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

Jacob Dorn

October 21, 2024

Problem Setup: Treatment Effects Under Weak Overlap

- **Goal**: estimate treatment effects
  - Focus on APO $\psi_0 = E[E[Y \mid X, D = 1]] = E[\mu(X, 1)]$ for ease
- **Setup**: Outcomes Y, controls X, treatment D, propensity $e(X) = E[D \mid X]$
- **Usual assumption**: strict overlap $\inf_x e(x) > 0$, or at least $E[1/e(X)]$ exists
- **This paper**: what if strict overlap fails?

Paper's Focus: (Augmented) Inverse Propensity Estimators

- Key estimators are Inverse Propensity Weighting and Augmented IPW

$$\hat{\psi}^{(IPW)} = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\hat{e}(X_i)}, \quad \hat{\psi}^{(AIPW)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}(X_i,1) + \frac{D_i(Y_i - \hat{\mu}(X_i,1))}{\hat{e}(X_i)}$$

  - $E\left[\frac{DY}{\hat{e}} \mid X\right] = \mu(X,1)e(X)/\hat{e}(X) \approx \mu(X,1)$
  - $E\left[\hat{\mu} + D\frac{Y-\hat{\mu}}{\hat{e}} \mid X\right] = \mu(X,1) + (\mu - \hat{\mu})\frac{e-\hat{e}}{\hat{e}} \approx \mu(X,1)$

- Under strict overlap, Wald $\hat{\psi} \pm 1.96\ \hat{SE}$ CIs cover with Prob $\to$ 95%

- Under weak overlap, both estimators nearly divide by zero

Under Weak Overlap, Usual Asymptotics Fail

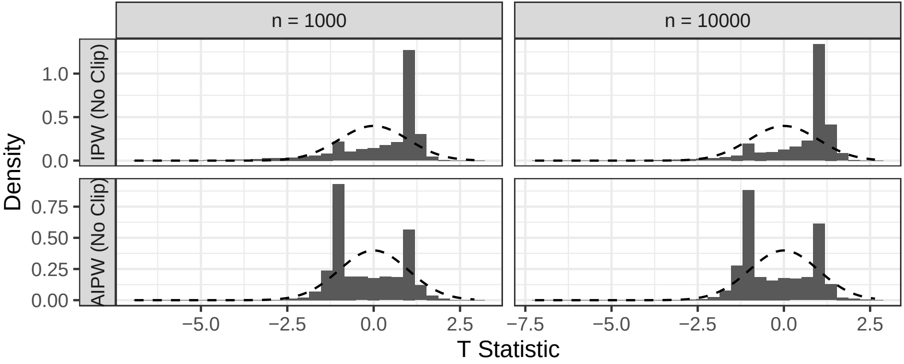

Figure: Weak overlap simulation (described later): IPW/AIPW t-statistics are far from $\mathcal{N}(0, 1)$ (dashed line). Under weak overlap, (A)IPW is consistent but not asymptotically Gaussian.

Existing Literature for Weak Overlap

- Theory: target standard ψ_0 with nonstandard estimators
  - New estimators like using $E[DY \mid e(X)]$ (Chaudhuri and Hill, 2016; Ma and Wang, 2020; Sasaki and Ura, 2022)
  - Confidence intervals like self-normalized subsampling (Ma and Wang, 2020; Heiler and Kazak, 2021)
- Practice: nonstandard estimands by weighting, trimming, or <u>clipping</u>

  Crump et al. (2009); Yang and Ding (2018); Li et al. (2018); Ma and Wang (2020), ...

$$\hat{\psi}^{(IPW)}(b_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{\max\{\hat{e}(X_i), b_n\}}, \quad \hat{\psi}^{(AIPW)}(b_n) = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}(X_i,1) + \frac{D_i(Y_i - \hat{\mu}(X_i,1))}{\max\{\hat{e}(X_i), b_n\}}$$

This Paper

Under clipped AIPW with the right rates, then standard $\hat{\psi} \pm 1.96\hat{SE}$ CIs cover the target ψ_0 even under weak overlap

Intuition for Asymptotic Normality

- Ma and Wang (2020): clipped IPW $\rightarrow^{d} \mathcal{N}(\cdot, \cdot)$, but bias even if e known
- Standard arguments: AIPW debiases $\hat{e}$ propensity error with $\hat{\mu}$
- Key insight: AIPW also debiases e clipping with $\hat{\mu}$
- But what debiases $\hat{\mu}$ with clipping?

Intuition for Asymptotic Normality

- Ma and Wang (2020): clipped IPW $\to^d \mathcal{N}(\cdot,\cdot)$, but bias even if e known
- Standard arguments: AIPW debiases $\hat{e}$ propensity error with $\hat{\mu}$
- Key insight: AIPW also debiases e clipping with $\hat{\mu}$
- But what debiases $\hat{\mu}$ with clipping? Clipping $\to 0$

Contributions

- Inference under weak overlap Khan and Tamer (2010); Ma and Wang (2020); Ma et al. (2023), ...
  - Uniform coverage of Wald $\hat{\mu} \pm 1.96\hat{SE}$ CIs using standard clipped AIPW
  - Clipped AIPW is NOT the optimal estimator — dominated by smarter things
- Two useful tricks for nonstandard estimands similar ideas in e.g. Semenova (2024)
  - Neyman orthogonal debiasing can apply to shifts away from ID failure
  - Points very near ID failure cannot be too common under margin conditions
- Regression with degenerate designs Hall et al. (1997); Gaïffas (2005); Pathak et al. (2023), some others
  - New results for local polynomial regression under weak overlap
  - Should this be a different paper(s)??

Today's Talk

1. Clipped AIPW asymptotics: main result + proof overview
2. Achieving regression rates under weak overlap
3. Simulations

Setting the Stage: Usual Strict Overlap Framework

- Usual strict overlap assumptions
  - Propensity scores bounded away from 0 and 1
  - Cross-fitting estimates of nuisances $\hat{\mu}, \hat{e}$ (will maintain)
  - Product of errors $\|\hat{e} - e\|_{P,2} = O_P(r_e)$, $\|\hat{\mu} - \mu\|_{P,2} = O_P(r_\mu)$, and $r_e * r_\mu = o(n^{-1/2})$
- Then t-statistics are well-calibrated: $\frac{\hat{\psi}^{AIPW} - \psi_0}{\hat{\sigma}} \rightarrow^d \mathcal{N}(0,1)$

Starting Point for Weak Overlap: Ma and Wang (2020)

- $P(e(X) \leq \pi) \sim \pi^{\gamma_0 - 1}$
  - $\gamma_0 > 2$: $E[1/e(X)]$ exists and standard asymptotics hold
  - $\gamma_0 < 2$: $E[1/e(X)] = \infty$ and IPW is not asymptotically normal **even if** $e(X)$ **is known**
  - $\gamma_0 \leq 1$: $E[DY/e(X)]$ may not be identified
- Heavy trimming (or clipping) $\Rightarrow$ asymptotic normality with bias (Ma and Wang, 2020)

I Consider a Uniform Family

Assumption 1

Observe data $(X, D, Y) \sim P \in \mathscr{P}$, *where* $\mathscr{P}$ *requires regularity conditions* *and* $P(e(X) \leq \pi) \leq C\pi^{\gamma_0 - 1}$ *for some fixed* $C > 0, \gamma_0 > 1$.

- Uniform: overlap can be stronger or nonsmooth under $\mathscr{P}$
- I also use $\mathscr{P}^{(cts)}$ for distributions that have "continuous" weak overlap

Weak Overlap Will Require Stronger Rates

- Sup-norms $\sup_x |\hat{\eta}(x) - \eta(x)| = o_P(r_\eta)$ to ensure rates near $e(x) = 0$
- Stronger rates needed on r_e, r_μ too
  - *Product of errors*: $r_\mu r_e(1 + b_n^{(\gamma_0-2)/2}) = o(n^{-1/2})$ ($b_n^{(\gamma_0-2)/2} \to \infty$ for $\gamma_0 < 2$)
  - *Debiased* $\hat{\mu}$: $r_\mu b_n^{(\gamma_0-2)*2/\gamma_0} = o(n^{-1/2})$, though laxer if $P \in \mathscr{P}^{(cts)}$
  - *Consistency* ($b_n \to 0$) and *asymp. known thresholding* ($r_e = o(b_n)$)

Main Result of the Paper: T-Statistics are Well-Calibrated

Theorem 1

Suppose $n^{-1/2} \ll b_n \ll 1$ and the rate conditions above hold. Then clipped AIPW t-statistics are well-calibrated under weak overlap:

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

Overview of Proof of T-Stat $\frac{\hat{\psi}(b_n)-\psi}{\hat{\sigma}} \to^d \mathcal{N}(0,1)$

1. Oracle CLT: $\frac{\hat{\psi}_{(orcl)}(b_n)-\psi}{\sigma_n} \to^d \mathcal{N}(0,1)$, where σ_n is oracle SE
2. Oracle equivalence: $\frac{\hat{\psi}-\hat{\psi}_{(orcl)}}{\sigma_n} \to_P 0$
3. Standard error consistency: $\frac{\hat{\sigma}_n}{\sigma_n} \to_P 1$ uniformly

Overview of Proof of T-Stat $\frac{\hat{\psi}(b_n)-\psi}{\hat{\sigma}} \to^d \mathcal{N}(0,1)$

1. Oracle CLT: $\frac{\hat{\psi}_{(orcl)}(b_n)-\psi}{\sigma_n} \to^d \mathcal{N}(0,1)$, where σ_n is oracle SE
   - Ma and Wang (2020): $\frac{\hat{\psi}^{(IPW)}_{(orcl)}(b_n)-\psi-\theta_n}{\sigma_n} \to^d \mathcal{N}(0,1)$
   - Uniform CLT with Berry–Esseen Theorem
   - $\frac{\hat{\psi}^{(AIPW)}_{(orcl)}(b_n)-\psi}{\sigma_n} \to^d \mathcal{N}(0,1)$ by A of AIPW
   - Convergence rate may be $\sigma_n^2 \sim n^{-1} b_n^{\gamma_0 - 3}$
2. Oracle equivalence: $\frac{\hat{\psi}-\hat{\psi}_{(orcl)}}{\sigma_n} \to_P 0$
3. Standard error consistency: $\frac{\hat{\sigma}_n}{\sigma_n} \to_P 1$ uniformly

Overview of Proof of T-Stat $\frac{\hat{\psi}(b_n)-\psi}{\hat{\sigma}} \to^d \mathcal{N}(0,1)$

1. Oracle CLT: $\frac{\hat{\psi}_{(orcl)}(b_n)-\psi}{\sigma_n} \to^d \mathcal{N}(0,1)$, where σ_n is oracle SE
2. Oracle equivalence: $\frac{\hat{\psi}-\hat{\psi}_{(orcl)}}{\sigma_n} \to_P 0$
   - Follows by the bias intuition from earlier + lots of algebra
   - Behavior driven by $e(X) \in [0, b_n(1+\epsilon)]$
3. Standard error consistency: $\frac{\hat{\sigma}_n}{\sigma_n} \to_P 1$ uniformly

Overview of Proof of T-Stat $\frac{\hat{\psi}(b_n)-\psi}{\hat{\sigma}} \to^d \mathcal{N}(0,1)$

1. Oracle CLT: $\frac{\hat{\psi}_{(orcl)}(b_n)-\psi}{\sigma_n} \to^d \mathcal{N}(0,1)$, where σ_n is oracle SE
2. Oracle equivalence: $\frac{\hat{\psi}-\hat{\psi}_{(orcl)}}{\sigma_n} \to_P 0$
3. Standard error consistency: $\frac{\hat{\sigma}_n}{\sigma_n} \to_P 1$ uniformly: easier!

Interpreting Rate Requirements ($P \in \mathscr{P}^{(Cts)}$)

- $\gamma_0 > 2$ (strict overlap): usual rates basically enough
- For fixed $\gamma_0 > 1$, $b_n \to 0$ exists if...
  - $r_\mu, r_e = o(n^{-1/3})$ OR
  - $r_e = O(n^{-1/2})$ and $r_\mu = o\left(n^{-1/4}\right)$ OR
  - $r_e = o(1)$ and $r_\mu = O(n^{-1/2})$
- A "curse of weak overlap" if $\hat{\mu}$ is nonparametric: $r_\mu * r_e = o(n^{-1/2})$ not enough
  - Intuition: If r_μ is parametric, we are done, but if r_e is parametric, still need to debias $\hat{\mu}$

But Can Those Rates Be Achieved?

- Need stronger rates for $\hat{e}(X)$ and $\hat{\mu}(X, 1)$ for $P(D = 1 \mid X) \approx 0$
- AND weak overlap makes $\hat{\mu}(X, 1)$ harder for $P(D = 1 \mid X) \approx 0$
- Can we estimate $E[Y \mid X, D = 1]$ when $P(D = 1 \mid X) \approx 0$?

But Can Those Rates Be Achieved?

- Need stronger rates for $\hat{e}(X)$ and $\hat{\mu}(X, 1)$ for $P(D = 1 \mid X) \approx 0$
- AND weak overlap makes $\hat{\mu}(X, 1)$ harder for $P(D = 1 \mid X) \approx 0$
- Can we estimate $E[Y \mid X, D = 1]$ when $P(D = 1 \mid X) \approx 0$? Second part of talk
  - Pointwise rates: optimal if $e(X)$ smooth, inconsistency if $e(X)$ degenerate
  - Global rates: may have a new property, but not sure that belongs in this paper

Usual Outcome Regression Rates

- Standard: strict overlap + $X \in \mathbb{R}^d$ compact + $\mu(x,1) \in$Hölder(β_μ)
  - Hölder: $\ell_\mu = \lfloor \beta_\mu \rfloor$-order derivatives are $(\beta_\mu - \ell_\mu)$-smooth
- Then optimal rates via local polynomial regression with bandwidth $h_n \to 0^+$
  - Regress Y on $U((X-x)/h)$, the $0, \ldots, \ell_\mu$-order interactions of $(X-x)/h_n$
  - Local: weight observations by $D * K\left(\frac{X-x}{h_n}\right) \sim D * \left\|\frac{X-x}{h_n}\right\|$
  - Best pointwise rate is $n^{-\beta_\mu/(2\beta_\mu+d)}$, global is $(n/\log(n))^{-\beta_\mu/(2\beta_\mu+d)}$
- Two key ingredients: neighbor probability + full-rank
  1. Neighbor observation probability $P\left(D=1, \|X-x\| \leq h\right) \sim h^d$
  2. Gram (?) matrix $E[UU' \mid D=1, \|X-x\| \leq h]$ is full rank

Weak Overlap Challenges for Local Polynomial Regression

1. Fewer neighboring observations when $e(X) \approx 0$
   - Weak overlap $\Rightarrow P(D = 1, \|X - x\| \leq h)$ can be smaller than h^d
   - Turns out, key parameter is Mou et al. (2023)'s $\alpha_{(Mou)} \equiv d/(\gamma_0 - 1)$
   - Now neighbor observation probability $P(D = 1, \|X - x\| \leq h) \gtrsim h^{d+\alpha_{(Mou)}}$
2. Potential degeneracy of $E[UU' \mid D = 1, \|X - x\| \leq h]$
   - Helps to assume $e(X) = E[D \mid X]$ is β_e-smooth ▸ not quite Hölder
   - Now Gram matrix behavior has a phase transition around $\beta_e = \alpha_{(Mou)}$

What I Have 1: Pointwise, Smooth Propensity Function

Proposition 1

Suppose $\mathscr{P}^{(rates)}$ is the set of distributions $P \in \mathscr{P}$ such that regularity conditions hold and either $\beta_\mu < 1$ (NW) or $\beta_e > \alpha_{(Mou)}$ (smooth propensities).

Then under local polynomial regression with optimal bandwidth,
$\sup_x \sup_{P \in \mathscr{P}^{(rates)}} E_P \left[\| \hat{\mu}(x,1) - \mu(x,1) \| \right] = O\left(n^{-\beta_\mu/(2\beta_\mu + \alpha_{(Mou)} + d)} \right)$.

Discussion: Local Rates in Good Case

$$\sup_{x} \sup_{P \in \mathscr{P}^{(rates)}} E_P \left[\|\hat{\mu}(x,1) - \mu(x,1)\| \right] = O\left(n^{-\beta_\mu / (2\beta_\mu + \alpha_{(Mou)} + d)} \right)$$

- Weak overlap parameter $\alpha_{(Mou)}$ plays the role of added covariate dimension
- By extending Gaïffas (2005), will be the optimal pointwise rate
- Intuition: Gram matrix is full rank under NW (automatic) or smooth propensities (local expansion), so $\alpha_{(Mou)}$ just harms neighbor probability

What I have 2: Pointwise, Degenerate Propensity Function

Proposition 2

Suppose $\mathscr{P}^{(rates)}$ is the set of distributions $P \in \mathscr{P}$ such that regularity conditions hold, $\beta_\mu > 1$ (no NW), $\alpha_{(Mou)} > \beta_e$ (degenerate), and $d \geq 2$ (multivariate).

Then if b_n is the optimal local polynomial consistency rate,

$$n^{-\bar{\beta}_\mu/(2\bar{\beta}_\mu+\alpha_{(Mou)}+d)} \lesssim \sup_{x,P} E_P\left[\|\hat{\mu}(x,1)-\mu(x,1)\|\right] \lesssim n^{-\underline{\beta}_\mu/(2\underline{\beta}_\mu+\alpha_{(Mou)}+d)},$$

where $\underline{\beta}_\mu \leq \bar{\beta}_\mu \leq \beta_\mu - 1$ are defined on the next slide. ▸ Construction

Discussion: Local Rates in Bad Case

$$n^{\overbrace{-\bar{\beta}_\mu/(2\bar{\beta}_\mu+\alpha_{(Mou)}+d)}^{\bar{\beta}_\mu=\beta_\mu-1-\frac{\alpha(Mou)^{-\beta_e}}{\beta_e+3}}} \lesssim \sup_{x,P} E_P\left[\|\hat{\mu}(x,1)-\mu(x,1)\|\right] \lesssim n^{\overbrace{-\underline{\beta}_\mu/(2\underline{\beta}_\mu+\alpha_{(Mou)}+d)}^{\underline{\beta}_\mu=\beta_\mu-1-\frac{\alpha(Mou)^{-\beta_e}}{3}}}$$

- Slower than β_μ for $\beta_e \to \alpha^-_{(Mou)}$, potential inconsistency for $\beta_e \ll \alpha_{(Mou)}$
- Can achieve better rates with other estimators (Gaïffas, 2005; Pathak et al., 2023)
- Univariate $d = 1$ is a black hole of mystery to me (Hall et al., 1997)

The Upshot: Sufficient Conditions for T-Statistics

Proposition 3

Suppose $E[Y \mid X, D = 1]$ *is* β_μ*-smooth and* $E[D \mid X]$ *is* β_e*-smooth. Define* $\underline{\beta}_\mu = \beta_\mu - 1\{\alpha_{(Mou)} > \beta_e, \beta_\mu > 1\} - \max\{(\alpha_{(Mou)} - \beta_e)/3, 0\}$. *Suppose*

$$\frac{\underline{\beta}_\mu}{2\underline{\beta}_\mu + d\frac{\gamma_0}{\gamma_0 - 1}} + \frac{\beta_e}{(2\beta_e + d)\frac{\gamma_0}{\gamma_0 - 1}} > 1/2.$$

There there is a set of feasible nuisance estimators and a $b_n \to 0$ *such that the clipped AIPW t-statistics cover with probability tending to 95%.*

What I'm Working On: Global Rates Under Weak Overlap

- Usual optimal global rate $(n/\log(n))^{-\beta_\mu/(2\beta_\mu+d)}$ has a polylog penalty
- May avoid polylog penalty under weak overlap + smooth propensities
  - Split X into singularities ($E[D \mid \|X - x\| \leq h] \sim h^{\alpha(Mou)+d}$) and non-singularities
  - Non-singularities: pointwise rate is better, so can pay a log cost
  - Singularities: cannot be too close while respecting weak overlap

What I'm Working On: Global Rates Under Weak Overlap

- Usual optimal global rate $(n/\log(n))^{-\beta_\mu/(2\beta_\mu+d)}$ has a polylog penalty
- May avoid polylog penalty under weak overlap + smooth propensities
  - Split X into singularities ($E[D \mid \|X - x\| \leq h] \sim h^{\alpha(M_{Ou})+d}$) and non-singularities
  - Non-singularities: pointwise rate is better, so can pay a log cost
  - Singularities: cannot be too close while respecting weak overlap
- This has become a nightmare to formalize
  - Singularities can be degenerate: good news for rates, bad news for Jacob
  - Is this a different paper? Log penalty won't show up in AIPW rate requirements
  - Is lack of polylog penalty even interesting? If it is to you, LET'S TALK

What I'm Working On: Global Rates Under Weak Overlap

- Usual optimal global rate $(n/\log(n))^{-\beta_\mu/(2\beta_\mu+d)}$ has a polylog penalty
- May avoid polylog penalty under weak overlap + smooth propensities
  - Split X into singularities ($E[D \mid \|X-x\| \leq h] \sim h^{\alpha(Mou)+d}$) and non-singularities
  - Non-singularities: pointwise rate is better, so can pay a log cost
  - Singularities: cannot be too close while respecting weak overlap
- This has become a nightmare to formalize
  - Singularities can be degenerate: good news for rates, bad news for Jacob
  - Is this a different paper? Log penalty won't show up in AIPW rate requirements
  - Is lack of polylog penalty even interesting? If it is to you, LET'S TALK

Next: simulations!

Simulated DGP is Inspired By Ma and Wang (2020)

- DGP: weak overlap with $\gamma_0 = 1.5$
  - $P(e(X) \leq \pi) = \pi^{1.5-1}$, $Y = (1 - e(X)) + (\varepsilon - 4)/\sqrt{8}$, $\varepsilon \sim \xi_4^2$ i.i.d.
- $\hat{e}(X)$ superparametric, $\hat{\mu}(X)$ nonparametric & biased
  - $\hat{e}(X) = \max\{e(X) - n^{-0.6}, n^{-4}\}$, $\hat{\mu}(X) = \mu(X)(1 + n^{-3/8})$
- Clip at rate b_n to solve $b_n^2 P_n(\hat{e}(X) \leq b_n) = 1/(2n)$
- Saw earlier: unclipped/untrimmed IPW & AIPW t-statistics fail badly

T-Statistics Are Nearly Standard Under Clipped AIPW

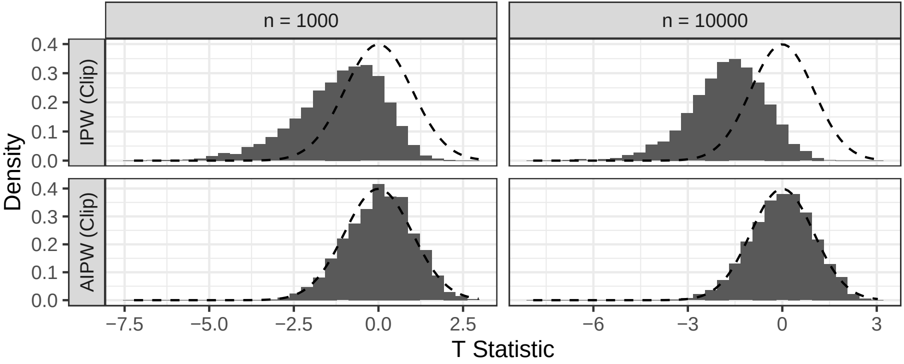

Figure: Distribution of simulated T-statistics for clipped IPW (left) and AIPW (right). Clipped AIPW T-statistics are close to $\mathcal{N}(0,1)$ (dashed line). ▸ Trimmed

P-Values Are Nearly Uniform Under Clipped AIPW

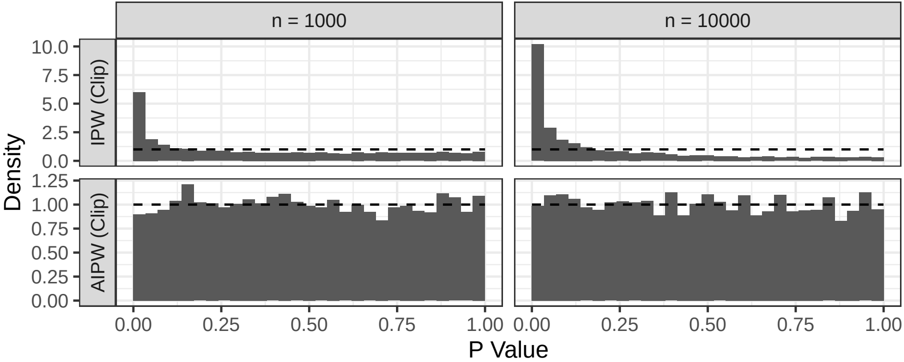

Figure: Distribution of simulated p-values on the null of the true APO for clipped IPW (left) and AIPW (right). Clipped AIPW p-values are close to uniform (dashed line). ▸ Trimmed

Conclusion

- Under even weak overlap, clipped AIPW 1.96 $\hat{SE}$ CIs can be well-calibrated
- Weak overlap makes regression rates harder, but not impossible
- Weak overlap global consistency rates may avoid usual polylog penalty
- Potential for future work to apply this approach to other ID failures?

Conclusion

- Under even weak overlap, clipped AIPW 1.96 $\hat{SE}$ CIs can be well-calibrated
- Weak overlap makes regression rates harder, but not impossible
- Weak overlap global consistency rates may avoid usual polylog penalty
- Potential for future work to apply this approach to other ID failures?

Let's chat! `jdorn@upenn.edu`

Bibliography I

Saraswata Chaudhuri and Jonathan B Hill. Heavy tail robust estimation and inference for average treatment effects, 2016.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1): 187–199, 01 2009. ISSN 0006-3444. doi: 10.1093/biomet/asn055. URL `https://doi.org/10.1093/biomet/asn055`.

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

Peter Hall, J. S. Marron, M. H. Neumann, and D. M. Titterington. Curve estimation when the design density is low. *The Annals of Statistics*, 25(2):756 – 770, 1997. doi: 10.1214/aos/1031833672. URL `https://doi.org/10.1214/aos/1031833672`.

Bibliography II

Phillip Heiler and Ekaterina Kazak. Valid inference for treatment effect parameters under irregular identification and many extreme propensity scores. *Journal of Econometrics*, 222(2):1083–1108, 2021. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2020.03.025. URL `https://www.sciencedirect.com/science/article/pii/S0304407620303377`.

Shakeeb Khan and Elie Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010. doi: https://doi.org/10.3982/ECTA7372. URL `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7372`.

Fan Li, Kari Lock Morgan, and Alan M Zaslavsky. Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113 (521):390–400, 2018. doi: 10.1080/01621459.2016.1260466.

Bibliography III

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

Xinwei Ma, Yuya Sasaki, and Yulong Wang. Testing limited overlap. *Econometric Theory*, 2023.

Wenlong Mou, Peng Ding, Martin J. Wainwright, and Peter L. Bartlett. Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency, 2023.

Reese Pathak, Martin J. Wainwright, and Lin Xiao. Noisy recovery from random linear observations: Sharp minimax rates under elliptical constraints, 2023. URL `https://arxiv.org/abs/2303.12613`.

Bibliography IV

Yuya Sasaki and Takuya Ura. Estimation and inference for moments of ratios with robustness against large trimming bias. *Econometric Theory*, 38(1):66–112, 2022. doi: 10.1017/S0266466621000025.

Vira Semenova. Aggregated intersection bounds and aggregated minimax values, 2024. URL `https://arxiv.org/abs/2303.00982`.

S Yang and P Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 03 2018. ISSN 0006-3444. doi: 10.1093/biomet/asy008. URL `https://doi.org/10.1093/biomet/asy008`.

Regularity Conditions for DGPs

Let $\mathscr{P} \equiv \mathscr{P}(M, q, \sigma_{\text{min}}, \pi_{\text{min}}, C, \gamma_0, \{r_{\mu,n}\}, \{r_{e,n}\})$ for $M > 3\sigma_{\text{min}}^4$ be the set of distributions P satisfying the following conditions:

1. *Conditional moments.* $\mathbb{E}[|Y - E[Y \mid X, D]|^q \mid X, D] \leq M^q < \infty$ almost surely for some $q > 3$.
2. *Residuals.* $\text{Var}(Y \mid X, D) \geq \sigma_{\text{min}}^2 > 0$ almost surely.
3. *Treated fraction.* $P(D = 1) \geq \pi_{\text{min}} > 0$.
4. *Propensity tail.* $P(e(X) \leq \pi) \leq C\pi^{\gamma_0 - 1}$ for all $\pi \in [0, 1]$ and some $\gamma_0 > 1$.

Continuous Weak Overlap

Definition 1

Let $\mathscr{P}^{(Cts)}(\rho)$ be the set of distributions $P \in \mathscr{P}$ such that for all $\pi \in [0, 1]$, $P(e(X) \leq \pi/2) \leq (1 - \rho)P(e(X) \leq \pi)$.

- "We cannot coincidentally have strict overlap with $\inf_x e(x) = b_n$"

Propensity Smoothness Definition

- Challenge: $e(X) = X^{3/2}$ for $X \sim Unif([0,1])$ ($\alpha_{(Mou)} = 3/2$)
  - Zero-order expansion around $x_0 = 0$: 0
  - First-order expansion around $x_0 = 0$: $0 + (X - 0) * 0$
  - Second-order expansion around $x_0 = 0$: $0 + 0 + \frac{(X-0)^2}{2} * \infty$
- But $e(X)^{4/3} = X^2$ is arbitrarily smooth

Propensity Smoothness Definition

Assumption 2

There is a fixed $M_{(prop)} \geq 1$ s.t. $e(X)^{M_{(prop)}} \in \Sigma(\beta_e M_{(prop)}, L^{\beta_e M_{(prop)}})$.

- Challenge: $e(X) = X^{3/2}$ for $X \sim Unif([0,1])$ $(\alpha_{(Mou)} = 3/2)$
  - Zero-order expansion around $x_0 = 0$: 0
  - First-order expansion around $x_0 = 0$: $0 + (X - 0) * 0$
  - Second-order expansion around $x_0 = 0$: $0 + 0 + \frac{(X-0)^2}{2} * \infty$
- But $e(X)^{4/3} = X^2$ is arbitrarily smooth: measure as $\beta_e * 4/3$
- Could generalize using homogeneous functions

Intuition: Pointwise Inconsistency

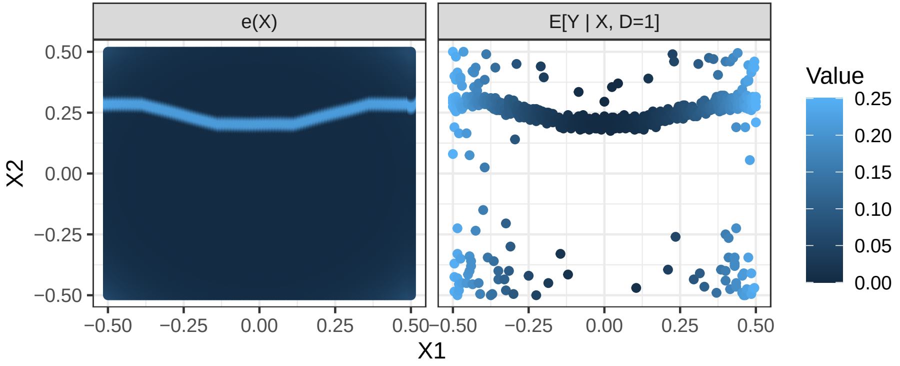

Figure: Bad DGP: $e(X)$ is larger near a curve ($\alpha_{(Mou)} - \beta_e$ in numerator) of sufficient area (1/3 of denominator) to drive $Var_{KD}(X_2)$ (2/3) and $Cov_{KD}(X_2, \mu)$ (bias), and may need disappearing shoulder width (β_e in denominator).

T-Statistics Are Nearly Standard Under Clipped AIPW

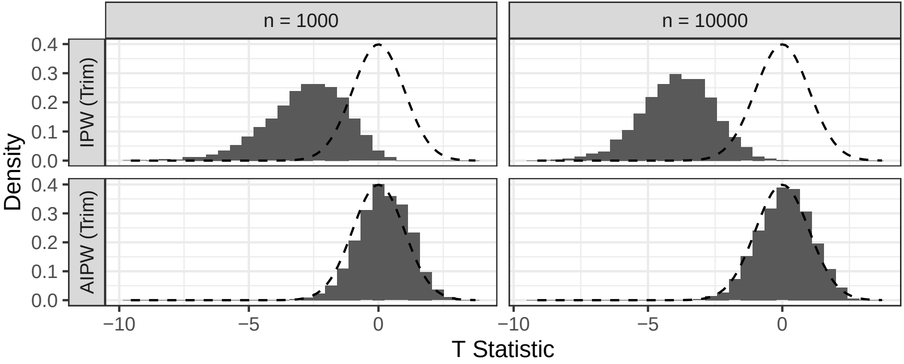

Figure: Distribution of simulated T-statistics for trimmed IPW (left) and AIPW (right). Trimmed AIPW T-statistics are close to $\mathcal{N}(0,1)$ (dashed line).

P-Values Are Nearly Uniform Under Trimmed AIPW

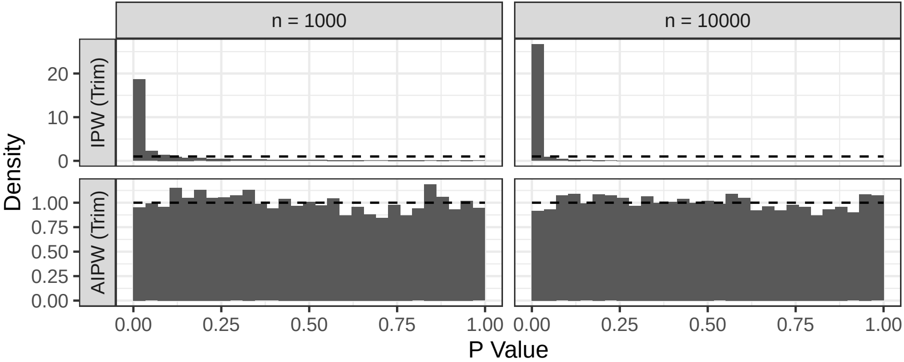

Figure: Distribution of simulated p-values on the null of the true APO for clipped IPW (left) and AIPW (right). Clipped AIPW p-values are close to uniform (dashed line).