E_{P(n)}\left[\frac{D\sigma^2_{\min}}{\max\{e(X),b_n\}^2}\right] \not\to 0$.

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

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

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

Proof of Corollary 2. First, note that for any sequence of $P(n) \in \mathscr{P}$ and any sequence of $b_n \to 0$, $E_{P(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right] = O\left(nv_n(b_n)\right)$ by Lemma 12.

Now suppose Assumption 4(i) holds, so that $P(e(X) \leq \pi) \geq C'\pi^{\gamma_0-1}$ for all $P \in \mathscr{P}$. Then

$$\begin{aligned}
E_{P(n)}\left[\frac{D}{\max\{e,b_n\}^2}\right] &= b_n^{-2} E_{P(n)}[e1\{e \leq b_n\}] + E_{P(n)}\left[\frac{1\{e \geq b_n\}}{e}\right] \\
&= b_n^{-2} E_{P(n)}[e1\{e \leq b_n\}] + \int_0^\infty P\left(b_n \leq e \leq t^{-1}\right) dt
\end{aligned}$$

$$
\begin{aligned}
&= b_n^{-2} E_{P(n)}[e1\{e \le b_n\}] + 1 + \int_1^{b_n^{-1}} P\left(e \le t^{-1}\right) dt \\
&\ge b_n^{-2} C' \int_0^{b_n} t^{\gamma_0 - 1} dt + 1 + C' \int_1^{b_n^{-1}} t^{1-\gamma_0} dt \\
&= \Omega\left(1 + b_n^{\gamma_0 - 2} + \log\left(1/b_n\right)^{1\{\gamma_0 = 2\}}\right) = \Omega\left(n v_n\right),
\end{aligned}
$$

so that by Proposition 2, $\inf_{P \in \mathscr{P}} \sigma_n^2 = \Omega\left(v_n(b_n)\right)$, completing the proof. □

C.4 Asymptotic Normality and Rates

Proof of Corollary 5. It is clear that if $\|\hat{\mu}^{(-k)} - \mu\|^2_{L^\infty(P(n))} = o(1)$, then A3'(a) holds, so I proceed under the $L^{p_\mu}(P(n))$ outcome rate conditions.

First, suppose Assumption 2 and Assumption 4(ii) hold. For the fourth implication, note that if $\min\{p_\mu, p_e\}$ is finite or $\gamma_0 \neq 2$, then $0 \le \zeta = (\gamma_0 - 1)\left(1 - \frac{1}{p_e} - \frac{1}{p_\mu}\right) \le \min\{\gamma_0 - 1, 1\}\left(1 - \frac{1}{\min\{p_\mu, p_e\}}\right)$ and $b_n^{\min\{\gamma_0 - 1, 1\}\frac{p_\mu p_e - p_\mu - p_e}{ell_\mu p_e}} \gg \log(1/b_n)^{1\left\{(\gamma_0 - 1)\frac{p_\mu p_e - p_\mu - p_e}{ell_\mu p_e} = 1\right\}}$. As a result, it suffices to show:

$$
\begin{aligned}
&\text{A3'(a):} \quad \left(b_n E_{P(n)}[R]\right)^{\frac{1}{p_\mu}} \gtrsim b_n^{1/p_\mu} \gg r_{\mu,n} \\
&\text{A3'(b):} \quad b_n^{1+\frac{1}{p_e}} E_{P(n)}[R]^{\frac{1}{p_e}} \gtrsim b_n^{\frac{p_e+1}{p_e}} \gg r_{e,n} \\
&\text{A3'(c):} \quad r_{\mu,n} \frac{P(n)(e(X) \le b_n)^{\frac{p_\mu - 1}{p_\mu}}}{\sqrt{E_{P(n)}[R]}} \lesssim r_{\mu,n} b_n^{(\gamma_0 - 1)\frac{p_\mu - 1}{p_\mu}} \lesssim r_{\mu,n} b_n^{\min\{\gamma_0 - 1, 1\}\frac{p_\mu - 1}{p_\mu}} \ll n^{-1/2} \\
&\text{A3'(d):} \quad r_{\mu,n} r_{e,n} \frac{b_n^{\zeta - 1}}{\sqrt{E_{P(n)}[R]}} \lesssim n^{-1/2} b_n^{\min\{\gamma_0, 1\}\left(\frac{1}{p_\mu} - 1\right) + 1 + \frac{1}{p_e} + \min\{\gamma_0 - 1, 1\}\left(1 - \frac{1}{p_e} - \frac{1}{p_\mu}\right) - 1} \ll n^{-1/2} \\
&\qquad \text{and } r_{\mu,n} r_{e,n} \frac{\log\left(\frac{1}{b_n}\right) 1\{p_\mu = p_e = \infty, \gamma_0 = 2\}}{\sqrt{E_{P(n)}[R]}} \lesssim n^{-1/2} r_{e,n} b_n^{-1} \ll n^{-1/2}.
\end{aligned}
$$

Thus, every condition of Assumption 3' holds.

Now, instead proceed assuming Assumption 4(i) holds. Then $E_{P(n)}[R] = \Theta\left(1 + C'(\gamma_0 - 1)\int_{b_n}^{2b_n} \pi^{\gamma_0 - 3} d\pi\right) = \Theta\left(b_n^{\min\{\gamma_0 - 2, 0\}} \log(1/b_n)^{1\{\gamma_0 = 2\}}\right)$, where $R = \frac{D}{\max\{e, b_n\}^2}$. In the A3(a)(ii) case, take some δ_n satisfying $r_{\mu,n}^2 \ll \delta_n \ll 1$. Then, for the first two terms:

$$
\begin{aligned}
&\text{A3'(a):} \quad \left(b_n E_{P(n)}[R]\right)^{\frac{1}{p_\mu}} \gtrsim \left(b_n b_n^{\min\{\gamma_0 - 2, 0\}}\right)^{\frac{1}{p_\mu}} = \left(b_n^{\min\{\gamma_0 - 1, 1\}}\right)^{\frac{1}{p_\mu}} \gg r_{\mu,n}. && \text{(A3(a))} \\
&\text{A3'(b):} \quad b_n^{1+\frac{1}{p_e}} E_{P(n)}[R]^{\frac{1}{p_e}} \gtrsim b_n^{1 + \frac{\min\{\gamma_0 - 1, 1\}}{p_\mu}} \gg r_{e,n}. && \text{(A3(b))}
\end{aligned}
$$

A 3(c) implies A3'(c) by inspection. A 3(d) similarly implies A3'(d) by inspection. □

Lemma 14 (Oracle consistency)**.** *If $n^{-1/2} \ll b_n \ll 1$, then $|P_n[\phi_n] - E_{P(n)}[\mu(x)]| \xrightarrow{\mathscr{P}} 0$.*

Proof of Lemma 14. Let $P(n)$ be a sequence of distributions in $\mathscr{P}$. For any $t > 0$, I have:

$$\begin{aligned}
P(n)\left(|P(n)_n[\phi_n] - E_{P(n)}[\mu(x)]| > t\right) &\leq \frac{E[|\phi_n - E_{P(n)}[\mu(x)]|^2]}{nt^2} && \text{(Chebyshev's inequality)} \\
&\leq \frac{E[|\phi_n - E_{P(n)}[\phi_n]|^q]^{2/q}}{nt^2} && \text{(Jensen's inequality)} \\
&\leq \frac{[(4M)^q E[e(X)/\{e(X) \vee b_n\}^2]]^{2/q}}{t^2 n b_n^{2(q-2)/q}} && \text{(Lemma 3.(iii))} \\
&\leq \frac{(4M)^2}{t^2}\frac{1}{nb_n^2}.
\end{aligned}$$

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

Lemma 15 (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 $n\sigma_n^2/\mathrm{Var}_{P(n)}(\phi_n) \xrightarrow{\mathscr{P}} 1$.*

Proof of Lemma 15. Let $P(n)$ be a sequence of distributions in $\mathscr{P}$.

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

$$\begin{aligned}
P\left(\left|\frac{P(n)_n[\phi_n^2] - P[\phi_n^2]}{\mathrm{Var}_{P(n)}(\phi_n)}\right| > t\right) &\leq \frac{E\{|P(n)_n[\phi_n^2] - P[\phi_n^2]|^{q/2}\}}{t^{q/2}\mathrm{Var}_{P(n)}(\phi_n)^{q/2}} && \text{(Markov inequality)} \\
&\leq \frac{2}{t^{q/2}n^{q/2-1}}\frac{E\{|\phi_n^2 - P[\phi_n^2]|^{q/2}\}}{Var_{P(n)}(\phi_n)^{q/2}} && \text{(von Bahr-Esseen inequality)} \\
&\leq \frac{2^{q/2+1}}{t^{q/2}n^{q/2-1}}\frac{E[|\phi_n|^q]}{Var_{P(n)}(\phi_n)^{q/2}} && \text{(Jensen's inequality)} \\
&\leq \frac{2^{q/2+1}}{t^{q/2}n^{q/2-1}}\frac{(8M)^q E[e(X)/\{e(X) \vee b_n\}^2]}{b_n^{q-2}(\mathrm{Var}_{P(n)}(\phi_n))^{q/2}} && \text{(Lemma 3.(iv))} \\
&\leq \frac{2^{q/2+1}}{t^{q/2}n^{q/2-1}}\frac{(8M)^q E[e(X)/\{e(X) \vee b_n\}^2]}{b_n^{q-2}\sigma_{\min}^q E[e(X)/\{e(X) \vee b_n\}^2]^{q/2}} && \text{(Lemma 5)} \\
&\leq \frac{(8M)^q 2^{q/2+1}}{t^{q/2}\sigma_{\min}^q(\pi_{\min}/2)^{q/2-1}}\frac{1}{n^{q/2-1}b_n^{q-2}}. && \text{(Lemma 5)}
\end{aligned}$$

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

Then, by the triangle inequality:

$$\begin{aligned}
\left|\frac{n\sigma_n^2}{\mathrm{Var}_{P(n)}(\phi_n)} - 1\right| &\leq \left|\frac{P(n)_n[\phi_n]^2 - P[\phi_n]^2}{\mathrm{Var}_{P(n)}(\phi_n)}\right| + \left|\frac{P(n)_n[\phi_n^2] - P[\phi_n^2]}{\mathrm{Var}_{P(n)}(\phi_n)}\right| \\
&\leq |P(n)_n[\phi_n] + P[\phi_n]| \times \frac{|P(n)_n[\phi_n] - P[\phi_n]|}{\sigma_{\min}^2\pi_{\min}/2} + o_{P(n)}(1) && \text{(Lemma 5 + above)} \\
&\leq (2M + o_{P(n)}(1))o_{P(n)}(1) + o_{P(n)}(1) = o_{P(n)}(1), && (P[\phi_n] \leq M + \text{Lemma 14})
\end{aligned}$$

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

Lemma 16 (Variance component bound)**.** *Suppose b_n satisfies $n^{-1/2} \ll b_n \ll 1$, the conditions of Assumption 2 hold, and $P(n)$ is a sequence of distributions in $\mathscr{P}$. Write $R = \frac{D}{\max\{e(X), b_n\}^2}$. Then with probability tending to one,*

$$E_{P(n)}\left[(\hat{\phi}_n - \phi_n)^2\right] = O\left(E_{P(n)}\left[R\right]\left(r_{\mu,n}^2 E_{P(n)}\left[b_n R\right]^{\frac{-2}{p_\mu}} + \left(r_{e,n}/b_n\right)^2 E_{P(n)}\left[b_n R\right]^{\frac{-2}{p_e}}\right)\right),$$

where $\hat{\phi}_n$ is the estimated influence function and ϕ_n is the oracle influence function with clipped nuisances. If in addition there is a sequence of $\delta_n \gg r_{\mu,n}^2$ such that $\|\hat{\mu} - \mu\|_{L^\infty(P(n))}^2 = o(1)$, then $E_{P(n)}\left[(\hat{\phi}_n - \phi_n)^2\right] = O\left(E_{P(n)}\left[R\right]\left(r_{e,n}/b_n\right)^2 E_{P(n)}\left[b_n R\right]^{\frac{-2}{p_e}}\right) + o\left(E_{P(n)}\left[R\right]\right)$.

Proof of Lemma 16. Define $\tilde{R} = \frac{e}{\max\{e, b_n\}^2}$. Take the high-probability event under Assumption 2. Note that $E[\tilde{R}] = E[R]$, so that for $t > 1$ finite:

$$\|\tilde{R}\|_{L^t(P(n))} \leq \|\tilde{R}\|_{L^1(P(n))}^{1/t} \|\tilde{R}\|_{L^\infty(P(n))}^{(t-1)/t} = \|\tilde{R}\|_{L^1(P(n))} \left(\frac{\|\tilde{R}\|_{L^\infty(P(n))}}{\|\tilde{R}\|_{L^1(P(n))}}\right)^{(t-1)/t} \leq E_{P(n)}[R]\left(E_{P(n)}[b_n R]\right)^{(1-t)/t},$$

and for t infinite,

$$\|\tilde{R}\|_{L^t(P(n))} \leq b_n^{-1} = E_{P(n)}[R]\left(E_{P(n)}[b_n R]\right)^{-1} = E_{P(n)}[R]\left(E_{P(n)}[b_n R]\right)^{(1-t)/t}.$$

In particular, when $t = \frac{p}{p-2}$, then $\frac{1-t}{t} = \frac{-2}{p}$.

Recall that $|\hat{\mu} - \mu|$ is uniformly bounded under Assumption 2, so that $|Y - \hat{\mu}| = |Y - \mu| + O(1)$, $\|\hat{\mu} - \mu\|_{L^p(P(n))} = O(r_{\mu,n}) = O\left(E_{P(n)}[R]\right)$ for all $p \in [1, p_\mu]$, and $\|\hat{e} - e\|_{L^p(P(n))} = O(r_{e,n})$ for all $p \in [1, p_e]$. Then:

$$\begin{aligned}\left(\hat{\phi}_n - \phi_n\right)^2 &= \left((\hat{\mu} - \mu)\frac{\max\{e, b_n\} - D}{\max\{e, b_n\}} + (Y - \hat{\mu})\left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}}\right)\right)^2 \\ &\leq 2(\hat{\mu} - \mu)^2\left(1 + \frac{D}{\max\{e, b_n\}^2}\right) + 4\left((Y - \mu)^2 + (\hat{\mu} - \mu)^2\right)\left(\frac{D}{\max\{\hat{e}, b_n\}} - \frac{D}{\max\{e, b_n\}}\right)^2.\end{aligned}$$

For the first term, if the sup-norm consistency assumption holds, then $E_{P(n)}\left[(\hat{\mu} - \mu)^2\left(1 + \frac{D}{\max\{e, b_n\}^2}\right)\right] \leq \|\hat{\mu} - \mu\|_{L^\infty(P(n))} E_{P(n)}[1 + R] = o(E_{P(n)}[R])$. Otherwise:

$$\begin{aligned}E_{P(n)}\left[(\hat{\mu} - \mu)^2\left(1 + \frac{D}{\max\{e, b_n\}^2}\right)\right] &= O\left(r_{\mu,n}^2\left(1 + \left\|\tilde{R}\right\|_{L^{\frac{p_\mu}{p_\mu - 2}}(P(n))}\right)\right) = O\left(r_{\mu,n}^2 E_{P(n)}\left[b_n R\right]^{\frac{-2}{p_\mu}} E_{P(n)}\left[R\right]\right) \\ &= O\left(E_{P(n)}\left[R\right] r_{\mu,n}^2 E_{P(n)}\left[b_n R\right]^{\frac{-2}{p_\mu}}\right).\end{aligned}$$

For the remaining term, note that:

$$E_{P(n)}\left[((Y-\mu)^2+(\hat{\mu}-\mu)^2)\left(\frac{D}{\max\{\hat{e},b_n\}}-\frac{D}{\max\{e,b_n\}}\right)^2\right] = O\left(E_{P(n)}\left[\tilde{R}\left(\frac{\max\{\hat{e},b_n\}-\max\{e,b_n\}}{\max\{\hat{e},b_n\}}\right)^2\right]\right)$$
$$= O\left(\left\|\tilde{R}\right\|_{L^{\frac{p_e}{p_e-2}}(P(n))} E_{P(n)}\left[\left|\frac{\max\{\hat{e},b_n\}}{\max\{\hat{e},b_n\}}-1\right|^{p_e}\right]^{2/p_e}\right)$$
$$= O\left(E_{P(n)}[R]\, E_{P(n)}[b_n R]^{\frac{-2}{p_e}}\left(\frac{r_{e,n}}{b_n}\right)^2\right),$$

with the final line holding by Lemma 10. □

Lemma 17. *Recall the definitions of ϕ_n as the oracle clipped influence function and $\hat{\phi}_n$ the estimated influence function. Suppose b_n satisfies $n^{-1/2} \ll b_n \ll 1$ and the conditions of Assumption 2 hold. Then (i) $E_{P(n)}[\phi_n^2] = \mathrm{Var}_{P(n)}(\phi_n) + O(1)$ and $P(n)_n\left[\hat{\phi}_n^2 - \phi_n^2\right] = O_{P(n)}\left(E_{P(n)}[(\hat{\phi}_n-\phi_n)^2]\right) + o_{P(n)}\left(Var_{P(n)}(\phi_n)\right)$. If in addition Assumption 3'(a),(b) hold, then (ii) $P(n)_n\left[\hat{\phi}_n^2 - \phi_n^2\right] = o_{P(n)}\left(\mathrm{Var}_{P(n)}(\phi_n)\right)$. If in addition the conditions of Theorem 1' hold, then (iii) $E\left[\tilde{\psi}_{clip}^{AIPW,(k)}(b_n) - \hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right]^2 + Var\left(\tilde{\psi}_{clip}^{AIPW,(k)}(b_n) - \hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right) = o_{P(n)}\left(n^{-1}\mathrm{Var}_{P(n)}(\phi_n)\right)$.*

Proof of Lemma 17. Take the high-probability event under Assumption 2. Note that:

$$E_{P(n)}[\phi_n^2] = \mathrm{Var}_{P(n)}(\phi_n) + E_{P(n)}[\phi_n]^2 = \mathrm{Var}_{P(n)}(\phi_n) + O(1)^2. \qquad \text{(Assumption 1(a))}$$

Note that by Lemma 2 and that $b_n \to 0$, eventually $b_n \leq 2\pi_{\min}$ so that $Var_{P(n)}(\phi_n) = \Omega(1)$. Write $\tilde{a}_n = \frac{P(n)_n[(\hat{\phi}_n-\phi_n)^2]}{Var_{P(n)}(\phi_n)}$ and $a_n = \frac{E_{P(n}[(\hat{\phi}_n-\phi_n)^2]}{Var_{P(n)}(\phi_n)}$. Note that $a_n = \tilde{a}_n + O_{P(n)}\left(n^{-1}b_n^{-2}\right) = \tilde{a}_n + o_{P(n)}(1)$ by Markov's inequality. Then:

$$\left|P(n)_n\left[\hat{\phi}_n^2 - \phi_n^2\right]\right| = \left|P(n)_n[(\hat{\phi}_n-\phi_n)^2] + 2P(n)_n[(\hat{\phi}_n-\phi_n)\phi_n]\right|$$
$$\leq P(n)_n[(\hat{\phi}_n-\phi_n)^2] + 2\sqrt{P(n)_n[(\hat{\phi}_n-\phi_n)^2]P(n)_n[\phi_n^2]} \qquad \text{(Cauchy-Schwarz)}$$
$$= O_{P(n)}\left(Var_{P(n)}(\phi_n)\left(\tilde{a}_n + \sqrt{\tilde{a}_n}\right)\right) \qquad (b_n \gg n^{\frac{-1}{2}}, \text{L15})$$
$$= O_{P(n)}\left(Var_{P(n)}(\phi_n)\left(a_n + \sqrt{a_n}\right)\right) + o\left(Var_{P(n)}(\phi_n)\right)$$
$$= O_{P(n)}\left(E_{P(n)}[(\hat{\phi}_n-\phi_n)^2]\right) + o_{P(n)}\left(Var_{P(n)}(\phi_n)\right).$$

To complete the result (ii) if Assumption 3'(a),(b) hold, recall the definition $R = \frac{D}{\max\{e,b_n\}^2}$. Then by inspection of Lemma 16,

$$a_n = \frac{E_{P(n)}[(\hat{\phi}_n-\phi_n)^2]}{\mathrm{Var}_{P(n)}(\phi_n)} = o\left(E_{P(n)}[R]/\mathrm{Var}_{P(n)}(\phi_n)\right) = o(1), \qquad \text{(Proposition 2)}$$

so that $\left|P(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right| = o\left(Var_{P(n)}(\phi_n)\right)$. Thus, $Var\left(\tilde{\psi}_{clip}^{AIPW,(k)}(b_n)-\hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right) = o_{P(n)}\left(n^{-1}\text{Var}_{P(n)}(\phi_n)\right)$ for any fixed fold. But then any weighted average of $o_{P(n)}\left(n^{-1}\text{Var}_{P(n)}(\phi_n)\right)$ terms is $o_{P(n)}\left(n^{-1}\text{Var}_{P(n)}(\phi_n)\right)$, so the full result holds.

Finally, I show the result (iii) under the conditions of Theorem 1'. By Lemma 11,

$$E\left[\tilde{\psi}_{clip}^{AIPW,(k)}(b_n)-\hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right] = O\left(\tilde{r}_{\mu,n}P(e(X)\leq b_n)^{\frac{p_\mu-1}{p_\mu}} + r_{\mu,n}r_{e,n}\left(b_n^{\zeta-1}+\log\left(\frac{1}{b_n}\right)^{1\{\zeta=1\}}+r_{e,n}^{\zeta\frac{p_e}{p_e+\gamma_0-1}}b_n^{-1}\right)\right),$$

where $\zeta = (\gamma_0-1)\frac{p_\mu p_e - p_\mu - p_e}{p_e p_\mu}$. Write this as $O\left(I+II+III+IV\right)$ and $\tilde{r}_{\mu,n} = \min\left\{r_{\mu,n}P(e(X)\leq b_n)^{1/p_\mu}\right\}$. By Assumption 3'(c) and Assumption 3'(d), $I+II+III = o\left(n^{-1}E_{P(n)}[\phi_n^2]\right)$. For $IV = r_{\mu,n}r_{e,n}^{1+\zeta\frac{p_e}{p_e+\gamma_0-1}}b_n^{-1}$, partition n as follows. If $r_{e,n}\leq b_n^{1+\frac{\gamma_0-1}{p_e}}$, then $IV \leq r_{\mu,n}r_{e,n}b_n^1 b_n^{-1} = o\left(n^{-1}E_{P(n)}[\phi_n^2]\right)$ by Assumption 3'(d). If $r_{e,n}\geq b_n^{1+\frac{\gamma_0-1}{p_e}}$, then $IV \leq r_{\mu,n}r_{e,n}b_n^{\zeta-1} = o\left(n^{-1}E_{P(n)}[\phi_n^2]\right)$ by Assumption 3'(d). Thus, for any fixed fold, $E\left[\tilde{\psi}_{clip}^{AIPW,(k)}(b_n)-\hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right] = o\left(n^{-1/2}\sqrt{E_{P(n)}\left[\frac{D}{\max\{e(X),b_n\}^2}\right]}\right)$ for any fixed fold. But by iterated expectations, $E\left[\tilde{\psi}_{clip}^{AIPW,(k)}(b_n)-\hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right]^2 = o_{P(n)}\left(n^{-1}\text{Var}_{P(n)}(\phi_n)\right)$ when averaging across folds.

□

Lemma 18 (Estimated variance consistency)**.** *Suppose conditions of of Proposition 1 hold. Let $P(n)$ be a sequence of distributions in $\mathscr{P}$. Then $\left|\frac{\hat{\sigma}_n^2-\bar{\sigma}_n^2}{\bar{\sigma}_n^2}\right| = \left|P(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right| O_{P(n)}\left(1/\text{Var}_{P(n)}(\phi_n)\right)+o_{P(n)}(1)$. If in addition Assumption 3'(a),(b) hold, then $\hat{\sigma}_n^2/\sigma_n^2 \xrightarrow{\mathscr{P}} 1$.*

Proof of Lemma 18. Take the high-probability event under Assumption 2. Recall the definition $\bar{\sigma}_n^2 = n^{-1}\text{Var}_{P(n)}(\phi_n)$.

Let $\sigma_n^2 = n^{-1}(P(n)_n[\phi_n^2]-P(n)_n[\phi_n]^2)$ be the oracle sample variance. By Lemma 15, $\sigma_n^2/\bar{\sigma}_n^2 \xrightarrow{\mathscr{P}} 1$. Therefore it suffices to show that $(\hat{\sigma}_n^2-\sigma_n^2)/\bar{\sigma}_n^2 = \frac{P(n)_n[\hat{\phi}_n^2]-P(n)_n[\phi_n^2]-\hat{\psi}_{clip}^{AIPW}(b_n)^2+\tilde{\psi}_{(Oracle)}^{AIPW}(b_n)^2}{Var_{P(n)}(\phi_n)} \xrightarrow{\mathscr{P}} 0$.

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

By the triangle inequality:

$$\begin{aligned}
\left|\frac{\hat{\sigma}_n^2-\sigma_n^2}{\bar{\sigma}_n^2}\right| &\leq \left|\frac{P(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]}{Var_{P(n)}(\phi_n)}\right| + \left|\frac{\hat{\psi}_{clip}^{AIPW}(b_n)^2-\tilde{\psi}_{(Oracle)}^{AIPW}(b_n)^2}{Var_{P(n)}(\phi_n)}\right| \\
&\precsim \left|P(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right| O\left(\frac{1}{E_{P(n)}\left[D/\max\{e(X),b_n\}^2\right]}\right) && \text{(Lemma 6)} \\
&+ O_{P(n)}\left(\left|\hat{\psi}_{clip}^{AIPW}(b_n)-\tilde{\psi}_{(Oracle)}^{AIPW}(b_n)\right|\right) && \text{(Lemma 5)} \\
&= \left|P(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right| O_{P(n)}\left(\frac{1}{E_{P(n)}\left[D/\max\{e(X),b_n\}^2\right]}\right)+o_{P(n)}(1) && \text{(Proposition 1)} \\
&= \left|P(n)_n\left[\hat{\phi}_n^2-\phi_n^2\right]\right| O_{P(n)}\left(1/\text{Var}_{P(n)}(\phi_n)\right)+o_{P(n)}(1) && \text{(Proposition 2)} \\
&= o_{P(n)}(\text{Var}_{P(n)}(\phi_n))O_{P(n)}\left(1/\text{Var}_{P(n)}(\phi_n)\right)+o_{P(n)}(1) && \text{(Lemma 17)}
\end{aligned}$$

$= o_{P(n)}(1).$

Therefore $(\hat{\sigma}_n^2 - \sigma_n^2)/\bar{\sigma}_n^2 \to_{P(n)} 0$. By Lemma 15, $\sigma_n^2/\bar{\sigma}_n^2 \to_{P(n)} 1$. As a result, $(\hat{\sigma}_n^2 - \bar{\sigma}_n^2)/\bar{\sigma}_n^2 \to_{P(n)} 0$ and $\hat{\sigma}_n^2/\bar{\sigma}_n^2 \xrightarrow{\mathscr{P}} 1$. □

Lemma 19. *Suppose the conditions of Theorem 1' hold. Then* $\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Oracle)}^{AIPW}(b_n)\right) = o_{P(n)}(1)$.

Proof of Lemma 19. Take the high-probability event under Assumption 2. 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:

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

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

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 $\sum_k \frac{n_k}{n}\hat{r}_k = o_{P(n)}(1)$. But any weighted average of a finite number of $o_{P(n)}(1)$ terms is $o_{P(n)}(1)$, so it suffices to show $\hat{r}_1 = o_{P(n)}(1)$. By Markov's inequality, it suffices to show that with probability tending to one, $E\left[\hat{r}_1^2 \mid \{X, D, Y\}_{r(i)>1}\right] = o_{P(n)}(1)$. $E\left[\hat{r}_1^2 \mid \{X, D, Y\}_{k(i)>1}\right]$ only depend on the $k(i) > 1$ data via $\hat{\mu}$ and $\hat{e}$. Recall that by assumption, $\hat{\mu}$ and $\hat{e}$ given this data satisfy the rate requirements Assumption 3', so I leave the conditioning implicit. Write $\tilde{r}_1 = \sigma_n \hat{r}$. By Lemma 5, it suffices to show that $E[\tilde{r}_1^2] = o_{P(n)}\left(n^{-1}E_{P(n)}\left[\frac{D}{\max\{e(X), b_n\}^2}\right]\right)$. Then:

$$\begin{aligned} E\left[\tilde{r}_1^2\right] &= E\left[\tilde{\psi}_{clip}^{AIPW,(k)}(b_n) - \hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right]^2 + Var\left(\tilde{\psi}_{clip}^{AIPW,(k)}(b_n) - \hat{\psi}_{clip}^{AIPW,(k)}(b_n)\right) \\ &= o_{P(n)}\left(\sigma_n^{-2}\right) + n^{-1}E_{P(n)}\left[(\hat{\phi}_n - \phi_n)^2\right] = o_{P(n)}\left(\sigma_n^{-2}\right). \qquad \text{(Lemma 17)} \end{aligned}$$

As a result:

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

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

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

Proof of Theorem 1'. By Lemma 19, $\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n) - \tilde{\psi}_{(Oracle)}^{AIPW}(b_n)\right) = o_{P(n)}(1)$. Therefore, by Proposi-

tion 3, $\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n)-\psi_n\right)=\sigma_n^{-1}\left(\hat{\psi}_{clip}^{AIPW}(b_n)-\tilde{\psi}_{(Oracle)}^{AIPW}(b_n)\right)+\sigma_n^{-1}\left(\tilde{\psi}_{(Oracle)}^{AIPW}(b_n)-\psi_n\right) \overset{P(n)}{\rightsquigarrow} N(0,1)$.

□

Proof of Theorem 1. For either claim, let $P(n)$ be a sequence of distributions P in the relevant set. Note that in either case, the assumptions of Theorem 1' hold by Corollary 5. Therefore, by Theorem 1',

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

where $P(n)_n$ denotes the empirical average under distribution $P(n)$ and σ_n is defined in Proposition 3.

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

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

with the final result holding by Slutsky's theorem. herefore,

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

by properties of a cumulative distribution function.

The remaining claim for Wald confidence intervals is now standard. For simplicity of exposition, I prove the result for the class $\mathscr{P}$ under Assumption 4(ii):

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

$$= \limsup_{n\to\infty} \sup_{P\in\mathscr{P}} \left| P^n \left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n) - \psi(P)}{\sigma_n} < z_{1-\frac{\alpha}{2}} + o(1) \right) - \Phi(z_{1-\frac{\alpha}{2}}) \right|$$
$$+ \limsup_{n\to\infty} \sup_{P\in\mathscr{P}} \left| P^n \left(\frac{\hat{\psi}^{AIPW}_{clip}(b_n) - \psi(P)}{\sigma_n} < z_{\frac{\alpha}{2}} + o(1) \right) - \Phi(z_{\frac{\alpha}{2}}) \right|$$
(Lemma 18)

$$= 0.$$ (Theorem 1)

□

Proof of Corollary 1. If $\limsup_{n\to\infty} n^{1/2} r_{\mu,n} \precsim 1$, then there is a $\bar{r}_{\mu,n}$ satisfying $\max\{n^{-1/2}, r_{\mu,n}\} \ll \bar{r}_{\mu,n} \ll \min\left\{ n^{-1/(2p_\mu)}, n^{-1/2} r_{e,n}^{\left(1-\frac{1}{p_e+1}\right)\left(1-\frac{1}{p_\mu}\right)} \right\}$, so that I can without loss of generality replace $r_{\mu,n}$ with this $\bar{r}_{\mu,n}$.

I make an additional substitution in the $r_{e,n} \ll n^{\frac{-(p_e+\gamma_0-1)}{p_e\gamma_0}} \ll n^{\frac{-(p_e+1)}{p_e\gamma_0}}$ case, which implies:

$$n^{\frac{(2-\gamma_0)p_\mu-2}{2\gamma_0 p_\mu}} = n^{\frac{-1}{2p_\mu}\frac{(\gamma_0-2)(p_\mu-1)+\gamma_0}{\gamma_0}} \ll n^{\frac{-1}{2p_\mu}} \text{ and } n^{\frac{(2-\gamma_0)p_\mu-2}{2\gamma_0 p_\mu}} = n^{\frac{2(p_\mu-1)-\gamma_0 p_\mu}{2\gamma_0 p_\mu}} = n^{\frac{p_\mu-1}{p_\mu\gamma_0}-\frac{1}{2}} \ll r_{e,n}^{\frac{p_\mu-1}{p_\mu}\frac{-p_e}{p_e+1}} n^{\frac{-1}{2}},$$

so that there is a larger $\bar{r}_{\mu,n}$ satisfying:

$$n^{-1/2} \ll \max\left\{ r_{\mu,n}, n^{\frac{(2-\gamma_0)p_\mu-2}{2\gamma_0 p_\mu}} \right\} \ll \bar{r}_{\mu,n} \ll \min\left\{ n^{\frac{-1}{2p_\mu}}, n^{-1/2} r_{e,n}^{-\left(1-\frac{1}{p_e+1}\right)\left(1-\frac{1}{p_\mu}\right)} \right\}.$$

Without loss of generality replace $r_{\mu,n}$ with this $\bar{r}_{\mu,n}$ in this case.

By Corollary 5, which was proved earlier in this document, it suffices to show that there is a sequence of b_n satisfying $r_{\mu,n}^{p_\mu} + r_{e,n}^{1-\frac{1}{p_e+1}} + n^{-1/2} \ll b_n \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}} \ll 1$, so that Assumption 2 and Assumption 4(ii) hold.

By assumption, $r_{\mu,n}^{p_\mu} \ll n^{-1/2}$. By construction, $r_{\mu,n} \gg n^{-1/2}$, so $\left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}} \to 0$. It therefore only remains to show that $r_{e,n}^{1-\frac{1}{p_e+1}} + n^{-1/2} \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}}$. Let n be large enough that $r_{e,n} \leq 1$. Then:

$$r_{e,n}^{\left(1-\frac{1}{p_e+1}\right)\left(\frac{p_\mu-1}{p_\mu}\right)} \leq r_{e,n}^{\left(1-\frac{1}{p_e+1}\right)\left(1-\frac{1}{p_\mu}\right)} \ll n^{-1/2} r_{\mu,n}^{-1} \Rightarrow r_{e,n}^{1-\frac{1}{p_e+1}} \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{p_\mu-1}}$$
$$n^{-1/2} = \left(n^{\frac{1-p_\mu}{2p_\mu}}\right)^{\frac{p_\mu}{(p_\mu-1)}} = \left(n^{-1/2} n^{\frac{1}{2p_\mu}}\right)^{\frac{p_\mu}{(p_\mu-1)}} \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}}.$$

Therefore there is a sequence of b_n satisfying $r_{e,n}^{1-\frac{1}{p_e+1}} + n^{-1/2} \ll b_n \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}}$. But by construction, for any such b_n, $r_{\mu,n}^{p_\mu} \ll n^{-1/2} \ll b_n \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}} \ll 1$.

Therefore the constraints of Assumption 2 and Assumption 4(ii) hold, and by Theorem 1, for this sequence of b_n, $\lim_{n\to\infty} \sup_{P\in\mathscr{P}} \sigma_n^{-2} E_P\left[\left(\hat{\psi}^{AIPW}_{clip}(b_n) - \tilde{\psi}^{AIPW}_{(Oracle)}(b_n)\right)^2\right] = 0$. But by the finite semiparametric efficiency bound under somewhat weak overlap (Newey, 1994; Hahn, 1998; Khan and Tamer, 2010),

$\lim_{n\to\infty} \sup_{P\in\mathscr{P}} \sigma_n^{-2} E_P\left[\left(\tilde{\psi}^{AIPW}_{(Oracle)}(b_n) - \tilde{\psi}^{AIPW}_{(Oracle)}(0)\right)^2\right] = 0$ and $\sup_{P\in\mathscr{P}} \left|n\sigma_n^2/AV_n - 1\right| = \sup_{P\in\mathscr{P}} \left|Var\left(\tilde{\psi}^{AIPW}_{(Oracle)}(b_n)\right)/V\right.$ 0.

Now consider the results for unthresholded AIPW under the assumption that $r_{e,n} \ll n^{\frac{-(p_e+1)}{p_e\gamma_0}}$. Recall that I imposed $r_{\mu,n} \gg n^{\frac{(2-\gamma_0)p_\mu-2}{2\gamma_0 p_\mu}}$, so that:

$$b_n \ll \left(n^{-1/2} r_{\mu,n}^{-1}\right)^{\frac{p_\mu}{(p_\mu-1)}} \ll \left(n^{\frac{(\gamma_0-2)p_\mu+2}{2\gamma_0 p_\mu} - \frac{1}{2}}\right)^{\frac{p_\mu}{(p_\mu-1)}} = n^{\frac{-1}{\gamma_0}}.$$

I first show that the probability of the thresholding mattering is small. Let P^* be a distribution in $\mathscr{P}$ satisfying $P^*(e(X) \le \pi) = \min\{1, \max\{0, C\pi^{\gamma_0-1}1\{\pi \ge b_n\}\}\}$. If p_e is finite, let π^* solve $r_{e,n}^{p_e} = \int_{b_n}^{\pi^*} C(\gamma_0 - 1)t^{\gamma_0-2}(t-b_n)^{p_e}dt$. Let $\hat{\pi}$ solve $\int_{b_n}^{\hat{\pi}} C(\gamma_0-1)(t-b_n)^{p_e+\gamma_0-2}dt = r_{e,n}^{p_e}$, i.e. $\hat{\pi} = \Theta\left(b_n + r_{e,n}^{p_e/(p_e+\gamma_0-1)}\right)$. (If p_e is infinite, take $\hat{\pi} = b_n + r_{e,n}$ so that the same statement holds.) Note that:

$$P(\hat{e} \le b_n, D=1) \le \sup_{\|\hat{e}-e\|_{L^{p_e}(P^*)} \le r_{e,n}} P^*(\hat{e} \le b_n, D=1) \le \pi^* P^*(e \le \pi^*) \le C(\pi^*)^{\gamma_0} \le C\hat{\pi}^{\gamma_0} = o(1/n).$$

Therefore with probability tending to one, the expected number of observations with $D=1, \hat{e} \le b_n$ is $o(n * n^{-1}) = o(1)$, and with probability tending to one, $\hat{\psi}^{AIPW}_{clip}(b_n) = \hat{\psi}^{AIPW}_{clip}(0)$. Therefore on an event with probability tending to one, $\sigma_n^{-2} E_P\left[\left(\hat{\psi}^{AIPW}_{clip}(0) - \tilde{\psi}^{AIPW}_{(Oracle)}(0)\right)^2\right] = o(1)$ and the feasible unthresholded estimator is semiparametric efficient. □

C.5 Degradation of Black-Box Nuisance Requirements

Proof of Corollary 3. Without loss of generality assume p_μ is infinite.

If $\gamma_0 < 2$, without loss of generality also assume p_e is infinite, and take some $r_{\mu,n}, r_{e,n} \to 0$ satisfying $r_{e,n} = o(b_n)$, and $r_{\mu,n}r_{e,n} \ll n^{-1/2} \ll r_{\mu,n}r_{e,n}r_{\mu,n}r_{e,n}(r_{e,n}/b_n)b_n^{(\gamma_0-2)/2}$, which is feasible for $r_{e,n}$ close enough to b_n because $\gamma_0 < 2$ and $b_n \to 0$.

If $\gamma_0 = 2$ so that p_e is finite, take some $r_{\mu,n}, r_{e,n} \to 0$ such that $b_n^{1+1/p_e} \gg r_{e,n}$ and such that

$$r_{\mu,n}r_{e,n} \ll n^{-1/2} \ll n^{-1/2}\left(\log(1/b_n) + \left(r_{e,n}b_n^{-(1+1/p_e)}\right)^{p_e}\right)^{1/2} \ll r_{\mu,n}r_{e,n}b_n^{-1}r_{e,n}^{p_e/(p_e+1)}r_{e,n}^{-1/(p_e+1)},$$

which is feasible for $r_{e,n}$ close enough to b_n^{1+1/p_e}.

If $\gamma_0 > 2$, take some $r_{\mu,n}, r_{e,n} \to 0$ such that $b_n^{1+1/p_e} \gg r_{e,n} \gtrsim b_n^{1+\frac{\gamma_0-1}{p_e}}$ and $r_{\mu,n}r_{e,n} \ll n^{-1/2} \ll r_{\mu,n}r_{e,n}b_n^{-1}r_{e,n}^{\frac{p_e}{p_e+\gamma_0-1}}$.

If p_e is finite, take $\pi_n > b_n$ to solve $(\pi_n^{\gamma_0-1} - b_n^{\gamma_0-1})^{1/p_e}\pi_n = r_{e,n}$. If $\gamma_0 = 2$, then $b_n \gg r_{e,n}^{p_e/(p_e+\gamma_0-1)}$, so that $\pi_n = \Theta(b_n)$ and $\pi_n^{\gamma_0-1} - b_n^{\gamma_0-1} = \Theta((r_{e,n}/b_n)^{p_e})$. Otherwise, $b_n \lesssim r_{e,n}^{p_e/(p_e+\gamma_0-1)}$ so that $\pi_n =$

$\Theta\left(r_{e,n}^{\frac{p_e}{p_e+\gamma_0-1}}\right)$ and $\pi_n^{\gamma_0-1} - b_n^{\gamma_0-1} = \Theta\left(r_{e,n}^{\frac{p_e(\gamma_0-1)}{p_e+\gamma_0-1}}\right)$. If p_e is infinite, take $\pi_n = b_n + r_{e,n}$.

Take P to be the distribution with $X \sim Unif([0,1])$, $D \mid X \sim Bern(X^{1/(\gamma_0-1)})$, and $Y \mid D, X \sim \mathcal{N}(0,1)$, and take $\mathscr{P} = \{P\}$. This family satisfies Assumption 1 for this γ_0, $C = 1$, and $\sigma_{\max} = 1$.

Construct the nuisance function estimates as $\hat{\mu}(X) = \mu(X) - r_{\mu,n}1\{e(X) \in (b_n, \pi_n]\}$ and $\hat{e}(X) = e(X) - (e(X) - b_n)1\{e(X) \in (b_n, \pi_n]\}$. These nuisance function estimates are deterministic and uniformly consistent, so they satisfy Assumption 2.

I next show that $\|\hat{\mu} - \mu\|_{L^{p_\mu}(P)} \leq r_{\mu,n}$ and $\|\hat{e} - e\|_{L^{p_e}(P)} \leq r_{e,n}$. This is clearly true for infinite p_μ and p_e; if p_e is finite, then these hold by construction after observing that $P(\hat{e} \neq e) = P(\hat{\mu} \neq \mu) = P(e \in (b_n, \pi_n]) = \pi_n^{\gamma_0-1} - b_n^{\gamma_0-1}$.

Note that by construction, $r_{e,n} \ll b_n^{1+\min\{\gamma_0-1,1\}/p_e}$, depending on the separate cases of $\gamma_0 < 2$ (and imposing p_e infinite) or $\gamma_0 \geq 2$ (and imposing $r_{e,n} \ll b_n^{1+1/p_e}$). Note also that the outcome regression estimate is sup-norm consistent. Therefore Assumptions 3(a),(b) hold.

Note also that by construction of $r_{\mu,n}, r_{e,n}$, $r_{\mu,n}r_{e,n} \ll n^{-1/2}$, so that $\sup_{P \in \mathscr{P}} \|\hat{\mu}_n^{(-k)} - \mu\|_{L^{p_\mu}(P)} \times \|\hat{e}_n^{(-k)} - e\|_{L^{p_\mu}(P)} \ll n^{-1/2}$.

It remains to show that Wald inference fails. I show that the squared bias is of a greater order than the oracle variance order $n^{-1}\left(b_n^{\gamma_0-2} + \log(1/b_n)^{1\{\gamma_0=2\}}\right)$, while the sample and estimated variance are of the same order as the oracle variance. That will imply that t-statistics fail to cover asymptotically, because estimate $\pm$ 1.96 standard errors will be of the same order as the bias plus a term of the same order as the oracle standard deviation.

Recalling that the true average potential outcome is zero. Then the bias is:

$$\begin{aligned} B_n = E_P[\hat{\phi}_n] &= (\gamma_0 - 1)\int_{b_n}^{\pi_n} r_{\mu,n}\pi^{\gamma_0-2}\left(\frac{\pi - b_n}{b_n}\right)d\pi \geq \frac{\gamma_0 - 1}{2} r_{\mu,n} b_n^{-1} \int_{(\pi_n+b_n)/2}^{\pi_n} \pi^{\gamma_0-2}(\pi_n - b_n)\, d\pi \\ &= \Omega\left(r_{\mu,n}b_n^{-1}(\pi_n - b_n)(\pi_n^{\gamma_0-1} - b_n^{\gamma_0-1})\right) \end{aligned}$$

If p_e is infinite, then $\gamma_0 < 2$ (otherwise $\zeta < 1$ would be impossible) and $B_n = \Omega\left(r_{\mu,n}b_n^{-1}r_{e,n}^2 b_n^{\gamma_0-3}\right) \gg n^{-1/2}b_n^{(\gamma_0-2)/2}$. If p_e is finite, then $\gamma_0 \geq 2$ by construction and $B_n = \Omega\left(r_{\mu,n}r_{e,n}b_n^{-1}r_{e,n}^{\frac{(p_e-1)(\gamma_0-1)}{p_e+\gamma_0-1}}\right)$. If $\gamma_0 = 2$, then $B_n = \Omega\left(r_{\mu,n}r_{e,n}b_n^{-1}r_{e,n}^{\frac{p_e}{p_e+1} - \frac{1}{p_e+1}}\right)$. If $\gamma_0 > 2$, $B_n \gg r_{\mu,n}r_{e,n}b_n^{-1}r_{e,n}^{\frac{p_e}{p_e+\gamma_0-1}} \gg n^{-1/2}$.

Note that

$$\begin{aligned} Var\left(\hat{\phi}_n\right) &\leq 2Var\left(\phi_n\right) + 2Var\left(\hat{\phi}_n - \phi_n\right) \\ &= O\left(b_n^{\gamma_0-2} + \log(1/b_n)^{1\{\gamma_0=2\}}\right) + O\left(\pi_n(\pi_n^{\gamma_0-1} - b_n^{\gamma_0-1})b_n^{-2}\right). \end{aligned}$$

If $\gamma_0 < 2$, then $\pi_n = O(b_n)$, so that this term is $O(b_n^{\gamma_0 - 2})$. If $\gamma_0 = 2$, then $\pi_n^{\gamma_0 - 1} - b_n^{\gamma_0 - 1} = \Theta\left((r_{e,n}/b_n)^{p_e}\right)$, so that this term is $O\left(\log(1/b_n) + \left(r_{e,n} b_n^{-(1+1/p_e)}\right)^{p_e}\right)$. If $\gamma_0 > 2$, then this term is $O(1)$. Therefore $B_n^2 \gg n^{-1} Var\left(\hat{\phi}_n\right)$ for any value of γ_0.

Note also that $1 \gg b_n \gg n^{-1/2}$, $1 \gg r_{\mu,n}$, and $r_{e,n} \ll b_n \ll b_n^{\frac{\gamma_0 - 1 + p_e}{(\gamma_0 - 1)(p_\mu p_e + 1 - p_\mu - p_e) + p_e}}$ for $p_\mu = p_e = \infty$, so that by the proof of Corollary 5, the conditions of Proposition 1 and Assumption 3'(a),(b) hold. Therefore by Lemma 18, $\hat{\sigma}_n^2 / \sigma_n^2 \xrightarrow{\mathscr{P}} 1$.

As a result, the Wald confidence intervals fail asymptotically:

$$\begin{aligned} P(\psi(P) \in \hat{\mathcal{C}}_n(\alpha)) &= P\left(\hat{\sigma}_n z_{\alpha/2} \leq \hat{\psi}_{clip}^{AIPW}(b_n) - \psi_0 \leq \hat{\sigma}_n z_{1-\alpha/2}\right) \\ &\leq P\left(\hat{\psi}_{clip}^{AIPW}(b_n) - E\left[\hat{\psi}_{clip}^{AIPW}(b_n)\right] + B_n \leq \hat{\sigma}_n z_{1-\alpha/2}\right) \\ &= P\left(o_P(B_n) + B_n \leq o_P(B_n)\right) = o(1). \end{aligned}$$

Therefore Wald confidence intervals fail to cover asymptotically, despite the usual product rate holding. □

Proof of Example 1. For simplicity, I proceed assuming $r_{e,n} \gg n^{-1/2}$. By Theorem 1 and the sup-norm bounds, it remains to show that there is a $b_n \to 0$ such that 3(c) ($r_{\mu,n} b_n^{\gamma_0/2} \ll n^{-1/2}$) and 3(b) ($r_{e,n} \ll b_n$).

Because $r_{\mu,n} r_{e,n} \ll n^{-1/2}$, there exists some $\delta_n \to \infty$ such that $r_{\mu,n} r_{e,n} \delta_n \ll n^{-1/2}$. Choose some b_n such that $1 \gg b_n$ and $r_{e,n} \ll b_n \ll (r_{e,n}\delta_n)^{2/\gamma_0}$. This is feasible because $\gamma_0 > 2$, so that $r_{e,n}^{2/\gamma_0} \gg r_{e,n}$ and $\delta_n^{2/\gamma_0} \to \infty$. Then $r_{e,n} \ll b_n$ and $r_{\mu,n} b_n^{\gamma_0/2} \ll r_{\mu,n} r_{e,n} \delta_n \ll n^{-1/2}$. Thus, both conditions hold. □

Proof of Example 2. For simplicity, I proceed assuming $r_{e,n} \gg n^{-1/2}$. By Theorem 1, it remains to show that there is a $b_n \to 0$ such that $r_{\mu,n} r_{e,n} \log(1/b_n) \ll n^{-1/2}$, $r_{\mu,n} b_n \ll n^{-1/2}$, and $r_{e,n} \ll b_n$.

Because $r_{\mu,n} r_{e,n} \log(1/r_{e,n}) \ll n^{-1/2}$, there exists a b_n such that $r_{e,n} \ll b_n \ll 1$ and $r_{\mu,n} b_n \log(1/b_n) \ll n^{-1/2}$. For this b_n, all three conditions hold by inspection. □

Proof of Example 3. For simplicity, I proceed assuming $r_{e,n}, r_{\mu,n} \gg n^{-1/2}$.

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

Now suppose that $\gamma_0 < 2$. Take $b_n = r_{e,n}\left(\max\left\{r_{e,n}, n^{1/2} r_{\mu,n} r_{e,n}^{\gamma_0/2}\right\}\right)^{-1/2}$. Note that $n^{-1/2} \ll r_{e,n} \ll b_n$ and $r_{\mu,n} b_n^{\frac{\gamma_0}{2}} \precsim n^{-1/2}\left(n^{1/2} r_{\mu,n} r_{e,n}^{\gamma_0/2}\right)^{1/2} \ll n^{-1/2}$, so that all conditions of Assumption 3 hold by inspection.

□

Proof of Example 4. If $\gamma_0 \geq 2$, the claim holds by Example 1 and Example 2.

Now suppose that $\gamma_0 < 2$, and consider both cases.

(i) $r_{\mu,n} = O(n^{-1/2})$. Take $b_n \to 0$ such that $b_n \gg r_{e,n}$. Then $b_n^{(\gamma_0 - 2)/2} \ll r_{e,n}^{(\gamma_0 - 2)/2} \ll r_{e,n}^{-1/2}$ and $r_{\mu,n} b_n^{\gamma_0/2} \ll n^{-1/2}$, so that all conditions of Assumption 3 hold by inspection.

(ii) $r_{e,n} = O(n^{-1/2})$. For simplicity, assume $r_{\mu,n} \gtrsim n^{-1/2}$. Take $b_n = n^{-1/2}\left(n^{-1/8} r_{\mu,n}^{-1/2}\right)$, so that $1 \gg b_n \gg n^{-1/2} \gtrsim r_{e,n}$. Note that $r_{\mu,n} b_n^{\frac{\gamma_0}{2}} \ll r_{\mu,n} b_n^{\frac{1}{2}} = n^{-1/4} r_{\mu,n} \frac{3}{4} n^{\frac{-1}{16}} \ll n^{-1/2}$, so that all conditions of Assumption 3 hold by inspection. □

C.6 Necessary Smoothness Conditions

Lemma 20 (Minimal expected nearby observations)**.** *Suppose the conditions of Theorem 2 hold. Define* $A(x_0 \mid h) = \{x : x \in [-1,1]^d, \|x - x_j\| \leq h\}$. *Then*

$$\inf_{x_0 \in [-1,1]^d, P \in \mathscr{P}, h > 0} E_P\left[D \mid X \in A(x_0 \mid h)\right] \geq 2^{-(\gamma_0+1)/(\gamma_0-1)} C^{-1/(\gamma_0-1)} h^{\frac{d}{\gamma_0-1}}.$$

Proof of Lemma 20. Proof by contradiction. Suppose not, and there is a $P \in \mathscr{P}$, $x_0 \in [-1,1]^d$, $h > 0$ with

$$E_P\left[D \mid X \in A(x_0 \mid h)\right] < 2^{-(\gamma_0+1)/(\gamma_0-1)} C^{-1/(\gamma_0-1)} h^{\frac{d}{\gamma_0-1}}.$$

Define $\pi = 2^{-(\gamma_0+1)/(\gamma_0-1)} C^{-1/(\gamma_0-1)} h^{\frac{d}{\gamma_0-1}}$ and $B = \{X \in A(x_0 \mid h) : e(X) \leq \pi\}$. Then $P(X \in B) < P(X \in A(x_0 \mid h))/2 \geq h^d/2$. As a result:

$$P\left(e(X) \leq \pi\right) > \frac{h^d}{2} = C\pi^{\gamma_0-1} \frac{C^{-1} h^d \pi^{1-\gamma_0}}{2} = C\pi^{\gamma_0-1} C^{-1} 2^{-1-\gamma_0} h^d \left(2^{\frac{\gamma_0+1}{1-\gamma_0}} C^{\frac{-1}{\gamma_0-1}} h^{\frac{d}{\gamma_0-1}}\right)^{1-\gamma_0} = C\pi^{\gamma_0-1}.$$

Contradiction. □

Lemma 21 (Minimal coefficient)**.** *Suppose Assumption 6 holds and* $e(X) \in \Sigma(\beta_e, L)$ *for some* $\beta_e > \frac{d}{\gamma_0-1}$. *For* $0 \leq j < \beta_e$, *Let* $c_j(v \mid x_0)$ *be the coefficient in the* j^{th}*-order Taylor expansion of* $e(x)$ *around* x_0 *applied in the direction of* v. *Then there is a* $c^* > 0$ *such that for all* x_0, *there exists an* $x \neq x_0$ *and a* $0 \leq j \leq \alpha^{(Mou)}$ *such that* $x_0 + \frac{x - x_0}{\|x - x_0\|} \in [-1,1]^d$ *and* $c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right) \geq c^*$.

Proof of Lemma 21. If $\beta_e \leq 1$, the claim is immediate by Lemma 20. I therefore proceed assuming $\beta_e > 1$.

Proof by contradiction. Suppose for all $c > 0$, there exists a P and an x_0 such that $c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right) < c$ for all $j \leq \alpha^{(Mou)}$.

Define:

$$h^* = \text{argsup}_{h \in (0,1]} \sum_{\alpha^{(Mou)} < j \leq \ell_e} \bar{c}_j (h)^j + L_e (h)^{\beta_e} \leq \frac{C^{1/d}}{\lceil \alpha^{(Mou)} + 2 \rceil} (h)^{\frac{d}{\gamma_0-1}}.$$

This h^* is well-defined because as h tends to zero from above, the left-hand side is of a lower order than the right-hand side. By continuity, $h_n^* \in (0,1]$ and satisfies this weak inequality.

Take $c^* = \min_{0\leq j\leq \alpha^{(Mou)}} \frac{C^{\frac{1}{1-\gamma_0}}}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}-j}$. I claim that for this c^*, the the content of Lemma 21 holds.

Proof by contradiction. Suppose, not, and there is an $x_0 \in [-1,1]^d$ and an x as above such that $c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right) < c^*$ for all $j \leq \alpha^{(Mou)}$. Then for all $x \in [-1,1]^d/x_0$ such that $\|x-x_0\| \leq h^*$ and $x_0 + \frac{x-x_0}{\|x-x_0\|} \in [-1,1]^d$:

$$\begin{aligned}
e(x) &= f(x \mid x_0) + g(x \mid x_0) \\
&= \sum_{0\leq j\leq \alpha^{(Mou)}} c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right)\|x-x_0\|^j + \sum_{\alpha^{(Mou)}<j\leq \ell_e} c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right)\|x-x_0\|^j + L\|x-x_0\|^{\beta_e} \\
&\leq \sum_{0\leq j\leq \alpha^{(Mou)}} \frac{C^{\frac{1}{1-\gamma_0}}}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}-j}\|x-x_0\|^j + \frac{C^{\frac{1}{1-\gamma_0}}}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}} \leq C^{\frac{1}{1-\gamma_0}}\frac{\lceil \alpha^{(Mou)}+1\rceil}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}}.
\end{aligned}$$

So that:

$$\begin{aligned}
P\left(e(X) \leq C^{\frac{1}{1-\gamma_0}}\frac{\lceil \alpha^{(Mou)}+1\rceil}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}}\right) &\geq P\left(x_0 + \frac{X-x_0}{\|X-x_0\|} \in [-1,1]^d, \|X-x_0\| \geq h^*\right) \\
&\geq (h^*)^d = C\left(C^{\frac{1}{1-\gamma_0}}(h^*)^{d/(\gamma_0-1)}\right)^{\gamma_0-1} \\
&= C\left(\frac{\lceil \alpha^{(Mou)}+2\rceil}{\lceil \alpha^{(Mou)}+1\rceil}\right)^{\gamma_0-1}\left(C^{\frac{1}{1-\gamma_0}}\frac{\lceil \alpha^{(Mou)}+1\rceil}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}}\right)^{\gamma_0-1} \\
&> C\left(C^{\frac{1}{1-\gamma_0}}\frac{\lceil \alpha^{(Mou)}+1\rceil}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}}\right)^{\gamma_0-1}.
\end{aligned}$$

But by Assumption 1(d), $P\left(e(X) \leq C^{\frac{1}{1-\gamma_0}}\frac{\lceil \alpha^{(Mou)}+1\rceil}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}}\right) \leq C\left(C^{\frac{1}{1-\gamma_0}}\frac{\lceil \alpha^{(Mou)}+1\rceil}{\lceil \alpha^{(Mou)}+2\rceil}(h^*)^{\frac{d}{\gamma_0-1}}\right)^{\gamma_0-1}$. Contradiction.

Therefore, for this $c^* > 0$ and all x_0, there exists an $x \neq x_0$ and a $0 \leq j \leq \alpha^{(Mou)}$ such that $c_j\left(\frac{x-x_0}{\|x-x_0\|} \mid x_0\right) \geq c^*$. □

Proof of Proposition 4. Proof by contradiction. Suppose not, and for all $\rho, \nu > 0$, there is an

$$h \in \left(0, (C^{1/d}/(2L))^{\frac{1}{\beta_e}-\frac{d}{\gamma_0-1}}\right],$$

$x_0 \in [-1,1]^d$, and $P \in \mathscr{P}$ such that $P(e(X) \geq \rho \sup_{\|x-x_0\|\leq h} e(x) \mid D=1, \|X-x_0\| \leq h) \leq \nu$.

Take some $\rho > 0$ and some sequence of $\gamma_n \to 0^+$, with associated bandwidths h_n, such that

$$P(n)\left(e(X) \geq \rho\left(\sup_{\|x-x_0\|\leq h_n} e(x)\right) \mid D=1, \|X-x_0\| \leq h_n\right) \leq \gamma_n.$$

Because $[-1,1]^d$ is compact and the coefficients are in a compact space, there is a subsequence of n for which x_0 and the local polynomial coefficients of $e(x)$ around x_0 are convergent. Without loss of generality, suppose this is the sequence. Write $g_n(x \mid x_0) = P(n)(D=1 \mid X=x) - f_n(x \mid x_0)$ be the local polynomial propensity residuals. Also write $f^*(\cdot \mid x_0)$ for the local polynomial coefficients at the limiting coefficients.

First, I show that $h_n \to 0$. Suppose not, and $\limsup_{n\to\infty} h_n(x) = h^* > 0$. Without loss of generality, suppose $\lim_{n\to\infty} h_n(x) = h^*$. Write $A = \{x \mid \|x - x_0\| \leq h^*\}$. Then by continuity of densities and bounds on derivatives, for all n large enough,

$$
\begin{aligned}
\gamma_n &\geq P(n)\left(e(X) \geq \frac{\rho}{2}\left(\sup_{\|x-x_0\|\leq h^*} e(x)\right) \mid D=1, \|X-x_0\| \leq h^*\right) \\
&\geq P(n)\left(e(X) \geq \frac{3\rho}{8}\left(\sup_{\|x-x_0\|\leq h^*} f^*(x \mid x_0)\right) \mid D=1, \|X-x_0\| \leq h^*\right) \\
&\quad - 1\left\{\sup_{\|x-x_0\|\leq h^*} |e(x) - f^*(x \mid x_0)| \geq \frac{\rho}{8} \sup_{\|x-x_0\|\leq h^*} f^*(x \mid x_0)\right\} \\
&\geq P(n)\left(e(X) \geq \frac{3\rho}{8}\left(\sup_{\|x-x_0\|\leq h^*} f^*(x \mid x_0)\right) \mid D=1, \|X-x_0\| \leq h^*\right) \\
&\quad - 1\left\{\sup_{x\in A} \|f_n(x \mid x_0) - f^*(x \mid x_0)\| + |g_n(x \mid x_0)| \geq \frac{\rho}{16} \sup_{\|x-x_0\|\leq h^*} f^*(x \mid x_0)\right\} \\
&= P(n)\left(e(X) \geq \frac{3\rho}{8}\left(\sup_{\|x-x_0\|\leq h^*} f^*(x \mid x_0)\right) \mid D=1, \|X-x_0\| \leq h^*\right) - 1\left\{O((h^*)^{\beta_e}) \geq \Omega((h^*)^{\frac{-d}{\gamma_0-1}}\right\} \\
&\geq P(n)\left(e(X) \geq \frac{\rho}{4}\left(\sup_{\|x-x_0\|\leq h^*} f^*(x \mid x_0)\right) \mid D=1, \|X-x_0\| \leq h^*\right),
\end{aligned}
$$

which is a positive constant. Therefore $\gamma_n \not\to 0$. Contradiction.

I therefore proceed assuming $h_n \to 0$. Let the lowest-order nonzero coefficient in f^* be of order j^*. j^* is defined and finite by Lemma 21. Define

$$
G_n = P(n)\left(e(X) \geq \frac{\rho}{2}\left(\sup_{\|x-x_0\|\leq h^*} e(x)\right) \mid D=1, \|X-x_0\| \leq h^*\right) - \gamma_n.
$$

I wish to show that G_n does not converge to zero. By the Bolzano-Weierstrass Theorem, it suffices to show that there is a nonconvergent subsequence of n.

Write $m_{n,j} \equiv \sup_{\|v\|=1} |\tilde{c}_j(v)|\, h_n^{j-j^*}$ and $m_n \equiv \max_{0\leq j\leq j^*} m_{n,j}$. I consider two cases: (i) $m_n \to 0$ or (ii) $\liminf_{n\to\infty} m_n > 0$ (and potentially infinite).

Case (i): suppose $m_n \to 0$. Then for all $x \in A_n$,

$$
e(x) = f^*(x \mid x_0) + \tilde{f}_n(x \mid x_0) + g_n(x \mid x_0)
$$

$$
\begin{aligned}
&= \sup_{\|\alpha\|=j^*} D^\alpha e(x_0)(x-x_0)^\alpha + O(h_n^{\min\{j^*+1,\beta_e\}}) + o(h_n^{j^*}) + O(h_n^{\beta_e}) \\
&= h_n^{j^*} c_{j^*}\left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_n}\right)^{j^*} + o(h_n^{j^*}).
\end{aligned}
$$

Let $n \geq n'$ imply that the $o(h_n^{j^*})$ term is at most half as large as $\sup_{x\in A_n} h_n^{j^*} c_{j^*}\left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_n}\right)^{j^*}$, as well as to imply that $x_0 + h_n v \in [-1,1]^d$ if and only if there is an $h > 0$ such that $x_0 + hv \in [-1,1]^d$. Then:

$$
\begin{aligned}
&P(n)\left(e(X) \geq \sup_{x\in A_n} e(x) \mid D=1, X \in A_n\right) \\
\geq &P(n)\left(h_n^{j^*} c_{j^*}\left(\frac{X-x_0}{\|X-x_0\|}\right)\left(\frac{\|X-x_0\|}{h_n}\right)^{j^*} \geq \frac{\rho}{2}\sup_{x\in A_n} h_n^{j^*} c_{j^*}\left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_n}\right)^{j^*} \mid D=1, X\in A_n\right) \\
= &P(n')\left(c_{j^*}\left(\frac{X-x_0}{\|X-x_0\|}\right)\left(\frac{\|X-x_0\|}{h_{n'}}\right)^{j^*} \geq \frac{\rho}{2}\sup_{x\in A_{n'}} c_{j^*}\left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_{n'}}\right)^{j^*} \mid D=1, X\in A_{n'}\right) > 0.
\end{aligned}
$$

Contradiction.

Case (ii): suppose $\liminf_{n\to\infty} m_n > 0$. Write

$$
\tilde{f}_n(x \mid x_0) = h_n^{j^*} m_n \sum_{j=0}^{\ell_e} \underbrace{\frac{\sum_{\|\alpha\|=j} D^\alpha e(x_0)\left(\frac{x-x_0}{\|x-x_0\|}\right)^\alpha h_n^{j-j^*}}{m_n}}_{\text{“}\tilde{d}_{j,n}\left(\frac{x-x_0}{\|x-x_0\|}\right)} \left(\frac{\|x-x_0\|}{h_n}\right)^j.
$$

By construction, $\tilde{d}_{j,n}\left(\frac{x-x_0}{\|x-x_0\|}\right)$ is bounded between -1 and 1, so that there is a convergent subsequence for all α. Without loss of generality I proceed assuming this is the full sequence.

Write $\tilde{d}_j^*(v) = \lim_{n\to\infty} \tilde{d}_{j,n}(v)$ for all v. Write $\tilde{d}_{j^*}^*(v) = c_{j^*}(v)/m_n$, where $c_{j^*}(v)$ is the j^*-order coefficient in f^*. By construction, for all $\|x-x_0\| \leq h_n$,

$$
\begin{aligned}
e(x \mid x_0) &= f^*(x \mid x_0) + \tilde{f}_n(x \mid x_0) + g_n(x \mid x_0) \\
&= m_n h_n^{j^*}\left(\sum_{j=0}^{j^*} \tilde{d}_j^*\left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_n}\right)^j + o(1) + O\left(h_n^{\min\{1,\beta_e-j^*\}}\right)\right) \\
&= m_n h_n^{j^*} \sum_{j=0}^{j^*} \tilde{d}_j^*\left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_n}\right)^j + o(m_n h_n^{j^*}).
\end{aligned}
$$

Let $n \geq n'$ imply that the $o(m_n h_n^{j^*})$ term is at most half as large as the largest value of the first term over $x \in A_n$, as well as that $x_0 + h_n v \in [-1,1]^d$ if and only if there is an $h > 0$ such that $x_0 + hv \in [-1,1]^d$.

Then:

$$
\begin{aligned}
&P(n)\left(e(X) \geq \sup_{x \in A_n} e(x) \mid D=1, X \in A_n\right) \\
\geq &P(n)\left(m_n h_n^{j^*} \sum_{j=0}^{j^*} \tilde{d}_j^* \left(\frac{X-x_0}{\|X-x_0\|}\right)\left(\frac{\|X-x_0\|}{h_n}\right)^j \geq \frac{\rho}{2} \sup_{x \in A_n} m_n h_n^{j^*} \sum_{j=0}^{j^*} \tilde{d}_j^* \left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_n}\right)^j \mid D=1, X \in A_n\right) \\
= &P(n')\left(\sum_{j=0}^{j^*} \tilde{d}_j^* \left(\frac{X-x_0}{\|X-x_0\|}\right)\left(\frac{\|X-x_0\|}{h_{n'}}\right)^j \geq \frac{\rho}{2} \sup_{x \in A_{n'}} \sum_{j=0}^{j^*} \tilde{d}_j^* \left(\frac{x-x_0}{\|x-x_0\|}\right)\left(\frac{\|x-x_0\|}{h_{n'}}\right)^j \mid D=1, X \in A_{n'}\right) > 0.
\end{aligned}
$$

Contradiction.

Therefore there is a $\rho, \nu > 0$ such that for all $h > 0$ small enough, for all $P \in \mathscr{P}$ and $x_0 \in [-1,1]^d$ $P(e(X) \geq \rho \sup_{\|x-x_0\| \leq h} e(x) \mid D=1, \|X-x_0\| \leq h) > \nu$. □

Lemma 22 (Minimal eigenvalue technical result)**.** *Suppose the conditions of Theorem 2 hold. There is an $h' > 0$ such that*

$$
\lim_{\varepsilon \to +0} \sup_{P \in \mathscr{P}, x_0 \in [-1,1]^d, h \in (0,h'], \|v\|=1} P\left(\|v^T U\| \leq \varepsilon \mid D=1, \|X-x_0\| \leq h\right) = 0.
$$

Proof of Lemma 22. Take some sequence of $\varepsilon_n \to^+ 0$. Take h', ρ, ν from Assumption 7.

Let $P(n)$ be a sequence of distributions in $\mathscr{P}$, let v_n be a sequence of vectors with $\|v\| = 1$, let h_n be a sequence in $(0, h']$, and write $A_n = \{x : \|x - x_0\| \leq h_n\}$. Then:

$$
\begin{aligned}
P(n)\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid D=1, X \in A_n\right) &= \frac{E_{P(n)}\left[e(X) 1\left\{\|v_n^T U\| \leq \sqrt{\varepsilon_n}\right\} \mid X \in A_n\right]}{E_{P(n)}\left[e(X) \mid X \in A_n\right]} \\
&\leq \frac{\left(\sup_{x \in A_n} e(x)\right) P(n)\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid X \in A_n\right)}{P(n)\left(e(X) \geq \rho\left(\sup_{x \in A_n} e(x)\right)\right) \rho\left(\sup_{x \in A_n} e(x)\right)} \\
&\leq \frac{\left(\sup_{x \in A_n} e(x)\right) P(n)\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid X \in A_n\right)}{\nu \rho \sup_{x \in A_n} e(x)} \\
&= \frac{P(n)\left(\|v_n^T U\| \leq \sqrt{\varepsilon_n} \mid X \in A_n\right)}{\rho \nu} = O\left(\sqrt{\varepsilon_n}\right) = o(1).
\end{aligned}
$$

□

Lemma 23 (Uniform minimal expected eigenvalue)**.** *Suppose the conditions of Theorem 2 hold, and let $U(v)$ be the vector of zero-through-$\lfloor \beta_e \rfloor$-order interactions of v. Define:*

$$
\begin{aligned}
\lambda(\bar{h}) &\equiv \inf_{h \in (0,\bar{h}], \|v\|=1, x_0 \in [-1,1]^d, P \in \mathscr{P}} v^T E_P\left[U\left(\frac{X-x_0}{h}\right) U\left(\frac{X-x_0}{h}\right)^T \mid D=1, \|X-x_0\| \leq h\right] v, \\
\lambda^* &\equiv \liminf_{\bar{h} \to +0} \lambda(\bar{h}),
\end{aligned}
$$

where $U(\cdot)$ is the $\lfloor \beta_\mu \rfloor$-order local polynomial interaction matrix. Then $\lambda^* > 0$.

Proof of Lemma 23. Let h' be from Lemma 22. Fix some $\bar{h} \in (0, h']$. Let h_n be a sequence in $(0, \bar{h}]$, let v_n be a sequence of vectors with $\|v\| = 1$, let $x_{0,n}$ be a sequence of points in $[-1,1]^d$, and let $P(n)$ be a sequence of distributions in $\mathscr{P}$. Let $A_n = \{x : \|x - x_{0,n}\| \leq h_n\}$.

By Lemma 22, there is a $\varepsilon > 0$ such that:

$$\inf_{P \in \mathscr{P}, x_0 \in [-1,1]^d, h \in (0, h_n], \|v\|=1} P\left(\|v^T U\| \geq \varepsilon \mid D = 1, \|X - x_0\| \leq h\right) > 0.$$

Call this infimum $\delta > 0$. $\|v^T U\|$ is bounded, so δ is finite.

Define:

$$\begin{aligned}
\lambda_n &\equiv v_n^T E_{P(n)}\left[U\left(\frac{X - x_{0,n}}{h_n}\right) U\left(\frac{X - x_{0,n}}{h_n}\right)^T \mid D = 1, X \in A_n\right] v_n \\
&= E_{P(n)}\left[v_n^T U\left(\frac{X - x_{0,n}}{h_n}\right)\left(v_n^T U\left(\frac{X - x_{0,n}}{h_n}\right)\right)^T \mid D = 1, X \in A_n\right] \\
&= E_{P(n)}\left[\left(v_n^T U\left(\frac{X - x_{0,n}}{h_n}\right)\right)^2 \mid D = 1, X \in A_n\right] \\
&\geq \varepsilon^2 P(n)\left(\left|v_n^T U\left(\frac{X - x_{0,n}}{h_n}\right)\right| \geq \varepsilon \mid D = 1, X \in A_n\right) \geq \varepsilon^2 \delta > 0.
\end{aligned}$$

Therefore, for all $\bar{h} \leq h'$, $\lambda(\bar{h}) \geq \varepsilon^2 \delta$. Therefore $\lambda^* \geq \varepsilon^2 \delta > 0$. □

Lemma 24 (Uniform functional approximation across multiple gridpoints)**.** *Suppose the conditions of Theorem 2 hold. Let $f_{k_n,n}(v) : [-1,1]^d \to \mathbb{R}$ be a set of k_n functions that are uniformly bounded. Let $\mathcal{S}(m, \underline{h})$ be the set of sets of m tuples (x, h) with $x \in [-1,1]^d$ and $h \in [\underline{h}_n, h']$, where h' comes from Lemma 23's notion of h small enough and where for all $(x_1, h_1), (x_2, h_2) \in S \in \mathcal{S}(m, \underline{h})$ with $x_2 \neq x_1$, there are no points $x \in [-1,1]^d$ with $\|x - x_1\| \leq h_1$ and $\|x - x_2\| \leq h_2$. Write $\tilde{c} = 2^{\frac{2(1+\gamma_0)}{1-\gamma_0}} C^{\frac{2}{1-\gamma_0}} (\sup_v |f(v)|)^{-2}$. Suppose (i) $\underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}} \gg n^{-1}$; (ii) m_n be a sequence tending to infinity such for all fixed $a > 0$, $m_n \ll exp\left(-an\underline{h}_n^{\frac{d\gamma_0}{\gamma_0-1}}\right)$; and (iii) and $k_n \left(1 - \left(\frac{e-1}{e}\right)^{\frac{m_n}{exp\left(\log(1/2) + \tilde{c}^{-1} n^{-1} \underline{h}_n^{-\frac{d\gamma_0}{\gamma_0-1}}\right)}}\right) \to 0$. Then there is a sequence of $\varepsilon_n \to^+ 0$ such that for all $k = 1, \ldots, k_n$:*

$$\sup_{P \in \mathscr{P}, S \in \mathcal{S}(m_n, \underline{h}_n)} P\left(\max_j \left| \frac{\sum_i D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right) f_{k,n}\left(\frac{X_i - x_j}{h_{n,j}}\right) - E\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right) f_{k,n}\left(\frac{X - x_j}{h_{n,j}}\right)\right]}{nE_{P(n)}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \right| \geq \varepsilon_n \right) = o(1).$$

Proof of Lemma 24. This proof will get hairy. For simplicity, I proceed assuming K is the uniform bandwidth

scaled downwards by one-half; the proof would require far more care if the kernel took on nonzero values for an unbounded set. Also, for simplicity, assume that $e(X)$ is bounded above by one-half — the difficulty here comes from small propensity scores, and is already substantial. Let the upper bound of $|f|$ be $b \geq 0$.

Define $R_n \equiv \tilde{c} n \underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}$ and

$$V_{n,j,k} \equiv \sum_i D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right) \frac{f_{k,n}\left(\frac{X_i - x_j}{h_{n,j}}\right) - E\left[f_{k,n}\left(\frac{X - x_j}{h_{n,j}}\right) \mid DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right) = 1\right]}{nE_{P(n)}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]}.$$

By construction, there is a sequence of $\varepsilon_n^+ \to 0$ such that $k_n\left(1 - (1 - 1/e)^{\frac{m_n}{exp\left(\log(1/2) + R_n \varepsilon_n^2\right)}}\right) \to 0$.

Note that the events $D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)$ are mutually disjoint for a given i across j; therefore write $j(i)$ for the j such that $D_i K\left(\frac{\|X_i - x_{j(i)}\|}{h_{n,j}}\right) = 1$ if feasible, and write $j(i) = 0$ if no such j exists.

Let $P(n)$ be a sequence of distributions in $\mathscr{P}$, let S_n be a sequence of sets in $\mathcal{S}(m_n, \underline{h}_n)$, and let $\{(x_{n,j}, h_{n,j})\}$ be a sequence of associated points and bandwidths. By Lemma 20, for all j, n,

$$E_{P(n)}\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right] \geq C^{1/(1-\gamma_0)} 2^{\frac{1+\gamma_0}{1-\gamma_0}} \underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}.$$

Note that I use a laxer bound for $\underline{h}_n$ because the polynomial order here is found elsewhere, and the polynomial order in Lemma 20 is not found elsewhere. Thus, the argument will continue to hold in the presence of certain typos.

Consider the event A of $\{i, j(i)\}$. Note that for any given j, by the Chernoff bound for binomial random variables,

$$P(n)\left(\frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \geq 2\right] \leq exp\left(\frac{-nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]}{3}\right) \leq exp\left(-\tilde{c}\frac{b^2}{3} n \underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}\right),$$

$$\begin{aligned}\text{So that } P(n)\left(\max_j \frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \leq 2\right) &\leq 1 - \sum_j P(n)\left(\frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \geq 2\right] \\ &\leq 1 - O\left(m_n exp\left(-\tilde{c}\frac{b^2}{3} n \underline{h}_n^{\frac{d\gamma_0}{\gamma_0 - 1}}\right)\right) = 1 - o(1).\end{aligned}$$

I therefore proceed under the high probability event that A is such that $\max_j \frac{\sum D_i K\left(\frac{\|X_i - x_j\|}{h_{n,j}}\right)}{nE\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]} \leq 2$.

I now apply the Hoeffding inequality to the $\sum DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right) \leq 2E\left[DK\left(\frac{\|X - x_j\|}{h_{n,j}}\right)\right]$ elements of $V_{n,j}$

conditional on A, for all k, n, j

$$P(n)\left(|V_{n,j,k}| \geq \varepsilon_n\right) \leq 2exp\left(\frac{-2\left(nE_{P(n)}\left[DK\left(\frac{\|X-x_j\|}{h_{n,j}}\right)\right]\right)^2\varepsilon_n^2}{\left(\sum D_i K\left(\frac{\|X_i-x_j\|}{h_{n,j}}\right)\right)b^2}\right) \leq 2exp\left(-R_n\varepsilon_n^2\right).$$

Then:

$$\begin{aligned}
P(n)\left(\max_{k=1}^{k_n}\max_{j=1}^{m_n}|V_{n,j,k}| \geq \varepsilon_n \mid A\right) &\leq \sum_{k=1}^{k_n}\left(1-\prod_{j=1}^{m_n}\left(1-P(n)\left(|V_{n,j,k}| \geq \varepsilon_n\right)\right)\right) \\
&\leq k_n\left(1-\prod_{j=1}^{m_n}\left(1-2exp\left(-R_n\varepsilon_n^2\right)\right)\right) = k_n\left(1-\left(1-2exp\left(\frac{-1}{16b^2}n\underline{h}_n^{d\frac{\gamma_0}{\gamma_0-1}}\varepsilon_n^2\right)\right)^{m_n}\right) \\
&= k_n\left(1-\left(\left(1-\frac{1}{exp\left(\log(1/2)+R_n\varepsilon_n^2\right)}\right)^{exp\left(\log(1/2)+R_n\varepsilon_n^2\right)}\right)^{\frac{m_n}{exp\left(\log(1/2)+R_n\varepsilon_n^2\right)}}\right) \\
&= k_n\left(1-(1-1/e-o(1))^{\frac{m_n}{exp\left(\log(1/2)+R_n\varepsilon_n^2\right)}}\right) = o(1).
\end{aligned}$$

Therefore $P\left(\max_{k=1}^{k_n}\max_{j=1}^{m_n}|V_{n,j,k}| \geq \varepsilon_n\right) = o(1)$. □

Lemma 25 (Nondegeneracy of local polylnomial eigenvalues at estimated bandwidths over gridpoints)**.** *Suppose the conditions of Theorem 2 hold. Fix some $k > 0$ and let $\mathcal{S}_n$ be a set of g_n points $x_{n,j} \in [-1,1]^d$, with $g_n \leq k'n$ for some fixed $k' > 0$. For each $x_{n,j}$, let $h_{n,j} = suph : n\sum_i D_i 1\{\|X_i - x_{n,j}\|\} \leq kh^{-2\beta_\mu}$ and let $h^*_{n,j}$ solve $nE[D\mathbf{1}\{\|X - x_{n,j}\|\}] = kh^{-2\beta_\mu}$. Then there is a $\delta_n \to^+ 0$ such that*

$$\limsup_{n\to\infty}\sup_{P\in\mathscr{P}} P\left(\max_j\left|\frac{\lambda_{\min}\left(\frac{\sum_i D_i K\left(\frac{\|X_i-x_j\|}{h_{n,j}}\right)U\left(\frac{X_i-x_j}{h_{n,j}}\right)U\left(\frac{X_i-x_j}{h_{n,j}}\right)^T}{\sum_i D_i K\left(\frac{\|X_i-x_j\|}{h_{n,j}}\right)}\right)}{\lambda_{\min}\left(\frac{E_P\left[DK\left(\frac{\|X-x_j\|}{h^*_{n,j}}\right)U\left(\frac{X-x_j}{h^*_{n,j}}\right)U\left(\frac{X-x_j}{h^*_{n,j}}\right)^T\right]}{E_P\left[DK\left(\frac{\|X-x_j\|}{h^*_{n,j}}\right)\right]}\right)} - 1\right| \geq \delta_n\right) = o(1).$$

Proof of Lemma 25. For convenience, I proceed assuming K is the uniform bandwidth, scaled downwards by one-half. Let $P(n)$ be a sequence of distributions in $\mathscr{P}$.

Apply Lemma 24 to the sequence $\underline{h}_n = n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}/\log(n)$ and $m_n = 3g_n \ll exp\left(an\underline{h}_n^{d\frac{\gamma_0}{\gamma_0-1}}\right)$ for all $a >$, to yield a sequence of $\varepsilon_n^{(a)} \to^+ 0$.

Let $h_{n,j}$ solve $\min_h\left|N_n(h \mid x_{n,j}) - kh^{-2\beta_\mu}\right|$, and let $h^*_{n,j}$ solve $\min_h\left|N^*_n(h \mid x_{n,j}) - kh^{-2\beta_\mu}\right|$, where $N^*_n(h \mid x_{n,j}) = nE\left[e(X)\mathbf{1}\{\|X - x_{n,j}\| \leq h\}\right]$. Further, let $[\underline{h}_{n,j}, \bar{h}_{n,j}]$ be the convex hull of the set of h that solve $N^*_n(h \mid x_{n,j})(1+\varepsilon_n^{(a)})^{-1} = kh^{-2\beta_\mu}$ or $N^*_n(h \mid x_{n,j})(1+\varepsilon_n^{(a)}) = kh^{-2\beta_\mu}$. By Lemma 24, with probability tending to one, $h_{n,j} \in [\underline{h}_{n,j}, \bar{h}_{n,j}]$.

Write:

$$A_{n,j}(h,h') = \frac{E_{P(n)}\left[DK\left(\frac{\|X-x_{n,j}\|}{h}\right)U\left(\frac{X-x_{n,j}}{h}\right)U\left(\frac{X-x_{n,j}}{h}\right)^T\right]}{E_{P(n)}\left[DK\left(\frac{\|X-x_{n,j}\|}{h'}\right)\right]}$$

$$B_{n,j}(h,h') = \frac{\sum_{i=1}^n D_i K\left(\frac{\|X_i-x_{n,j}\|}{h}\right)U\left(\frac{X_i-x_{n,j}}{h}\right)U\left(\frac{X_i-x_{n,j}}{h}\right)^T}{\sum_{i=1}^n D_i K\left(\frac{\|X_i-x_{n,j}\|}{h'}\right)}$$

The claim is that there is a $\delta_n \to^+ 0$ such that

$$P(n)\left(\left|\frac{\lambda_{\min}\left(B_{n,j}(h_{n,j},h_{n,j})\right)}{\lambda_{\min}\left(A_{n,j}(h^*_{n,j},h^*_{n,j})\right)} - 1\right| \geq \delta_n\right) = o(1).$$

$A_{n,j}^{-1}$ is symmetric, so that $\|A_{n,j}^{-1}\|_{(op)}^2$ is equal to the squared largest eigenvalue of $A_{n,j}^{-1}$, where $\|\cdot\|_{(op)}$ is equal to the operator norm. By Lemma 23, $\|A_{n,j}^{-1}\|_{(op)}^2 = \lambda_{\min}(A_{n,j})^{-2}$ is bounded above.

Take $m_n = 3g_n \leq 3k'n$, $k_n = (n+1)$, and $\underline{h}_n = n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} / \log(n)^{\frac{\gamma_0 - 1}{d\gamma_0}}$, so that the first condition of Lemma 24 holds by Lemma 23. The second condition holds because $m_n = O(n)$. The third condition holds by L'Hopital's Rule applied to $n\left(1 - \left(\frac{e-1}{e}\right)^{\frac{an}{exp(b+cn^d/\log(n)}}\right) \to 0$ for any fixed $a, b, c, d > 0$. Thus, I may apply Lemma 24 to the $k_n = n+1$ bounded functions $f(v) = U_k(v)U_p(v)$ and $1\left\{\frac{\underline{h}_{n,j}}{h^*_{n,j}} \leq v \leq \frac{\bar{h}_{n,j}}{h^*_{n,j}}\right\}$ at the bandwidths $\underline{h}_n$. Let the associated ε_n terms be $\varepsilon_n^{(b)}$ Then:

$$\begin{aligned}
&\max_j \left|A_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} - B_{n,j}(h_{n,j},h_{n,j})_{k,p}\right| \\
\leq &\max_j \left| \begin{array}{c} \\ - \end{array} \begin{array}{c} A_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} \\ B_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} \end{array} \right| + \max_j \left| \begin{array}{c} \\ - \end{array} \begin{array}{c} B_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p} \\ B_{n,j}(h_{n,j},h_{n,j})_{k,p} \end{array} \right| + \max_j \left| \begin{array}{c} \\ - \end{array} \begin{array}{c} B_{n,j}(h^*_{n,j},h_{n,j})_{k,p} \\ B_{n,j}(h_{n,j},h_{n,j})_{k,p} \end{array} \right| \\
\leq &o_{P(n)}(1) + O_{P(n)}\left(\max_j \left|A_{n,j}(h^*_{n,j},h^*_{n,j})_{k,p}\right|\right) \max_j \left|\left(1 - \frac{\sum D_i K\left(\frac{\|X_i-x_{n,j}\|}{h^*_{n,j}}\right)}{\sum D_i K\left(\frac{\|X_i-x_{n,j}\|}{h_{n,j}}\right)}\right)\right| + O\left(\max_j \frac{\sum D_i 1\left\{\|X_i - x_{n,j}\| \in [\underline{h}_{n,j}, \bar{h}_{n,j}]\right\}}{\sum D_i 1\left\{\|X_i - x_{n,j}\| \leq \underline{h}_{n,j}\right\}}\right) \\
= &o_{P(n)}(1) + O_{P(n)}\left(\varepsilon_n^{(a)}\right) + o\left(\left(1+\varepsilon_n^{(a)}\right)^2(1+\varepsilon_n^{(b)})\right) = o_{P(n)}(1).
\end{aligned}$$

Therefore $\max_j \|A_{n,j} - B_{n,j}\|_{(op)} = o(1)$.

Then, by well-known arguments (Horn and Johnson, 2013, p. 381):

$$\begin{aligned}
\max_j \left|\lambda_{\min}(B_{n,j}) - \lambda_{\min}(A_{,nj})\right| = \max_i \left|\lambda_{\max}(B_{n,j}^{-1}) - \lambda_{\max}(A_{,nj}^{-1})\right| &\leq \max_j \|B_{n,j}^{-1} - A_{n,j}^{-1}\|_{(op)} \\
&\leq \max_j \frac{\|A_{n,j}^{-1}\|_{(op)}^2 \|A_{n,j} - B_{n,j}\|_{(op)}}{1 - \|A_{n,j}^{-1}(A_{n,j} - B_{n,j})\|_{(op)}} = \max_j \frac{O_{P(n)}(1)o_{P(n)}(1)}{1 - o_{P(n)}(1)} = o_{P(n)}(1),
\end{aligned}$$

so that

$$\max_j \left| \frac{\lambda_{\min}(B_{n,j})}{\lambda_{\min}(A_{n,j})} - 1 \right| = \max_j \left| \frac{\lambda_{\min}(B_{n,j}) - \lambda_{\min}(A_{n,j})}{\lambda_{\min}(A_{n,j})} \right| \leq \max_j \frac{|\lambda_{\min}(B_{n,j}) - \lambda_{\min}(A_{n,j})|}{\min_j \lambda_{\min}(A_{n,j})} = o_{P(n)}(1),$$

so that the full claim holds. □

Lemma 26 (KL divergence)**.** *For any given $L' > 0$ and $x_0 \in [-1,1]^d$, construct distributions $P_{n,m}$ for $m = 1,2$ as follows. Draw $X \sim U([-1,1]^d)$. Draw $D \mid X \sim Bern\left(C^{-(\gamma_0-1)} P_{n,m}\left(\|X - x_0\| \leq \|x - x_0\|\right)^{1/(\gamma_0-1)}\right)$. Finally, draw $Y \mid X, D \sim \mathcal{N}(D\mu_{n,m}(X), \sigma^2_{\min})$ where $\mu_{n,1}(X) = 0$ and*

$$\mu_{n,2}(X) = \frac{L'}{exp\left(\frac{-4}{3}\right)} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} exp\left(\frac{-1}{1 - \left(\frac{2\|X - x_0\|}{n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}}\right)^2}\right) 1\left\{\|X - x_0\| \leq \frac{n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}}{2}\right\}.$$

Finally, define $\mathscr{P} = \{P_{n,1}\}_{n=1^\infty, m=1,2}$. Then (i) if there exists a $\mathscr{P}_0$ satisfying Assumptions 6 and 7, then there exists a fixed L' such that $\mathscr{P}$ satisfies Assumptions 6 and 7. (ii) there is an $\alpha > 0$ finite such that $KL\left(P_{n,1}, P_{n,2}\right) \leq \alpha$.

Proof of Lemma 26. First, I show (i) that such an L' exists. Because there exists a $P \in \mathscr{P}_0$ in this set and every $P_{n,m}$ has the smallest possible range of $Y - \mu(X) \mid X, D = 1$, it must be that for every $P_{n,m} \in \mathscr{P}$, Assumption 1(a) and Assumption 1(c) hold. Also, if I define $V(x) = P_{n,m}(\|X - x_0\| \leq \|x - x_0\|)$ which is distributed $Unif([0,1])$, then for all $P_{n,m}$:

$$P_{n,m}\left(e(X) \leq \pi\right) = P_{n,m}\left(C^{-(\gamma_0-1)} V(X)^{\frac{1}{\gamma_0-1}} \leq \pi\right) = P_{n,m}\left(V(X) \leq C\pi^{\gamma_0-1}\right) = C\pi^{\gamma_0-1}.$$

Therefore, Assumption 1(d) holds and Assumption 4(i) holds with $C = C'$ and $\bar{\pi}$ small enough. It is also clear that $E_{P_{n,m}}[Y \mid X, D = 0] = E_{P_{n,0}}[Y \mid X, D = 1] \in \Sigma(\beta_\mu, L)$, since these functions are a constant zero.

It remains to show that there is an $L' > 0$ such that for all n, m, $Var_{P_{n,2}}(\mu_{n,2}(X)) \leq M$ (Assumption 1(b) and completing Assumption 6) and $\mu_{n,2}(X) \in \Sigma(\beta_\mu, L)$ (Assumption 7). For the variance upper bound:

$$\text{Var}_{P_{n,2}}\left(\mu_{n,2}(X)\right) \leq (L')^2 \left(n^{\frac{-2\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}\right) exp\left(\frac{8}{3} - 2\right) \leq (L')^2 exp(2/3).$$

so that it suffices to take $L' \leq \sqrt{Mexp(-2/3)}$. For Hölder continuity, write $\mu_{n,2}(X) = \frac{L'}{a} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}} g_a\left(\frac{\|X - x_0\|}{n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}}}\right)$. If $a > 0$ is small enough, then g_a is infinitely differentiable and in $\Sigma(\beta_\mu, 1/2)$. Thus, by standard arguments (Tsybakov, 2009), if L' is small enough, $\mu_{n,2}(X) \in \Sigma(\beta_\mu, L)$. Thus, there is an $L' > 0$ such that $\mathscr{P}$ satisfies

Assumptions 6 and 7.

Second, I show the main claim (ii). It is useful to write $h_n = n^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$. Then:

$$\begin{aligned}
KL(P_{n,1}, P_{n,2}) &= nP_{n,1}\left(\log \frac{dP_{n,1}}{dP_{n,2}}\right) \\
&\leq nP_{n,1}\left(\|X - x_0\| \leq \frac{h_n}{4}\right) P_{n,1}\left(D = 1 \mid \|X - x_0\| \leq \frac{h_n}{4}\right) \frac{\left(\frac{L'}{exp\left(\frac{-4}{3}\right)} h_n^{\beta_\mu} exp\left(\frac{-1}{1-1/4}\right)\right)^2}{2} \\
&= nk2^{-2d-1}\left(L'\right)^2 \left(\frac{k}{C}\right)^{\frac{1}{\gamma_0 - 1}} h_n^{d + \frac{d}{\gamma_0 - 1} + 2\beta_\mu} = k2^{-2d-1}\left(L'\right)^2 \left(\frac{k}{C}\right)^{\frac{1}{\gamma_0 - 1}} = \text{“}\alpha\text{.”}
\end{aligned}$$

□

Proof of Proposition 5. First, I construct such a $\underline{\delta}$. Let δ_I be the smallest value of x such that $(x/\log(x))^{1/(2\beta_\mu)} \geq 4^{1/(2\beta_\mu)}$, $x/\log(x) \geq 4$, and $x \geq 4$. Then take:

$$\underline{\delta} = \sup d \geq \delta_I \text{ s.t. } (d/\log(d))^{\frac{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}{2\beta_\mu(\gamma_0 - 1)}} \geq d^{d/(2\log(d)) - 1} \text{ or } d = \delta_I$$

This is defined because as d tends to infinity, the inequality does not hold.

Now I show the claim holds. Take $\pi = (\log(\underline{\delta})/\underline{\delta})^{1/(2\beta_\mu)}$ and $\underline{r} = (\underline{\delta}/\log(\underline{\delta}))^{\frac{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}{2\beta_\mu(\gamma_0 - 1)}} = \pi^{-\frac{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}{\gamma_0 - 1}}$. Note that by construction, $\pi \in (0, 2^{-1/\beta_\mu}]$.

Let $\delta \geq \underline{\delta}$ and the sequence as above be given.

Take $t_n^{(0)} = (\log(\delta)/\delta)^{1/(2\beta_\mu)}$, so that $\log\left(m_n^{(k)}\right) = \delta 2^{-(k)}\left(t_n^{(k)}\right)^{2\beta_\mu}$ for all $k = 0, 1, \ldots$. Note as a result that $\log\left(m_n^{(k)}\right) = \log(\delta)\left(\prod_{0 \leq j < k} \frac{t_n^{(j+1)}/t_n^{(j)}}{2^{1/(2\beta_\mu)}}\right)^{2\beta_\mu}$. In particular, $\log\left(m_n^{(k+1)}/m_n^{(k)}\right) = \log\left(m_n^{(k)}\right)\left(\left(t_n^{(k+1)}/t_n^{(k)}\right)^{2\beta_\mu}/2\right)$ for all $k = 0, 1, \ldots$.

Next, I show by induction that for all $k = 0, 1, \ldots$, $t_n^{(k+1)} \geq t_n^{(k)}/\pi$ and $m_n^{(k+1)} \geq \underline{\rho} m_n^{(k)}$. In the base case, $k = 0$, $t_n^{(k+1)}/t_n^{(k)} = \pi^{-1}$ by construction of π, and $m_n^{(1)}/m_n^{(0)} = \delta^{\delta/(2\log(\delta)) - 1} \geq \underline{r}$ by construction of $\underline{r}$. In the inductive case,

$$\begin{aligned}
t_n^{(k+1)}/t_n^{(k)} &= \left(m_n^{(k)}/m_n^{(k-1)}\right)^{(\gamma_0 - 1)/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))} \geq (\underline{r})^{(\gamma_0 - 1)/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))} \\
m_n^{(k+1)}/m_n^{(k)} &= \left(m_n^{(k)}\right)^{\left(t_n^{(k+1)}/t_n^{(k)}\right)^{2\beta_\mu}/2 - 1} \geq \delta^{\delta/(2\log(\delta)) - 1} \geq \underline{r}.
\end{aligned}$$

Thus, for all $k = 0, 1, \ldots$, $t_n^{(k+1)} \geq t_n^{(k)} \geq \pi$ and $m_n^{(k+1)} \geq \underline{r} m_n^{(k)}$ for some $\underline{r} \geq 4$ and $\pi \in (0, 2^{-1/\beta_\mu}]$. The remaining claims hold by inspection. □

Lemma 27 (k_n^* characterization)**.** *Let $\underline{\delta}$, π, $m_n^{(k)}$, and $h_n^{(k)}$ be as in Proposition 5. Let k_n^* be the smallest k for which $\pi^{(k-1)\beta_\mu} \leq 1/\log(n)$. Then $k_n^* = O(\log(\log(n)))$.*

Proof of Lemma 27. $(k_n^* - 1) \leq \frac{\log(1/\pi)}{\beta_\mu} \log(\log(n)) = \Theta(\log(\log(n)))$. □

Lemma 28 (Minimal small-propensity points)**.** *Suppose the conditions of Theorem 2 hold, and there are m_n points $x_j \in [-1,1]^d$ such that $\inf_{j \neq j'} \|x_j - x_{j'}\| \geq h$ and $\max_j nE[D\mathbf{1}\{\|X - x_j\| \leq h/d\}] \leq s(h/d)^{-2\beta_\mu}$. Then there is a universal constant $B \geq 2$ that depends on $C, \gamma_0, \beta_\mu, d, s$ such that $m_n \leq B\left(nh^{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}\right)^{1-\gamma_0}$.*

Proof of Lemma 28. Note that the set of points with $\mathbf{1}\{\|X - x_j\| \leq h/d\}$ are mutually exclusive across x. Note also that for any set A, $E[D1\{X \in A\}] \leq \pi$ implies $P(X \in A)/2 \leq P(X \in A, e(X) \leq 2\pi) \leq P(e(X) \leq 2\pi)$. As a result, for j as in the statement,

$$E[e(X) \mid \|X - x_j\| \leq h/d] \leq \frac{nE[e(X)\mathbf{1}\{\|X - x_j\| \leq h/d\}]}{nP(\|X - x_j\| \leq h/d)} \leq \frac{sd^{d+2\beta_\mu}}{nh^{d+2\beta_\mu}}$$
$$m_n (h/d)^d / 2 \leq P\left(e(X) \leq \frac{2sd^{d+2\beta_\mu}}{nh^{d+2\beta_\mu}}\right) \leq C\left(\frac{2sd^{d+2\beta_\mu}}{nh^{d+2\beta_\mu}}\right)^{\gamma_0 - 1}$$
$$m_n \leq C 2^{\gamma_0} s^{\gamma_0 - 1} d^{2\beta_\mu(\gamma_0 - 1) + d\gamma_0} \left(nh^{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}\right)^{1-\gamma_0}.$$

Finally, if the constant is below 2, without loss of generality set $B = 2$. □

Lemma 29 (Grid width construction)**.** *Suppose the conditions of Theorem 2 hold, and define $N_n(h \mid x) = \sum_{i=1}^n 1\{D_i = 1, \|X_i - x\| \leq h\}$. Then there is a $c > 0$ and an algorithm based only on β_μ, d, and the (X, D) data that generates a grid width w_n such that with probability tending to one, $n^{-1/(2\beta_\mu + d)} \precsim w_n \leq dcn^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$ and for all $x \in [-1,1]^d$, $N_n\left(cn^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} \mid x_{n,j}\right) \geq 2\left(cn^{\frac{-1}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}\right)^{-2\beta_\mu}$.*

Proof of Lemma 29. First, I construct a $\tilde{h}_n$ that will be used to construct w_n. Let $z = 2$. By Lemma 20, there is a $k'' \geq 0$ such that $E[D1\{\|X - x\| \leq h] \geq k'' h^{d\gamma_0/(\gamma_0 - 1)}$ for all $h \in (0,1]$ and all $x \in [-1,1]^d$. Take $k' = (k''2)^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}$ and take $\tilde{h}_n = k' n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}$, so that $k'' n \tilde{h}_n^{d\gamma_0/(\gamma_0 - 1)} = 2^z \tilde{h}_n^{-2\beta_\mu}$.

To construct a grid without use of γ_0, define a pseudo-grid of width $\bar{w}_n = d/\lfloor (5n)^{1/(2\beta_\mu + d)} \rfloor$. Let n be large enough such that $\bar{w}_n \geq d(n/4)^{-1/(2\beta_\mu + d)}$ for all n. For each pseudo-gridpoint $\bar{x}_{n,j}$, take $\bar{h}_{n,j} = \inf_{h \leq 1} h : N_n(h \mid \bar{x}_{n,j}) \geq 2h^{-2\beta_\mu}$. Take the true grid width as $w_n = \lfloor 1/\left(d \sup \bar{h}_{n,j}\right) \rfloor$.

To show that w_n works, take another pseudogrid with pseudo-grid-width $\tilde{w}_n = 1/\lfloor 1/(2d\tilde{h}_n) \rfloor$ and consider hypothetical gridpoints $\tilde{x}_{n,j}$. For simplicity assume $1/(2d\tilde{h}_n)$ is an integer, and in particular that $\tilde{w}_n \leq \tilde{h}_n/d$. Notice also that $n^{-1/d} \ll n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))} = \Theta(\tilde{h}_n) = \Theta(w_n)$, so that $n^{-1/d} \ll \tilde{h}_n \leq k' d n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}$ for all n large enough.

Note also that $\|x - \tilde{x}_{n,j}\| \leq \tilde{h}_n$ and $\|x - \tilde{x}_{n,k}\| \leq \tilde{h}_n$ for $x \in [-1,1]^d$ implies $k = j$, because the gridpoints are separated by at least $d\tilde{h}_n$. There are also $O(\tilde{h}_n^{-d}) = o(n)$ gridpoints. Therefore by Lemma 24, with probability tending to one, $N_{n,j}(\tilde{h}_n \mid \tilde{x}_{n,j}) \geq nE[D1\{\|X - \tilde{x}_{n,j} \leq \tilde{h}_n\}]/2 \geq \frac{nk''}{2} \tilde{h}_n^{d\gamma_0/(\gamma_0 - 1)} = 2^{z-1} (\tilde{h}_n)^{-2\beta_\mu}$.

Note that for every $x \in [-1,1]^d$, there is some $\tilde{x}_{n,j}$ with $\|x - \tilde{x}_{n,j}\| \leq d\tilde{w}_n$. Therefore, with probability tending to one, $N_n((2d+1)\tilde{h}_n \mid x) \geq N_n(d\tilde{w}_n + \tilde{h}_n \mid x) \geq \inf_j N_n(\tilde{h}_n \mid \tilde{x}_{n,j}) \geq 2^{z-1}(\tilde{h}_n)^{-2\beta_\mu} = (2d+1)^{2\beta_\mu} 2^{z-1}((2d+1)\tilde{h}_n)^{-2\beta_\mu} \geq 2^z((2d+1)\tilde{h}_n)^{-2\beta_\mu}$.

Finally, take $c = (2d+1)k'$. With probability tending to one, $w_n \leq d(2d+1)\tilde{h}_n = dcn^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0-1))}$. Note that with probability tending to one, $w_n \leq (2+1/d)d\tilde{h}_n$. Note also that by Lemma 24 applied to $h_{n,0} = (n/4)^{-1/(2\beta_\mu+d)}$ at the pseudo-gridpoints $\bar{x}_{n,j}$, with probability tending to one, $N_n(h_{n,0} \mid \bar{x}_{n,j}) \geq n(h_{n,0})^{d\gamma_0}(k''^{\gamma_0} h_{n,0}) - d/(\gamma_0 - 1) \gg 4(h_{n,0})^{-2\beta_\mu}$, so that with probability tending to one, $dcn^{-1/(2\beta_\mu+d\gamma_0/(\gamma_0-1))} \geq w_n \geq h_{n,0} \gtrsim n^{-1/(2\beta_\mu+d)}$.

Finally, note that because each $x_{n,j}$ is in $[-1,1]^d$, by the argument above, $N_n\left(cn^{-1/(2\beta_\mu+d\gamma_0/(\gamma_0-1))} \mid x_{n,j}\right) = N_n\left((2d+1)\tilde{h}_n \mid x_{n,j}\right) \geq 2\left((2d+1)\tilde{h}_n\right) - 2\beta_\mu$. □

Lemma 30 (Bounding inclusion probability)**.** *Take c as in Lemma 29; let $\underline{\delta}, \pi$ be as in Proposition 5; take some $\delta \geq \underline{\delta}$; construct $m_n^{(k)}$ and $t_n^{(k)}$ as in Proposition 5; take $h_n^{(k)} = cn^{-1/(2\beta_\mu+d\gamma_0/(\gamma_0-1))}/t_n^{(k)}$; and take k_n^* as in Lemma 27. Let $\{(x_{n,j}, h_{n,j})\}$ be a set of gridpoints $x_{n,j}$ covering $[-1,1]^d$ with edge lengths w_n such that (i) $h_{n,j} = \inf_{h \leq 1} h : N_n(h \mid x_{n,j}) \geq 2h^{-2\beta_\mu}$, (ii) $h_{n,j} \leq h_n^{(1)} \leq w_n/d$, and (iii) $N_n\left(h_n^{(k)} \mid x_{n,j}\right) \geq \frac{n}{2} E\left[D1\left\{\|X - x_{n,j}\| \leq h_n^{(k)}\right\}\right]$ for all j,k and $h_{n,j} \leq h_n^{(1)} \leq w_n/d$ for all j. Then for every $k = 1, \ldots, k_n^* + 1$, there are at most $m_n^{(k)}$ gridpoints j with $h_{n,j} > h_n^{(k)}$.*

Proof of Lemma 30. Suppose j satisfies $h_{n,j} \geq h_n^{(k)}$. Then by (i) and (ii),

$$\frac{n}{2} E\left[D1\left\{\|X - x_{n,j}\| \leq h_n^{(k)}\right\}\right] \geq \frac{n\left(h_n^{(k)}\right)^d}{2} E\left[D \mid \|X - x_{n,j}\| \leq h_n^{(k)}\right],$$

so that $nE\left[D1\left\{\|X - x_{n,j}\| \leq h_n^{(k)}\right\}\right] \leq 4\left(h_n^{(k)}\right)^{-2\beta_\mu}$. Let $\tilde{m}_n^{(k)}$ be the number of gridpoints with $h_{n,j} \geq h_n^{(k)}$. Then by Lemma 28,

$$\tilde{m}_n^{(k)} \leq B\left(n(h_n^{(k)})^{2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}}\right)^{1-\gamma_0} = B\left(\frac{h_n^{(k)}}{h_n^{(1)}}\right)^{-(\gamma_0-1)\left(2\beta_\mu + d\frac{\gamma_0}{\gamma_0-1}\right)} \leq exp\left(2^{-k}\delta\left(\frac{h_n^{(k)}}{h_n^{(1)}}\right)^{-2\beta_\mu}\right) = m_n^{(k)}.$$

□

Lemma 31 (Grid characteristics)**.** *Take $c, \pi, \delta, m_n^{(k)}, h_n^{(k)}, m_n^{(k)}, k_n^*$ as in Lemma 30. Then there is a construction of a grid width w_n, gridpoints $x_{n,j}$, and bandwidths $h_{n,j}$ constructed using only β_μ, d, and the (X, D) data, such that with probability tending to one, (i) $N_n\left(h_n^{(1)}\right) \geq 2(h_n^{(1)})^{-2\beta_\mu}$ for all j, (ii) $h_{n,j} \leq h_n^{(1)}$ for all j, (iii) $\|x - x_{n,j}\| \leq h_{n,j}$ and $\|x - x_{n,k}\| \leq h_{n,k}$ implies $k = j$, (iv) $N_n\left(h_n^{(k)} \mid x_{n,j}\right) \geq \frac{n}{2} E\left[D1\{\|X - x_{n,j}\| \leq h_n^{(k)}\}\right]$ for all j and all $k = 1, \ldots, k_n^* + 1$, (v) the smallest eigenvalue of $\frac{\sum D1\{\|X - x_{n,j}\| \leq h_{n,j}\} U\left(\frac{X - x_{n,j}}{h_{n,j}}\right) U\left($*

is at least $\lambda^*/2$ *for every* j*, and (vi) for all* $k = 1, \ldots, k_n^* + 1$*, there are at most* $m_n^{(k)}$ *gridpoints* j *with* $h_{n,j} > h_n^{(k)}$.

Proof of Lemma 31. Let $w_n, x_{n,j}$ be constructed as in Lemma 29. For for every j, take $h_{n,j} = \inf_{h \leq 1} h : N_n(h \mid x_{n,j}) \geq 2h^{-2\beta_\mu}$. Notice that this construction is well-defined and does not use the outcome data and only depends on d, β_μ, and the (X, D) data.

By Lemma 29, $n^{-1/(2\beta_\mu + d)} \precsim w_n \leq dcn^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))} = dh_n^{(1)}$.

(i) holds by Lemma 29. But then (ii) holds by this construction of $h_{n,j}$.

(iii) holds because $h_{n,j} \leq h_n^{(1)} \leq w_n/d$ by Lemma 29.

(iv) will hold by Lemma 24. Consider the $k_n^* + 1$ functions $1\{D = 1, \|X - x\| \leq h_n^{(k)}\}$ for $k = 1, \ldots, k_n^* + 1$ evaluated at the $O_p(n)$ gridpoints $x_{n,j}$ with global bandwidth $h = h_n^{(1)} = \max_k h_n^{(k)}$. I continue on event that the number of gridpoints is $O(1)$ and the grid width is at least $h_n^{(1)}$, which were shown to be arbitrarily high probability earlier in this proof, so that the preconditions of Lemma 24 hold. Recall by Lemma 27 that $k_n^* = O(\log(\log(n)))$. Note that $h_n^{(1)} = \Theta^{n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}}$; $n \ll exp\left(-an(h_n^{(1)})^{d\gamma_0/(\gamma_0 - 1)}\right)$ by inspection; and $k_n^* \left(1 - \left(\frac{e-1}{e}\right)^{n/\left(exp\left(\log(1/2) + \tilde{c}^{-1} n^{-1} (h_n^{(1)})^{-d\gamma_0/(\gamma_0 - 1)}\right)\right)}\right) \to 0$ by L'Hopital's Rule. Therefore by Lemma 24, with probability tending to one, $N_n\left(h_n^{(k)} \mid x_{n,j}\right) \geq \frac{n}{2} E\left[D1\{\|X - x_{n,j}\| \leq h_n^{(k)}\}\right]$ for all j and all $k = 1, \ldots, k_n^* + 1$.

(v) by strict monotonicty of $h^{-2\beta}$ and weak monotonicty of $N_n(h \mid x)$ in the opposite direction, $h_{n,j} = \inf_{h \leq 1} h : N_n(h \mid x_{n,j}) \geq 2h^{-2\beta_\mu} = \sup_{h \leq 1} h : N_n(h \mid x_{n,j}) \leq 2h^{-2\beta_\mu}$. Thus on the high probability event that the number of gridpoints is $O(n)$, by Lemma 25 and Lemma 23, if n is large enough so that $h_n^{(1)}$ is small enough, then the smallest eigenvalue of the induced design matrices is at least $\lambda^*/2$.

(vi) holds by Lemma 30, with each condition either holding by construction or by claim above. □

Lemma 32 (Maximal local polynomial residual)**.** *Let* w_n *and* $\{x_{n,j}, h_{n,j}\}$ *be constructed as in Lemma 31, let* Z *be the* (X, D) *data, and define* $\bar{\mu}_j = E[\hat{\mu}_j \mid Z]$*. Then* $E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mid Z\right] = O_P\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$.

Proof of Lemma 32. Let $\bar{\mu}_{n,j} = E[\hat{\mu}(x_{n,j}) \mid Z]$. Recall by Lemma 31 that with probability tending to one, the smallest eigenvalue of the realized local polynomial design matrices $\sum DKUU^T / \sum DK$ is at least $\lambda^*/2 > 0$. Note that there are no x with $\|x - x_{n,j}\| \leq h_{n,j}$ and $\|x - x_{n,k}\| \leq h_{n,k}$ by (Lemma 31(iii)), so that $\hat{\mu}(x_{n,j}) \mid Z$ are independent normal draws. Therefore the set of $\hat{\mu}(x_{n,j}) - \bar{\mu}(x_{n,j})$ random variables are independent normal, with variance bounded above by $4\lambda^{-2} N_n\left(h_{n,j} \mid x_{n,j}\right)^{-1} \sigma_{\max}^2 = cN_n\left(h_{n,j} \mid x_{n,j}\right)^{-1}$ for some universal constant c.

Let $c, \delta, t_n^{(k)} h_n^{(k)}, m_n^{(k)}, k_n^*$ be defined as in Lemma 27, so that $h_n^{(k_n^*)} \leq c' h_n^{(1)} \pi^{(k-1)} \precsim n^{-1/(2\beta_\mu + d)} \left(1/\log(n)\right)^{1/\beta_\mu}$

by Proposition 5 and the definition of k_n^*. For any fixed $k = 1, \ldots, k_n^*$, by standard arguments,

$$\begin{aligned}
\text{“}M_k\text{”} &= E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 1\{h_n^{(k+1)} \leq h_{n,j} \leq h_n^{(k)}\} \mid Z\right] \\
&\leq c' \left(\min_{j: h_n^{(k+1)} \leq h_{n,j} \leq h_n^{(k)}} N_n\left(h_{n,j} \mid x_{n,j}\right)\right)^{-1} \log\left(\sum_j 1\{h h_n^{(k+1)} \leq h_{n,j} \leq h_n^{(k)}\}\right)
\end{aligned}$$

for some universal constant c'. Thus, with probability tending to one,

$$\begin{aligned}
M_k &\leq c' \left(\min_{j: h_n^{(k+1)} \leq h_{n,j} \leq h_n^{(k)}} h_{n,j}^{-2\beta_\mu}\right)^{-1} \log\left(m_n^{(k)}\right) \leq c' \left(h_n^{(k)}\right)^{2\beta_\mu} \log\left(m_n^{(k)}\right) && \text{(L30,31(vi))} \\
&\leq c'\delta \left(h_n^{(k)}\right)^{2\beta_\mu} 2^{-(k+1)} \left(t_n^{(k)}\right)^{2\beta_\mu} && \text{(Proposition 5)} \\
&= c'\delta \left(h_n^{(k)}\right)^{2\beta_\mu} 2^{-(k+1)} \left(h_n^{(k)} / h_n^{(1)}\right)^{-2\beta_\mu} = c'\delta \left(h_n^{(1)}\right)^{2\beta_\mu} 2^{-k}.
\end{aligned}$$

Also note that $h_{n,j} < h_n^{(k_n^*)}$ implies $N_n\left(h_{n,j} \mid x_{n,j}\right) = (h_{n,j})^{-2\beta_\mu} \geq \left(h_n^{(k_n^*)}\right)^{-2\beta_\mu}$, so that analogously, on the high probability event expressed above that there are $O(n)$ gridpoints and $\max_j h_{n,j} \leq h_n^{(1)}$),

$$\text{“}M_{k+1}\text{”} = E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 1\left\{h_{n,j} < h_n^{(k_n^*)}\right\} \mid Z\right] \leq c' \left(h_n^{(k_n^*)}\right)^{2\beta_\mu} \log\left(\frac{1}{n}\right) \precsim n^{\frac{-2\beta_\mu}{2\beta_\mu + \frac{d\gamma_0}{\gamma_0 - 1}}} \frac{\log(n)}{\log(n)^2} = o\left(n^{\frac{-2\beta_\mu}{2\beta_\mu + \frac{d\gamma_0}{\gamma_0 - 1}}}\right).$$

Thus, on an event with probability tending to one,

$$\begin{aligned}
E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mid Z\right] &\leq \sum_{k=1}^{k_n^*} M_k + M_{k+1} \leq \sum_{k=1}^{k_n^*} c'\delta \left(h_n^{(1)}\right)^{2\beta_\mu} 2^{-k} + o\left(n^{\frac{-2\beta_\mu}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}}\right) \\
&\leq c'\delta \left(h_n^{(1)}\right)^{2\beta_\mu} + o\left(n^{\frac{-2\beta_\mu}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}}\right) = O\left(n^{\frac{-2\beta_\mu}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}}\right).
\end{aligned}$$

□

Proof of Theorem 2. Recall that $\psi_n = n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$. There are two main directions to show.

Lower bound pointwise rate. Define $x_0 = (0, \ldots, 0)$. Let $\mathscr{P}$ be as in Lemma 26, with associated distributions $P_{n,m}$ for $m = 1, 2$. By Lemma 26, $K(P_{n,1}, P_{n,2}) \leq \alpha$. Define the seminorm $d(P, Q) = |E_P[Y \mid X = x_0, D = 1] - E_Q[Y \mid X = x_0, D = 1]|$. By construction,

$$d(P_{n,1}, Q_{n,2}) = \overbrace{L' exp\left(1/3\right)}^{\text{“}A/2\text{”}} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}},$$

where $L' > 0$ is fixed. Therefore $d(P_{n,1}, P_{n,d}) \geq 2s_n$, where $s_n = An^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$. Thus, standard arguments

(Tsybakov, 2009) show that for any fixed estimator $\hat{\mu}$ of $E[Y \mid X = x_0, D = 1]$ and all n large enough,

$$\sup_{P \in \mathscr{P}} P\left(|\hat{\mu}(x_0) - \mu(x_0)| \geq s_n\right) \geq \max_{j=1,2} P_j\left(|\hat{\mu}(x_0) - \mu_{n,j}(x_0)| \geq s_n\right) \geq \max\left(\frac{exp(-\alpha)}{4}, \frac{1-\sqrt{\alpha/2}}{2}\right) > 0.$$

Thus, for all $t > 1$ (see Tsybakov Theorem 2.3),

$$\liminf_{n\to\infty} \inf_{\hat{\mu}} \sup_{P \in \mathscr{P}} P\left(n^{\frac{\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}} |\hat{\mu}(x_0) - \mu(x_0)| \geq t^{\frac{\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}\right) \geq \max\left(\frac{exp(-ct)}{4}, \frac{1-\sqrt{ct/2}}{2}\right),$$

where c is a constant that only depends on the parameters of the problem. Thus, $n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$ is a lower bound on the *pointwise* rate of convergence.

Achievable uniform rate. This is the more difficult and interesting direction. Thankfully, several lemmas above make the remaining task relatively simple.

By Lemma 31 and Lemma 32, there is a construction of gridpoints $x_{n,j}$ separated by edges of length $w_n = O_P\left(n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ and bandwidths $h_{n,j}$ depending only on β_μ, d, and the (X, D) data Z such that with probability tending to one, $\max_j h_{n,j} = O\left(n^{-1/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$, the smallest eigenvalue of the design matrix used to construct $\hat{\mu}(x_{n,j})$ based on regression of Y on $U\left(\frac{X - x_{n,j}}{h_{n,j}}\right)$ with weights $D1\{\|X - x_{n,j}\| \leq h_{n,j}\}$ is at least $\lambda^*/2$ for some fixed $\lambda^* > 0$, and $E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \bar{\mu}_{n,j}\right)^2 \mid Z\right] = O\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ where $\bar{\mu}_{n,j} = E[\hat{\mu}(x_{n,j}) \mid Z]$. But then on this event, $\max_j \hat{\mu}(x_{n,j}) - \mu(x_{n,j})^2 = O\left((\lambda^*)^{-2} \max_j h_{n,j}^{2\beta_\mu}\right) = O\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$. Therefore for this construction, with probability tending to one, $E\left[\max_j \left(\hat{\mu}(x_{n,j}) - \mu(x_{n,j})\right)^2 \mid Z\right] = O\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$. Therefore $\max_j \left(\hat{\mu}(x_{n,j}) - \mu(x_{n,j})\right)^2 = O_P\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ by Markov's inequality.

Now consider predictions within the grid. For a given point $x \in [-1, 1]^d$, if $x = x_{n,j}$ for some j, take $\hat{\mu}(x) = \hat{\mu}(x_{n,j})$. Otherwise, construct $\hat{\mu}(x)$ via local polynomial regression of $\hat{\mu}(x_{n,j})$ on $U\left(\frac{x_{n,j} - x}{(\beta_\mu + 1)w_n}\right)$ for gridpoints $x_{n,j}$ with $\|x_{n,j} - x\| \leq (\beta_\mu + 1)w_n$. By standard arguments, each such design matrix is nondenerate so that $(\hat{\mu}(x) - \mu(x))^2 = O\left(w_n^{2\beta_\mu} + \max_j |\hat{\mu}(x_{n,j}) - \mu(x_{n,j})|\right) = O_P\left(n^{-2\beta_\mu/(2\beta_\mu + d\gamma_0/(\gamma_0 - 1))}\right)$ for all $x \in [-1, 1]^d$ simultaneously. Therefore $\hat{\mu}$ achieves the uniform rate ψ_n, and with no polylogarithmic penalty.

Completing the proof. An achievable uniform rate of convergence is also an achievable pointwise rate of convergence. Thus, $n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$ is the optimal pointwise rate of convergence. A lower bound on the pointwise rate of convergence is also a lower bound on the uniform rate of convergence. Thus, $n^{\frac{-\beta_\mu}{2\beta_\mu + d\frac{\gamma_0}{\gamma_0 - 1}}}$ is also the optimal uniform rate of convergence. □

Proof of Corollary 4. Write $\alpha_\mu = \frac{\beta_\mu}{2\beta_\mu + d\gamma_0/(\gamma_0 - 1)}$ and $\alpha_e = \frac{\beta_e}{2\beta_e + d}$.

By standard arguments (Stone, 1982) and Theorem 2, there are cross-fit estimators $\hat{\mu}(X)$ and $\hat{e}(X)$ such that $\|\hat{\mu} - \mu\|_{L^\infty(P)} \precsim_P r_{\mu,n} \precsim (n/\log(n))^{-\alpha_\mu}$ and $\|\hat{e} - e\|_{L^\infty(P)} \precsim_P r_{e,n} \precsim (n/\log(n))^{-\alpha_e}$, none of which depend on γ_0. As a result, Assumption 2 holds.

Take $b_n = r_{e,n}\log(n)$. Because $\alpha_e > 0$, $1 \gg b_n \gg r_{e,n}$. By Theorem 1, it only remains to show that the conditions of Assumption 3 hold.

(a) *Outcome Consistency.* $r_{\mu,n} \ll 1$.

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

(c) *Regression error with singularities.* If $\gamma_0 \geq 2$, the claim holds by inspection. If not, then:

$$\begin{aligned} r_{\mu,n} b_n^{\frac{\gamma_0}{2}} &\ll r_{\mu,n} r_{e,n}^{\gamma_0/2} \log(n)^{\gamma_0/2} \precsim (n/\log(n))^{-\alpha_\mu - \alpha_e\gamma_0/2} \log(n)^{\gamma_0/2} \\ &= \log(n)^{\alpha_\mu + \alpha_e\gamma_0/2 + \gamma_0/2} n^{-\alpha_\mu - \alpha_e\gamma_0/2} \\ &= n^{-1/2} \log(n)^{\alpha_\mu + \alpha_e\gamma_0/2 + \gamma_0/2} n^{1/2 - \alpha_\mu - \alpha_e\gamma_0/2} \ll n^{-1/2}. \qquad \text{(Equation (6))} \end{aligned}$$

(d) *Product of errors.* If $\gamma_0 \geq 2$, the claim holds by inspection. If not, then $r_{\mu,n} r_{e,n} b_n^{(\gamma_0-2)/2} \ll r_{\mu,n} b_n^{\gamma_0/2} \ll n^{-1/2}$.

As a result, by Theorem 1, the result for Wald confidence interval validity holds. □

C.7 Choice of Threshold

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

Recall the equation:

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

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 $b = 1$:

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

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 $\hat{g}_n(b) \equiv \frac{b\frac{1}{n}\sum \mathbf{1}\{\hat{e}(X)\leq b\}}{\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}} + b^2\sqrt{\frac{1}{n}\sum \frac{D}{\max\{\hat{e},b\}^2}}$ 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. □