Heavy Tail Robust Estimation and Inference for Average Treatment Effects

Saraswata Chaudhuri* and Jonathan B. Hill†
Dept. of Economics, McGill University
Dept. of Economics, University of North Carolina

March 15, 2016

Abstract

We study the probability tail properties of Inverse Probability Weighting (IPW) estimators of the Average Treatment Effect (ATE) when there is limited overlap between the covariate distributions of the treatment and control groups. Under unconfoundedness of treatment assignment conditional on covariates, such limited overlap is manifested in the propensity score for certain units being very close (but not equal) to 0 or 1. This renders IPW estimators possibly heavy tailed, and with a slower than $\sqrt{n}$ rate of convergence. Most existing estimators are either based on the assumption of strict overlap, i.e. the propensity score is bounded away from 0 and 1; or they truncate the propensity score; or trim observations based on a variety of techniques based on covariate or propensity score values. Trimming or truncation is ultimately based on the covariates, ignoring important information about the inverse probability weighted random variable Z that identifies ATE by $E[Z] =$ ATE. We propose a tail-trimmed IPW estimator whose performance is robust to limited overlap. In terms of the propensity score, which is generally unknown, we plug-in its parametric estimator in the infeasible Z, and then negligibly trim the resulting feasible Z adaptively by its large values. Trimming leads to bias if Z has an asymmetric distribution and an infinite variance, hence we estimate and remove the bias using important improvements on existing theory and methods. Our estimator sidesteps dimensionality, bias and poor correspondence properties associated with trimming by the covariates or propensity score. Monte Carlo experiments demonstrate that trimming by the covariates or the propensity score requires the removal of a substantial portion of the sample to render a low bias and close to normal estimator, while our estimator has low bias and mean-squared error, and is close to normal, based on the removal of very few sample extremes.

JEL Classification: C12; C13; C30.

AMS Classification: 62F12; 62F35.

Keywords: average treatment effect; limited overlap; tail trimming; robust estimation

*Dept. of Economics, McGill University, Montreal, Quebec, saraswata.chaudhuri@mcgill.ca.

†Corresponding author. Dept. of Economics, University of North Carolina at Chapel Hill, www.unc.edu/∼jbhill, jbhill@email.unc.edu.

We thank participants at the Cowles Foundation 2014 Conference on Econometrics at Yale University. In particular, we thank Xiaohong Chen and Shakeeb Khan for helpful comments. We also thank two referees and an associate editor for helpful feedback. The authors, of course, are solely responsible for all errors.

1 Introduction

We propose a tail-trimmed Inverse Probability Weighting (IPW) estimator of the average treatment effect (ATE) in observational studies. The estimator is robust to heavy tails that arise due to *limited overlap* in the distribution of the observed covariates X for the treatment and the control groups.

The strong ignorability assumption of Rosenbaum and Rubin (1983) can (nonparametrically) point identify the ATE. It requires the existence of a set of observed covariates X satisfying *unconfoundedness* of the treatment assignment conditional on the observed covariates, and *strict overlap*. We maintain the assumption of perfect compliance, that is the treatment is taken *if and only if* it is assigned.

We focus on *strict overlap* which requires the propensity score, the probability of taking the treatment conditional on the observed covariates X, to be bounded away from zero and one. We slacken the strict overlap assumption by allowing for *limited overlap*: the propensity score can be arbitrarily close to zero or one (Khan and Tamer, 2010).[1] Limited overlap accommodates conventional models where the treatment assignment depends on a latent variable crossing some threshold (e.g. Busso, DiNardo, and McCrary (2009)). While limited overlap still allows for point identification, this may result in *irregular* identification (Khan and Tamer, 2010). Consequently the tails of IPW estimators of the ATE may get thicker causing instability in estimation and inference, and a breakdown of the standard asymptotic properties such as $\sqrt{n}$-convergence and asymptotic normality. Identification is irregular precisely because Z may not belong to the domain of attraction of a normal law. Hence conventional estimators can have non-Gaussian limits when properly scaled (cf. Ibragimov and Linnik, 1971) and robust estimators can have a slower than $\sqrt{n}$ convergence rate (Khan and Tamer, 2010).[2] This is discussed in the supplemental material Chaudhuri and Hill (2014, Part I); see also Khan and Nekipelov (2013).

Our main contribution is a tail-trimmed parametric IPW estimator of the ATE. Our estimator is *robust* in the sense that it is consistent, asymptotically unbiased and normally distributed *even* under limited overlap, irrespective of heavy tails, and irrespective of the (finite) number of covariates in X. Our estimator is parametric because it plugs in a parametric estimator for the generally unknown propensity score in the infeasible Z that point identifies ATE. We trim the resulting feasible Z adaptively by a vanishing sample portion of large values, which results in asymptotic bias in the limit distribution when Z has an infinite variance and an asymmetric distribution. Using important improvements to bias correction theory developed in Peng (2001) and Hill (2015), we estimate an approximation of the bias based on a power law assumption on Z. Our resulting estimator is asymptotically unbiased in its limit distribution even if Z has distribution tails that decay faster than any power law (cf. Hill, 2015). Although our presentation can be easily extended beyond ATE estimation to general parametric IPW M-estimation as in Wooldridge (2007), we focus on ATE estimation for brevity.

As a second contribution, in Chaudhuri and Hill (2014, Part I) we provide a detailed characterization of the effect of the relative tail behavior of the covariates X and the unobserved errors on subsequent estimation and inference based on IPW estimators. In the conventional threshold crossing models for treatment assignment, we characterize when Z has a power law distribution tail, and possibly an infinite variance. Although an infinite variance does not guarantee a standard ATE estimator will have a non-Gaussian limit,[3] this nevertheless suggests

[1] Crump, Hotz, Imbens, and Mitnik (2009) use limited overlap in a broader empirical sense, in particular "parts of the covariate space with limited numbers of observations for either the treatment or control group". See p. 188.

[2] Location estimators' sensitivity to heavy tailed data in general is well known. See Bahadur (1960) and Jureckova (1981).

[3] See Chapter 9 in Feller (1971), and recently Chritsopeit and Werner (2001).

the need for an estimator that is robust to the possibility of heavy tails, and therefore ensures standard inference.

Three features of our estimator are worth noting. First, if overlap is strict or limited overlap is not significant enough to render heavy tails, our estimator is asymptotically equivalent to the untrimmed parametric IPW estimator. Second, if limited overlap results in an infinite variance, trimming based on either feasible or infeasible Z yield the same asymptotic results: the power law properties of the infeasible Z and the trimming mechanism are all that matter for explaining why our estimator works. This is, however, an asymptotic result. In general, we still achieve the well known property that estimation based on the feasible Z promotes an estimator variance that is no larger than if the infeasible Z were used (see Wooldridge, 2007). The inequality holds even asymptotically if Z has a finite variance. Third, we use Karamata Theory for power law tails to motivate a model for, and to estimate, bias. The power law decay rate, however, neither needs to be known *nor even true* (e.g. tails may decay exponentially fast) for our bias corrected tail-trimmed estimator to be valid for standard inference (cf. Hill, 2015).[4]

Although our estimator is based closely on bias correction theory developed in Peng (2001) and expanded in Hill (2015), we make several key contributions that apply in general to robust mean estimation. First, by re-centering for the trimming criterion and re-scaling by the number of non-trimmed observations we ensure both an unbiased estimator when Z has a symmetric distribution, and otherwise diminished bias making our bias estimator more accurate. Second, we use a slight variation on the bias formula in Hill (2015) which promotes a bias correction that does not affect the limit distribution of our ATE estimator, and greatly simplifies inference. Third, we use the bias correction only when it helps.

Khan and Nekipelov (2013) provide an array of results showing the failure of pivotal and bootstrap inference for conventional IPW estimators with a plug-in. Our robust ATE estimator with bias correction and corresponding estimator of the asymptotic scale results in pivotal inference by construction, whether tails decay according to a power law or not. This occurs precisely because we remove a vanishing fractile of tail observations of Z that erode regular identification under substantial limited overlap.

Self-standardized untrimmed IPW estimators, however, are not pivotal (Busso, DiNardo, and McCrary, 2009; Khan and Tamer, 2010; Khan and Nekipelov, 2013). We present a unique set of results that verify this in Chaudhuri and Hill (2014, Part I). Using a latent variable treatment selection framework we show Z has power law tails, with monotonically heavier tails as the degree of limited overlap increases. Thus, regular and irregular identification hinge on the exact degree of limited overlap in that framework.

It is important to recognize that our goal is fundamentally different from that of the conventional use of trimming in the ATE literature. The focus there is either to put bounds on the ATE (e.g. Lechner (2008)) or to locate a suitable region of common support to point identify the ATE for a subpopulation (that may or may not be the population of interest) defined by the common support and achieve internal validity of the ATE estimator. See Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Crump, Hotz, Imbens, and Mitnik (2009), Lee, Lessler, and Stuart (2011), and Traskin and Small (2011). In contrast, the ATE is already point identified under limited overlap. Our tail-trimmed IPW estimator overcomes the problems of the existing IPW estimators that are associated with irregular identification.

It is because of our different goal that our trimming strategy is based on Z itself, rather than the otherwise sensible conventional strategies of trimming or truncating either directly on the conditioning covariates (involved

[4]Valid inference could possibly be made without trimming by using a bootstrap or subsampling method, although sharp caveats exist in the heavy tailed case. See Hall (1990) and Khan and Nekipelov (2013).

in the ignorability assumption) or the propensity score. See Section 2.2 for a broad review. Since our problem concerns dealing with a possible infinite variance of feasible or infeasible Z, trimming based on feasible Z is our natural strategy. By doing so, we use all the available information about the causes of extremes in feasible Z, and sidestep the issues related to the dimensionality of the covariates, and the poor correspondence between the covariates or propensity score and Z. By trimming *negligibly* we ensure asymptotic normality in general, without a model of treatment assignment or assumptions on the covariates.

The rest of the paper is organized as follows. In Section 2 we motivate our estimator by describing the framework, discussing the problem of ATE estimation under limited overlap, and detailing existing methods to deal with it. We then introduce our tail-trimmed estimator in Section 3 and present its asymptotic properties under a general set of high level assumptions. Finally, we perform Monte Carlo experiments in Section 4 and in Chaudhuri and Hill (2014, Part II) in order to compare our robust and asymptotically unbiased estimator with existing estimators. Our estimator performs best overall within a simulation design that allows for multiple covariates and possibly asymmetrically distributed Z (and therefore bias due to trimming): it exhibits small bias and mean-squared-error, and is close to normal, based on a remarkably small amount of trimming. If limited overlap is severe then other estimators considered either exhibit bias and are far from normal, both leading to poor inference, or require a substantial amount of trimming and therefore waste observations in order to be competitive.

Throughout $a_n \sim b_n$ implies $a_n/b_n \to 1$ as $n \to \infty$. K is a positive finite constant, the value of which may change from line to line. $\iota > 0$ is a tiny number that may be different in different places. $[z]$ is the integer part of z. $I(A)$ denotes an indicator variable for the event A.

2 Framework and Literature Review

2.1 IPW Estimators under Limited Overlap

Let D be a binary variable such that $D = 1$ if the treatment is taken and $D = 0$ otherwise. Let $Y_1 \equiv Y(D = 1)$ and $Y_0 \equiv Y(D = 0)$ denote the potential outcomes. See Rubin (1974). Our object of interest is the population ATE:

$$\theta \equiv E[Y_1 - Y_0]. \tag{1}$$

Y_1 and Y_0 cannot be simultaneously observed for the same unit: we only observe the realized outcome

$$Y = DY_1 + (1 - D)Y_0.$$

This causes a problem in observational studies with not (completely-at-) random treatment assignment, because the difference in the expected realized outcome for the treatment and the control groups $E[Y|D = 1] - E[Y|D = 0]$ cannot identify the ATE θ in general.

Identification of θ can, however, be achieved by the following *strong ignorability* (unconfoundedness and strict overlap) assumption (Rosenbaum and Rubin, 1983), cf. Crump, Hotz, Imbens, and Mitnik (2009). Assume there exists a set of observed covariates X (throughout $\perp$ expresses independence).

Assumption A1 (Unconfoundedness): $Y_1, Y_0 \perp D|X$.

Assumption A2 (Strict Overlap): $0 < p_* \leq p(X) \equiv P(D = 1|X) \leq 1 - p_* < 1$ $a.s.$ for a constant p_*.

Assumptions A1 and A2 immediately imply identification:

$$E\left[\frac{D}{p(X)}Y - \frac{1-D}{1-p(X)}Y\right] = E\left[E\left[\frac{D}{p(X)}Y_1 - \frac{1-D}{1-p(X)}Y_0\middle| X\right]\right] = E\left(E[Y_1|X] \; - E[Y_0|X]\right) = \theta.$$

An IPW estimator of the ATE is a sample analog of the left-hand-side, with a plug-in for unknown $p(X)$ (see, e.g., Hirano, Imbens, and Ridder, 2003, and their references).

Notice $p(X) = 0$ or $p(X) = 1$ with positive probability imply an absence of strict overlap, or even limited overlap defined below. This violates A2 and is a well recognized problem with ATE identification and estimation. We abstract from such severe, albeit realistic, non-overlap possibilities, and instead focus on the case of *limited overlap* that may indeed be difficult to rule out even after careful balancing of the covariates X by the analyst. The terminology is borrowed from Khan and Tamer (2010).

Assumption A2′ (Limited Overlap): $0 < p(X) \equiv P(D = 1|X) < 1 \; a.s.$

Although trivially A2′ nests strict overlap A2, the problem is far more subtle under A2′. The ATE θ is point identified but, as Khan and Tamer (2010) showed, under A1 and A2′ the efficiency bound is infinity. In practice, this can lead to instability due to a slower than standard rate of convergence for IPW estimators, and a large or unbounded variance. A similar problem arises in IPW estimators of $E[Y_0]$.[5]

2.2 Existing IPW Methods to Handle Limited Overlap

If the proportion of units with *small or large* $p(X)$ is not sufficiently low to prevent instability, but low enough to guarantee the identification of θ, one could possibly remove some or all of these units and thus trim the tails of the distribution of the IPW estimator to restore the standard asymptotic properties. We discuss four strands of the literature.

Weight Capping Capping the weights involves truncating extreme observations of $p(X)$ by percentile cutpoints like 1 and 99 or by fixed cutpoints p_* and $1 - p_*$, thereby mimicking strict overlap A2. See, Lee, Lessler, and Stuart (2011) and Chaudhuri and Min (2012) respectively. The method is ad hoc and can increase bias substantially, although Lee, Lessler, and Stuart (2011) give simulation evidence supporting percentile cutpoints, while Frolich (2004) finds capping works better than removing the concerned units altogether as is done by the conventional trimming rules with the IPW estimators. Potter (1993) explores different cutpoint selection methods based on minimizing a suitably chosen mean squared error function. The asymptotic properties of such estimators are apparently not completely known.

Unit Removal A more conventional strand involves the removal of units from the treated and the control groups for which there is no comparable units in the opposite group. See, for example, Heckman, Ichimura, and Todd (1998), Dehejia and Wahba (1999), Crump, Hotz, Imbens, and Mitnik (2009), and Traskin and Small (2011). These trimming rules were designed in the context of matching estimators to obtain internal validity of the estimates, while Crump, Hotz, Imbens, and Mitnik (2009), where the focus is primarily on identifying the subpopulation (in terms of the covariates) for which ATE can be estimated with maximum precision, applies generally. However, the resulting estimator may not identify the ATE for the original population, unless the treatment effect

[5]The limited overlap problem is due to the tail behavior of the true propensity score $p(X)$. This is fundamentally different from the problem associated with parametric mis-specification of the propensity score model, cf. Kang and Schafer (2007).

is homogeneous.

Tail Trimming A third strand exploits a classic tail-trimmed estimator. Studies that are closest in spirit to the present study are due to Khan and Tamer (2010) and Yang (2015). (Also see Crump, Hotz, Imbens, and Mitnik (2009) who, as noted above, have a slightly different focus and also work with a different definition of limited overlap.) Khan and Tamer (2010) assume $D = I(\alpha + \beta X - U \geq 0)$ where X is a scalar covariate/index, and U is a random error independent of X. They show asymptotic normality is assured by removing units Z with $|X| > \nu_n$, where $\{\nu_n\}$ is a sequence of non-random numbers, and $\nu_n \to \infty$ as the sample size $n \to \infty$. The proposed estimator based on the observed sample $\{Y_i, D_i, X_i\}_{i=1}^n$ trims by X_i (tx):

$$\theta_n^{(tx)} \equiv \frac{1}{n}\sum_{i=1}^n h(X_i)Y_i I\left(|X_i| \leq \nu_n\right) \text{ where } h(X_i) \equiv \frac{D_i}{p(X_i)} - \frac{1-D_i}{1-p(X_i)} = \frac{D_i - p(X_i)}{p(X_i)(1-p(X_i))}. \tag{2}$$

Several features of their method are worth noting:

(1). The propensity score is assumed known for ease of presentation.

(2). The rate of convergence of $\theta_n^{(tx)}$ is studied under the normalization $\beta = 1$ when X_i and U_i are iid logistic. The convergence rate, when ν_n minimizes the mean-squared-error, is $(n/\ln(n))^{1/2}$, aligning identically with a sample mean of an iid random variable with power law distribution tails with index exactly 2, hence the variance of $h(X_i)Y_i$ is infinite. See, e.g., the textbook treatments of Leadbetter, Lindgren, and Rootzen (1983) and Resnick (1987). In their second example where X_i is logistic and U_i is normal, the convergence rate is even slower, aligning with a tail index less than 2, hence heavier tails in X_i imply heavier tails for $h(X_i)Y_i$. That the rate of convergence appears to suggests rates of tail decay are neither shown nor discussed in the literature to the best of our knowledge.

(3). By fixing $Var(U_i) = 1$ and letting $\beta > 0$ vary, we demonstrate Chaudhuri and Hill (2014, Part I) that the tail decay rate for $h(X_i)Y_i$ is monotonic in β, with heavier tails and infinite variance occurring with $\beta \geq 1$. The converse is true if, equivalently, we fix $\beta = 1$, as in Lewbel (1997) and Khan and Tamer (2010), and let $Var(U_i)$ or $Var(X_i)$ vary: heavier tails align with larger $Var(X_i)/Var(U_i)$. This points to a natural signal-noise property: heavier tails align with a stronger signal (i.e. large β or large $Var(X_i)$) and smaller noise (i.e. small $Var(U_i)$), which can have a dramatic impact on IPW estimators of the ATE. As far as we know, a complete characterization of the rate of convergence or asymptotic distribution for $\theta_n^{(tx)}$ in this more general setting, where either β or $Var(U_i)$ is arbitrary, is not available.

(4). It is not clear how a covariate trimming rule should be modified when *multiple* covariates are required to ensure that Assumption A1 holds. Possible solutions could be trimming based on $p(X_i)$, as in Crump, Hotz, Imbens, and Mitnik (2009) when $V(Y_1|X) = V(Y_0|X)$ is a constant X-$a.s.$, or based on the weight $h(X_i)$. Both are related to the literature on weight capping discussed above. However, $h(X_i)Y_i$, and not $h(X_i)$, identifies θ. Hence, if $E[h^2(X_i)Y_i^2] = \infty$ then *in general* only trimming sufficiently many of the largest realizations of $|h(X_i)Y_i|$ will *guarantee* asymptotic normality irrespective of the relationship between covariate X, propensity score $p(X)$ and realized outcome Y, cf. Csörgo, Horváth, and Mason (1986); Hahn, Weiner, and Mason (1991); Hill (2015).

(5). Estimators like $\theta_n^{(tx)}$ may be asymptotically biased. Indeed and somewhat trivially, unless $\theta = 0$ and $h(X_i)Y_i$ has a symmetric distribution around θ, we do not have $E[h(X_i)Y_iI(|X_i| \leq \nu_n)] = \theta$ in general. Moreover under limited overlap when tails are heavy, bias may converge too slowly such that $(n/\sigma_n^2)^{1/2}\{E[h(X_i)Y_iI(|X_i| \leq$

$\nu_n)] - \theta\} \to (0, \infty]$ in which case there is asymptotic bias in the limit distribution, where $\sigma_n^2 \equiv E[(h(X_i)Y_i I(|X_i| \leq \nu_n) - E[h(X_i)Y_i I(|X_i| \leq \nu_n)])^2]$. See especially Csörgo, Horváth, and Mason (1986), and see Khan and Tamer (2010) and Hill (2015).

Yang (2015) studies estimators of the type $\hat{\mu}_n \equiv 1/n \sum_{i=1}^n W_i I(-\tilde{\nu}_n \leq V_i \leq -\nu_n)$, where W_i and V_i are random variables, $(\nu_n, \tilde{\nu}_n) > 0$ and $(\nu_n, \tilde{\nu}_n) \to \infty$. Let $x_{n,i} \equiv W_i I(-\tilde{\nu}_n \leq V_i \leq -\nu_n)$, $\sigma_n^2 \equiv E[(x_{n,i} - E[x_{n,i}])^2]$, and bias is $\mathcal{B}_n \equiv E[x_{n,i}] - E[W_i]$. Under an iid assumption, Yang (2015) gives necessary and sufficient conditions for the existence of $(\tilde{\nu}_n, v_n)$ such that the Lindeberg condition for $(n^{1/2}/\sigma_n)(\hat{\mu}_n - E[W_i] - \mathcal{B}_n)$ holds, an optimal convergence rate is achieved, and $(n^{1/2}/\sigma_n)\mathcal{B}_n = O(1)$. Yang (2015) only tackles inverse density weighted cases $W_i = Y_i/f_V(v)$ where $f_V(v)$ is the density function for V_i, thus W_i is trimmed by some covariate as in Khan and Tamer (2010). Theory is only developed for endogenous selection models where one-sided trimming is used: $\tilde{\nu}_n$ is fixed while $v_n \to \infty$, thus only one threshold sequence is chosen. Yang's (2015) goal is a set of theoretical statements that characterize the existence of an optimal $\{v_n\}$ in terms of rate of convergence, but not inference itself. Indeed, there is possible asymptotic bias in the limit distribution $(n^{1/2}/\sigma_n)(\hat{\mu}_n - E[W_i]) \xrightarrow{d} N(\mathfrak{B}, 1)$ where $\mathfrak{B} \equiv \lim_{n\to\infty}(n^{1/2}/\sigma_n)\mathcal{B}_n < \infty$, and an estimator of $\mathfrak{B}$ is not given. Moreover, there is no guarantee that the chosen $\{v_n\}$ for a given sample will actually lead to trimming, and generally the estimator results in bias making it sub-optimal relative to competing estimators (see Section 5 in Yang, 2015).

Our estimator seeks to address the above issues. It trims by a plug-in version of $Z_i = h(X_i)Y_i$ allowing for parametric estimation of $p(X_i)$.[6] Asymptotic normality is assured whether limited overlap implies $h(X_i)Y_i$ has an infinite variance or not. Indeed, the power law decay rate needs neither be known, nor even true, for a standard asymptotic theory to be valid and for our bias correction approach to be valid (see Hill, 2015). We demonstrate by simulation that trimming $h(X_i)Y_i$ when $h(X_i)Y_i$ is a sample extreme leads to a sharp and approximately normal estimator when only a few sample extremes are removed, which makes the bias correction in small samples fairly sharp. On the other hand, a computation experiment in Chaudhuri and Hill (2014, Part I: Appendix G) reveals that the link between scalar X_i, $p(X_i)$ or Y_i, and $h(X_i)Y_i$, can be fairly weak in a latent variable treatment selection framework, hence trimming by X_i, $p(X_i)$ or Y_i can lead to unstable estimators. A similar Monte Carlo experiment in Section 4 shows that, when trimming by X_i or $p(X_i)$, a substantially greater number of observations need to be trimmed to ensure approximate normality in small samples, and therefore accurate asymptotic inference.

Finally, recall that the ATE is already identified under limited overlap and hence our focus is beyond internal stability. Thus, the approach of Crump, Hotz, Imbens, and Mitnik (2009) of not involving the outcome Y_i in the trimming rule in order to avoid deliberate bias with respect to the treatment effects being analyzed is not necessary for our purpose. Our simulation experiment shows trimming by Y_i leads to poor inference when limited overlap is severe enough for Z_i to have an infinite variance.

Small Sample Inference Lastly, Rothe (2015) exploits exact small sample inference methods in the statistics literature to produce robust intervals of the ATE. The data, however, must be distributed according to a scale mixture of normals. We only require a power law assumption on tail decay to justify a model of bias, while Hill (2015) shows the bias model leads to valid inference even if tails decay faster than a power law.

[6] A non-parametric estimator of $p(X)$ can in principle be used for efficient estimation of ATE under when the overlap is indeed strict (see Hirano, Imbens, and Ridder (2003)), but aspects of our limit theory will be different and consume unnecessary space for development.

3 Tail-Trimmed IPW Estimator

We present our core trim-by-Z IPW estimator $\hat{\theta}_n^{(tz)}$ and then discuss asymptotic bias. We then present an optimally fitted bias-corrected estimator $\hat{\theta}_n^{(tz:o)}$. We complete the section by summarizing how to implement our estimator based on logical fractile choices for the tail-trimmed estimator and bias estimator.

3.1 The Tail-Trimmed Estimator

Our goal is IPW estimation and inference of θ using the observed sample $\{Y_i, D_i, X_i\}_{i=1}^n$ on n units drawn at random from the population of interest. We work with a postulated parametric model $p(X, \gamma)$, where $\gamma \in \mathbb{R}^q$ is unknown with finite dimension $q \geq 1$. The model is assumed correct: there exists a unique γ_0 such that $p(X) = p(X, \gamma_0)$ *a.e.* $\sigma(X_i)$. See Assumption B1 below for the precise statement of the assumption.

Write

$$h_i(\gamma) \equiv h(X_i, \gamma) \equiv \frac{D_i}{p(X_i, \gamma)} - \frac{1 - D_i}{1 - p(X_i, \gamma)} \text{ with } h_i = h_i(\gamma_0) \text{, and } Z_i(\gamma) \equiv h_i(\gamma) Y_i \text{ with } Z_i \equiv Z_i(\gamma_0).$$

Define sample order statistics of mean centered $Z_i(\gamma)$:

$$\hat{Z}_{n,i}(\gamma) \equiv Z_i(\gamma) - \frac{1}{n}\sum_{j=1}^n Z_j(\gamma),\ \hat{Z}_{n,i}^{(a)}(\gamma) \equiv \left|\hat{Z}_{n,i}(\gamma)\right| \text{ and } \hat{Z}_{n,(1)}^{(a)}(\gamma) \geq \hat{Z}_{n,(2)}^{(a)}(\gamma) \geq \cdots \geq \hat{Z}_{n,(n)}^{(a)}(\gamma), \tag{3}$$

and let $\{k_n\}$ be an *intermediate order* sequence: $k_n \in \{1, ..., n\}$, $k_n \to \infty$ and $k_n/n \to 0$. Let $\hat{\gamma}_n$ be an estimator for γ_0. The tail-trimmed IPW estimator is

$$\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) \equiv \frac{1}{n - k_n}\sum_{i=1}^n Z_i(\hat{\gamma}_n) I\left(\left|Z_i(\hat{\gamma}_n) - \frac{1}{n}\sum_{j=1}^n Z_j(\hat{\gamma}_n)\right| < \hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)\right). \tag{4}$$

There are several features of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ that demand clarification. First, we scale by $n - k_n$ and use the mean-centered variable $Z_i(\hat{\gamma}_n) - 1/n\sum_{j=1}^n Z_j(\hat{\gamma}_n)$ as the trimming criterion in order to achieve an *asymptotically unbiased estimator* when Z_i is symmetrically distributed about θ. This is seemingly never exploited in the literature, but improves upon bias control when Z_i is asymmetrically distributed. Second, $k_n \to \infty$ and $k_n/n \to 0$ imply trimming matters for asymptotics, but is *negligible*. The threshold $\hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)$ is therefore an *intermediate order* statistic hence $\hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n) \xrightarrow{p} \infty$ (Leadbetter, Lindgren, and Rootzen, 1983; Galambos, 1987). Negligibility ensures $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ is consistent since Z_i may be asymmetrically distributed, it allows us to use extreme value theory for bias estimation, and it promotes asymptotic normality.

Third, $\hat{Z}_{n,i}(\hat{\gamma}_n)$ exploits two plug-ins: one for the propensity score via $\hat{\gamma}_n$, and one for mean centering via $Z_i(\hat{\gamma}_n) - 1/n\sum_{j=1}^n Z_j(\hat{\gamma}_n)$. Neither plug-in impacts the asymptotic properties of tail estimators like $\hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)$, as long as $k_n \to \infty$ slower than the plug-in $\hat{\gamma}_n$ rate of convergence (cf. Hill, 2014), and a moment bound on $h_i(\gamma_0)(\partial/\partial\gamma)p(X_i, \gamma_0)$ holds. The latter is standard in a maximum likelihood setting. The former easily achieved when $k_n \to \infty$ no faster than a slowly varying function, and $\hat{\gamma}_n = \gamma_0 + O_p(1/n^{\varphi})$ for some $\varphi > 0$, including nonparametric (typically where $\varphi \in (0, 1/2)$) and parametric ($\varphi = 1/2$) estimation, since then $1/n\sum_{i=1}^n Z_i(\hat{\gamma}_n)$ $= \theta + O_p(1/n^{\iota})$ for some $\iota > 0$ by classic arguments. We shorten theory details by only considering parametric estimators of γ_0 under Assumption B2 below.

We now restrict probability tail decay and the rate of increase $k_n \to \infty$. First, distribution properties.

Assumption A3 (Distribution Properties):

i. All random variables lie in a complete probability measure space $(\Omega, \mathcal{F}, \mathcal{P})$. $(Y_i, D_i, X_i)'$ are iid.

ii. If $E[Z_i^2] = \infty$ then Z_i has power law distribution tails:

$$P(Z_i - \theta \leq -c) \sim d_1 c^{-\kappa_1} \text{ and } P(Z_i - \theta \geq c) \sim d_2 c^{-\kappa_2}, \text{ where } \kappa_i > 1 \text{ and } d_i \in (0, \infty). \tag{5}$$

iii. Define $\xi \equiv [\gamma', \theta]' \in \mathbb{R}^{q+1}$ and $\mathcal{Z}_i(\xi) \equiv Z_i(\gamma) - \theta$, let ξ_0 be the true value of ξ, and let Ξ be a compact subset of $\mathbb{R}^{q+1}$ containing ξ_0. Let $\{c_n(\xi)\}$ be any sequence of mappings $c_n : \Xi \to (0, \infty)$ that satisfy $P(|\mathcal{Z}_i(\xi)| > c_n(\xi))$ $= k_n/n$.

a. $\mathcal{Z}_i(\xi)$ has for each ξ a continuous distribution with a continuous density function $f_{\mathcal{Z}(\xi)}$, and $E[\sup_{\xi\in\Xi} |\mathcal{Z}_i(\xi)|^{\iota}]$ $< \infty$ for some $\iota > 0$.

b. $c_n(\xi)$ is continuously differentiable with $\inf_{\xi\in\Xi}\{c_n(\xi)\} \to \infty$, $\sup_{\xi\in\Xi}\{c_n(\xi)\} = O(n^{\varpi})$ for some $\varpi > 0$, and $(\partial/\partial\xi)c_n(\xi_0) = O(c_n \mathring{\mathcal{L}}_n)$ for some slowly varying function $\mathring{\mathcal{L}}_n \to (0, \infty]$.

c. There exists a continuously differentiable mapping $\mathcal{K} : \Xi \to (0, \infty)$ with $\inf_{\xi\in\Xi} \mathcal{K}(\xi) > 0$, $\sup_{\xi\in\Xi} \mathcal{K}(\xi)$ $< \infty$ and $\sup_{\xi\in\Xi} ||(\partial/\partial\xi)\mathcal{K}(\xi)|| < \infty$, such that $\forall u \in \mathbb{R}$:

$$\lim_{n\to\infty} \sup_{\xi\in\Xi} \left| \frac{n}{k_n} c_n(\xi) \left\{ f_{\mathcal{Z}(\xi)} \left(-c_n(\xi) e^{u/k_n^{1/2}} \right) + f_{\mathcal{Z}(\xi)} \left(c_n(\xi) e^{u/k_n^{1/2}} \right) \right\} - \mathcal{K}(\xi) \right| = 0. \tag{6}$$

Remark 1 A complete measure space ensures majorants and integrals are measurable, and probabilities where applicable are outer probability. See Pollard (1984, Appendix C) and Dudley (1978).

Remark 2 Under (ii), in Chaudhuri and Hill (2014, Part I) we show that if the treatment assignment D_i satisfies a latent variable threshold crossing model, then (5) holds for some $(\kappa_1, \kappa_2) > 1$. The two-tailed representation is:

$$P(|Z_i - \theta| \geq c) = dc^{-\kappa}(1 + o(1)), \text{ where } \kappa \equiv \min\{\kappa_1, \kappa_2\}, \ d \equiv d_1 I(\kappa_1 \leq \kappa_2) + d_2 I(\kappa_1 \geq \kappa_2). \tag{7}$$

The tail index is identically the moment supremum $\kappa \equiv \arg\sup_{\alpha>0}\{E|Z_i|^{\alpha} < \infty\}$ (Resnick, 1987), hence $\kappa > 1$ ensures the ATE $\theta = E[Z_i]$ is finite.

We use parametric power law (5) to verify the Lindeberg condition for asymptotic normality when $\kappa \leq 2$, and to support a model of bias due to trimming. If tails decay faster than a power law, e.g. when limited overlap is not severe or strict overlap holds, then asymptotic normality and unbiasedness in the limit distribution are automatic, cf. Theorem 3.1, below. Model (5) is a special case of regularly varying tails $P(|Z_i - \theta| \geq c) = \mathcal{L}(c)c^{-\kappa}$ where $\mathcal{L}(c)$ is slowly varying, and here we use $\mathcal{L}(c) = d(1 + o(1))$ for simplicity. Other parametric models are possible both for verifying the Lindeberg condition and modeling bias, including logarithmic $\mathcal{L}(c)$. See Haeusler and Teugels (1985) amongst others. Moreover, the bias model need not be correct when tails are thinner than any power law (see Hill, 2015, Theorem 2.3).

Remark 3 (iii) is used to derive expansions of the trimming indicator $I(|\hat{Z}_{n,i}(\hat{\gamma}_n)| < \hat{Z}^{(a)}_{n,(k_n)}(\hat{\gamma}_n))$ around the two plug-ins $\hat{\gamma}_n$ and $1/n\sum_{i=1}^n Z_i(\hat{\gamma}_n)$. Distribution continuity A3(iii.a) implies $c_n(\xi)$ exists for each n. Property (6) is essentially a uniform tail balance condition for $\mathcal{Z}_i(\xi) \equiv Z_i(\gamma) - \theta$, and it holds when $\mathcal{Z}_i(\xi)$ has a power law tail for each ξ, with scale and tail index parameters that are uniformly bounded functions of ξ.

Next, we bound k_n to ensure the plug-ins $\hat{\gamma}_n$ and $1/n\sum_{j=1}^n Z_j(\hat{\gamma}_n)$ do not impact asymptotics.

Assumption A4 (Trimming Rate): $k_n \to \infty$ and $k_n = o(\ln(n))$.

The next three assumptions impose restrictions on the propensity score and its estimation. Obviously they are not required if $p(X_i)$ is assumed known.

Assumption B1 (parametric function): Let $\mathbb{X} \subseteq \mathbb{R}^k$ denote the support of $X_i \in \mathbb{R}^k$, and let $\Gamma \subset \mathbb{R}^q$. There exists a known mapping $p : \mathbb{X} \times \Gamma \to (0,1)$ such that $p(x,\gamma_0) = P(D_i = 1|x)\ \forall x \in \mathbb{X}$ for a unique interior point $\gamma_0 \in \Gamma$. $p(\cdot,\gamma)$ is Borel measurable for each $\gamma \in \Gamma$. $p(X_i,\gamma)$ is continuous and differentiable on Γ, $\sigma(X_i)$-*a.e.*

Assumption B2 (plug-in): $\hat{\gamma}_n$ satisfies $\sqrt{n}(\hat{\gamma}_n - \gamma_0) = 1/\sqrt{n}\sum_{i=1}^n w_i(1 + o_p(1))$ where $w_i \in \mathbb{R}^q$ is iid, $\sigma(X_i, D_i)$-measurable, it has a continuous distribution, $E[w_i] = 0$, $E[w_i^2] > 0$, and $E|w_i|^{2+\iota} < \infty$ for some $\iota > 0$.

Assumption B3 (moment bounds):

i. $\sup_{\gamma\in\Gamma}\{|h_i(\gamma)Z_i(\gamma)| \times ||(\partial/\partial\gamma)p_i(\gamma)||\}$ is L_p-bounded for some $p > 0$.

ii. $h_i(\gamma_0)(\partial/\partial\gamma)p(X_i,\gamma_0)$ is $L_{2+\iota}$-bounded for some $\iota > 0$.

Remark 4 We assume a parametric function to focus ideas, and due to its popularity. Common examples are logit $p(x,\gamma) = 1/(1 + \exp\{-x'\gamma\})$, and probit $p(x,\gamma) = \Phi(x'\gamma)$, where Φ is the standard normal cdf. Another example, which we will use in this paper, is Laplace: $p(x,\gamma) = .5\exp\{\sqrt{2}x'\gamma\}$ if $x'\gamma \leq 0$ and $p(x,\gamma) = 1 - .5\exp\{-\sqrt{2}x'\gamma\}$ if $x'\gamma > 0$.[7] Consider the additively separable threshold crossing model for treatment assignment is $D = I(g(X) - U \geq 0)$ for some measurable function $g(X)$. Then $p(X,\gamma_0) = F_{U|X}(g(X))$, hence a parametric form $p(X,\gamma_0)$ follows from the conditional distribution of the unobserved idiosyncratic component U.

Remark 5 B2 obviously implies $\sqrt{n}(\hat{\gamma}_n - \gamma_0) = O_p(1)$, while the standard method for achieving B2 is maximum likelihood. Other methods can be used, but are never used in practice because they do not offer any advantage over the maximum likelihood estimator (MLE) under Assumption B1. If $p(\cdot,\gamma)$ is continuously differentiable, with square integrable $h_i(\gamma)(\partial/\partial\gamma)p(X_i,\gamma_0)$, then under Assumption B1, the MLE

$$\hat{\gamma}_n \equiv \underset{\gamma\in\Gamma}{\arg\max}\left\{\sum_{i=1}^n l(D_i,X_i,\gamma)\right\} \text{ with } l(D_i,X_i,\gamma) \equiv \ln\left(p(X_i,\gamma)^{D_i}(1-p(X_i,\gamma))^{1-D_i}\right) \tag{8}$$

satisfies B2 with $w_i = (E[S_i(\gamma_0)S_i(\gamma_0)'])^{-1}S_i(\gamma_0)$ where $S_i(\gamma) \equiv (\partial/\partial\gamma)l(D_i,X_i,\gamma) = h_i(\gamma)(\partial/\partial\gamma)p(X_i,\gamma)$ satisfies $E[S_i(\gamma)] = 0$ *if and only if* $\gamma = \gamma_0$. Functions $p(x,\gamma)$ that are not everywhere differentiable on Γ are also allowed, provided primitive stochastic differentiability conditions hold (see, e.g. Pakes and Pollard, 1989, Section 3). This covers, for example, Laplace $p(x,\gamma)$ provided $\inf_{\gamma'\gamma}|X_i'\gamma| > 0\ a.s.$

[7] In the Laplace case, as long as X_i has linearly independent components and therefore $\inf_{\gamma'\gamma}|X_i'\gamma| > 0\ a.s.$, then $p(X_i,\gamma)$ is continuous and almost surely differentiable on Γ, in which case B1 holds.

Remark 6 In the heavy tail case $E[Z_i^2] = \infty$, as long as $\hat{\gamma}_n \xrightarrow{p} \gamma_0$ faster than the trimming fractile $k_n \to \infty$, then $\hat{\gamma}_n$ does not asymptotically affect our core estimator $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$, nor the bias estimator in Section 3.2. This is assured when $\hat{\gamma}_n \xrightarrow{p} \gamma_0$ faster than a slowly varying function coupled with Assumption A4. We assume here $\sqrt{n}$-convergence to reduce technical arguments since a slower rate in the thin tail case $E[Z_i^2] < \infty$ will naturally govern asymptotics (e.g. nonparametric estimators of $p(x)$).

Remark 7 B3(i) is used to extract an asymptotic expansion for the trimming indicator $I(|Z_i(\hat{\gamma}_n) - 1/n \sum_{j=1}^n Z_j(\hat{\gamma}_n)| < \hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n))$ around γ_0. B3(ii) implies the rate of convergence of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ is determined by the order of the tail-trimmed second moment of $Z_i - \theta$, effectively as if γ_0 were known. In the maximum likelihood case B3(ii) follows instantly from B2 since $E|w_i|^{2+\iota} < \infty$ implies $h_i(\gamma)(\partial/\partial\gamma)p(X_i, \gamma)$ is $L_{2+\iota}$-bounded.

The limit distribution of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ requires a deterministic sequence that the thresholds $\hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)$ approximate, identically $c_n = c_n(\xi_0)$ in A3(iii):

$$P\left(|Z_i - \theta| \geq c_n\right) = \frac{k_n}{n}. \tag{9}$$

The proper standardization for $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ requires the following constructions:

$$\begin{aligned}
\mathcal{D}_n &\equiv -E\left[\frac{\partial}{\partial\gamma}p(X_i,\gamma_0)h_i Z_i I\left(|Z_i - \theta| < c_n\right)\right] \\
\vartheta_{n,i} &\equiv \left(Z_i - \theta\right) I\left(|Z_i - \theta| < c_n\right) - E\left[\left(Z_i - \theta\right) I\left(|Z_i - \theta| < c_n\right)\right] + \mathcal{D}_n' w_i.
\end{aligned}$$

Now define variance and bias terms:

$$\sigma_n^2 \equiv E\left[\left\{\left(Z_i - \theta\right) I\left(|Z_i - \theta| < c_n\right) - E\left[\left(Z_i - \theta\right) I\left(|Z_i - \theta| < c_n\right)\right]\right\}^2\right] \tag{10}$$

$$\mathcal{V}_n^2 \equiv E\left[\vartheta_{n,i}^2\right] = \sigma_n^2 + 2E\left[\left\{Z_i I\left(|Z_i| < c_n\right) - E\left[Z_i I\left(|Z_i| < c_n\right)\right]\right\} w_i'\right]\mathcal{D}_n + \mathcal{D}_n' E\left[w_i w_i'\right]\mathcal{D}_n \tag{11}$$

$$\mathcal{B}_n \equiv \frac{n}{n - k_n} E\left[\left(Z_i - \theta\right) I\left(|Z_i - \theta| \geq c_n\right)\right].$$

In the maximum likelihood case, $S_i(\gamma) \equiv h_i(\gamma) \times (\partial/\partial\gamma)p(X_i, \gamma)$ is the score hence $E[S_i(\gamma_0)] = 0$. Thanks to the expression of $h(X_i)$, this implies $-\mathcal{D}_n$ is identically the covariance of $Z_i I(|Z_i - \theta| < c_n)$ and the score $S_i(\gamma_0)$, hence $\vartheta_{n,i}$ retains its conventional interpretation as the residual from an L_2 metric projection of the demeaned infeasible $Z_i I(|Z_i - \theta| < c_n)$ on the score. Recall that, when the infeasible untrimmed IPW estimator has a finite variance this interpretation is key to understanding why the asymptotic variance of the infeasible untrimmed IPW estimator cannot be smaller than that of the feasible untrimmed IPW estimator (see Graham, 2011). This beneficial attribute of feasible IPW estimation therefore remains valid even under trimming, irrespective of heavy tails: the variance of the infeasible $\hat{\theta}_n^{(tz)}(\gamma_0)$ cannot be smaller than the variance of the feasible $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ for any tail index $\kappa > 1$, hence there is no price to pay for trimming. There is, of course, a price to pay for not trimming: the untrimmed feasible and infeasible IPW estimators do not have a finite variance when Z_i has an infinite variance, hence the classic L_2 efficiency benefit of using a propensity score plug-in is unknown.

We show in the appendices that

$$\frac{n^{1/2}}{\mathcal{V}_n}\left(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)+\mathcal{B}_n-\theta\right)=\frac{1}{\mathcal{V}_n}\frac{1}{n^{1/2}}\sum_{i=1}^{n}\vartheta_{n,i}\left(1+o_p(1)\right),$$

where the right hand side is a self-standardized sum of independent (and for each n identically distributed) $\vartheta_{n,i}$. The term $\mathcal{V}_n^2$ captures dispersion in the tail-trimmed Z_i, and the influence of the propensity score plug-in $\hat{\gamma}_n$ on that dispersion. A standard requirement is $\liminf_{n\to\infty}\mathcal{V}_n^2>0$. This is only key when $E[Z_i^2]<\infty$: by Theorem 3.1, $\mathcal{V}_n^2\sim K\sigma_n^2$ with $K=1$ when $E[Z_i^2]=\infty$, while $\liminf_{n\to\infty}\sigma_n^2>0$ is assured by distribution non-degeneracy and trimming negligibility $c_n\to\infty$.

Assumption A5 (positive scale). $\liminf_{n\to\infty}\mathcal{V}_n^2>0$.

Unless otherwise stated, all proofs are presented in Appendix B. The estimator $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ is asymptotically normal, and asymptotically biased in its limit distribution when $\kappa<2$.

Theorem 3.1 *Let Assumptions A1, A2′, A3-A5, and B1-B3 hold.*

a. $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)\xrightarrow{p}\theta$ *and* $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)+\mathcal{B}_n-\theta)\xrightarrow{d}N(0,1)$.

b. $\mathcal{V}_n^2\sim K\sigma_n^2$ *for some* $K\in(0,1]$. *If* $\kappa>2$ *then* $\mathcal{V}_n=O(1)$, *and if* $\kappa\leq 2$ *then* $\mathcal{V}_n^2\sim\sigma_n^2\to\infty$.

c. If Z_i *has a symmetric distribution and/or* $\kappa\geq 2$ *then* $(n^{1/2}/\mathcal{V}_n)(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)-\theta)\xrightarrow{d}N(0,1)$. *If* Z_i *has an asymmetric distribution and* $\kappa<2$ *then* $(n^{1/2}/\mathcal{V}_n)|\mathcal{B}_n|\to\infty$ *for* any *intermediate order sequence* $\{k_n\}$.

Remark 8 $\mathcal{V}_n^2\sim K\sigma_n^2$ for some $K\in(0,1]$ follows from the efficiency benefit of feasible IPW estimation. If $E[Z_i^2]$ $=\infty$ then the benefit is lost and $\mathcal{V}_n^2\sim\sigma_n^2$. This follows from $\sqrt{n}$ convergence of the plug-in $\hat{\gamma}_n$, while $\hat{\theta}_n^{(tz)}(\cdot)$ has a slower than $\sqrt{n}$ rate when $E[Z_i^2]=\infty$.

Remark 9 The rate of convergence $n^{1/2}\mathcal{V}_n^{-1}$ is determined entirely by the use of trimming since $\mathcal{V}_n^2\sim K\sigma_n^2=KE[(Z_i-\theta)^2I(|Z_i-\theta|<c_n)]$. This is trivial when $E[Z_i^2]<\infty$, but if $E[Z_i^2]=\infty$ then the plug-in $\hat{\gamma}_n\xrightarrow{p}\gamma_0$ *faster* than the trimmed mean converges, hence $\hat{\gamma}_n$ does not affect asymptotics: $\mathcal{V}_n^2\sim K\sigma_n^2$.

The rate of convergence is easily characterized since $\mathcal{V}_n^2\sim K\sigma_n^2$ and σ_n^2 can be approximated by Karamata's Theorem when $E[Z_i^2]=\infty$.

Lemma 3.2 *Let Assumptions A1, A2′, A3-A5, and B1-B3 hold.*

a. If $E[Z_i^2]<\infty$ $(\kappa>2)$ *then asymptotics are the same as if trimming were not used, and the propensity score plug-in impacts asymptotics:*

$$n^{1/2}\left(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)-\theta\right)\xrightarrow{d}N\left(0,\sigma^2+E\left[(Z_i-\theta)w_i'\right]\mathcal{D}+\mathcal{D}'E\left[w_iw_i'\right]\mathcal{D}\right),$$

where $\mathcal{D}\equiv E[(\partial/\partial\gamma)p(X_i,\gamma_0)h_iZ_i]$ *and* $\sigma^2\equiv E[(Z_i-\theta)^2]$.

b. If $E[Z_i^2]=\infty$ $(\kappa\leq 2)$ *then trimming, but not the propensity score plug-in, impacts asymptotics. If* $\kappa=2$ *then* $\{n/\ln(n/k_n)\}^{1/2}\times(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)-\theta)\xrightarrow{d}N(0,d)$, *where* d *is the power law scale in (7). If* $\kappa\in(1,2)$ *then:*

$$\frac{n^{1/2}}{(n/k_n)^{1/\kappa-1/2}}\left(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)+\mathcal{B}_n-\theta\right)\xrightarrow{d}N\left(0,\frac{2}{2-\kappa}d^{2/\kappa}\right)\text{ where }\frac{n^{1/2}}{(n/k_n)^{1/\kappa-1/2}}|\mathcal{B}_n|\to\infty.$$

Remark 10 Tail trimming has no impact on first order efficiency if $E[Z_i^2] < \infty$, and hence with the MLE plug-in $\hat{\gamma}_n$ the asymptotic variance of our tail trimmed estimator takes the standard form:

$$\mathcal{V}_n^2 \to E\left[\left((Z_i - \theta) - E[Z_i S_i'(\gamma_0)] \left(E[S_i(\gamma_0)S_i'(\gamma_0)]\right)^{-1} S_i(\gamma_0)\right)^2\right],$$

which is simply the variance of the residual from the population least squares projection of the (demeaned) infeasible Z_i (based on the true $p(X_i)$) on the score $S_i(\gamma_0)$ for the parametric model of $p(X_i) = p(X_i, \gamma_0)$. If $\kappa < 2$ then trimming impacts asymptotics, but $\hat{\gamma}_n$ does not because $\hat{\gamma}_n$ has an order $1/n^{1/2}$ while the order of $1/n \sum_{i=1}^n (Z_i - \theta) I(|Z_i - \theta| < c_n)$ is $\sigma_n / n^{1/2}$, hence $\mathcal{V}_n^2 \sim \sigma_n^2$. The convergence rate in this case $n^{1/2}/\sigma_n$ can be increased by increasing the rate of trimming $k_n \to \infty$.

Remark 11 The rate of convergence of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ is affected by the number of trimmed observations k_n only in the infinite variance case $\kappa < 2$. The rate $n^{1/2}/\left(n/k_n\right)^{1/\kappa - 1/2} = k_n^{1/\kappa - 1/2} n^{1-1/\kappa}$ increases monotonically as $k_n \nearrow Kn$. Sample extremes in mean estimation add noise and therefore dampen the rate of convergence, hence removing more of them increases the convergence rate. In practice, however, removing more sample extremes augments bias.[8] In Chaudhuri and Hill (2014, Part I: Lemma D.1) we show that bias dominates the first order mean squared error of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ when $\kappa \neq 2$, and dominates for all κ if the Assumption A4 trimming bound $k_n = o(\ln(n))$ were not invoked (recall $k_n = o(\ln(n))$ ensures $\hat{\gamma}_n$ and $Z_i(\hat{\gamma}_n) - 1/n \sum_{j=1}^n Z_j(\hat{\gamma}_n)$ do not impact asymptotics). Thus, optimizing the convergence rate in general comes at a cost of a diminished mse and therefore higher bias. Further, Hill and Prokhorov (2016) prove that the second order bias of a tail-trimmed mean is also lower for smaller k_n. In terms of inference, using a small k_n that slowly increases promotes the least bias. This is natural since the untrimmed estimator is unbiased (in its limit distribution). This is also useful since our bias estimator exploits a tail approximation of bias based on Karamata theory, and by construction that approximation is better farther out in the tails, and therefore if fewer observations are trimmed. Finally, we do not explore higher order asymptotics in this paper, but an interesting (and unresolved) question is whether a unique k_n exits which minimizes a higher order mean-squared-error.

3.2 Bias-Corrected Tail-Trimmed Estimation

We now estimate and remove bias. As opposed to Peng (2001) and Hill (2015), we exploit a bias formula that leads to an estimator that does not affect the limit distribution of the bias corrected ATE estimator.

3.2.1 Bias-Correction

We exploit a key approximation of the bias term $\mathcal{B}_n$ under power law (5). We focus on the general case here, and leave for Chaudhuri and Hill (2014, Part I) formulas under tail symmetry.

Lemma 3.3 *Under power law (5):*

$$\mathcal{B}_n \sim \frac{n}{n-k_n}\left\{d_2^{1/\kappa_2}\left(\frac{\kappa_2}{\kappa_2 - 1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_2} - d_1^{1/\kappa_1}\left(\frac{\kappa_1}{\kappa_1 - 1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_1}\right\}. \tag{12}$$

[8]In regression model estimation, sample extremes in regressors have a well known leverage effect, which increases the rate of convergence when the regressors have an infinite variance. See, e.g., Hill (2012b) for theory and references.

Under a second order power law property imposed below, the approximation error in (12) vanishes at a $\sqrt{n}$ rate (which is no slower than the convergence rate $\sqrt{n}/\mathcal{V}_n$ of our estimators), hence it suffices to estimate the right hand side of (12). This was first noted in Peng (2001) for iid data. Hill (2015) allows for dependence, generalizes how bias is estimated in order to simplify asymptotics, and optimally fits an estimator of an expression similar to the right hand side of (12) to reduce bias further.

We now improve upon Hill's (2015) estimator in several key ways explained below, leading to a bias corrected estimator with the same limit distribution as $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$. Define tail specific versions of $\hat{Z}_{n,i}(\gamma) \equiv Z_i(\gamma) - 1/n \sum_{j=1}^n Z_j(\gamma)$, and their order statistics: $\hat{Z}_{n,i}^{(a)}(\gamma) \equiv |\hat{Z}_{n,i}(\gamma)|$ and

$$\hat{Z}_{n,i}^{(-)}(\gamma) \equiv -\hat{Z}_{n,i}(\gamma) I\left(\hat{Z}_{n,i}(\gamma) < 0\right) \text{ and } \hat{Z}_{n,i}^{(+)}(\gamma) \equiv \hat{Z}_{n,i}(\gamma) I(\hat{Z}_{n,i}(\gamma) > 0) \text{ with } \hat{Z}_{n,(j)}^{(\cdot)}(\gamma) \geq \hat{Z}_{n,(j+1)}^{(\cdot)}(\gamma).$$

Now let $\{m_n\}$ be an intermediate order sequence: $m_n \in \{1, ..., n\}$, $m_n \to \infty$ and $m_n = o(n)$. We estimate the two-tailed κ and tail specific (κ_1, κ_2) with Hill's (1975) seminal tail index estimator:[9]

$$\hat{\kappa}_{m_n,1}^{-1}(\gamma) = \frac{1}{m_n - 1} \sum_{j=1}^{m_n - 1} \ln \left(\frac{\hat{Z}_{n,(j)}^{(-)}(\gamma)}{\hat{Z}_{n,(m_n)}^{(-)}(\gamma)} \right) \text{ and } \hat{\kappa}_{m_n,2}^{-1}(\gamma) = \frac{1}{m_n - 1} \sum_{j=1}^{m_n - 1} \ln \left(\frac{\hat{Z}_{n,(j)}^{(+)}(\gamma)}{\hat{Z}_{n,(m_n)}^{(+)}(\gamma)} \right).$$

Hall (1982) proposes estimators of the scales (d_1, d_2):

$$\hat{d}_{m_n,1}(\gamma) \equiv \frac{m_n}{n} \left(\hat{Z}_{n,(m_n)}^{(-)}(\gamma) \right)^{\hat{\kappa}_{m_n,1}(\gamma)} \text{ and } \hat{d}_{m_n,2}(\gamma) \equiv \frac{m_n}{n} \left(\hat{Z}_{n,(m_n)}^{(+)}(\gamma) \right)^{\hat{\kappa}_{m_n,2}(\gamma)}.$$

We therefore estimate bias as follows:[10]

$$\begin{aligned} \hat{\mathcal{B}}_n(\gamma) \quad = \quad & \frac{n}{n - k_n} \left\{ \hat{d}_{m_n,2}^{1/\hat{\kappa}_{m_n,2}(\gamma)}(\gamma) \left(\frac{\hat{\kappa}_{m_n,2}(\gamma)}{\hat{\kappa}_{m_n,2}(\gamma) - 1} \right) \left(\frac{k_n}{n} \right)^{1 - 1/\hat{\kappa}_{m_n,2}(\gamma)} \right. \\ & \left. - \hat{d}_{m_n,1}^{1/\hat{\kappa}_{m_n,1}(\gamma)}(\gamma) \left(\frac{\hat{\kappa}_{m_n,1}(\gamma)}{\hat{\kappa}_{m_n,1}(\gamma) - 1} \right) \left(\frac{k_n}{n} \right)^{1 - 1/\hat{\kappa}_{m_n,1}(\gamma)} \right\}. \end{aligned} \tag{13}$$

The bias-corrected tail-trimmed ATE estimator is therefore

$$\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n). \tag{14}$$

The estimator $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$ is non-trivially different from estimators in Peng (2001) and Hill (2015). First, unlike Peng (2001), it allows for estimation of $\{\hat{\kappa}_{m_n,i}(\gamma), \hat{d}_{m_n,i}(\gamma)\}$ with a different fractile m_n than k_n used for trimming. If $\{\hat{\kappa}_{m_n,i}(\gamma), \hat{d}_{m_n,i}(\gamma)\}$ are $m_n^{1/2}$-consistent, and

$$m_n / k_n \to \infty, \tag{15}$$

then $\{\hat{\kappa}_{m_n,i}(\hat{\gamma}_n), \hat{d}_{m_n,i}(\hat{\gamma}_n)\}$ do not affect the limit distribution of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n)$ (cf. Hill, 2015). Second, Hill

[9] Many alternative estimators of κ are available: see Hill (2010) for references.

[10] Different order sequences $\{m_{1,n}, m_{2,n}\}$ can used to estimate κ_1 and κ_2, but in practice there will not be a convenient way to determine all three sequences $\{k_n, m_{1,n}, m_{2,n}\}$. For practical simplicity we therefore use one sequence $\{m_n\}$ for all tail estimators. Our simulations suggest this does not hinder the performance of our estimator.

(2015) uses a reduced version of the bias approximation in (12) for a one-tailed estimation problem that results in a one-tailed version of the threshold c_n appearing in the bias approximation. Thus, the reduction requires using the trimming threshold, here $\hat{Z}^{(a)}_{n,(k_n)}(\hat{\gamma}_n)$, in the bias estimator $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$. This unnecessarily complicates limit theory since $\hat{Z}^{(a)}_{n,(k_n)}(\hat{\gamma}_n)$ appears both in $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n)$ *and* $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$. We bypass the simplification, hence the threshold c_n does not appear in (12) and therefore $\hat{Z}^{(a)}_{n,(k_n)}(\hat{\gamma}_n)$ does not appear in (13). This is a key improvement over estimators in Peng (2001) and Hill (2015) since, under fractile rule (15), the estimator $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$ does not affect asymptotics: $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n) - \theta) \xrightarrow{d} N(0,1)$. See Theorem 3.4 below.

A shortcoming of $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n)$ is its use of one fractile m_n for tail exponent estimation, while $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$ is well defined only when $\hat{\kappa}_{m_n,i} > 1$, and when m_n is not greater than the number of negative or positive $\hat{Z}_{n,i}(\hat{\gamma}_n)$. Further, it seems desirable to choose m_n such that $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n)$ is close to an unbiased estimator, for example the untrimmed $1/n\sum_{i=1}^n Z_i(\hat{\gamma}_n)$.

Consider $m_n(\phi) = [\phi m_n]$ where $\phi \in \Phi^* = [\underline{\phi}, \bar{\phi}]$ for some chosen $0 < \underline{\phi} < \bar{\phi}$, and let $\hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi)$ be bias (13) computed with $m_n(\phi)$. Similar to an estimator in Hill (2015), the new bias-corrected estimator is

$$\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n) \text{ where } \phi^*_n = \underset{\phi\in\Phi^*}{\arg\min}\left|\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi) - \frac{1}{n}\sum_{i=1}^n Z_i(\hat{\gamma}_n)\right| \tag{16}$$

where

$$\Phi^* = \left\{\phi \in \left[\underline{\phi}, \bar{\phi}\right] : \left[\hat{\kappa}_{m_n(\phi),i}\right]_{i=1}^2 > 1 \text{ and } m_n(\phi) > \min\left\{\sum_{i=1}^n I\left(\hat{Z}_{n,i}(\hat{\gamma}_n) < 0\right), \sum_{i=1}^n I\left(\hat{Z}_{n,i}(\hat{\gamma}_n) > 0\right)\right\}\right\}. \tag{17}$$

Notice $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n)$ merely fixes $\phi = 1$. In view of the form $m_n(\phi) = [\phi m_n]$ with $\phi > 0$, as long as $m_n/k_n \to \infty$ then $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)$ has the same limit distribution as $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n)$.

Even though $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)$ corrects for bias, sampling error can render it farther from the untrimmed $\tilde{\theta}_n(\hat{\gamma}_n) \equiv 1/n\sum_{i=1}^n Z_i(\hat{\gamma}_n)$ than the non-bias-corrected $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n)$. In practice, we therefore use whichever estimator is closest to an unbiased estimator:

$$\begin{aligned}\hat{\theta}^{(tz:o)}_n(\hat{\gamma}_n) \equiv &\left\{\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)\right\} I\left(\left|\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n) - \tilde{\theta}^{(tz)}_n(\hat{\gamma}_n)\right| < \left|\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) - \tilde{\theta}^{(tz)}_n(\hat{\gamma}_n)\right|\right) \\ &+ \hat{\theta}^{(tz)}_n(\hat{\gamma}_n) I\left(\left|\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n) - \tilde{\theta}^{(tz)}_n(\hat{\gamma}_n)\right| \geq \left|\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) - \tilde{\theta}^{(tz)}_n(\hat{\gamma}_n)\right|\right).\end{aligned} \tag{18}$$

As long as $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n)$ is biased asymptotically in its limit distribution, then $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)$ will be chosen with probability approaching one. Small sample experiments reveal $\hat{\theta}^{(tz:o)}_n(\hat{\gamma}_n)$ has a tangible advantage over $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)$ precisely due to sampling error in bias estimation. Since $\hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)$ does not affect asymptotics, each $\hat{\theta}^{(tz:o)}_n(\hat{\gamma}_n)$, $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi^*_n)$ and $\hat{\theta}^{(tz)}_n(\hat{\gamma}_n) + \mathcal{B}_n$ has the same scale $\mathcal{V}_n$ and limit distribution, as we show below.

3.2.2 Large Sample Properties

A second order tail property and restricted $m_n \to \infty$ ensure $\{\hat{\kappa}_{m_n,i}(\gamma), \hat{d}_{m_n,i}(\gamma)\}$ are $m_n^{1/2}$-convergent.

Assumption A3′ (Second Order Power Law): A3(i) and A3(iii) hold. Further, (ii) for some $d_i > 0$, $\eta_i > 0$, and

$\kappa_i > 1$:

$$P\left(Z_i - \theta < -c\right) = d_1 c^{-\kappa_1}\left(1 + O(c^{-\eta_1})\right) \text{ and } P\left(Z_i - \theta > c\right) = d_2 c^{-\kappa_2}\left(1 + O(c^{-\eta_2})\right). \quad (19)$$

Further, $m_n \to \infty$, $m_n = o(n^{2\eta/(2\eta+\kappa)})$ and $m_n/k_n \to \infty$ where $\eta \equiv \min\{\eta_1, \eta_2\}$ and $\kappa \equiv \min\{\kappa_1, \kappa_2\}$.

Remark 12 Decay (19) is a popular assumption in the literature, dating to Hall (1982). Many higher order tail forms, with a restriction on $m_n \to \infty$, are similarly viable (see Haeusler and Teugels, 1985, Section 5), but we limit ourselves to just one for simplicity of notation.

Remark 13 The fractile bound $m_n = o(n^{2\eta/(2\eta+\kappa)})$ reflects the need to use observations strictly from the tails when Z_i deviates from an exact Pareto law (cf. Hall, 1982; Haeusler and Teugels, 1985). An exact Pareto law has $\eta = \infty$, in which case we need only bound $m_n = o(n)$.

Remark 14 The A3$'$ and A4 requirements $m_n = o(n^{2\eta/(2\eta+\kappa)})$, $m_n/k_n \to \infty$, and $k_n = o(\ln(n))$ are satisfied when $k_n = [\lambda_k(\ln(n))^{\delta_k}]$ and $m_n = [\lambda_m(\ln(n))^{\delta_m}]$ for any $0 < \delta_k < 1$, $\delta_m > \delta_k$, and $\lambda_k, \lambda_m > 0$. The discussion of Section 3.3 implies the use of first or higher order asymptotics does not lead to interior solutions for trimming parameters (λ_k, δ_k), but implies bias reduction requires small (λ_k, δ_k) for trimming. Conversely, larger (λ_m, δ_m) for bias estimation augments the rate of convergence of the bias estimators. Our simulation study gives some guidance for choosing these parameters.

The bias corrected estimators are asymptotically normal and unbiased, with the same normalization due to $m_n/k_n \to \infty$.

Theorem 3.4 *Under Assumptions A1, A2$'$, A3$'$, A4, A5, B1-B3 and (15)* $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n) - \theta)$, $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n, \phi_n^*) - \theta)$ *and* $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz:o)}(\hat{\gamma}_n) - \theta)$ *are asymptotically* $N(0,1)$.

Estimation of the scale $\mathcal{V}_n^2$, defined in (11), is straightforward. In the expansion $\sqrt{n}(\hat{\gamma}_n - \gamma_0) = 1/\sqrt{n}\sum_{i=1}^n w_i(1 + o_p(1))$ w_i is generally unobserved. Consider MLE: $w_i = (E[S_i(\gamma_0)S_i(\gamma_0)'])^{-1}S_i(\gamma_0)$ where $S_i(\gamma) = h_i(\gamma)(\partial/\partial\gamma)p(X_i, \gamma)$. Define

$$\begin{aligned}
\hat{w}_{n,i} &\equiv \left(\frac{1}{n}\sum_{i=1}^n S_i(\hat{\gamma}_n)S_i(\hat{\gamma}_n)'\right)^{-1} S_i(\hat{\gamma}_n) \\
\hat{\mathcal{D}}_n &\equiv -\frac{1}{n}\sum_{i=1}^n S_i(\hat{\gamma}_n)Z_i(\hat{\gamma}_n)I\left(\left|\hat{Z}_{n,i}(\hat{\gamma}_n)\right| < \hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)\right) \\
\hat{\mathcal{V}}_n^2 &\equiv \frac{1}{n-k_n}\sum_{i=1}^n \left\{\left(\hat{Z}_{n,i}(\hat{\gamma}_n)I\left(\left|\hat{Z}_{n,i}(\hat{\gamma}_n)\right| < \hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)\right) + \left(\frac{n-k_n}{n}\right)\hat{\mathcal{B}}_n(\hat{\gamma}_n)\right) + \hat{\mathcal{D}}_n'\hat{w}_{n,i}\right\}^2.
\end{aligned}$$

Notice $\hat{Z}_{n,i}(\hat{\gamma}_n)I(|\hat{Z}_{n,i}(\hat{\gamma}_n)| < \hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)) + ((n-k_n)/n)\hat{\mathcal{B}}_n(\hat{\gamma}_n)$ approximates the demeaned $(Z_i - \theta)I(|Z_i - \theta| < c_n) - E[(Z_i - \theta)I(|Z_i - \theta| < c_n)]$ since $((n-k_n)/n)\hat{\mathcal{B}}_n(\hat{\gamma}_n)$ estimates $((n-k_n)/n)\mathcal{B}_n = E[(Z_i - \theta)I(|Z_i - \theta| \geq c_n)] = -E[(Z_i - \theta)I(|Z_i - \theta| < c_n)]$.

In order to handle the mapping $S_i(\hat{\gamma}_n)$, we strengthened B1 smoothness properties of $p(X_i, \gamma)$, and the B3 moment conditions.

Assumption B1$'$ (parametric function). B1 holds, and $p(X_i, \gamma)$ is twice continuously differentiable, $\sigma(X_i)$-*a.e.*

Assumption B3$'$ (moment bounds):

i. $\sup_{\gamma \in \Gamma}\{||S_i(\gamma)Z_i(\gamma)|\}$, $\sup_{\gamma \in \Gamma}||S_i(\gamma)S_i(\gamma)'Z_i(\gamma)||$ and $\sup_{\gamma \in \Gamma}||h_i(\gamma)(\partial^2/\partial\gamma\partial\gamma')p_i(\gamma) \times Z_i(\gamma)||$ are L_p-bounded for some $p > 0$.

ii. $\sup_{\gamma \in \Gamma}||S_i(\gamma)||$ is L_4-bounded, and $||h_i(\gamma)(\partial^2/\partial\gamma\partial\gamma')p_i(\gamma)||$ is L_2-bounded.

Remark 15 Twice differentiability under B1$'$ of the propensity score is used to handle the plug-in in $S_i(\hat{\gamma}_n)$. We can replace it with a Lipschitz property on the first derivative at the cost of heavier notation. B3$'$ is used to derive limits for $1/n\sum_{i=1}^n S_i(\hat{\gamma}_n)Z_i(\hat{\gamma}_n)I(|\hat{Z}_{n,i}(\hat{\gamma}_n)| < \hat{Z}^{(a)}_{n,(k_n)}(\hat{\gamma}_n))$ and $1/n\sum_{i=1}^n S_i(\hat{\gamma}_n)S_i(\hat{\gamma}_n)'$. Bounding moments on the envelopes $\sup_{\gamma \in \Gamma}\{\cdot\}$ simplifies probability limit arguments. The B3$'$(ii) envelope bounds can be replaced with pointwise bounds and higher order smoothness properties that suffice for uniform laws of large numbers.

The proof of the following is lengthy and therefore relegated to Chaudhuri and Hill (2014, Part I).

Theorem 3.5 *Under Assumptions A1, A2$'$, A3$'$, A4, A5, B1$'$, B2, and B3$'$* $\hat{\mathcal{V}}_n^2/\mathcal{V}_n^2 \xrightarrow{p} 1$.

3.3 Implementation

The bias corrected estimator requires choices of the trimming fractile k_n and the fractile m_n for computing tail indices used for bias estimation. We discuss fractile choice based on first order asymptotics involving the rate of convergence and mean squared error, and higher order bias. We omit most technical details in order to simplify the discussion. See Hill and Prokhorov (2016) for related theory details.

3.3.1 First Order Asymptotics

If we optimize the rate of convergence $n^{1/2}/\sigma_n$ of our estimators by minimizing the variance σ_n^2, then it is always optimal to trim more in the heavy tailed case, a well known result demonstrated here by Lemma 3.2, and elsewhere (e.g. Hahn, Kuelbs, and Samur, 1987; Hill, 2012a,b, 2015). Trimming more sample extremes, however, necessarily augments first order bias when Z is not symmetrically distributed, and it augments higher order bias as we discuss below, which necessarily distorts (asymptotic) inference.

Khan and Tamer (2010) use the mean-squared-error to justify their thresholds choice. In our case, since the scale satisfies $\mathcal{V}_n^2 \sim K\sigma_n^2$ for some characterizable $K \in (0, 1]$, the asymptotic first order mean-squared-error of $\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)$ is $\mathcal{MSE}_n \equiv K\sigma_n^2/n + \mathcal{B}_n^2$. Since we use negligible trimming, minimizing $\mathcal{MSE}_n$ with respect to k_n always leads to a corner solution that depends on κ. A small k_n and slow $k_n \to \infty$ diminishes $\mathcal{MSE}_n$ when $\kappa \neq 2$ because bias dominates. Conversely, because $k_n = o(\ln(n))$, a larger k_n and faster $k_n \to \infty$ diminishes $\mathcal{MSE}_n$ when $\kappa = 2$ due to a dominant dispersion. See Chaudhuri and Hill (2014, Part I). Thus, except for the hairline infinite variance case $\kappa = 2$, mean-squared-error and bias minimization are identical, and imply we should remove few observations per sample, and increase the number removed very slowly, e.g. $k_n = \max\{1, \lambda_k(\ln(n))^{\delta_k}\}$ for $\lambda_k > 0$ and $\delta_k \in (0, 1)$. Choosing (δ_k, λ_k) by reducing bias or mean-squared-error generally leads to corner solutions, but small values are optimal when $\kappa \neq 2$. If we are free to choose $k_n \to \infty$ then for non-slowly varying k_n bias always dominates mse and small k_n is optimal.

3.3.2 Higher Order Bias

Hill and Prokhorov (2016, Section 4) show that trimming more tail observations augments small sample bias in a higher order expansion of a trimmed mean, irrespective of the values of (κ_1, κ_2). Moreover, recall that we do not estimate bias $\mathcal{B}_n$ per se, but asymptotic approximation (12) based on Karamata theory. Hence, at least in the power law case, trimming more observations moves us farther from the tails, making it more difficult to approximate, and therefore estimate, bias $\mathcal{B}_n$. A poor bias approximation leads to a poor estimator of bias, and therefore poor asymptotic inference.[11] Thus, in terms of higher order bias and inference, it seems desirable to use a small k_n and slow $k_n \to \infty$. Similarly, using a higher order expansion of the tail exponent estimators in $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$ it can be shown that using a large m_n diminishes higher order bias of $\hat{\mathcal{B}}_n(\hat{\gamma}_n)$.

In order to satisfy $k_n = o(\ln(n))$, $m_n \to \infty$ no faster than a slowly varying rate, and $k_n/m_n \to \infty$, a convenient choice is $k_n = \max\{1, [\lambda_k(\ln(n))^{1-\iota}]\}$ and $m_n = \max\{1, [\lambda_m \ln(n)]\}$ with $\lambda_k < \lambda_m$ and infinitesimal $\iota > 0$. In our simulation study we use $\lambda_k = .25$, $\lambda_m \in [2, 16]$ and $\iota = 10^{-10}$ which implies very few observations are trimmed relative to n, and far more tail observations are used for bias estimation. This results in a superb estimator $\hat{\theta}_n^{(tz:o)}(\hat{\gamma}_n)$ with small bias and mean-squared-error, and is approximately normal.

4 Monte Carlo Study

We present several Monte Carlo experiments in order to study IPW estimators of θ. We initially use one covariate and the treatment assignment model $D = I(\alpha + \beta X - U \geq 0)$ with $\alpha = 0$, and we assume the propensity score is known. Under the distributional assumptions of this simulation study, this serves as a benchmark since (i) having one covariate allows for strict control of limited overlap, and leads to symmetrically distributed Z and therefore unbiased estimation when trimming by Z, X, $p(X)$, or Y (see below); (ii) the power law properties of Z are fully characterized in Chaudhuri and Hill (2014, Part I); (iii) we omit the possibility of sampling error due to estimation of $p(X)$; and (iv) it provides a case where trimming by X and $p(X)$ are equivalent.

In the remaining experiments we relax symmetry by letting $\alpha \neq 0$; we use a parametric model $p(X, \gamma_0)$ for $p(X)$ and a plug-in estimator for γ_0; we use multiple covariates; and we consider trimming by Y. Including information on Y in the trimming criterion can lead to bias (see Crump, Hotz, Imbens, and Mitnik, 2009, p. 188). It would be interesting to see the extent of this bias in a controlled experiment.[12]

4.1 One Covariate, Known $p(X)$, and Symmetric Z

We begin with $D = I(\alpha + \beta X - U \geq 0)$ for choices $\alpha = 0$ and $\beta \in \{.25, 1, 2\}$, and $Y_j \perp X, U$, and we use the true propensity score.

4.1.1 Simulation Design

Initially we draw all variables from the same distribution: $(Y_{0,i}, Y_{1,i}, X_i, U_i)$ are iid standard normal, or Laplace with cdf $F(r) = .5e^{\sqrt{2}r}$ if $r \leq 0$ and $F(r) = 1 - .5e^{-\sqrt{2}r}$ if $r > 0$. We then draw $(Y_{0,i}, Y_{1,i}, X_i) \sim$ Laplace with $U_i \sim$ normal, and $(Y_{0,i}, Y_{1,i}, X_i) \sim$ normal with $U_i \sim$ Laplace. Under distribution symmetry, and $\alpha = 0$ and $Y_{j,i}$

[11] The same type of higher order expansion can be characterized for the bias-corrected tail-trimmed mean $\hat{\theta}_n^{(tz:bc)} \equiv \hat{\theta}_n^{(tz)} + \hat{\mathcal{B}}_n$ by expanding $\hat{\theta}_n^{(tz)}$ and the tail exponents in $\hat{\mathcal{B}}_n$. Although we do not provide the results in this paper since they are tediously long, the same essential findings arise as in Hill and Prokhorov (2016, Section 4). Trimming fewer observations leads to smaller higher order bias in $\hat{\theta}_n^{(tz)}$ *and* $\hat{\mathcal{B}}_n$, and increasing the tail exponent fractile m_n diminishes higher order bias in $\hat{\mathcal{B}}_n$.

[12] We thank a referee for suggesting the demonstration of this bias.

$\perp X_i, U_i$, in all cases the ATE $\theta = 0$ and Z_i has a symmetric distribution about 0, hence $\hat{\theta}_n^{(tz)}$, $\theta_n^{(tx)}$ and $\hat{\theta}_n^{(tx)}$ are asymptotically unbiased in their limit distribution. The sample sizes are $n \in \{100, 250, 500, 1000\}$.

We compute the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$, and the optimal bias-corrected version $\hat{\theta}_n^{(tz:o)}$ in (18). We use fractiles $k_n = [.25(\ln(n))^{1-\iota}]$ and $m_n(\phi_n^*) = [\phi_n^* \ln(n)]$, where $\iota = 10^{-10}$, and ϕ_n^* minimizes $|\hat{\theta}_n^{(tz)} + \hat{\mathcal{B}}_n(\phi_n^*) - \tilde{\theta}_n|$ over $\phi \in [2, 16]$ subject to the constraint in (16) and (17).

In this study we trim $k_n = [.25 \ln(n)] \in \{1, 1, 2, 2\} = \{1\%, .4\%, .4\%, .2\%\}$ observations when $n \in \{100, 250, 500, 1000\}$. These fractiles work well for heavy tail robustness, but work quite poorly for estimating the tail exponents required for bias-correction. We therefore allow for larger values for m_n, in particular up to $64k_n$.

Our choice of $\{k_n, m_n(\phi)\}$ is theoretically justified by Theorem 3.4, since Z_i has a second order tail form $P(|Z_i| > c) = dc^{-\kappa}(1 + O(c^{-\eta}))$ with $\eta \geq \kappa$ in either Laplace or Normal cases (cf. Chaudhuri and Hill, 2014, Part I: Theorems F.3 and F.4). Hence, $m_n = O(\ln(n))$ with $m_n/k_n \to \infty$ is always valid. See also Section 3.3 for the logic behind forcing k_n to be small and $k_n \to \infty$ slow, with a larger m_n, based on first and higher order asymptotic arguments.

We compare $\hat{\theta}_n^{(tz)}$ and $\hat{\theta}_n^{(tz:o)}$ to the untrimmed estimator $\tilde{\theta}_n \equiv 1/n \sum_{i=1}^n Z_i(\hat{\gamma}_n)$, the trim-by-$X$ estimator $\theta_n^{(tx)} = 1/n \sum_{i=1}^n Z_i(\hat{\gamma}_n)\ I(|X_i| \leq \nu_n)$ with threshold $\nu_n = \ln(\ln(n))$, and the adaptive version $\hat{\theta}_n^{(tx)} = 1/n \sum_{i=1}^n Z_i(\hat{\gamma}_n) I(|X_i| \leq X_{(k_n^{(x)})}^{(a)})$ discussed in Chaudhuri and Hill (2014, Part I: Appendix G) based on the order statistics of $X_i^{(a)} \equiv |X_i|$ with $k_n^{(x)} = [2n/\ln(n)] \in \{43, 91, 161, 290\} = \{43\%, 36\%, 32\%, 29\%\}$ when $n \in \{100, 250, 500, 1000\}$. The choice ν_n for $\theta_n^{(tx)}$ is based on the fact that by design $\theta_n^{(tx)}$ is unbiased, while a small and slow $\nu_n \to \infty$ implies heavier trimming which augments the convergence rate when $\beta > 1$, and $\nu_n = \ln(n)$ need not lead to any trimming for a particular sample. See Chaudhuri and Hill (2014, Part I: Appendix G) for discussion. Further, with $\nu_n = \ln(\ln(n))$ about $\{13, 22, 34, 53\}$ observations are typically trimmed for $\theta_n^{(tx)}$ when $n \in \{100, 250, 500, 1000\}$. The choice $k_n^{(x)} = [2n/\ln(n)]$ for $\hat{\theta}_n^{(tx)}$ implies comparatively heavy trimming, while $k_n^{(x)}$ is much larger than k_n to ensure extreme $Z_i's$ are trimmed as discussed in Chaudhuri and Hill (2014, Part I: Appendix G). As a control, we also use the much smaller $k_n^{(x)} = k_n$.

We also compute the trim-by-$p(X)$ estimator defined as follows. Let $p_i(\gamma) \equiv p(X_i, \gamma)$, define order statistics $p_{(1)}(\gamma) \geq \cdots \geq p_{(n)}(\gamma)$, and an intermediate order sequence $\{k_n^{(p)}\}$. The estimator is

$$\hat{\theta}_n^{(tp)}(\hat{\gamma}_n) \equiv \frac{1}{n} \sum_{i=1}^n Z_i(\hat{\gamma}_n) I\left(p_{(n-k_n^{(p)}+1)}(\hat{\gamma}_n) \leq p_i(\hat{\gamma}_n) \leq p_{(k_n^{(p)})}(\hat{\gamma}_n)\right).$$

In this case $k_n^{(p)}$ observations are trimmed from each tail, hence a total of $2k_n^{(p)}$ observations are trimmed with probability one. We therefore use either $k_n^{(p)} = [.0125 \ln(n)]$, in order to match $2k_n^{(p)} = k_n$ with respect to $\hat{\theta}_n^{(tz:o)}$; or $k_n^{(p)} = [\lambda_p n/\ln(n)]$ where $\lambda_p \in \{.25, .5, 1, 2\}$, while $\lambda_p = 1$ matches $2k_n^{(p)} = k_n^{(x)}$.

Under our maintained assumptions $n^{1/2}\mathcal{S}_n^{-1}(\hat{\theta}_n^{(tx)}(\hat{\gamma}_n) - \theta) \xrightarrow{d} N(0,1)$ in the heavy tail case $E[Z_i^2] = \infty$, and $n^{1/2}\mathcal{S}_n^{-1}(\hat{\theta}_n^{(tx)}(\hat{\gamma}_n) - \theta) \xrightarrow{d} N(0, K)$ for some $K \in (0, \infty)$ that depends on $p(X_i, \gamma_0)$. In the threshold crossing model $D_i = I(\beta X_i - U_i \geq 0)$ where U_i and X_i are independent, and U_i has a symmetric distribution about zero, then it can be shown that $(n^{1/2}/\tilde{\mathcal{S}}_n)(\hat{\theta}_n^{(tp)}(\hat{\gamma}_n) - \theta) \xrightarrow{d} N(0,1)$ for some sequence of positive constants $\{\tilde{\mathcal{S}}_n\}$, where $\tilde{\mathcal{S}}_n \to \infty$ if $E[Z_i^2] = \infty$.

4.1.2 Results

Let $\check{\theta}_{n,r}$ be the r^{th} sample value of any estimator, over $r = 1, ..., R$ samples, $R = 10,000$. Table 1 contains the simulation mean $1/R\sum_{r=1}^{R}\check{\theta}_{n,r}$, median, root mean squared error [mse] $s_n \equiv (1/R\sum_{r=1}^{R}\check{\theta}_{n,r}^2)^{1/2}$, and the percent of observations that are trimmed on average per sample. We also use the standardized ratio $\check{\theta}_{n,r}/s_n$ to test for normality by the Kolmogorov-Smirnov test. We report the KS statistic divided by its 5% critical value: values above one imply rejection of standard normality at the 5% level. In Table 2 we report rejection frequencies for an asymptotic test of $\theta = 0$ against $\theta \neq 0$ at the $\{1\%, 5\%, 10\%\}$ levels based on the statistic $\check{\theta}_{n,r}/s_n$ and critical values taken from a standard normal distribution. We only report results for sample sizes $n \in \{100, 250\}$ since the remaining results are similar, and we do not tabulate here the adaptive trim-by-$p(X)$ results since it performs on par with the adaptive trim-by-X estimator. See Chaudhuri and Hill (2014, Part II) for all compiled results.

The untrimmed $\tilde{\theta}_n$ is very sensitive to limited overlap $\beta \geq 1$. The presence of large values influences the sign of $\tilde{\theta}_n$, giving the appearance of bias. It is exceptionally heavy tailed when $\beta > 1$, and $\{U_i, X_i\}$ are iid or X_i is heavier tailed than U_i, and therefore $\tilde{\theta}_n$ is far from normally distributed. Empirical size for the t-test is therefore highly distorted, especially when $n \geq 250$ where the degree of heavy tailedness is better observed.

Overall the tail-trimmed $\{\hat{\theta}_n^{(tz)}, \hat{\theta}_n^{(tz:o)}\}$ are best across all measures: low bias, median close to θ, low mse, approximate normality, and rejection frequencies near the nominal test sizes. The adaptive trim-by-X estimator $\hat{\theta}_n^{(tx)}$ with a much larger trimming fractile $k_n^{(x)} > k_n$ is on par with $\{\hat{\theta}_n^{(tz)}, \hat{\theta}_n^{(tz:o)}\}$ in most cases; in some cases it has a smaller mse; while it deviates from normality in the very heavy tailed case where $(Y_{0,i}, Y_{1,i}, X_i) \sim$ normal with $U_i \sim$ Laplace and $\beta > 1$. The performance of $\hat{\theta}_n^{(tx)}$ comes at a substantial cost since we must trim far more observations than for the trim-by-Z estimators: $k_n^{(x)}/k_n \in \{43, 91, 80.5, 145\}$ for $n \in \{100, 250, 500, 1000\}$. This is staggering: we must trim 145 times as many observations when $n = 1000$ in order to achieve an estimator that compares well with $\{\hat{\theta}_n^{(tz)}, \hat{\theta}_n^{(tz:o)}\}$.

If we simply set $k_n^{(x)} = k_n$ then $\hat{\theta}_n^{(tx)}$ performs roughly on par with the untrimmed estimator due to the weak correspondence between X_i and Z_i: it exhibits small sample bias, larger mse, and deviates from normality when $\beta \geq 1$, where the deviation is profound in the heaviest tail cases. Similarly, the trim-by-X estimator $\theta_n^{(tx)}$ with our chosen threshold ν_n also compares closely to the untrimmed $\tilde{\theta}_n$, even though on average it removes far more observations than $\hat{\theta}_n^{(tx)}$ with k_n.

The trim-by-$p(X)$ estimator $\hat{\theta}_n^{(tp)}$ is similar to $\hat{\theta}_n^{(tx)}$. It generally works best when $k_n^{(p)} = [\lambda_p n/\ln(n)]$ and $\lambda_p \in \{1, 2\}$. This is ultimately due to a weak correspondence between $p(X_i)$ and Z_i.

The above findings verify by simulation the weak probabilistic link between $(X, p(X))$ and Z in a latent variable treatment assignment framework with a linear threshold crossing mechanism. These also provide strong support of the computational experiment in Chaudhuri and Hill (2014, Part I: Appendix G). Conversely, trimming by Z necessarily removes the most damaging observation(s), resulting in approximately normal estimators $\{\hat{\theta}_n^{(tz)}, \hat{\theta}_n^{(tz:o)}\}$, and sharp asymptotic inference, with very little trimming.

4.2 Asymmetric Z, Multivariate X, Unknown $p(X)$

We repeat the experiment in Section 4.1, except we now allow for multivariate X, a constant term, e.g. in the scalar X case $D = I(\alpha + \beta X - U \geq 0)$ with $\alpha \neq 0$, and we allow for estimation of the propensity score. When $\alpha \neq 0$, by repeating arguments in Chaudhuri and Hill (2014, Part I: Appendix F) it is straightforward to show that Z has asymmetric power law tails with symmetric tail indices: $\kappa_1 = \kappa_2$.

We only report results for sample sizes $n \in \{100, 250\}$ for estimators with non-trimming, trim-by-Z with optimal bias correction, and adaptive trim-by-X with $k_n^{(x)} > k_n$, since trim-by-$p(X)$ is similar, and the remaining are suboptimal under limited overlap. We omit reporting t-test rejection rates since these mimic findings from Sections 4.1: an estimator closer to normal has rejection rates closer to the nominal size of the test under the null. See Chaudhuri and Hill (2014, Part II) for test results for each $n \in \{100, 250, 500, 1000\}$; for t-test rejection rates; and for the trim-by-$p(X)$ estimator with fractiles $k_n^{(p)} = [\lambda_p n / \ln(n)]$ and $\lambda_p \in \{1, 2\}$ since only these in Section 4.1 lead to estimates that are robust to limited overlap.

4.2.1 One Covariate, Known $p(X)$, and Asymmetric Z

Let $D = I(.25 + \beta X - U \geq 0)$. Although $\kappa_1 = \kappa_2$, we still generalize bias estimation by using the general formula (13). See Table 3 for results. The estimators perform about the same as when $\alpha = 0$ (Z has a symmetric distribution). One difference is apparent: when $\beta > 1$ then the trim-by-Z and adaptive trim-by-X estimators are slightly farther from normal in some cases. Overall, however, the asymmetric bias correction for $\hat{\theta}_n^{(tz:o)}$ works well.

4.2.2 Unknown $p(X)$

We now estimate a parametric propensity score function with possibly multivariate X_i. The treatment assignment is $D_i = I(\gamma_0' X_i - U_i \geq 0)$, so we use the model $p(X_i, \gamma) \equiv F_U(\gamma' X_i)$ for the given distribution F_U described above, and we compute $\hat{\gamma}_n$ by maximum likelihood (8). We now drop the argument $\hat{\gamma}_n$ and simply write, e.g., $\hat{\theta}_n^{(tz)}$.

There are four cases. Let $\tilde{X}_i$ be stochastic covariates, and $\beta \in \{.25, 1, 2\}$ as in Section 4.1. The first two cases are the same as those in Sections 4.1 and 4.2.1, except that an estimate of $p(X_i)$ is used.

Case 1. The covariate is scalar $X_i = \tilde{X}_i$, and $(Y_{0,i}, Y_{1,i}, \tilde{X}_i, U_i)$ have the various distributions in Section 4.1. We include a constant term for estimation, hence $[1, \tilde{X}_i]$ is used for estimating $\gamma_0 = [0, \beta]'$.

Case 2. We now add and estimate a constant term. The covariate is $X_i = [1, \tilde{X}_i]$ for scalar $\tilde{X}_i$; $\gamma_0 = [.25, \beta]$ as in Section 4.1; $(Y_{0,i}, Y_{1,i}, \tilde{X}_i, U_i)$ are as above; and $[1, \tilde{X}_i]$ is used for estimating γ_0.

The last two cases have multiple stochastic covariates.

Case 3. Stochastic covariates are $\tilde{X}_i = [\tilde{X}_{j,i}]_{j=1}^3$, where $\tilde{X}_{1,i}$ is Bernoulli with $P(\tilde{X}_{1,i} = 1) = .3$, $\tilde{X}_{3,i} = \tilde{X}_{2,i}^2$, and $(Y_{0,i}, Y_{1,i}, \tilde{X}_{2,i}, U_i)$ are as above; $\gamma_{0,1} = .5$, $\gamma_{0,2} = \beta$ and $\gamma_{0,3} = \beta/2$. We include a constant term for estimating γ_0 $= [0, .5, \beta, \beta/2]'$.

Case 4. We now add and estimate a constant term. The covariates are $\tilde{X}_i = [\tilde{X}_{j,i}]_{j=1}^4$, $\tilde{X}_{1,i} = 1$, $\tilde{X}_{2,i}$ is Bernoulli with $P(\tilde{X}_{2,i} = 1) = .3$, $\tilde{X}_{4,i} = \tilde{X}_{3,i}^2$, and $(Y_{0,i}, Y_{1,i}, \tilde{X}_{3,i}, U_i)$ are as above; the constant term is $\gamma_{0,1} = .25$, and the remaining parameters are $(\gamma_{0,2}, \gamma_{0,3}, \gamma_{0,4}) = (.5, \beta, \beta/2)$.

The general bias estimator (13) is again used, although Z has symmetric tail indices. The heaviest tailed covariate in Case 3 (and 4) is $\tilde{X}_{3,i}$ (and $\tilde{X}_{4,i}$), the square of the scalar regressor used in Section 4.1. Thus, $\tilde{X}_{3,i}$ (and $\tilde{X}_{4,i}$) and U_i drive the tail properties of Z_i. The trim-by-X estimator uses just one covariate for trimming: we naturally use $\tilde{X}_i$ in Cases 1 and 2, $\tilde{X}_{2,i}$ in Case 3, and $\tilde{X}_{3,i}$ in Case 4. We follow standard practice and include a constant term for estimation in all cases.

Since there is essentially no difference between using the true or estimated propensity score, the results are placed in Chaudhuri and Hill (2014, Part II). The only noticeable difference, however, is the slightly smaller mse of $\hat{\theta}_n^{(tz:o)}(\hat{\gamma}_n)$ relative to $\hat{\theta}_n^{(tz:o)}(\gamma_0)$ when $E[Z_i^2] < \infty$, for larger sample sizes $n \in \{500, 1000\}$. Recall that $\mathcal{V}_n^2/\sigma_n^2$

$\to (0,1)$ is predicted by Theorem 3.1 when $E[Z_i^2] < \infty$, where $\mathcal{V}_n^2$ and σ_n^2 are the respective mse's of $\hat{\theta}_n^{(tz:o)}(\hat{\gamma}_n)$ and $\hat{\theta}_n^{(tz:o)}(\gamma_0)$, hence it is not surprising that we only see the difference with a larger sample size. As an example, when $n = 500$, X_i is scalar, all variables are Gaussian, and $\beta > 1$, then the mse's of $(\hat{\theta}_n^{(tz:o)}(\hat{\gamma}_n), \hat{\theta}_n^{(tz:o)}(\gamma_0))$ are $(.0905, .0913)$, and when $n = 1000$ then the mse's are $(.0625, .0651)$. If all variables are Laplace, then the mse's are $(.0941, .0942)$ and $(.0647, .0663)$ respectively when n is 500 and 1000. See Tables H.1(c) and H.9(b) in Chaudhuri and Hill (2014, Part II).

4.3 Trim-by-Y

We now consider trimming by Y. We work in the benchmark setting of Section 4.1, and with $D = I(.25 + \beta X - U \geq 0)$ as in Section 4.2.1 to obtain an asymmetrically distributed Z. We want simply to focus on the pure effects of trimming on bias. The estimator is $\hat{\theta}_n^{(ty)} = 1/n \sum_{i=1}^n Z_i I(|Y_i| \leq Y_{(k_n^{(y)})}^{(a)})$. Under a suitable normalization, $\hat{\theta}_n^{(ty)}$ is asymptotically unbiased in its limit distribution by the benchmark design.[13] Crump, Hotz, Imbens, and Mitnik (2009), however, argue that removing units based on the outcome values Y can introduce bias. This will logically materialize in small samples here due to the presence of a few extreme values under the limited overlap case $\beta \geq 1$, even though asymptotically bias vanishes in the benchmark setting. Bias, however, occurs even asymptotically in the limit distribution when Z has an asymmetric distribution because trimming is symmetric.

First, Figure G.2 in Chaudhuri and Hill (2014, Part I: Appendix G) plots an estimate of $P(|Z_i| > c_z \mid |Y_i| > c_y)$ by using the methods presented there. It reveals essentially a perfect correspondence of extremes values of Y and Z in that simple setting when $\beta < 1$ ($E[Z^2] < \infty$). That correspondence, however, erodes monotonically in $\beta > 1$ ($E[Z^2] = \infty$). We therefore use the same thresholds for trimming Y as we do for Z: $k_n^{(y)} = k_n$, and expect $\hat{\theta}_n^{(ty)}$ to work well when $\beta < 1$. Tables 1-2 verify this intuition: compared to $\hat{\theta}_n^{(tz)}$ and $\hat{\theta}_n^{(tz:o)}$, $\hat{\theta}_n^{(ty)}$ has larger bias, it is farther from normally distributed, and exhibits larger empirical size distortions when $\beta \in \{1, 2\}$, with the worst performance at $\beta = 2$. If Z has asymmetric tails then $\hat{\theta}_n^{(ty)}$ logically is more biased, with higher dispersion, and is more deviated from normality.

5 Conclusion

Under assumptions of unconfoundedness and limited overlap, the ATE can be point identified as the mean of a random variable Z that depends on the realized outcome and the propensity score for each sample unit. Small and even large sample performance of robust IPW estimators of the ATE crucially depend on the number of extreme observations of Z that are trimmed. As a primary contribution we use information from Z itself to determine when to trim, and we correct for the resulting possible bias with a new estimator that does not impact asymptotics as to opposed to previous attempts in the literature. We allow for a plug-in estimator for the propensity score and show it also does not impact asymptotics when limited overlap is severe enough that Z has an infinite variance, and in all cases our trimmed estimator's mean-squared-error cannot be larger when the propensity score plug-in is used. We show in a controlled experiment that our estimator works exceptionally well when only a few observations are trimmed, while estimators that trim based on covariates, or the propensity score, require a far greater amount of trimming for comparable results. We explicitly ignore the topic of an optimal amount of trimming, aside from

[13] Let $c_n^{(y)}$ satisfy $P(|Y_i| \geq c_n^{(y)}) = k_n/n$. In the benchmark case $D_i Y_{1,i} + (1 - D_i) Y_{0,i}$ is symmetrically distributed about zero for any fixed value of D_i. Hence, by independence: $E[\{D_i Y_{1,i} + (1 - D_i) Y_{0,i}\} I(|D_i Y_{1,i} + (1 - D_i) Y_{0,i}| \leq c_n^{(y)}) | X_i, U_i] = 0$ $a.s.$, thus $E[Z_i I(|Y_i| \leq c_n^{(y)})] = 0 = \theta$. Since estimators with threshold $Y_{(k_n)}^{(a)}$ or $c_n^{(y)}$ are asymptotically equivalent in their limit distribution (see, e.g., Lemma A.4 in Appendix A), $\hat{\theta}_n^{(ty)}$ will be asymptotically unbiased in its limit distribution in this benchmark case.

showing that very little trimming works very well. A future topic of interest therefore concerns a data-adaptive technique for selecting the number of observations to trim in a way that leads to sharp inference in small samples.

A Appendix: Expansions

Define the moment supremum

$$\kappa \equiv \arg\sup \left\{ \alpha > 0 : E\left|Z_i\right|^{\alpha} < \infty \right\}.$$

In the infinite variance case $\kappa \leq 2$ this is identically the tail index in A3. Throughout we drop γ_0, e.g. $Z_i = Z_i(\gamma_0)$. Recall $p_i(\gamma) \equiv p(X_i, \gamma)$ hence $p_i = p_i(\gamma_0)$. Let $K > 0$ be a finite constant whose value may change from place to place. $\iota > 0$ is a tiny constant whose value may change.

We need to expand trimming indicators and order statistics in order to handle a plug-in estimator for γ_0 and for the ATE. Denote by θ_0 the true ATE and let θ be an arbitrary scalar, and assume without loss of generality

$$\theta_0 = 0.$$

Since there are two plug-ins $\hat{\gamma}_n$ and $1/n \sum_{i=1}^n Z_i(\gamma)$ it is helpful to write $Z_i(\gamma) - 1/n \sum_{j=1}^n Z_j(\gamma)$ compactly as a function of one vector parameter. Define

$$\xi \equiv \left[\gamma', \theta\right]' \text{ and } \hat{\xi}_n \equiv \left[\hat{\gamma}_n', \frac{1}{n}\sum_{i=1}^n Z_i(\hat{\gamma}_n)\right]', \; \mathcal{Z}_i(\xi) \equiv Z_i(\gamma) - \theta \text{ and } \mathcal{Z}_i \equiv Z_i - \theta_0 = Z_i, \tag{A.1}$$

and write

$$\mathcal{Z}_i^{(a)}(\xi) \equiv \left|\mathcal{Z}_i(\xi)\right|, \text{ and } \mathcal{Z}_{(1)}^{(a)}(\xi) \geq \mathcal{Z}_{(2)}^{(a)}(\xi) \geq \cdots \geq \mathcal{Z}_{(n)}^{(a)}(\xi).$$

Thus, $\mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)$ is simply the threshold $\hat{Z}_{n,(k_n)}^{(a)}(\hat{\gamma}_n)$ defined by (3) and (4).

The two dimensional plug-in estimator is $\hat{\xi}_n$. Let $\{c_n(\xi)\}_{n \geq 1}$ be a sequence of mappings $c_n : \Xi \to (0, \infty)$ that satisfy:

$$P\left(\left|\mathcal{Z}_i(\xi)\right| > c_n(\xi)\right) = k_n/n.$$

By construction and A3(ii) the threshold $c_n(\xi_0)$ satisfy:

$$c_n = c_n(\xi_0) = K(n/k_n)^{1/\kappa}. \tag{A.2}$$

Together $\theta_0 = 0$, the fact that Z_i is iid, and distribution tail property A3 yield

$$\frac{1}{n}\sum_{i=1}^n Z_i = O_p\left(\mathcal{L}_n / n^{1-1/\min\{\kappa,2\}}\right), \tag{A.3}$$

where $\mathcal{L}_n$ is slowly varying and $\kappa > 1$ is the A3 power law tail index. By case $\mathcal{L}_n = 1$ if $\kappa \neq 2$ and $\mathcal{L}_n = \ln(n)$ if $\kappa = 2$ (see Ibragimov and Linnik, 1971). Combine $\hat{\gamma}_n = \gamma_0 + O_p(1/n^{1/2})$ under B2, $n^{1-1/\min\{\kappa,2\}}\mathcal{L}_n/n^{1/2} = O(1)$ and (A.3) to deduce the plug-in estimator satisfies:

$$\hat{\xi}_n - \xi_0 = O_p\left(\mathcal{L}_n / n^{1-1/\min\{\kappa,2\}}\right). \tag{A.4}$$

Finally, recall that by the definition of a derivative, any differentiable $f : \mathbb{R}^k \to \mathbb{R}$ satisfies

$$f(x_1) - f(x_0) = \frac{\partial}{\partial x'} f(x_1) \times (x_1 - x_0) + o\left(\|x_1 - x_0\|\right), \tag{A.5}$$

where $o(||x_1 - x_0||) \to 0$ faster than $||x_1 - x_0|| \to 0$. We first characterize the thresholds used for trimming.

Lemma A.1 *Under Assumptions A3, B1, and B2:*

a. $\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)/c_n = \mathcal{Z}^{(a)}_{(k_n)}/c_n + o_p(1/k_n^{1/2})$ *and* $\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n) \neq \mathcal{Z}^{(a)}_{(k_n)}$ *a.s.*

b. $\mathcal{Z}^{(a)}_{(k_n)}/c_n = 1 + O_p(1/k_n^{1/2})$.

Proof.

Claim (a). The almost sure inequality follows from distribution continuity. We will show $\ln(\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)/c_n) = \ln(\mathcal{Z}^{(a)}_{(k_n)}/c_n) + o_p(1/k_n^{1/2})$. The claim then follows by the mean value theorem. Let *iff* = *if and only if*.

Define

$$\mathcal{I}_n(u, \xi) \equiv \frac{1}{k_n} \sum_{i=1}^{n} \left\{ I\left(|\mathcal{Z}_i(\xi)| > c_n(\xi) e^{u/k_n^{1/2}}\right) - P\left(|\mathcal{Z}_i(\xi)| > c_n(\xi) e^{u/k_n^{1/2}}\right) \right\}.$$

By construction $k_n^{1/2} \ln(\mathcal{Z}^{(a)}_{(k_n)}(\xi)/c_n(\xi)) \leq u$ *iff* $1/k_n \sum_{i=1}^{n} I(|\mathcal{Z}_i(\xi)| > c_n(\xi) e^{u/k_n^{1/2}}) \leq 1$ *iff*

$$\begin{aligned} k_n^{1/2} \mathcal{I}_n(u, \xi) &\leq k_n^{1/2} \left\{ 1 - \frac{P\left(|\mathcal{Z}_i(\xi)| > c_n(\xi) e^{u/k_n^{1/2}}\right)}{P\left(|\mathcal{Z}_i(\xi)| > c_n(\xi)\right)} \right\} \\ &= k_n^{1/2} \left(1 - \frac{n}{k_n} \left\{ 1 + F_{\mathcal{Z}_i(\xi)}\left(-c_n(\xi) e^{u/k_n^{1/2}}\right) - F_{\mathcal{Z}_i(\xi)}\left(c_n(\xi) e^{u/k_n^{1/2}}\right) \right\} \right). \end{aligned}$$

Under A3(iii.a) $\mathcal{Z}_i(\xi)$ has a continuous density function $f_{\mathcal{Z}(\xi)}$. Then by $P(|\mathcal{Z}_i(\xi)| > c_n(\xi)) = k_n/n$, the A3(iii.c) tail balance property (6), and the mean value theorem, there exists u_*, $|u_*| \leq |u$, such that

$$\begin{aligned} & k_n^{1/2} \left(1 - \frac{n}{k_n} \left\{ 1 + F_{\mathcal{Z}_i(\xi)}\left(-c_n(\xi) e^{u/k_n^{1/2}}\right) - F_{\mathcal{Z}_i(\xi)}\left(c_n(\xi) e^{u/k_n^{1/2}}\right) \right\} \right) \\ &= \frac{n}{k_n} c_n(\xi) \left\{ f_{\mathcal{Z}(\xi)}\left(-c_n(\xi) e^{u_*/k_n^{1/2}}\right) + f_{\mathcal{Z}(\xi)}\left(c_n(\xi) e^{u_*/k_n^{1/2}}\right) \right\} u = \mathcal{K}(\xi) u \left(1 + o(1)\right), \end{aligned} \tag{A.6}$$

where $o(1) \to 0$ as $n \to \infty$ does not depend on ξ, $\mathcal{K}(\xi)$ is continuous, $\inf_{\xi \in \Xi} \mathcal{K}(\xi) > 0$ and $\sup_{\xi \in \Xi} \mathcal{K}(\xi) < \infty$. Thus $k_n^{1/2} \ln(\mathcal{Z}^{(a)}_{(k_n)}(\xi)/c_n(\xi)) \leq u$ *iff* $k_n^{1/2} \mathcal{I}_n(u, \xi) \leq \mathcal{K}(\xi) u (1 + o(1))$. Now, $\mathcal{K}(\hat{\xi}_n) = \mathcal{K} + o_p(1)$ in view of $\hat{\xi}_n \xrightarrow{p} \xi$ and continuity. This yields by Cramer's theorem:

$$\lim_{n \to \infty} P\left(k_n^{1/2} \ln\left(\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)/c_n(\hat{\xi}_n)\right) \leq u\right) = \lim_{n \to \infty} P\left(\frac{k_n^{1/2}}{\mathcal{K}\left(1 + o(1)\right)} \mathcal{I}_n(u, \hat{\xi}_n) \leq u\right). \tag{A.7}$$

By the same argument

$$\lim_{n \to \infty} P\left(k_n^{1/2} \ln\left(\mathcal{Z}^{(a)}_{(k_n)}/c_n\right) \leq u\right) = \lim_{n \to \infty} P\left(\frac{k_n^{1/2}}{\mathcal{K}\left(1 + o(1)\right)} \mathcal{I}_n(u, \xi_0) \leq u\right). \tag{A.8}$$

Combine (A.7) with Lemma A.2.b, below, to deduce:

$$\lim_{n\to\infty} P\left(k_n^{1/2}\ln\left(\mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)/c_n(\hat{\xi}_n)\right)\leq u\right) = \lim_{n\to\infty} P\left(\frac{k_n^{1/2}}{\mathcal{K}\left(1+o(1)\right)}\mathcal{I}_n(u,\xi_0)\leq u\right).$$

Hence, for each $u\in\mathbb{R}$: $\lim_{n\to\infty}P(k_n^{1/2}\ln(\mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)/c_n(\hat{\xi}_n))\leq u) = \lim_{n\to\infty}P(k_n^{1/2}\ln(\mathcal{Z}_{(k_n)}^{(a)}/c_n)\leq u)$. Finally, $k_n^{1/2}\ln(c_n(\hat{\xi}_n)/c_n) = o_p(1)$ by Lemma A.2.a. The claim $\lim_{n\to\infty}P(k_n^{1/2}\ln(\mathcal{Z}_{(k_n)}^{(a)}/c_n)\leq u) = \lim_{n\to\infty}P(k_n^{1/2}\ln(\mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)/c_n)\leq u)$ now follows from by Cramer's theorem.

Claim (b). In view of (A.8) we need only show $E[(k_n^{1/2}\mathcal{I}_n(u,\xi_0))^2]=O(1)$. By independence and $c_n\to\infty$:

$$E\left[\left(k_n^{1/2}\mathcal{I}_n(u,\xi_0)\right)^2\right] = \frac{n}{k_n}P\left(|\mathcal{Z}_i|>c_ne^{u/k_n^{1/2}}\right)P\left(|\mathcal{Z}_i|\leq c_ne^{u/k_n^{1/2}}\right) = \frac{n}{k_n}P\left(|\mathcal{Z}_i|>c_ne^{u/k_n^{1/2}}\right)(1+o(1)).$$

The argument leading to (A.6) implies $(n/k_n)P(|\mathcal{Z}_i|>c_ne^{u/k_n^{1/2}}) = 1+O(1/k_n^{1/2})$. $\mathcal{QED}$.

Lemma A.2 *Let Assumptions A3, A4, and B1-B3 hold.*

a. For slowly varying functions $\mathcal{L}_n$ defined by (A.4) and $\mathring{\mathcal{L}}_n$ defined under A3(iii.b), $|c_n(\hat{\xi}_n)/c_n - 1| = O_p(\mathcal{L}_n\mathring{\mathcal{L}}_n/n^{1-1/\min\{\kappa,2\}}) = o_p(1/k_n^{1/2})$, and $|\mathcal{K}(\hat{\xi}_n)-\mathcal{K}| = O_p(\mathcal{L}_n/n^{1-1/\min\{\kappa,2\}}) = o_p(1/k_n^{1/2})$.

b. Define $\mathcal{I}_n(u,\xi)\equiv 1/k_n\sum_{i=1}^n\{I(|\mathcal{Z}_i(\xi)|>c_n(\xi)e^{u/k_n^{1/2}}) - P(|\mathcal{Z}_i(\xi)|>c_n(\xi)e^{u/k_n^{1/2}})\}$. Then $k_n^{1/2}\{\mathcal{I}_n(u,\hat{\xi}_n)$ $-\,\mathcal{I}_n(u,\xi_0)\} = o_p(1)$.

Proof.

Claim (a). By tail properties A3(iii.b,c), plug-in order (A.4) and derivative property (A.5) applied to $c_n(\hat{\xi}_n)$:

$$\begin{aligned}\left|\frac{c_n(\hat{\xi}_n)}{c_n}-1\right| &= \left\|\frac{1}{c_n}\frac{\partial}{\partial\xi}c_n(\xi_0)\right\|\times\left\|\hat{\xi}_n-\xi_0\right\| + o_p\left(\left\|\hat{\xi}_n-\xi_0\right\|\right) = O_p\left(\frac{\mathcal{L}_n\mathring{\mathcal{L}}_n}{n^{1-1/\min\{\kappa,2\}}}\right)\\ \left|\mathcal{K}(\hat{\xi}_n)-\mathcal{K}\right| &\leq \sup_{\xi\in\Sigma}\left\|\frac{\partial}{\partial\xi}\mathcal{K}(\xi)\right\|\times\left\|\hat{\xi}_n-\xi_0\right\| = O_p\left(\frac{\mathcal{L}_n}{n^{1-1/\min\{\kappa,2\}}}\right).\end{aligned} \tag{A.9}$$

Since under A3 and A4 $\{k_n,\mathcal{L}_n,\mathring{\mathcal{L}}_n\}$ are at most slowly varying functions, the proof is complete.

Claim (b). Since $\mathcal{I}_n(u,\xi)$ is not everywhere differentiable on Ξ, we treat this ordinary function as a *generalized function*, defined as a *regular sequence of good functions* in the sense of Lighthill (1958: Chapter 2, Def.'s 3, 5 and 7; se especially Chapter 2.3).[14,15]

Step 1 (generalized indicator function). We begin by treating $\mathcal{I}(w)\equiv I(w>0)$ as a generalized function. $\mathcal{I}(w)$ has a smooth regular sequences $\{\mathfrak{I}_{\mathcal{N}}(w)\}_{\mathcal{N}\geq 1}$ defined by

$$\mathfrak{I}_{\mathcal{N}}(w)\equiv\int_{-\infty}^{\infty}\mathcal{I}(v)\,\mathbb{S}\left(\mathcal{N}(v-w)\right)\mathcal{N}e^{-v^2/\mathcal{N}^2}dv, \tag{A.10}$$

[14]Similar usage of generalized functions can be found in Phillips (1995), Zinde-Walsh (2014) and Hill (2015).

[15]A *good function* is infinitely differentiable on $\mathbb{R}$, and it and all its derivatives are $O(|y|^{-\mathcal{N}})$ as $|y|\to\infty$ for any $\mathcal{N}>0$ (Lighthill, 1958, Def. 1). A sequence of good functions $\{f_{\mathcal{N}}(x)\}_{\mathcal{N}\in\mathbb{N}}$ is *regular* if $\lim_{\mathcal{N}\to\infty}\int_{-\infty}^{\infty}f_{\mathcal{N}}(x)F(x)dx$ exists for any good function $F(x)$ (Lighthill, 1958, Def. 3). Since good functions are integrable on $\mathbb{R}$, clearly $\{f_{\mathcal{N}}(x)+a\}_{\mathcal{N}\in\mathbb{N}}$ is regular if $\{f_{\mathcal{N}}(x)\}_{\mathcal{N}\in\mathbb{N}}$ is regular.

where $\mathbb{S}$ is a function that blots out $\mathcal{I}(v)$ when $v \notin [w - 1/\mathcal{N}, w + 1/\mathcal{N}]$. $\mathbb{S}(y)$ is assumed to be a *good function* (Lighthill, 1958, Def. 1 and p. 22), and as in Lighthill (1958, eq. (24)) and Phillips (1995, eq. (12)), we use:

$$\mathbb{S}(y) = e^{-1/(1-y^2)} \left(\int_{-1}^{1} e^{-1/(1-z^2)} dz \right)^{-1} I\left(|y| < 1\right). \tag{A.11}$$

Then $\int_{-1}^{1} \mathbb{S}(y)dy = 1$, and $\lim_{\mathcal{N}\to\infty} \int_{-\infty}^{\infty} \mathfrak{I}_{\mathcal{N}}(v)F(v)dv = \int_{-\infty}^{\infty} \mathcal{I}(v)F(v)dv$ for any good function F (Lighthill, 1958, Def. 7 and p. 22). Moreover, by Lemma A.3.a, below,

$$\left|\mathfrak{I}_{\mathcal{N}}(w) - \mathcal{I}(w)\right| \leq K \left|w\right|^{\iota} / \mathcal{N}^{\iota} + K/\mathcal{N} \text{ for any } \iota \in (0, 1). \tag{A.12}$$

The derivative $\mathfrak{D}_{\mathcal{N}}(w)$ of $\mathfrak{I}_{\mathcal{N}}(w)$ is a regular sequence for the Dirac delta function (Lighthill, 1958, p. 17):

$$\mathfrak{D}_{\mathcal{N}}(w) \equiv \left(\mathcal{N}/\pi\right)^{1/2} e^{-\mathcal{N}w^2}.$$

Step 2 (expansion of generalized $\mathcal{I}_n(u, \xi)$). Define

$$\zeta_{n,i}^{(0)}(\xi, u) \equiv \mathcal{Z}_i(\xi) + c_n(\xi)e^{u/k_n^{1/2}} \text{ and } \zeta_{n,i}^{(1)}(\xi, u) \equiv \mathcal{Z}_i(\xi) - c_n(\xi)e^{u/k_n^{1/2}},$$

hence

$$\mathcal{I}_n(u, \xi) = \frac{1}{k_n} \sum_{i=1}^{n} \left(\left\{ \mathcal{I}\left(\zeta_{n,i}^{(1)}(\xi, u)\right) + \mathcal{I}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right) \right\} - \left\{ P\left(\zeta_{n,i}^{(1)}(\xi, u) > 0\right) + P\left(-\zeta_{n,i}^{(0)}(\xi, u) > 0\right) \right\} \right).$$

Let $\{\mathcal{N}_n\}$ be an arbitrary sequence of positive integers, $\mathcal{N}_n \to \infty$ as $n \to \infty$. Since $I(|w| > c) = \mathcal{I}(w - c) + \mathcal{I}(-w - c)$, we treat $\mathcal{I}_n(u, \xi)$ as a generalized function with the regular sequence:

$$\mathbb{I}_{\mathcal{N}_n,n}(u, \xi) = \frac{1}{k_n} \sum_{i=1}^{n} \left(\left\{ \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi, u)\right) + \mathfrak{I}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right) \right\} - \left\{ E\left[\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi, u)\right)\right] + E\left[\mathfrak{I}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right)\right] \right\} \right).$$

We first prove $\sup_{\xi\in\Xi}\{k_n^{1/2}|\mathbb{I}_{\mathcal{N}_n,n}(u, \xi) - \mathcal{I}_n(u, \xi)|\} \xrightarrow{p} 0$. It then suffices to work with $\mathbb{I}_{\mathcal{N}_n,n}(u, \xi)$. By subadditivity, for any $\varepsilon > 0$:

$$\begin{aligned} P\left(\sup_{\xi\in\Xi} k_n^{1/2} \left|\mathbb{I}_{\mathcal{N}_n,n}(u, \xi) - \mathcal{I}_n(u, \xi)\right| > \varepsilon\right) \leq\ & P\left(\sup_{\xi\in\Xi} \left| \frac{1}{k_n^{1/2}} \sum_{i=1}^{n} \left\{ \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi, u)\right) - \mathcal{I}\left(\zeta_{n,i}^{(1)}(\xi, u)\right) \right\} \right| > \varepsilon/4\right) \\ & + P\left(\sup_{\xi\in\Xi} \left| \frac{1}{k_n^{1/2}} \sum_{i=1}^{n} \left\{ \mathfrak{I}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right) - \mathcal{I}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right) \right\} \right| > \varepsilon/4\right) \\ & + P\left(\sup_{\xi\in\Xi} \left| \frac{n}{k_n^{1/2}} E\left[\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi, u)\right) - \mathcal{I}\left(\zeta_{n,i}^{(1)}(\xi, u)\right)\right] \right| > \varepsilon/4\right) \\ & + P\left(\sup_{\xi\in\Xi} \left| \frac{n}{k_n^{1/2}} E\left[\mathfrak{I}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right) - \mathcal{I}\left(-\zeta_{n,i}^{(0)}(\xi, u)\right)\right] \right| > \varepsilon/4\right). \end{aligned}$$

We will prove the first probability on the right side of the inequality is $o(1)$, the remaining terms being similar. Use regular sequence property (A.12), $|x + y|^{\iota} \leq |x|^{\iota} + |y|^{\iota}$ for tiny $\iota > 0$ and $(x, y) \geq 0$, and the A3(iii.b) property

$\sup_{\xi\in\Xi}\{c_n(\xi)\} = O(n^{\varpi})$ for some $\varpi > 0$, to yield for any tiny $\iota > 0$:

$$P\left(\sup_{\xi\in\Xi}\left|\frac{1}{k_n^{1/2}}\sum_{i=1}^n \left\{\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi,u)\right) - \mathcal{I}\left(\zeta_{n,i}^{(1)}(\xi,u)\right)\right\}\right| > \varepsilon/4\right)$$
$$\leq P\left(K\frac{1}{k_n^{1/2}}\sum_{i=1}^n \frac{1}{\mathcal{N}_n^\iota}\sup_{\xi\in\Xi}\left|\zeta_{n,i}^{(1)}(\xi,u)\right|^\iota + \frac{n}{k_n^{1/2}}\frac{K}{\mathcal{N}_n^\iota} > \varepsilon/4\right)$$
$$\leq P\left(K\frac{1}{k_n^{1/2}}\sum_{i=1}^n \frac{1}{\mathcal{N}_n^\iota}\left\{\sup_{\xi\in\Xi}|\mathcal{Z}_i(\xi)|^\iota + Kn^{\iota\varpi}e^{\iota u/k_n^{1/2}}\right\} + \frac{n}{k_n^{1/2}}\frac{K}{\mathcal{N}_n^\iota} > \varepsilon/4\right).$$

Now invoke Markov's inequality, and $E[\sup_{\xi\in\Xi}|\mathcal{Z}_i(\xi)|^\iota] < \infty$ by A3(iii.a), to deduce:

$$P\left(\sup_{\xi\in\Xi}\left|\frac{1}{k_n^{1/2}}\sum_{i=1}^n \left\{\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi,u)\right) - \mathcal{I}\left(\zeta_{n,i}^{(1)}(\xi,u)\right)\right\}\right| > \varepsilon/4\right) \leq K\left(\frac{nn^{\iota\varpi}e^{\iota\iota u/k_n^{1/2}} + n}{k_n^{1/2}\mathcal{N}_n^\iota}\right) \leq K\frac{n^{1+\iota\varpi}e^{\iota\iota u/k_n^{1/2}}}{k_n^{1/2}\mathcal{N}_n^\iota}.$$

We can always pick $\{\mathcal{N}_n\}$ to satisfy $n^{1+\iota\varpi}k_n^{-1/2}/\mathcal{N}_n^\iota \to 0$, which proves the required limit.

Now expand $\mathbb{I}_{\mathcal{N}_n,n}(u,\hat{\xi}_n)$ around ξ_0. By the definition of a derivative:

$$\left|\mathfrak{I}_{\mathcal{N}_n}\left(\pm\zeta_{n,i}^{(\cdot)}(\hat{\xi}_n,u)\right) - \mathfrak{I}_{\mathcal{N}_n}\left(\pm\zeta_{n,i}^{(\cdot)}(\xi_0,u)\right)\right| \leq \mathfrak{D}_{\mathcal{N}_n}(\pm\zeta_{n,i}^{(\cdot)}(\xi_0,u))\left|\zeta_{n,i}^{(\cdot)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(\cdot)}(\xi_0,u)\right| \times (1+\mathcal{R}_{n,i}),$$

where $\mathcal{R}_{n,i} \xrightarrow{p} 0$ as $|\zeta_{n,i}^{(\cdot)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(\cdot)}(\xi_0,u)| \xrightarrow{p} 0$. Hence:

$$\begin{aligned}
&\left|k_n^{1/2}\left\{\mathbb{I}_{\mathcal{N}_n,n}(u,\hat{\xi}_n) - \mathbb{I}_{\mathcal{N}_n,n}(u,\xi_0)\right\}\right| \qquad (A.13)\\
&\leq \frac{1}{k_n}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right) \times k_n^{1/2}\left|\zeta_{n,i}^{(1)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(1)}(\xi_0,u)\right| \times (1+\mathcal{R}_{n,i})\\
&\quad + \frac{1}{k_n}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\xi_0,u)\right) \times k_n^{1/2}\left|\zeta_{n,i}^{(0)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(0)}(\xi_0,u)\right| \times (1+\mathcal{R}_{n,i})\\
&\quad + \left|\frac{n}{k_n^{1/2}}E\left[\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u)\right) - \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right)\right]\right| + \left|\frac{n}{k_n^{1/2}}E\left[\mathfrak{I}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\hat{\xi}_n,u)\right) - \mathfrak{I}_{\mathcal{N}_n}\left(-\zeta_{n,i}^{(0)}(\xi_0,u)\right)\right]\right|.
\end{aligned}$$

We will show the first and third terms are $o_p(1)$ and $o(1)$ respectively, the remaining terms being similar.

Step 2.1. Recall $\zeta_{n,i}^{(1)}(\xi,u) \equiv \mathcal{Z}_i(\xi) - c_n(\xi)e^{u/k_n^{1/2}}$ and $\mathcal{Z}_i(\xi) = Z_i(\gamma) - \theta$. By the triangular inequality:

$$\begin{aligned}
&\left|\frac{1}{k_n}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right) \times k_n^{1/2}\left(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(1)}(\xi_0,u)\right) \times (1+\mathcal{R}_{n,i})\right|\\
&\leq \left|\frac{1}{k_n^{1/2}}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right) \times \left(\mathcal{Z}_i(\hat{\xi}_n) - \mathcal{Z}_i(\xi_0)\right) \times (1+\mathcal{R}_{1,n,i})\right|\\
&\quad + \left|\frac{c_n}{k_n^{1/2}}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right)\right| \times \left|\frac{c_n(\hat{\xi}_n)}{c_n} - 1\right| e^{u/k_n^{1/2}} \times (1+\mathcal{R}_{2,n}) = \mathcal{C}_{1,n}(u) + \mathcal{C}_{2,n}(u), \qquad (A.14)
\end{aligned}$$

where $\mathcal{R}_{1,n,i} \xrightarrow{p} 0$ as $|\mathcal{Z}_i(\hat{\xi}_n) - \mathcal{Z}_i(\xi_0)| \xrightarrow{p} 0$ and $\mathcal{R}_{2,n} \xrightarrow{p} 0$ as $|c_n(\hat{\xi}_n)/c_n - 1| \xrightarrow{p} 0$.

Consider $\mathcal{C}_{1,n}(u)$ and write

$$\mathcal{A}_i \equiv \sup_{\gamma\in\Gamma}\left\{|h_i(\gamma)Z_i(\gamma)| \times \left\|\frac{\partial}{\partial\gamma}p_i(\gamma)\right\|\right\}.$$

Under B1 $\hat{\gamma}_n - \gamma_0 = O_p(1/n^{1/2})$. Hence, by a first order expansion of $Z_i(\hat{\gamma}_n)$ around γ_0, and the triangle inequality:

$$\begin{aligned}\mathcal{C}_{1,n}(u) \quad \leq \quad & \left|\frac{1}{k_n^{1/2}}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right) \times \mathcal{A}_i \times (1+\mathcal{R}_{3,n,i})\right| \times O_p\left(1/n^{1/2}\right) && \text{(A.15)}\\ & + \left|\frac{1}{k_n^{1/2}}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right)\right| \times \left|\frac{1}{n}\sum_{i=1}^n \mathcal{A}_i\right| \times (1+\mathcal{R}_{4,n}) \times O_p\left(1/n^{1/2}\right) && \text{(A.16)}\\ & + \left|\frac{1}{k_n^{1/2}}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right)\right| \times \left|\frac{1}{n}\sum_{i=1}^n Z_i\right| \times (1+\mathcal{R}_{5,n}), && \text{(A.17)}\end{aligned}$$

where $\mathcal{R}_{3,n,i} \xrightarrow{p} 0$ as $\mathcal{A}_i \times ||\hat{\gamma}_n - \gamma_0|| \xrightarrow{p} 0$, $\mathcal{R}_{4,n} \xrightarrow{p} 0$ as $1/n\sum_{i=1}^n \mathcal{A}_i \times ||\hat{\gamma}_n - \gamma_0|| \xrightarrow{p} 0$, and $\mathcal{R}_{5,n} \xrightarrow{p} 0$ as $|1/n\sum_{i=1}^n Z_i| \times ||\hat{\gamma}_n - \gamma_0|| \xrightarrow{p} 0$.

We will show each component is $o_p(1)$, hence $\mathcal{C}_{1,n}(u) = o_p(1)$. The expression in (A.15) is $o_p(1)$ by Lemma A.3.b, and the fact that $\mathcal{A}_i$ is L_p-bounded for some $p > 0$ by B3(i).

Next, by Lemma A.3.b $k_n^{-1/2}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}(\zeta_{n,i}^{(1)}(\xi_0,u)) = o_p(1/\mathcal{N}_n^{\iota})$ for tiny $\iota > 0$. Further, by B3(i) and Loève's inequality $E[(1/n\sum_{i=1}^n \mathcal{A}_i)^{\iota}] \leq Kn^{1-\iota}$ for tiny $\iota > 0$, hence

$$\frac{1}{n}\sum_{i=1}^n \mathcal{A}_i = \frac{1}{n}\sum_{i=1}^n \sup_{\gamma\in\Gamma}\left\{|h_i(\gamma)Z_i(\gamma)| \times \left\|\frac{\partial}{\partial\gamma}p_i(\gamma)\right\|\right\} = O_p(n^{1/\iota-1}). \qquad \text{(A.18)}$$

Thus, the expression in (A.16) is $o_p(n^{1/\iota-1-1/2}/\mathcal{N}_n^{\iota}) = o_p(1)$ for any $\{\mathcal{N}_n\}$, $\mathcal{N}_n/n^{1/\iota^2-3/(2\iota)} \to \infty$. The same argument extends to (A.17) since $1/n\sum_{i=1}^n Z_i = o_p(1)$ by (A.3).

Now consider $\mathcal{C}_{2,n}(u)$ in (A.14). First, $k_n^{-1/2}\sum_{i=1}^n \mathfrak{D}_{\mathcal{N}_n}(\zeta_{n,i}^{(1)}(\xi_0,u)) = o_p(1/\mathcal{N}_n^{\iota})$. Second, by Lemma A.2.a $|c_n(\hat{\xi}_n)/c_n - 1| = O_p(\mathcal{L}_n\mathring{\mathcal{L}}_n/n^{1-1/\min\{\kappa,2\}}) = o_p(1)$. Third, $c_n = O(n^{1/\kappa})$ by threshold relation (A.2). Therefore $\mathcal{C}_{2,n}(u) = o_p(n^{1/\kappa}/\mathcal{N}_n^{\iota}) = o_p(1)$ for any $\{\mathcal{N}_n\}$, $\mathcal{N}_n/n^{1/(\iota\kappa)} \to \infty$.

Step 2.2. Now turn to the third term in A.13. Use the definition of a derivative, and expand $\mathfrak{I}_{\mathcal{N}_n}(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u))$ around $\zeta_{n,i}^{(1)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(1)}(\xi_0,u)$ and $\mathfrak{I}_{\mathcal{N}_n}(\zeta_{n,i}^{(1)}(\xi_0,u))$ around $\zeta_{n,i}^{(1)}(\xi_0,u) - \zeta_{n,i}^{(1)}(\hat{\xi}_n,u)$ to yield both:

$$\begin{aligned}\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u)\right) &= \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(1)}(\xi_0,u)\right) - \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(1)}(\xi_0,u)\right)\zeta_{n,i}^{(1)}(\xi_0,u)\left(1+o_p\left(1\right)\right)\\ \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right) &= \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u) - \zeta_{n,i}^{(1)}(\hat{\xi}_n,u)\right) - \mathfrak{D}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u) - \zeta_{n,i}^{(1)}(\hat{\xi}_n,u)\right)\zeta_{n,i}^{(1)}(\hat{\xi}_n,u)\left(1+o_p\left(1\right)\right).\end{aligned}$$

Write $w_{n,i} \equiv \zeta_{n,i}^{(1)}(\hat{\xi}_n,u) - \zeta_{n,i}^{(1)}(\xi_0,u)$. Use $\mathfrak{D}_{\mathcal{N}_n}(-w) = \mathfrak{D}_{\mathcal{N}_n}(w)$, and the triangle inequality to deduce:

$$\frac{n}{k_n^{1/2}}E\left[\mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\hat{\xi}_n,u)\right) - \mathfrak{I}_{\mathcal{N}_n}\left(\zeta_{n,i}^{(1)}(\xi_0,u)\right)\right] = \frac{n}{k_n^{1/2}}E\left[\mathfrak{D}_{\mathcal{N}_n}\left(w_{n,i}\right)w_{n,i}\left(1+o_p\left(1\right)\right)\right].$$

Let $\delta(\cdot)$ be the delta Dirac function, hence $\int_{-\infty}^{\infty}\delta(w)F(w)dw = F(0)$ for any continuous function $F : \mathbb{R} \to \mathbb{R}$.

Moreover, by the Laplace approximation $\int_{-\infty}^{\infty} \mathfrak{D}_{\mathcal{N}}(w)F(w)dw = F(0) + O(1/\mathcal{N})$ (e.g. Phillips, 1995, p. 920). Hence, by dominated convergence $(n/k_n^{1/2})E[\mathfrak{D}_{\mathcal{N}_n}(w_{n,i})\, w_{n,i}(1 + o_p(1))] = O(nk_n^{-1/2}\mathcal{N}_n^{-1}) = o(1)$ for any choice of $\{\mathcal{N}_n\}$ such that $\mathcal{N}_n/(n/k_n^{1/2}) \rightarrow \infty$. $\mathcal{QED}$.

Lemma A.3 *Let Assumptions A3 hold. Define* $\mathcal{I}(w) \equiv I(w > 0)$, $\mathfrak{I}_{\mathcal{N}}(w) \equiv \int_{-\infty}^{\infty} \mathcal{I}(v)\, \mathbb{S}(\mathcal{N}(v - w))\mathcal{N}e^{-v^2/\mathcal{N}^2}dv$ *where* $\mathbb{S}$ *is the function (A.11), and* $\mathcal{N} > 0$. *Let* $\mathfrak{D}_{\mathcal{N}}(w) \equiv (\mathcal{N}/\pi)^{1/2}e^{-\mathcal{N}w^2}$.

a. $|\mathfrak{I}_{\mathcal{N}}(w) - \mathcal{I}(w)| \leq K|w|^{\iota}\mathcal{N}^{-\iota} + K/\mathcal{N}$ *for any* $\iota \in (0, 1)$.

b. Let ϖ_i *be an* L_p*-bounded random variable, and let* $u_0, u_1 \in \mathbb{R}$. *Then* $\sum_{i=1}^{n} \varpi_i \mathfrak{D}_{\mathcal{N}_n}(|Z_i| + u_o c_n e^{u_1/k_n^{1/2}}) = O_p(1/\mathcal{N}_n^{\iota})$ *for some sequence* $\{\mathcal{N}_n\}$.

Proof.

Claim (a). By construction of $\mathbb{S}(\cdot)$ and a change of variables:

$$\mathfrak{I}_{\mathcal{N}}(w) = \int_{w-1/\mathcal{N}}^{w+1/\mathcal{N}} \mathcal{I}(v)\, \mathbb{S}(\mathcal{N}(v - w))\, \mathcal{N}e^{-v^2/\mathcal{N}^2}dv = \int_{-1}^{1} \mathcal{I}(w + u/\mathcal{N})\, \mathbb{S}(u)\, e^{-(w+u/\mathcal{N})^2/\mathcal{N}^2}du.$$

Apply the Laplace approximation to the final integral to deduce $\mathfrak{I}_{\mathcal{N}}(w) = \mathcal{I}(w)\, e^{-w^2/\mathcal{N}^2} + O(1/\mathcal{N})$. See also Phillips (1995, eq. (24)). Now expand $e^{-w^2/\mathcal{N}^2}$ around $1/\mathcal{N}^2 = 0$: use derivative property (A.5) to yield $e^{-w^2/\mathcal{N}^2} - 1 = -e^{-w^2/\mathcal{N}^2}w^2/\mathcal{N}^2 + o(1/\mathcal{N}^2)$. Further, $e^{-w^2/\mathcal{N}^2}w^2\mathcal{N}^{-2} \leq |w/\mathcal{N}|^{\iota}$ for any $\iota \in (0, 1)$.[16] Therefore $|\mathcal{I}(w)\, e^{-w^2/\mathcal{N}^2} - \mathcal{I}(w)| \leq K|w|^{\iota}\mathcal{N}^{-\iota} + o(1/\mathcal{N}^2)$. Combining results, we have shown $|\mathfrak{I}_{\mathcal{N}}(w) - \mathcal{I}(w)| \leq K|w|^{\iota}\mathcal{N}^{-\iota} + o(1/\mathcal{N}^2) + O(1/\mathcal{N}) \leq K|w|^{\iota}\mathcal{N}^{-\iota} + K/\mathcal{N}$ for any $\iota \in (0, 1)$ as claimed.

Claim (b). Define $\zeta_{n,i}(u) \equiv |Z_i| + u_o c_n e^{u_1/k_n^{1/2}}$. Assume $u_o = 1$, the general result having a nearly identical proof. Recall $\mathfrak{D}_{\mathcal{N}}(w) \equiv (\mathcal{N}/\pi)^{1/2}e^{-\mathcal{N}w^2}$. By supposition ϖ_i is L_p-bounded for some $p > 0$. We may therefore apply Loève and Cauchy-Schwartz inequalities to yield for any tiny $r \in (0, p/2]$:

$$E\left|\sum_{i=1}^{n} \varpi_i \mathfrak{D}_{\mathcal{N}_n}(\zeta_{n,i}(u))\right|^r \leq n\mathcal{N}_n^{r/2}E\left|\frac{\varpi_i}{\exp\left\{\mathcal{N}_n^2\zeta_{n,i}^2(u)\right\}}\right|^r \leq K\left(n^2\mathcal{N}_n^r E\left[\frac{1}{\exp\left\{2r\mathcal{N}_n\zeta_{n,i}^2(u)\right\}}\right]\right)^{1/2}.$$

Boundedness of $\exp\{-|x|\}$ and the Cauchy-Schwartz inequality imply:

$$\begin{aligned} E\left[\frac{1}{\exp\left\{2r\mathcal{N}_n\zeta_{n,i}^2(u)\right\}}\right] &= E\left[\frac{1}{\exp\left\{2r\mathcal{N}_n\zeta_{n,i}^2(u)\right\}}I\left(|\zeta_{n,i}(u)| > \frac{1}{\mathcal{N}_n^{1/4}}\right)\right] + E\left[\frac{1}{\exp\left\{2r\mathcal{N}_n\zeta_{n,i}^2(u)\right\}}I\left(|\zeta_{n,i}(u)| \leq \frac{1}{\mathcal{N}_n^{1/4}}\right)\right] \\ &\leq \frac{1}{\exp\left\{2r\mathcal{N}_n^{1/2}\right\}} + KP\left(|\zeta_{n,i}(u)| \leq \frac{1}{\mathcal{N}_n^{1/4}}\right)^{1/2}. \end{aligned}$$

The A3 distribution properties imply Z_i has a density function f_Z that satisfies $f_Z(x) \rightarrow 0$ as $|x| \rightarrow \infty$. By a first order expansion it therefore follows that there exists an $a_* \in [-1, 1]$ such that:

$$P\left(|\zeta_{n,i}(u)| \leq \frac{1}{\mathcal{N}_n^{1/4}}\right) = \left|P\left(Z_i \geq c_n e^{u_1/k_n^{1/2}} - \frac{1}{\mathcal{N}_n^{1/4}}\right) - P\left(Z_i \geq c_n e^{u_1/k_n^{1/2}} + \frac{1}{\mathcal{N}_n^{1/4}}\right)\right|$$

[16]Note $\ln(e^{-w^2/\mathcal{N}^2}w^2\mathcal{N}^{-2}/|w\mathcal{N}^{-1}|^{\iota}) = -w^2/\mathcal{N}^2 + (1 - \iota/2)\ln(w^2/\mathcal{N}^2)$. If the latter term is negative for $\iota \in (0, 1)$ then $e^{-w^2/\mathcal{N}^2}w^2\mathcal{N}^{-2} \leq |w/\mathcal{N}|^{\iota}$. The maximum of $-x + y\ln(x)$ with respect to x is achieved at $x = y$, while $-y + y\ln(y) \leq 0$ for $y \leq e$. Finally, $y = 1 - \iota/2 \leq 1$ for all $\iota \in (0, 1)$.

$$\leq \quad K\frac{1}{\mathcal{N}_n^{1/4}}f\left(Z_i \geq c_n e^{u_1/k_n^{1/2}} - a_*\frac{1}{\mathcal{N}_n^{1/4}}\right) = o\left(\frac{1}{\mathcal{N}_n^{1/4}}\right).$$

Therefore $E[\exp\{-2r\mathcal{N}_n\zeta_{n,i}^2(u)\}] = o(\mathcal{N}_n^{-1/4})$, which implies $E|\sum_{i=1}^n \varpi_i \mathfrak{D}_{\mathcal{N}_n}(\zeta_{n,i}(u))|^r = o(n/\mathcal{N}_n^{(1/8-r/2)})$. Since r is tiny, $n/\mathcal{N}_n^{(1/8-r/2)} = O(1/\mathcal{N}_n^{r\iota})$ for tiny $\iota > 0$ and an appropriate choice of $\{\mathcal{N}_n\}$. Therefore $\sum_{i=1}^n \varpi_i \mathfrak{D}_{\mathcal{N}_n}(\zeta_{n,i}(u))$ $= O_p(1/\mathcal{N}_n^{\iota})$ by Markov's inequality. $\mathcal{QED}$.

Lemma A.4 *Recall $\theta_0 = 0$, $\xi \equiv [\gamma', \theta]'$ and $\mathcal{Z}_i(\xi) \equiv Z_i(\gamma) - \theta$. Let Assumptions A3, and B1-B3 hold.*

a. For any L_p-bounded ζ_i, $p > 0$: $\sigma_n^{-1}n^{-1/2}\sum_{i=1}^n |\zeta_i| \times |I(|\mathcal{Z}_i(\hat{\xi}_n)| < \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)) - I(|Z_i| < c_n)| = o_p(1)$.

b. $\sigma_n^{-1}n^{-1/2}\sum_{i=1}^n Z_i(\hat{\gamma}_n)\{I(|\mathcal{Z}_i(\hat{\xi}_n)| < \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)) - I(|Z_i| < c_n)\} = o_p(1)$.

Proof.

Claim (a). Define $\mathcal{A}_n \equiv 1/n\sum_{i=1}^n |\zeta_i| \times |I(|\mathcal{Z}_i(\hat{\xi}_n)| < \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)) - I(|Z_i| < c_n)|$. We use the generalized function notation in the proof of Lemma A.2.b. Define $\mathcal{I}(w) \equiv I(w < 0)$. The regular sequence for $\mathcal{A}_n$ is

$$\mathcal{A}_{\mathcal{N}_n,n} = \frac{1}{n}\sum_{i=1}^n |\zeta_i| \times \left|\mathfrak{I}_{\mathcal{N}_n}\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right) - \mathfrak{I}_{\mathcal{N}_n}\left(|Z_i| - c_n\right)\right|$$

where $\{\mathcal{N}_n\}$ is a sequence of positive finite integers, $\mathcal{N}_n \rightarrow \infty$ as $n \rightarrow \infty$.

Step 1. We first prove $(n^{1/2}/\sigma_n)|\mathcal{A}_{\mathcal{N}_n,n} - \mathcal{A}_n| \xrightarrow{p} 0$, hence we can work with $\mathcal{A}_{\mathcal{N}_n,n}$. Observe:

$$\begin{aligned}\frac{n^{1/2}}{\sigma_n}|\mathcal{A}_{\mathcal{N}_n,n} - \mathcal{A}_n| \quad &\leq \quad \frac{1}{\sigma_n n^{1/2}}\sum_{i=1}^n \left|\zeta_i\left\{\mathfrak{I}_{\mathcal{N}_n}\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right) - \mathcal{I}\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right)\right\}\right| \\ &+\frac{1}{\sigma_n n^{1/2}}\sum_{i=1}^n |\zeta_i\left\{\mathfrak{I}_{\mathcal{N}_n}\left(|Z_i| - c_n\right) - \mathcal{I}\left(|Z_i| - c_n\right)\right\}| = \mathfrak{B}_{1,\mathcal{N}_n} + \mathfrak{B}_{2,\mathcal{N}_n}.\end{aligned}$$

Use Lemma A.3.a, and $||x| - |y|| \leq |x - y|$, to deduce for tiny $\iota > 0$:

$$\begin{aligned}\mathfrak{B}_{1,\mathcal{N}_n} \quad &= \quad \frac{1}{\sigma_n n^{1/2}}\sum_{i=1}^n \left|\zeta_i\left\{\mathfrak{I}_{\mathcal{N}_n}\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right) - \mathcal{I}\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right)\right\}\right| \\ &\leq \quad \frac{1}{\sigma_n n^{1/2}\mathcal{N}_n^{\iota}}\sum_{i=1}^n |\zeta_i| \times \left\{\left|\mathcal{Z}_i(\hat{\xi}_n) - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right|^{\iota}\right\} + K/\mathcal{N}_n.\end{aligned}$$

Observe by Minkowski's inequality:

$$\begin{aligned}\left(\sum_{i=1}^n |\zeta_i| \times \left|\mathcal{Z}_i(\hat{\xi}_n) - \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right|^{\iota}(1 + o_p(1))\right)^{1/\iota} \quad &\leq \quad \left(\sum_{i=1}^n |\zeta_i| \times |Z_i(\hat{\gamma}_n) - Z_i|^{\iota}\right)^{1/\iota} + \left(\sum_{i=1}^n |\zeta_i| \times |Z_i|^{\iota}\right)^{1/\iota} \\ &+ \left(\sum_{i=1}^n |\zeta_i|\right)^{1/\iota}\left\{\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| + \mathcal{Z}_{(k_n)}^{(a)}(\hat{\xi}_n)\right\}.\end{aligned}$$

By supposition $|\zeta_i| \times |Z_i|^{\iota}$ is L_p-bounded for tiny $p > 0$. Now apply Loève's inequality: $E[(\sum_{i=1}^n |\zeta_i| \times |Z_i|^{\iota})^p] \leq n$ and $E[(\sum_{i=1}^n |\zeta_i|)^p] \leq n$, hence $\sum_{i=1}^n |\zeta_i| \times |Z_i|^{\iota}$ and $\sum_{i=1}^n |\zeta_i|$ are $O_p(n^{1/p})$ by Markov's inequality. Further,

$1/n\sum_{i=1}^{n} Z_i = O_p(\mathcal{L}_n/n^{1-1/\min\{\kappa,2\}})$ by (A.3) for slowly varying $\mathcal{L}_n$, $\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n) = c_n(1 + O_p(1/k_n^{1/2}))$ by Lemma A.1, and $c_n = K(n/k_n)^{1/\kappa}$ by (A.2). Moreover, by a first order expansion around γ_0:

$$\sum_{i=1}^{n} |\zeta_i| \times |Z_i(\hat{\gamma}_n) - Z_i|^{\iota} \leq \sum_{i=1}^{n} |\zeta_i| \times \left|\sup_{\gamma\in\Gamma}\left\{|h_i(\gamma)Z_i(\gamma)| \times \left\|\frac{\partial}{\partial\gamma}p_i(\gamma)\right\|\right\}\right|^{\iota} \times \|\hat{\gamma}_n - \gamma_0\|^{\iota}.$$

Estimator property B2 implies $||\hat{\gamma}_n - \gamma_0||^{\iota} = O_p(1/n^{\iota/2})$, and B3(i) states $\sup_{\gamma\in\Gamma}\{|h_i(\gamma)Z_i(\gamma)| \times ||(\partial/\partial\gamma)p_i(\gamma)||\}$ is L_p-bounded for tiny $p > 0$. Apply Loève and Markov inequalities again to yield $\sum_{i=1}^{n} |\zeta_i| \times |Z_i(\hat{\gamma}_n) - Z_i|^{\iota} = O_p(n^{1/p-\iota/2})$. This proves

$$\sum_{i=1}^{n} |\zeta_i| \times \left|\mathcal{Z}_i(\hat{\xi}_n) - \mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)\right|^{\iota} = \left\{O_p\left(n^{1/(\iota p)-\iota}\right) + O_p\left(n^{1/(\iota p)}\right) + O_p\left(n^{1/(\iota p)}(n/k_n)^{1/\kappa}\right)\right\}^{\iota} = O_p\left(n^{1/p+\iota/\kappa}\right).$$

Now use $\liminf_{n\to\infty}\sigma_n > 0$ to deduce there exists some sequence $\{\mathcal{N}_n\}$, $\mathcal{N}_n/n^{(1/p-1/2+\iota/\kappa)/\iota} \to \infty$, such that: $\mathfrak{B}_{1,\mathcal{N}_n} = O_p(n^{1/p-1/2+\iota/\kappa}/\mathcal{N}_n^{\iota} + K/\mathcal{N}_n) = o_p(1)$. A similar argument can be applied to $\mathfrak{B}_{2,\mathcal{N}_n}$.

Step 2. It remains to show $(n^{1/2}/\sigma_n)\mathcal{A}_{\mathcal{N}_n,n} \xrightarrow{p} 0$. Observe $\mathcal{Z}_i(\hat{\xi}_n) = Z_i(\hat{\gamma}_n) - 1/n\sum_{i=1}^{n} Z_i(\hat{\gamma}_n)$, and $||x| - |y|| \leq |x - y|$. By the definition of a derivative, and triangle and Cauchy-Schwartz inequalities:

$$\begin{aligned}\left|\frac{n^{1/2}}{\sigma_n}\mathcal{A}_{\mathcal{N}_n,n}\right| \leq\ & \frac{1}{n^{1/2}\sigma_n}\sum_{i=1}^{n}|\zeta_i|\,\mathfrak{D}_{\mathcal{N}_n}\left(|Z_i| - c_n\right)|Z_i(\hat{\gamma}_n) - Z_i| \times (1 + \mathcal{R}_{1,n,i}) \\ & + \frac{1}{n^{1/2}\sigma_n}\sum_{i=1}^{n}|\zeta_i|\,\mathfrak{D}_{\mathcal{N}_n}\left(|Z_i| - c_n\right) \times \left|\frac{1}{n}\sum_{i=1}^{n}\{Z_i(\hat{\gamma}_n) - Z_i\}\right| \times (1 + \mathcal{R}_{2,n}) \\ & + \frac{1}{n^{1/2}\sigma_n}\sum_{i=1}^{n}|\zeta_i|\,\mathfrak{D}_{\mathcal{N}_n}\left(|Z_i| - c_n\right) \times \left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \times (1 + \mathcal{R}_{3,n}) \\ & + \frac{c_n}{n^{1/2}\sigma_n}\sum_{i=1}^{n}|\zeta_i|\,\mathfrak{D}_{\mathcal{N}_n}\left(|Z_i| - c_n\right) \times \left|\frac{\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)}{c_n} - 1\right| \times (1 + \mathcal{R}_{4,n}),\end{aligned}$$

where $\mathcal{R}_{1,n,i} \xrightarrow{p} 0$ as $|Z_i(\hat{\gamma}_n) - Z_i| \xrightarrow{p} 0$, $\mathcal{R}_{2,n} \xrightarrow{p} 0$ as $|1/n\sum_{i=1}^{n}\{Z_i(\hat{\gamma}_n) - Z_i\}| \xrightarrow{p} 0$, $\mathcal{R}_{3,n} \xrightarrow{p} 0$ as $|1/n\sum_{i=1}^{n} Z_i| \xrightarrow{p} 0$, and $\mathcal{R}_{4,n} \xrightarrow{p} 0$ as $|\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)/c_n - 1 \xrightarrow{p} 0$.

Define $\mathcal{A}_i \equiv \sup_{\gamma\in\Gamma}\{|h_i(\gamma)Z_i(\gamma)| \times ||(\partial/\partial\gamma)p_i(\gamma)||\}$. A first order expansion leads to $|Z_i(\hat{\gamma}_n) - Z_i| \leq \mathcal{A}_i \times ||\hat{\gamma}_n - \gamma_0||$, where $\mathcal{A}_i$ is L_p-bounded under B3(i) and $||\hat{\gamma}_n - \gamma_0|| = O_p(1/n^{1/2})$ by B2. Therefore each summand with $\mathfrak{D}_{\mathcal{N}_n}(|Z_i| - c_n)$ is $O_p(1/\mathcal{N}_n^{\iota})$ for small $\iota > 0$ by Lemma A.3.b. Further, $1/n\sum_{i=1}^{n} Z_i = o_p(1)$ by (A.3), and $|\mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)/c_n - 1| = O_p(1/k_n^{1/2})$ by Lemma A.1 and $c_n = O_p(n^{1/\kappa})$ from (A.2). Finally, $1/n\sum_{i=1}^{n}\mathcal{A}_i = O_p(n^{1/\iota-1})$ from (A.18). It follows that the first four terms are $o_p(1)$ for some choice of $\{\mathcal{N}_n\}$.

Claim (*b*). Write:

$$\begin{aligned}&\frac{1}{\sigma_n n^{1/2}}\sum_{i=1}^{n} Z_i(\hat{\gamma}_n)\left\{I\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| < \mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)\right) - I\left(|Z_i| < c_n\right)\right\} \\ &\quad = \frac{1}{\sigma_n n^{1/2}}\sum_{i=1}^{n}\{Z_i(\hat{\gamma}_n) - Z_i\}\left\{I\left(\left|\mathcal{Z}_i(\hat{\xi}_n)\right| < \mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n)\right) - I\left(|Z_i| < c_n\right)\right\}\end{aligned}$$

$$+ \frac{1}{\sigma_n n^{1/2}} \sum_{i=1}^{n} Z_i \left\{ I\left(\left| \mathcal{Z}_i(\hat{\xi}_n) \right| < \mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n) \right) - I\left(|Z_i| < c_n \right) \right\}.$$

The second term is $o_p(1)$ by claim (a). The first term is not larger than:

$$K \frac{1}{\sigma_n n^{1/2}} \sum_{i=1}^{n} \sup_{\gamma \in \Gamma} \left\{ |h_i(\gamma) Z_i(\gamma)| \times \left\| \frac{\partial}{\partial \gamma} p_i(\gamma) \right\| \right\} \left| I\left(\left| \mathcal{Z}_i(\hat{\xi}_n) \right| < \mathcal{Z}^{(a)}_{(k_n)}(\hat{\xi}_n) \right) - I\left(|Z_i| < c_n \right) \right| \times \|\hat{\gamma}_n - \gamma_0\|.$$

Since $\sup_{\gamma \in \Gamma}\{|h_i(\gamma)Z_i(\gamma)| \times ||(\partial/\partial\gamma)p_i(\gamma)||\}$ is L_p-bounded, the first term is $o_p(1)$ by claim (a). $\mathcal{QED}$.

Lemma A.5 *Under A3, B1, B2* $1/n \sum_{i=1}^{n} (\partial/\partial\gamma) Z_i(\hat{\gamma}_n) I(|Z_i| < c_n) = E[(\partial/\partial\gamma) Z_i I(|Z_i| < c_n)](1 + o_p(1))$.

Proof. By construction

$$\frac{\partial}{\partial \gamma} Z_i(\gamma) = \left(\frac{D_i(\gamma) - p_i(\gamma)}{p_i(\gamma)\left(1 - p_i(\gamma)\right)} \right)^2 \frac{\partial}{\partial \gamma} p_i(\gamma) Y_i(\gamma) = -h_i(\gamma) \frac{\partial}{\partial \gamma} p_i(\gamma) Z_i(\gamma) = -S_i(\gamma) Z_i(\gamma),$$

say. Define $a_{n,i}(\gamma) \equiv S_i(\gamma) Z_i(\gamma) I(|Z_i| < c_n)$. It suffices to prove $\sup_{\gamma \in \Gamma} |1/n \sum_{i=1}^{n} a_{n,i}(\gamma) - E[a_{n,i}(\gamma)](1 + o_p(1))| \xrightarrow{p} 0$ where $o_p(1)$ may be a function of Γ, and $E[a_{n,i}(\hat{\gamma}_n)] = E[a_{n,i}](1 + o(1))$. The latter follows from continuity of $E[a_{n,i}(\gamma)]$ on Γ, and $\hat{\gamma}_n \xrightarrow{p} \gamma_0$ under B2. Now turn to the required ULLN.

Step 1 (pointwise LLN). If $a_{n,j}(\gamma)$ is uniformly integrable then $1/n \sum_{j=1}^{n} a_{n,j}(\gamma) - E[a_{n,j}(\gamma)] \xrightarrow{p} 0$ by Theorem 2 in Andrews (1988). Otherwise, assume without loss of generality that $\liminf_{n \to \infty} |E[a_{n,j}(\gamma)]| > 0$. Then $z_{n,j}(\gamma) \equiv a_{n,j}(\gamma)/E[a_{n,j}(\gamma)] - 1$ is integrable, independent, and identically distributed over $1 \leq j \leq n$. Let $i \equiv -1$. The characteristic function of $1/n \sum_{j=1}^{n} z_{n,j}(\gamma)$ is $E[\exp\{i\lambda n^{-1} \sum_{i=1}^{n} z_{n,i}(\gamma)\}] = (E[\exp\{i\lambda z_{n,j}(\gamma)/n\}])^n$. Since $E[z_{n,j}(\gamma)] = 0$ it follows that $(\partial/\partial\lambda) E[\exp\{i\lambda z_{n,j}(\gamma)/n\}]|_{\lambda=0} = 0$. Therefore $E[\exp\{i\lambda n^{-1} \sum_{i=1}^{n} z_{n,i}(\gamma)\}]$ $= (1 + 0 + o(1/n))^n \to 1$ as $n \to \infty$, hence $n^{-1} \sum_{i=1}^{n} z_{n,i}(\gamma) \xrightarrow{d} 0$, which implies $n^{-1} \sum_{i=1}^{n} z_{n,i}(\gamma) \xrightarrow{p} 0$, Therefore $|1/n \sum_{j=1}^{n} a_{n,j}(\gamma) - E[a_{n,j}(\gamma)](1 + o_p(1))| \xrightarrow{p} 0$.

Step 2 (ULLN). We first need two preliminary ULLN's. $\mu^*_{n,i}(\gamma) \equiv |z_{n,i}(\gamma)|/\sup_{\gamma \in \Gamma}\{E|z_{n,i}(\gamma)|\}$ is uniformly L_1-bounded on compact Γ, hence it belongs to a separable Banach space. This implies the L_1-bracketing numbers satisfy $N_{[\,]}(\varepsilon, \Gamma, ||\cdot||_1) < \infty$ (see Proposition 7.1.7 in Dudley, 1999). By the Step 1 LLN, $1/n \sum_{i=1}^{n}(\mu^*_{n,i}(\gamma)$ $- E[\mu^*_{n,i}(\gamma)]) \xrightarrow{p} 0$. Hence the first ULLN $\sup_{\gamma \in \Gamma} |1/n \sum_{i=1}^{n}\{\mu^*_{n,i}(\gamma) - E[\mu^*_{n,i}(\gamma)]\}| \xrightarrow{p} 0$ follows from Theorem 7.1.5 of Dudley (1999). Now replace $z^*_{n,i}(\gamma)$ with $g^*_{n,i}(\gamma) \equiv |z^*_{n,i}(\gamma)|/E|z^*_{n,i}(\gamma)|$ and invoke the first ULLN to obtain the second ULLN: $\sup_{\gamma \in \Gamma} 1/n \sum_{i=1}^{n}\{g^*_{n,i}(\gamma) - E[g^*_{n,i}(\gamma)]\}| = o_p(\sup_{\gamma \in \Gamma} |E[g^*_{n,i}(\gamma)]|) \xrightarrow{p} 0$. Finally, for any $\delta > 0$ define

$$r_n(\gamma, \delta) \equiv \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{z^*_{n,i}(\gamma) - E\left[z^*_{n,i}(\gamma)\right]}{\left|E\left[z^*_{n,i}(\gamma)\right]\right| + \delta} \right\} \left(\frac{\left|E\left[z^*_{n,i}(\gamma)\right]\right| + \delta - 1}{\left|E\left[z^*_{n,i}(\gamma)\right]\right| + \delta} \right)$$

By a generalization of the second ULLN $\sup_{\gamma \in \Gamma} |r_n(\gamma, \delta)| = o_p(1)$. Hence, by construction:

$$\sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ z^*_{n,i}(\gamma) - E\left[z^*_{n,i}(\gamma)\right] - r_n(\gamma, \delta) \times \left(\left|E\left[z^*_{n,i}(\gamma)\right]\right| + \delta \right) \right\} \right| = \sup_{\gamma \in \Gamma} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{z^*_{n,i}(\gamma) - E\left[z^*_{n,i}(\gamma)\right]}{\left|E\left[z^*_{n,i}(\gamma)\right]\right| + \delta} \right\} \right| \xrightarrow{p} 0.$$

Now use $\sup_{\gamma\in\Gamma}|r_n(\gamma,\delta)| = o_p(1)$ to yield

$$\sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{z_{n,i}^{*}(\gamma) - E\left[z_{n,i}^{*}(\gamma)\right] - r_n(\gamma,\delta)\times\left(\left|E\left[z_{n,i}^{*}(\gamma)\right]\right| + \delta\right)\right\}\right|$$
$$= \sup_{\gamma\in\Gamma}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{z_{n,i}^{*}(\gamma) - E\left[z_{n,i}^{*}(\gamma)\right](1+o_p(1))\right\} - o_p(1)\right| \xrightarrow{p} 0$$

where each $o_p(1)$ depends on Γ. Hence $\sup_{\gamma\in\Gamma}|1/n\sum_{i=1}^{n}\{z_{n,i}^{*}(\gamma) - E[z_{n,i}^{*}(\gamma)](1+o_p(1))\}| \xrightarrow{p} 0$. $\mathcal{QED}$.

B Appendix: Proofs of Main Results

Recall $\theta = 0$, and:

$$Z_i^{(a)}(\gamma) \equiv |Z_i(\gamma)|, \text{ and } Z_{(1)}^{(a)}(\gamma) \geq Z_{(2)}^{(a)}(\gamma) \geq \cdots \geq Z_{(n)}^{(a)}(\gamma)$$
$$\hat{Z}_{n,i}(\gamma) \equiv Z_i(\gamma) - \frac{1}{n}\sum_{j=1}^{n} Z_j(\gamma),\ \hat{Z}_{n,i}^{(a)}(\gamma) \equiv \left|\hat{Z}_{n,i}(\gamma)\right|, \text{ and } \hat{Z}_{n,(1)}^{(a)}(\gamma) \geq \hat{Z}_{n,(2)}^{(a)}(\gamma) \geq \cdots \geq \hat{Z}_{n,(n)}^{(a)}(\gamma).$$

Proof of Theorem 3.1.

Claim (a) Recall $\theta_0 = 0$. By B2 $w_i \in \mathbb{R}^q$ is the zero mean, finite variance iid variable that satisfies $\sqrt{n}(\hat{\gamma}_n - \gamma_0)$ $= 1/\sqrt{n}\sum_{i=1}^{n} w_i(1+o_p(1))$. We use the following definitions from Section 3:

$$\mathcal{D}_n \equiv -E\left[\frac{\partial}{\partial\gamma}p_i h_i Z_i I\left(|Z_i| < c_n\right)\right] \text{ and } \mathcal{B}_n \equiv E\left[Z_i I\left(|Z_i| \geq c_n\right)\right]$$
$$\sigma_n^2 \equiv E\left[\left\{Z_i I\left(|Z_i| < c_n\right) - E\left[Z_i I\left(|Z_i| < c_n\right)\right]\right\}^2\right]$$
$$\vartheta_{n,i} \equiv Z_i I\left(|Z_i| < c_n\right) - E\left[Z_i I\left(|Z_i| < c_n\right)\right] + \mathcal{D}_n' w_i$$
$$\mathcal{V}_n^2 \equiv E\left[\vartheta_{n,i}^2\right] = \sigma_n^2 + 2E\left[\left\{Z_i I\left(|Z_i| < c_n\right) - E\left[Z_i I\left(|Z_i| < c_n\right)\right]\right\} w_i'\right]\mathcal{D}_n + \mathcal{D}_n' E\left[w_i w_i'\right]\mathcal{D}_n.$$

Apply Lemma A.4 with $\mathcal{V}_n \sim K\sigma_n$ by (b), and use $\mathcal{B}_n = E[Z_i(I|Z_i| \geq c_n)] = -E[Z_i(I|Z_i| < c_n)]$ to obtain:

$$\frac{n^{1/2}}{\mathcal{V}_n}\left(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \mathcal{B}_n\right) = \frac{n^{1/2}}{\mathcal{V}_n}\frac{1}{n-k_n}\sum_{i=1}^{n}\left\{Z_i(\hat{\gamma}_n)I\left(|Z_i| < c_n\right) - E\left[Z_i I\left(|Z_i| < c_n\right)\right]\right\} + o_p(1).$$

By the mean value theorem, $\hat{\gamma}_n \xrightarrow{p} \gamma_0$, and Lemma A.5:

$$\frac{n^{1/2}}{\mathcal{V}_n}\left(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \mathcal{B}_n\right) = \frac{n^{1/2}}{\mathcal{V}_n}\frac{1}{n-k_n}\sum_{i=1}^{n}\left\{Z_i I\left(|Z_i| < c_n\right) - E\left[Z_i I\left(|Z_i| < c_n\right)\right]\right\} + o_p(1)$$
$$+\frac{n^{1/2}}{\mathcal{V}_n}E\left[\frac{\partial}{\partial\gamma}Z_i I\left(|Z_i| < c_n\right)\right]'(\hat{\gamma}_n - \gamma_0) + o_p(1),$$

where

$$
\begin{aligned}
E\left[\frac{\partial}{\partial\gamma}Z_i I\left(|Z_i|<c_n\right)\right] &= -E\left[\left(\frac{D_i}{p_i^2}+\frac{1-D_i}{(1-p_i)^2}\right)\frac{\partial}{\partial\gamma}p_i Y_i I\left(|Z_i|<c_n\right)\right] \\
&= -E\left[\left(\frac{D_i-p_i}{p_i\left(1-p_i\right)}\right)^2\frac{\partial}{\partial\gamma}p_i Y_i I\left(|Z_i|<c_n\right)\right] = -E\left[h_i\frac{\partial}{\partial\gamma}p_i Z_i I\left(|Z_i|<c_n\right)\right] = \mathcal{D}_n.
\end{aligned}
$$

Now use asymptotic linearity B2 for $n^{1/2}(\hat{\gamma}_n-\gamma_0)$ to yield:

$$
\begin{aligned}
&\frac{n^{1/2}}{\mathcal{V}_n}\left(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n)+\mathcal{B}_n\right) \\
&\quad =\frac{n^{1/2}}{\mathcal{V}_n}\left(\frac{1}{n}\sum_{i=1}^n\left\{Z_i I\left(|Z_i|<c_n\right)-E\left[Z_i I\left(|Z_i|<c_n\right)\right]\right\}+\mathcal{D}_n'\frac{1}{n}\sum_{i=1}^n w_i\right)(1+o_p(1)) \\
&\quad =\frac{1}{\mathcal{V}_n}\frac{1}{n^{1/2}}\sum_{i=1}^n\vartheta_{n,i}\left(1+o_p(1)\right).
\end{aligned}
$$

$\vartheta_{n,i}/\mathcal{V}_n$ is iid across $i\in\{1,...,n\}$, $E[\vartheta_{n,i}/\mathcal{V}_n]=0$, and $E[(\vartheta_{n,i}/\mathcal{V}_n)^2]=1$. Thus, if we demonstrate $\mathcal{V}_n^{-1}n^{-1/2}\sum_{i=1}^n\vartheta_{n,i}$ satisfies the Lindeberg condition then the claim follows by the Lindeberg central limit theorem.

The iid property implies for $\varepsilon>0$:

$$
\sum_{i=1}^n E\left[\left(\frac{\vartheta_{n,i}}{\mathcal{V}_n n^{1/2}}\right)^2 I\left(\frac{|\vartheta_{n,i}|}{\mathcal{V}_n n^{1/2}}>\varepsilon\right)\right]=E\left[\frac{\vartheta_{n,i}^2}{\mathcal{V}_n^2}I\left(\frac{|\vartheta_{n,i}|}{\mathcal{V}_n}>\varepsilon n^{1/2}\right)\right]=\int_{\varepsilon^2 n}^{\infty}P\left(\frac{|\vartheta_{n,i}|}{\mathcal{V}_n}>u^{1/2}\right)du. \quad \text{(B.1)}
$$

Sub-additivity and $|\mathcal{D}_n' w_i|\leq||\mathcal{D}_n||\times||w_i||$ imply:

$$
\begin{aligned}
\int_{\varepsilon^2 n}^{\infty}P\left(\frac{|\vartheta_{n,i}|}{\mathcal{V}_n}>u^{1/2}\right)du \quad \leq \quad & \int_{\varepsilon^2 n}^{\infty}P\left(\frac{|Z_i|\,I\left(|Z_i|<c_n\right)}{\mathcal{V}_n}>\frac{u^{1/2}}{3}\right)du+\int_{\varepsilon^2 n}^{\infty}P\left(\frac{|E\left[Z_i I\left(|Z_i|<c_n\right)\right]|}{\mathcal{V}_n}>\frac{u^{1/2}}{3}\right)du \\
&+\int_{\varepsilon^2 n}^{\infty}P\left(\|w_i\|>\frac{\|\mathcal{V}_n\|}{3\,\|\mathcal{D}_n\|}u^{1/2}\right)du.
\end{aligned}
$$

Assumption A5 states $\liminf_{n\to\infty}\mathcal{V}_n>0$, while $|E[Z_i I(|Z_i|<c_n)]|\leq E|Z_i||<\infty$. Hence, for all $n\geq N_\varepsilon$ and some $N_\varepsilon\geq 1$ that depends on ε:

$$
\int_{\varepsilon^2 n}^{\infty}P\left(\frac{|E\left[Z_i I\left(|Z_i|<c_n\right)\right]|}{\mathcal{V}_n}>\frac{u^{1/2}}{3}\right)du\leq\int_{\varepsilon^2 n}^{\infty}P\left(E\,|Z_i|>Ku^{1/2}\right)du=\int_{\varepsilon^2 n}^{\infty}I\left(E\,|Z_i|>Ku^{1/2}\right)du=0. \quad \text{(B.2)}
$$

Next, $||\mathcal{D}_n||=O(\sigma_n)$ and therefore $\mathcal{V}_n^2\sim K\sigma_n^2$ are shown in (*b*), hence $\liminf_{n\to\infty}||\mathcal{V}_n||/||\mathcal{D}_n||>0$. Furthermore, $||w_i||$ satisfies the Lindeberg condition because it is iid and square integrable. Therefore, for any $\varepsilon>0$:

$$
\int_{\varepsilon^2 n}^{\infty}P\left(\|w_i\|>\frac{\|\mathcal{V}_n\|}{3\,\|\mathcal{D}_n\|}u^{1/2}\right)du\leq\int_{\varepsilon^2 n}^{\infty}P\left(\|w_i\|>Ku^{1/2}\right)du\to 0. \quad \text{(B.3)}
$$

Finally, in (*b*) we prove $\mathcal{V}_n^2\sim K\sigma_n^2$ for some $K>0$, with $K=1$ if $E[Z_i^2]=\infty$. If $E[Z_i^2]<\infty$ then $\mathcal{V}_n^2\sim$

$K\sigma_n^2 \to KE[Z_i^2] > 0$ and $E[Z_i^2 I(|Z_i| > \varepsilon n^{1/2})] = \int_{\varepsilon^2 n}^{\infty} P(Z_i^2 > u)du \to 0$ for any $\varepsilon > 0$ hence:

$$\int_{\varepsilon^2 n}^{\infty} P\left(\frac{|Z_i|\, I\left(|Z_i| < c_n\right)}{\mathcal{V}_n} > \frac{u^{1/2}}{3}\right) du \leq \int_{\varepsilon^2 n}^{\infty} P\left(Z_i^2 > KE\left[Z_i^2\right] u\right) du = \frac{1}{KE\left[Z_i^2\right]} \int_{KE[Z_i^2]\varepsilon^2 n}^{\infty} P\left(Z_i^2 > v\right) dv \to 0$$

If $E[Z_i^2] = \infty$ then use $\mathcal{V}_n^2 \sim \sigma_n^2$ and a change of variables to write

$$\int_{\varepsilon^2 n}^{\infty} P\left(\frac{|Z_i|\, I\left(|Z_i| < c_n\right)}{\mathcal{V}_n} > \frac{u^{1/2}}{3}\right) du \sim \int_{\varepsilon^2 n}^{9c_n^2/\sigma_n^2} P\left(Z_i^2 > \frac{\sigma_n^2}{9} u\right) du = \frac{9}{\sigma_n^2} \int_{\varepsilon^2 n}^{9c_n^2/\sigma_n^2} P\left(Z_i^2 > v\right) dv.$$

The variance σ_n^2 is characterized by Karamata's Theorem under A3(ii) (Resnick, 1987, Theorem 0.6):[17]

$$\begin{aligned} &E\left[|Z_i|^{\kappa} I\left(|Z_i| \leq c_n\right)\right] \sim d\left\{\ln(n) - \ln(k_n)\right\} \sim d\ln(n) && \text{(B.4)} \\ &E\left[|Z_i|^{p} I\left(|Z_i| \leq c_n\right)\right] \sim \frac{p}{p-\kappa} c_n^p P\left(|Z_i| > c_n\right) \sim \frac{p}{p-\kappa} d^{p/\kappa} \left(\frac{n}{k_n}\right)^{p/\kappa - 1} \quad \forall p > \kappa. \end{aligned}$$

The A3 power law property implies by construction $c_n^2 \sim K(n/k_n)^{2/\kappa}$ with tail index $\kappa \in (1, 2]$, and by Karamata's Theorem $\sigma_n^2 \sim (2/(2-\kappa))c_n^2 P(|Z_i| > c_n)$, hence $c_n^2/\sigma_n^2 \sim K/P(|Z_i| > c_n) = Kn/k_n = o(n)$. Therefore, $\int_{\varepsilon^2 n}^{9c_n^2/\sigma_n^2} P(Z_i^2 > v)dv = 0$ for all $n \geq N_\varepsilon$ and some $N_\varepsilon \geq 1$ that depends on ε, hence:

$$\int_{\varepsilon^2 n}^{\infty} P\left(\frac{|Z_i|\, I\left(|Z_i| < c_n\right)}{\mathcal{V}_n} > Ku^{1/2}\right) \to 0. \tag{B.5}$$

Together, (B.1)-(B.5) imply the Lindeberg condition holds:

$$\lim_{n\to\infty} E\left[\frac{\vartheta_{n,i}^2}{\mathcal{V}_n^2} I\left(\frac{|\vartheta_{n,i}|}{\mathcal{V}_n} > \varepsilon n^{1/2}\right)\right] = \lim_{n\to\infty} \int_{\varepsilon^2 n}^{\infty} P(\vartheta_{n,i}^2/\mathcal{V}_n^2 > u)du = 0\ \forall \varepsilon > 0. \tag{B.6}$$

Claim (b). By construction of $\mathcal{V}_n^2 \equiv E[\vartheta_{n,i}^2]$, the A5 bound $\liminf_{n\to\infty} \mathcal{V}_n^2 > 0$, and $\liminf_{n\to\infty} \sigma_n^2 > 0$ given non-degeneracy and $c_n \to \infty$, we need only prove $\mathcal{D}_n = O(\sigma_n)$, and $\mathcal{D}_n = o(\sigma_n)$ when $E[Z_i^2] = \infty$. This will prove $\mathcal{V}_n^2 \sim K\sigma_n^2$. Since $\vartheta_{n,i}$ is the L_2 metric projection residual of the demeaned infeasible $Z_i I(|Z_i - \theta| < c_n)$ on the score, it must be the case that $K \in (0, 1]$, cf. Graham (2011).

Under B3(ii) each $(\partial/\partial\gamma_i)p_i h_i$ is $L_{2+\iota}$-bounded for some tiny $\iota > 0$. Therefore, by Holder's inequality:

$$\begin{aligned} \left|E\left[\frac{\partial}{\partial\gamma_i} p_i h_i Z_i I\left(|Z_i| < c_n\right)\right]\right| &\leq \left(E\left[\left|\frac{\partial}{\partial\gamma_i} p_i h_i\right|^{2+\iota}\right]\right)^{\frac{1}{2+\iota}} \left(E\left[|Z_i|^{\frac{2+\iota}{1+\iota}} I\left(|Z_i| < c_n\right)\right]\right)^{\frac{1+\iota}{2+\iota}} \\ &= K_i \left(E\left[|Z_i|^{\frac{2+\iota}{1+\iota}} I\left(|Z_i| < c_n\right)\right]\right)^{\frac{1+\iota}{2+\iota}} \equiv m_{i,n}(\iota), \end{aligned}$$

say, where $K_i < \infty$. It suffices to prove $m_{i,n}(\iota) = O(\sigma_n)$, and $m_{i,n}(\iota) = o(\sigma_n)$ when $E[Z_i^2] = \infty$. In view of $(2 + \iota)/(1 + \iota) < 2$, Lyapunov's inequality suffices for $m_{i,n}(\iota) \leq (E[Z_i^2 I(|Z_i| < c_n)])^{1/2} = \sigma_n$.

[17] Note that for any finite $a > 0$ and some $K(a) > 0$ we have $E[|Z_i|^{\kappa} I(|Z_i| \leq c_n)] = K(a) + \int_a^{c_n^{\kappa}} P(|Z_i| \geq u^{1/\kappa})du \sim K(a) + d\int_a^{c_n^{\kappa}} u^{-1}du = K(a) + d(\ln(c_n^{\kappa}) - \ln(a))$. Now use $c_n^{\kappa} = d(n/k_n)$ and $k_n = o(n)$ to deduce $E[|Z_i|^{\kappa} I(|Z_i| \leq c_n)] \sim d\{\ln(n) - \ln(k_n)\} \sim d\ln(n)$.

Now suppose $E[Z_i^2] = \infty$ (i.e. $\kappa \leq 2$). If $\kappa > (2+\iota)/(1+\iota)$ then $\sigma_n^2 \to \infty$ and $m_{i,n}(\iota) = O(1) = o(\sigma_n)$. If $\kappa = (2+\iota)/(1+\iota)$ then use Karamata theory (B.4) to get $m_{i,n}(\iota) \sim K\ln(n)$ and $\sigma_n \sim K(n/k_n)^{1/\kappa-1/2}$, hence $m_{i,n}(\iota) = o(\sigma_n)$. Finally, if $\kappa < (2+\iota)/(1+\iota)$ then $m_{i,n}(\iota) \sim K(n/k_n)^{1/\kappa-(1+\iota)/(2+\iota)}$ and $\sigma_n \sim K(n/k_n)^{1/\kappa-1/2}$ by (B.4), hence $m_{i,n}(\iota) = o(\sigma_n)$.

Claim (c). Since $\mathcal{V}_n^2 \sim K\sigma_n^2$, it suffices to inspect $(n^{1/2}/\sigma_n)\mathcal{B}_n$. If Z_i is symmetric about zero then $\mathcal{B}_n = 0$, so let Z_i have an asymmetric distribution. Under power law A3(ii), and by threshold construction (9), we have

$$c_n \sim d^{1/\kappa}\,(n/k_n)^{1/\kappa}\,. \tag{B.7}$$

The claim follows from (B.4) in the infinite variance case, (B.7), and bias formula (12). Together, we have the following. If $\kappa > 2$ then $\sigma_n^2 \to (0,\infty)$ hence $(n^{1/2}/\sigma_n)\mathcal{B}_n \sim Kn^{1/2}(k_n/n)c_n \sim Kn^{1/2}(k_n/n)^{1-1/\kappa} = Kk_n^{1-1/\kappa}/n^{1/2-1/\kappa}$. Therefore as long as $k_n/\ln(n) \to 0$ then $k_n/n^{(\kappa-2)/(2(\kappa-1))} \to 0$ for any $\kappa > 2$, hence $n^{1/2}\mathcal{B}_n = Kk_n^{1-1/\kappa}/n^{1/2-1/\kappa} \to 0$. Similarly, if $\kappa = 2$ then $\sigma_n^2 \sim K\ln(n)$ hence $(n^{1/2}/\sigma_n)\mathcal{B}_n \sim k_n^{1-1/2}/((\ln(n))^{1/2}n^{1/2-1/2}) = (k_n/\ln(n))^{1/2} \to 0$. Finally, if $\kappa < 2$ then $\sigma_n^2 \sim Kc_n^2(k_n/n)$ hence $(n^{1/2}/\sigma_n)\mathcal{B}_n \sim Kn^{1/2}(k_n/n)c_n/(c_n^2(k_n/n))^{1/2} = Kk_n^{1/2} \to \infty$. $\mathcal{QED}$.

Proof of Lemma 3.2. Claim (a) follows from trimming negligibility, finite variance, and Theorem 3.1. Invoke (B.7) and (B.4) for (b). $\mathcal{QED}$.

Proof of Lemma 3.3. Define left and right tail quantile functions (where $0 \leq u \leq 1$):

$$Q_1(u) \equiv \inf\left\{c \geq 0 : P\left(Z_i \leq -c\right) \geq u\right\} \text{ and } Q_2(u) \equiv \inf\left\{c \geq 0 : P\left(Z_i \geq c\right) \geq u\right\}.$$

Under power law (5), $Q_i(u) = d_i^{1/\kappa_i}u^{-1/\kappa_i}$ as $u \to 0$. Now use threshold construction (B.7) to deduce:

$$\begin{aligned} E\left[Z_iI\left(|Z_i| > c_n\right)\right] &= E\left[Z_iI\left(|Z_i| > c_n\right)\right] = \left(\int_0^{k_n/n} Q_2(u)du - \int_0^{k_n/n} Q_1(u)du\right) \\ &\sim \int_0^{k_n/n} d_2^{1/\kappa_2}u^{-1/\kappa_2}du - \int_0^{k_n/n} d_1^{1/\kappa_1}u^{-1/\kappa_1}du \\ &= d_2^{1/\kappa_2}\left(\frac{\kappa_2}{\kappa_2-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_2} - d_1^{1/\kappa_1}\left(\frac{\kappa_1}{\kappa_1-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_1}. \end{aligned} \tag{B.8}$$

This proves bias approximation (12) given $\mathcal{B}_n \equiv (n/(n-k_n))E[Z_iI(|Z_i| > c_n)]$. $\mathcal{QED}$.

Proof of Theorem 3.4. Recall $\theta_0 = 0$. We will prove $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n)) \xrightarrow{d} N(0,1)$. Then $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \hat{\mathcal{B}}_n(\hat{\gamma}_n.\phi_n^*)) \xrightarrow{d} N(0,1)$ in view of $m_n(\phi) = [\phi m_n]$ and $m_n/k_n \to \infty$ by arguments in Hill (2015, Theorems 2.1 and 2.2). The proof of $n^{1/2}\mathcal{V}_n^{-1}\hat{\theta}_n^{(tz:o)} \xrightarrow{d} N(0,1)$ follows similarly.

In view of $n^{1/2}\mathcal{V}_n^{-1}(\hat{\theta}_n^{(tz)}(\hat{\gamma}_n) + \mathcal{B}_n) \xrightarrow{p} N(0,1)$ and $\mathcal{V}_n^2 \sim K\sigma_n^2$ by by Theorem 3.1.a,b, we need only prove

$$\frac{n^{1/2}}{\sigma_n}\left(\hat{\mathcal{B}}_n(\hat{\gamma}_n) - \mathcal{B}_n\right) \xrightarrow{p} 0. \tag{B.9}$$

Define

$$\mathring{\mathcal{B}}_n \equiv \frac{n}{n-k_n}\left\{d_2^{1/\kappa_2}\left(\frac{\kappa_2}{\kappa_2-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_2} - d_1^{1/\kappa_1}\left(\frac{\kappa_1}{\kappa_1-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_1}\right\}.$$

Under power law A3$'$, arguments in Peng (2001, proof of Theorem 1) verify that $n^{1/2}\sigma_n^{-1}(\mathring{\mathcal{B}}_n - \mathcal{B}_n) = o(1)$.

It remains to prove $n^{1/2}\sigma_n^{-1}(\hat{\mathcal{B}}_n(\hat{\gamma}_n) - \mathring{\mathcal{B}}_n) = o_p(1)$. Write $\hat{\kappa}_{m_n,i} = \hat{\kappa}_{m_n,i}(\hat{\gamma}_n)$ and $\hat{d}_{m_n,i} = \hat{d}_{m_n,i}(\hat{\gamma}_n)$. The left or right tail bias components of $n^{1/2}\sigma_n^{-1}(\hat{\mathcal{B}}_n(\hat{\gamma}_n) - \mathring{\mathcal{B}}_n)$ are, up to the scale $n/(n-k_n) \approx 1$:

$$\frac{n^{1/2}}{\sigma_n}\left\{\hat{d}_{m_n,i}^{1/\hat{\kappa}_{m_n,i}}\left(\frac{\hat{\kappa}_{m_n,i}}{\hat{\kappa}_{m_n,i}-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\hat{\kappa}_{m_n,i}} - d_i^{1/\kappa_i}\left(\frac{\kappa_i}{\kappa_i-1}\right)\left(\frac{k_n}{n}\right)^{1-1/\kappa_i}\right\}.$$

The tail exponent limit theory for a filtered process developed in Hill (2014, Theorem 2.1), and detailed in Step 1 of the proof of Lemma A.4, along with $\hat{\gamma}_n = \gamma_0 + O(1/n^{1/2})$ by B2, implies $\hat{\kappa}_{m_n,i} = \kappa_i + O_p(1/m_n^{1/2})$ and $\hat{d}_{m_n,i}^{1/\hat{\kappa}_{m_n,i}}/d_i^{1/\kappa_i} = 1 + O_p(1/m_n^{1/2})$. By Karamata theory if $\kappa \equiv \min\{\kappa_1,\kappa_2\} = 2$ then $\sigma_n^2 \sim d\ln(n)$ and if $\kappa < 2$ then $\sigma_n^2 \sim K(n/k_n)^{2/\kappa - 1}$, and by A3$'$ $k_n = o(\ln(n))$ and $m_n/k_n \to \infty$. By the mean-value-theorem, it therefore follows $(k_n/n)^{1-1/\hat{\kappa}_{m_n,i}} - (k_n/n)^{1-1/\kappa_i} = O_p((k_n/n)^{1-1/\kappa} m_n^{-1/2}\ln(n))$. Therefore

$$\frac{n^{1/2}}{\sigma_n}\left\{\left(\frac{k_n}{n}\right)^{1-1/\hat{\kappa}_{m_n,i}} - \left(\frac{k_n}{n}\right)^{1-1/\kappa_i}\right\} = \frac{n^{1/2}}{\sigma_n} \times O_p\left(\left(\frac{k_n}{n}\right)^{1-1/\kappa}\frac{\ln(n)}{m_n^{1/2}}\right)$$
$$= \begin{cases} O_p\left(n^{1/2}\frac{k_n^{1/2}}{n^{1/2}}\left(\frac{k_n}{n}\right)^{1/2-1/\kappa}\frac{\ln(n)}{m_n^{1/2}}\right) = O_p\left(\frac{k_n^{1/2}}{m_n^{1/2}}\left(\frac{k_n}{n}\right)^{1/2-1/\kappa}\ln(n)\right) = o_p(1) & \text{if } \kappa > 2 \\ O_p\left(\frac{n^{1/2}}{\ln(n)}\left(\frac{k_n}{n}\right)^{1-1/2}\frac{\ln(n)}{m_n^{1/2}}\right) = O_p\left(\left(\frac{k_n}{m_n}\right)^{1/2}\right) = o_p(1) & \text{if } \kappa = 2 \\ O_p\left(\frac{n^{1/2}(k_n/n)^{1-1/\kappa}}{(n/k_n)^{1/\kappa-1/2}}\frac{\ln(n)}{m_n^{1/2}}\right) = o_p(1) & \text{if } \kappa < 2 \end{cases}$$

The remainder of the proof simply repeats this logic for all terms in $n^{1/2}\sigma_n^{-1}(\hat{\mathcal{B}}_n(\hat{\gamma}_n) - \mathring{\mathcal{B}}_n)$. $\mathcal{QED}$.

References

ANDREWS, D. (1988): "Laws of Large Numbers for Dependent Non-Identically Distributed Random Variables," *Econometric Theory*, 4, 458–467.

BAHADUR, R. (1960): "Asymptotic Efficiency of Tests and Estimates," *Sankhya*, 22, 229–252.

BUSSO, M., J. DINARDO, AND J. MCCRARY (2009): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," Discussion paper, University of Michigan.

CHAUDHURI, S., AND J. HILL (2014): "Supplemental Appendices I and II for Robust Estimation for Average Treatment Effects," mimeo.

CHAUDHURI, S., AND H. MIN (2012): "Doubly-Robust Parametric Estimation in Moment Conditions Models with Missing Data," Discussion paper, University of North Carolina.

CHRITSOPEIT, N., AND H. WERNER (2001): "A Necessary and Sufficient Condition of a Sequence of Random Variables Converging to a Normal Distribution," *Econometric Theory*, 17, 278–281.

CRUMP, R., V. HOTZ, G. IMBENS, AND O. MITNIK (2009): "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199.

CSÖRGO, S., L. HORVÁTH, AND D. MASON (1986): "What Portion of the Sample Makes a Partial Sum Asymptotically Stable or Normal?," *Probability Theory and Related Fields*, 72, 1–16.

DEHEJIA, R., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of American Statistical Association*, 94, 1053–1062.

DUDLEY, R. M. (1978): "Central Limit Theorems for Empirical Measures," *Annals of Probability*, 6, 899–929.

——— (1999): *Uniform Central Limit Theorems*. Cambridge University Press, Cambridge.

FELLER, W. (1971): *An Introduction to Probability Theory and Its Applications (Vol. II)*. Wiley, New York.

FROLICH, M. (2004): "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, 86, 77–90.

GALAMBOS, J. (1987): *The Asymptotic Theory of Extreme Order Statistics*. Krieger: Malabar.

GRAHAM, B. S. (2011): "Efficiency Bounds for Missing Data Models with Semiparametric Restrictions," *Econometrica*, 79, 437 – 452.

HAEUSLER, E., AND J. TEUGELS (1985): "On Asymptotic Normality of Hill's Estimator for the Exponent of Regular Variation," *Annals of Statistics*, 13, 743–756.

HAHN, M., J. KUELBS, AND J. SAMUR (1987): "Asymptotic Normality of Trimmed Sums of ϕ-Mixing Random Variables," *Annals of Probability*, 15, 1395–1418.

HAHN, M., D. WEINER, AND D. MASON (1991): *Sums, Trimmed Sums and Extremes*. Birkhäuser: Berlin.

HALL, P. (1982): "On Some Simple Estimates of an Exponent of Regular Variation," *Journal of the Royal Statistical Societ Series B*, 44, 37–42.

——— (1990): "Asymptotic Properties of the Bootstrap for Heavy-Tailed Distributions," *Annals of Probability*, 18, 1342–1360.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

HILL, B. M. (1975): "A Simple General Approach to Inference about the Tail of a Distribution," *Annals of Statistics*, 3(5), 1163–1174.

HILL, J. B. (2010): "On Tail Index Estimation for Dependent, Heterogeneous Data," *Econometric Theory*, 26, 1398–1436.

——— (2012a): "Heavy-Tail and Plug-In Robust Consistent Conditional Moment Tests of Functional Form," in *Festschrift in Honor of Hal White*, ed. by X. Chen, and N. Swanson, pp. 241–274. Springer: New York.

——— (2012b): "Least Tail-Trimmed Squares for Infinite Variance Autoregressions," *Journal of Time Series Analysis*, 34, 168–186.

——— (2014): "Tail Index Estimation for a Filtered Dependent Time Series," *Statistica Sinica*, 25.

——— (2015): "Robust Expected Shorfall Estimation for Infinite Variance Time Series," *Journal of Financial Econometrics*, 13, 1–44.

HILL, J. B., AND A. PROKHOROV (2016): "GEL Estimation for Heavy-Tailed GARCH Models with Robust Empirical Likelihood Inference," *Journal of Econometrics*, 190, 18–45.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Scores," *Econometrica*, 71, 1161–1189.

IBRAGIMOV, I., AND I. LINNIK (1971): *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff.

JURECKOVA, J. (1981): "Tail-Behavior of Location Estimators," *Annals of Statistics*, 9, 578–585.

KANG, J., AND J. SCHAFER (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 523–539.

KHAN, S., AND D. NEKIPELOV (2013): "On Uniform Inference in Nonlinear Models with Endogeneity," Discussion paper, Duke University.

KHAN, S., AND E. TAMER (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

LEADBETTER, M., G. LINDGREN, AND H. ROOTZEN (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer-Verlag.

LECHNER, M. (2008): "A Note on the Common Support Problem in Applied Evaluation Studies," *Annals of Economic and Statistics*, 91/92, 217–235.

LEE, B., J. LESSLER, AND E. STUART (2011): "Weight Trimming and Propensity Score Weighting," *PLOS One*, 6.

LEWBEL, A. (1997): "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory*, 13, 32–51.

LIGHTHILL, M. (1958): *Introduction to Fourier Analysis and Generalized Functions*. Cambridge Univ. Press, Cambridge.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027–1057.

PENG, L. (2001): "Estimating the Mean of a Heavy Tailed Distribution," *Statistics and Probability Letters*, 52, 255–264.

PHILLIPS, P. C. B. (1995): "Robust Nonstationary Regression," *Econometric Theory*, 11, 912–951.

POLLARD, D. (1984): *Convergence of Stochastic Processes*. Springer, New York.

POTTER, F. (1993): "The Effect of Weight Trimming on Nonlinear Survey Estimates," in *Proceedings of the Section on Survey Research Methods & Research*. American Statistical Association.

RESNICK, S. (1987): *Extreme Values, Regular Variation and Point Processes*. Springer-Verlag: New York.

ROSENBAUM, P., AND D. RUBIN (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

ROTHE, C. (2015): "Robust Confidence Intervals for Average Treatment Effects under Limited Overlap," Discussion Paper 8758, Columbia University.

RUBIN, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

TRASKIN, M., AND D. SMALL (2011): "Defining the Study Population for an Observational Study to Ensure Sufficient Overlap: A Tree Approach," *Statistics in Biosciences*, 3, 94–118.

WOOLDRIDGE, J. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics*, 141, 1281–1301.

YANG, T. (2015): "Asymptotic Trimming and Rate Adaptive Inference for Endogenous Selection Estimates," Discussion paper, Dept. of Economics, Boston College.

ZINDE-WALSH, V. (2014): "Measurement Error and Decomvolution Spaces of Generalized Functions," *Econometric Theory*, 30, 1207–1246.

Table 1. (a) Estimator Properties (Symmetric Z, known $p(X)$, Normal *or* Laplace, $n = 100$, 250)

	$n = 100$									$n = 250$								
		$(Y_0, Y_1, X, U) \sim$ **Normal**				$(Y_0, Y_1, X, U) \sim$ **Laplace**					$(Y_0, Y_1, X, U) \sim$ **Normal**				$(Y_0, Y_1, X, U) \sim$ **Laplace**			
		$\beta = .25$ ($\kappa = 17$)				$\beta = .25$ ($\kappa = 5$)					$\beta = .25$ ($\kappa = 17$)				$\beta = .25$ ($\kappa = 5$)			
Estimator	Tr%	Mean	Med	MSE	$KS_{.05}$	Mean	Med	MSE	$KS_{.05}$	Tr%	Mean	Med	MSE	$KS_{.05}$	Mean	Med	MSE	$KS_{.05}$
No Trim	0	.0023	.0025	.2027	.6031	.0018	.0020	.2179	.5773	0	-.0004	-.0006	.1289	.4570	.0015	.0018	.1366	.4855
TT(Z)	1	.0013	-.0002	2058	.5469	.0012	.0003	.2145	.5760	.4	-.0007	-.0006	.1295	.6493	.0010	.0001	.1341	.5245
TT-BC(Z)	1	.0013	.0001	.2055	.4101	.0013	.0004	.2129	.8190	.4	-.0007	-.0006	.1294	.4697	.0010	.0005	.1332	.7832
TT(X)	13	.0021	.0017	.1989	.5868	.0012	.0024	.2068	.6970	8.7	-.0005	-.0011	.1275	.5491	.0021	.0037	.1309	.7071
TT(X,$k_n^{(x)}$)	43	.0020	.0019	.1513	.4316	-.0005	.0013	.1826	.4713	36	-.0002	.0002	.1003	.3879	.0010	.0022	.1190	.6126
TT(X,k_n)	1	.0026	.0024	.2014	.5039	-.0010	-.0018	.6870	.4945	.4	-.0004	-.0003	.1286	.5406	.0023	.0023	.1363	.4842
TT(Y)	1	.0060	.0082	.2061	.4500	.0064	-.0078	.2357	.8615	.4	.0019	.0071	.1267	.5243	.0013	.0006	.1397	.4245
		$\beta = 1$ ($\kappa = 2$)				$\beta = 1$ ($\kappa = 2$)					$\beta = 1$ ($\kappa = 2$)				$\beta = 1$ ($\kappa = 2$)			
Estimator	Tr%	Mean	Med	MSE	$KS_{.05}$	Mean	Med	MSE	$KS_{.05}$	Tr%	Mean	Med	MSE	$KS_{.05}$	Mean	Med	MSE	$KS_{.05}$
No Trim	0	.0071	.0047	.3376	5.751	.0013	.0017	.4556	8.912	0	.0001	-.0021	.2302	5.632	-.0041	-.0036	.3351	10.21
TT(Z)	1	.0038	.0032	.2126	.9237	.0022	.0043	.2387	1.209	.4	-.0002	-.0035	.1486	1.027	-.0029	-.0016	.1659	1.242
TT-BC(Z)	1	.0037	.0032	.2102	.5484	.0028	.0042	.2389	.6935	.4	-.0002	-.0034	.1469	.9469	-.0029	-.0017	.1622	.6239
TT(X)	13	.0042	.0046	.2809	2.211	.0052	.0051	.2837	.1551	8.7	-.0008	-.0018	.1900	2.159	-.0020	-.0030	.1854	1.246
TT(X,$k_n^{(x)}$)	43	.0023	-.0002	.1602	.6443	.0006	.0012	.1980	.8040	36	.0005	.0007	.1103	.5807	-.0017	-.0009	.1321	.7328
TT(X,k_n)	1	.0049	.0049	.3185	4.505	-.0055	-.0015	.4284	7.408	.4	-.0006	-.0022	.2158	4.424	-.0033	-.0021	.2960	7.317
TT(Y)	1	-.0166	-.0143	.3115	1.006	.0065	.0058	.4053	1.906	.4	-.0229	.0048	.6458	8.351	-.0025	-.0018	.3197	2.886
		$\beta = 2$ ($\kappa = 1.25$)				$\beta = 2$ ($\kappa = 1.5$)					$\beta = 2$ ($\kappa = 1.25$)				$\beta = 2$ ($\kappa = 1.5$)			
Estimator	Tr%	Mean	Med	MSE	$KS_{.05}$	Mean	Med	MSE	$KS_{.05}$	Tr%	Mean	Med	MSE	$KS_{.05}$	Mean	Med	MSE	$KS_{.05}$
No Trim	0	.0001	-.0010	.6623	16.54	-.0014	.0053	.7859	16.06	0	.0097	.0009	1.137	27.47	-.0021	-.0028	.7826	19.56
TT(Z)	1	-.0008	-.0018	.2063	2.382	.0023	.0013	.2514	2.062	.4	.0006	.0009	.1722	2.143	.0002	.0004	.1946	1.732
TT-BC(Z)	1	.0006	-.0015	.2474	1.425	.0009	.0013	.3012	1.352	.4	.0016	.0007	.2417	1.324	-.0014	.0002	.2409	1.232
TT(X)	13	.0001	-.0008	.6621	16.53	.0059	.0035	.5513	9.964	8.7	.0096	.0010	1.137	27.47	-.0025	-.0025	.3910	8.286
TT(X,$k_n^{(x)}$)	43	-.0008	-.0001	.2034	1.634	.0012	.0019	.2431	1.219	36	.0030	.0016	.1506	1.413	-.0027	-,0019	.1693	1.322
TT(X,k_n)	1	.0002	-.0006	.6623	16.54	.0022	.0005	.7200	13.85	.4	-.0025	-.0033	.7877	27.77	-.0102	-.0033	.6726	18.05
TT(Y)	1	.0366	-.0010	1.048	8.056	.0250	-.0027	.6488	3.909	.4	.0191	.0020	.6472	6.459	-.0116	.0122	.5812	5.217

The treatment assignment is $D = I(\alpha + \beta X > U)$ with $\alpha = 0$, hence Z has a symmetric distribution. The true propensity score $p(X)$ is used to compute Z. "No Trim" is the untrimmed estimator $\tilde{\theta}_n$; "TT(Z)" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "TT–BC(Z)" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming. "TT(X)" is $\theta_n^{(tx)}$; and "TT(X,k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$. "TT(Y)" is $\hat{\theta}_n^{(ty)}$. $KS_{.05}$ is the Kolmogorov-Smirnov test statistic divided by its 5% critical value: values above 1 indicate rejection of standard normality at the 5% level. Tr% is the percent of observations Z_i trimmed. κ is the tail index of $Z = h(X)Y$. Other than $KS_{.05}$, all values are averages over the randomly drawn 10,000 samples.

Table 1. (b) Estimator Properties (Symmetric Z, known $p(X)$, Normal *and* Laplace, $n = 100$, 250)

$n = 100$

		$(Y_0, Y_1, X) \sim$ **Norm,** $U \sim$ **Lap**				$(Y_0, Y_1, X) \sim$ **Lap,** $U \sim$ **Norm**			
		$\beta = .25$				$\beta = .25$			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	.0005	.0003	.2054	.7790	.0001	.0031	.2189	.7417
TT(Z)	1	.0001	.0010	.2068	.5409	-.0013	-.0007	.2099	.8640
TT-BC(Z)	1	.0002	.0009	.2066	.6817	-.0013	.0000	.2086	.9786
TT(X)	13	.0007	-.0005	.2009	.8564	-.0012	.0004	.2032	.8950
TT(X,$k_n^{(x)}$)	43	-.0003	.0001	.1524	.5368	-.0004	.0004	.1804	.7685
TT(X,k_n)	1	-.0014	-.0025	.2039	.6515	-.0029	-.0007	.2165	.5335
TT(Y)	1	.0057	.0019	.2085	.6133	.0004	.0030	.2357	.9575
		$\beta = 1$				$\beta = 1$			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0019	-.0031	.2637	1.190	-.0041	-.0050	.5865	14.20
TT(Z)	1	-.0025	-.0036	.2179	.6263	.0010	.0030	.2288	1.565
TT-BC(Z)	1	-.0025	-.0035	.2151	.5152	.0022	.0002	.2566	.6806
TT(X)	13	-.0027	-.0041	.2500	1.031	-.0027	-.0022	.3528	14.52
TT(X,$k_n^{(x)}$)	43	-.0013	-.0024	.1596	.5082	-.0002	.0020	.1959	.4923
TT(X,k_n)	1	-.0011	-.0008	.2543	1.234	.0048	.0001	.6123	15.07
TT(Y)	1	.000	.0026	.2607	.7002	.0133	.0211	.4418	2.847
		$\beta = 2$				$\beta = 2$			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0002	-.0012	.5898	12.41	.0075	.0030	.7762	18.55
TT(Z)	1	.0006	-.0030	.2385	1.765	.0035	.0016	.2191	2.995
TT-BC(Z)	1	.0002	-.0036	.2481	1.352	.0021	.0005	.2552	1.849
TT(X)	13	-.0039	-.0023	.4086	5.725	.0083	.0049	.7773	18.50
TT(X,$k_n^{(x)}$)	43	.0009	.0009	.1871	1.132	.0030	.0029	.2705	3.335
TT(X,k_n)	1	.0017	-.0009	.4611	7.794	.0063	.0008	.7721	18.59
TT(Y)	1	.0034	.0051	.4635	3.195	.0186	.0236	.5712	4.562

$n = 250$

		$(Y_0, Y_1, X) \sim$ **Norm,** $U \sim$ **Lap**				$(Y_0, Y_1, X) \sim$ **Lap,** $U \sim$ **Norm**			
		$\beta = .25$				$\beta = .25$			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0002	-.0009	.1296	.5263	.0015	.0017	.1388	.5094
TT(Z)	.4	-.0003	-.0002	.1299	.7953	.0022	.0027	..1907	.4206
TT-BC(Z)	.4	-.0003	-.0003	.1296	.4572	.0023	.0026	.1315	.6002
TT(X)	8.7	-.0002	.0004	.1283	.7471	.0018	.0029	.1209	.7311
TT(X,$k_n^{(x)}$)	36	.0005	.0001	.1017	.5726	.0008	.0019	.1183	.3993
TT(X,k_n)	.4	-.0005	-.0009	.1302	.6876	.0019	.0021	.1368	.4415
TT(Y)	.4	.0028	.0035	.1282	.5046	-.0007	-.0026	.1433	.5514
		$\beta = 1$				$\beta = 1$			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0035	-.0068	.1164	.8191	-.0108	-.0045	.5499	18.24
TT(Z)	.4	-.0036	-.0051	.1461	.7105	-.0026	-.0027	.1659	1.631
TT-BC(Z)	.4	-.0036	-.0050	.1444	.7608	-.0028	-.0027	.1880	.8340
TT(X)	8.7	-.0033	-.0043	.1581	.4808	-.0044	-.0050	.2516	4.461
TT(X,$k_n^{(x)}$)	36	-.0011	-.0013	.1092	.7234	-.0025	-.0043	.1323	.7633
TT(X,k_n)	.4	-.0013	-.0026	.1654	.5948	-.0001	.0018	.5299	17.83
TT(Y)	.4	-.0045	-.0003	.1656	.6793	-.0118	-.0008	.3896	4.226
		$\beta = 2$				$\beta = 2$			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0009	-.0025	.3393	8.148	-.0038	-.0039	.9481	24.55
TT(Z)	.4	-.0038	-.0051	.1747	1.354	-.0001	.0011	.1755	4.004
TT-BC(Z)	.4	-.0052	-.0050	.1907	.7279	-.0006	.0011	.2078	1.764
TT(X)	8.7	-.0025	-.0026	.2754	4.063	-.0043	-.0044	.9480	24.56
TT(X,$k_n^{(x)}$)	36	-.0020	-.0014	.1304	.9685	-.0039	-.0011	.2128	3.518
TT(X,k_n)	.4	-.0050	-.0031	.3117	6.578	.0059	-.0009	.8406	23.15
TT(Y)	.4	-.0033	.0006	.4149	3.886	-.0198	-.0135	.5921	5.772

The treatment assignment is $D = I(\alpha + \beta X > U)$ with $\alpha = 0$, hence Z has a symmetric distribution. The true propensity score $p(X)$ is used to compute Z. "No Trim" is the untrimmed estimator $\tilde{\theta}_n$; "TT(Z)" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "TT–BC(Z)" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming. "TT(X)" is $\theta_n^{(tx)}$; and "TT(X,k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$. "TT(Y)" is $\hat{\theta}_n^{(ty)}$. KS$_{.05}$ is the Kolmogorov-Smirnov test statistic divided by its 5% critical value: values above 1 indicate rejection of standard normality at the 5% level. Tr% is the percent of observations Z_i trimmed. κ is the tail index of $Z = h(X)Y$. Other than KS$_{.05}$, all values are averages over the randomly drawn 10,000 samples.

Table 2. Rejection Frequencies (Symmetric Z, known $p(X)$, n= 100, 250)

$n = 100$							
$(Y_0, Y_1, X, U) \sim$ **Normal**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.011, .052, .102	.013, .052, .099	.010, .053, .103	.011, .052, .101	.011, .051, .103	.012, .051, .104	.013,.048,.109
1	.017, .039, .068	.013, .053, .098	.011, .053, .104	.019, .055, .094	.011, .049, .100	.012, .045, .076	.019,.037,.083
2	.020, .031, .043	.018, .051, .087	.018, .052, .093	.021, .032, .044	.016, .052, .095	.021, .032, .044	.004,.004,.005
$(Y_0, Y_1, X, U) \sim$ **Laplace**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.010, .049, .096	.010, .052, .101	.008, .052, .104	.010, .051, .099	.011, .050, .100	.011, .048 .099	.009,.046,.103
1	.016, .034, .052	.017, .049, .090	.014, .053, .097	.016, .054, .098	.012, .051, .102	.018, .038, .058	.022,.045,.063
2	.022, .034, .045	.017, .048, .084	.017, .049, .089	.026, .046, .066	.015, .054, .098	.022, .037, .051	.025,.034,.042
$(Y_0, Y_1, X) \sim$ **Normal**, $U \sim$ **Laplace**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.011, .051, .100	.010, .050, .103	.008, .051, .106	.012, .051, .100	.011, .051, .097	.011, .052, .100	.006,.046,.101
1	.013, .050, .098	.013, .049, .099	.010, .050, .104	.013, .051, .101	.011, .051, .101	.014, .054, .097	.009,.047,.089
2	.013, .026, .041	.015, .050, .092	.014, .053, .099	.025, .054, .083	.012, .052, .099	.024, .045, .069	.021,.040,.061
$(Y_0, Y_1, X) \sim$ **Laplace**, $U \sim$ **Normal**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.010, .048, .093	.009, .048, .099	.008, .050, .104	.008, .049, .097	.012, .055, .098	.011, .051, .102	.011,.040,.088
1	.018, .028, .039	.014, .049, .088	.013, .052, .100	.023, .050, .081	.015, .050, .101	.017, .027, .038	.018,.039,.053
2	.020, .030, .040	.017, .047, .082	.018, .052, .093	.020, .030, .040	.018, .051, .088	.020, .030, .041	.021,.035,.043

$n = 250$							
$(Y_0, Y_1, X, U) \sim$ **Normal**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.001, .053, .100	.011, .052, .104	.001, .053, .107	.011, .051, .101	.010, .050, .103	.010, .053, .100	.006,.051,.100
1	.016, .036, .062	.014, .049, .096	.011, .052, .101	.018, .055, .092	.011, .048, .097	.018, .043, .075	.005,.007,.009
2	.007, .011, .014	.015, .038, .069	.018, .054, .092	.008, .011, .014	.016, .054, .095	.013, .020, .024	.012,.023,.029
$(Y_0, Y_1, X, U) \sim$ **Laplace**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.010, .050, .100	.009, .050, .104	.007, .051, .104	.010, .049, .104	.009, .052, .099	.001, .050, .100	.010,.055,.101
1	.013, .027, .042	.016, .050, .094	.013, .051, .099	.016, .053, .097	.011, .050, .100	.015, .034, .054	.018,.030,.044
2	.015, .022, .029	.017, .046, .081	.016, .053, .094	.025, .050, .070	.012, .054, .106	.017, .027, .036	.023,.034,.043
$(Y_0, Y_1, X) \sim$ **Normal**, $U \sim$ **Laplace**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.011, .052, .099	.012, .052, .100	.010, .054, .102	.011, .051, .098	.012, .048, .097	.011, .051, .104	.011,.032,.098
1	.012, .048, .098	.010, .052, .103	.007, .051, .107	.011, .049, .099	.012, .052, .100	.010, .051, .095	.015,.054,.095
2	.017, .034, .054	.016, .049, .095	.014, .051, .100	.023, .052, .089	.012, .049, .102	.020, .040, .065	.017,.033,.042
$(Y_0, Y_1, X) \sim$ **Laplace**, $U \sim$ **Normal**							
β	No Trim	TT(Z)	TT–BC(Z)	TT(X)	TT(X,$k_n^{(x)}$)	TT(X,k_n)	TT(Y)
.25	.01, .048, .097	.009, .050, .101	.007, .052, .105	.009, .051, .098	.011, .050, .101	.010, .049, .099	.010,.053,.103
1	.013, .021, .028	.013, .044, .084	.014, .054, .101	.022, .051, .084	.010, .051, .097	.014, .021, .029	.019,.030,.043
2	.014, .019, .023	.016, .041, .072	.018, .054, .093	.014, .019, .023	.021, .053, .090	.015, .026, .029	.022,.029,.039

The treatment assignment is $D = I(\alpha + \beta X > U)$ with $\alpha = 0$, hence Z has a symmetric distribution. The true propensity score $p(X)$ is used to compute Z. Values are rejection frequencies of the null hypothesis ATE = 0, at the 1%, 5%, 10% levels. "No Trim" is the untrimmed estimator $\tilde{\theta}_n$; "TT(Z)" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "TT–BC(Z)" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming. "TT(X)" is $\theta_n^{(tx)}$; and "TT(X,k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$. "TT(Y)" is $\hat{\theta}_n^{(ty)}$.

Table 3. Estimator Properties (Asymmetric Z, Known $p(X)$, $n = 100, 250$)

	$n = 100$										$n = 250$								
		$(Y_0, Y_1, X, U) \sim$ **Normal**				$(Y_0, Y_1, X, U) \sim$ **Laplace**					$(Y_0, Y_1, X, U) \sim$ **Normal**				$(Y_0, Y_1, X, U) \sim$ **Laplace**			
		$\beta = .25$ ($\kappa = 17$)				$\beta = .25$ ($\kappa = 5$)					$\beta = .25$ ($\kappa = 17$)				$\beta = .25$ ($\kappa = 5$)			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	.0019	.0016	.2074	.5265	.0009	.0033	.2254	.6493	0	-.0013	-.0019	.1315	.3786	-.0026	-.0036	.1439	.5307
TT-BC(Z)	1	.0019	.0033	.2058	.7577	.0010	.0006	.2175	.5957	.4	-.0011	.0001	.1303	.6102	-.0024	-.0018	.1383	.5104
TT(X,$k_n^{(x)}$)	43	.0023	.0023	.1549	.6590	-.0006	-.0009	.1894	.4123	36	-.0008	-.0011	.1020	.4841	-.0015	-.0022	.1237	.6083
TT(Y)	43	-.0088	-.0098	.2059	.3916	.0054	-.0002	.2261	.6080	36	-.0044	-.0028	.1301	.5234	-.0009	.0078	.1447	.8261
		$\beta = 1$ ($\kappa = 2$)				$\beta = 1$ ($\kappa = 2$)					$\beta = 1$ ($\kappa = 2$)				$\beta = 1$ ($\kappa = 2$)			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	.0045	.0029	.3581	6.455	.0041	.0026	.4771	.9200	0	-.0012	.0008	.2481	6.980	.0014	-.0017	.4005	13.72
TT-BC(Z)	1	.0050	.0074	.2155	.6294	.0037	.0006	.2378	.6198	.4	.0005	.0019	.1468	.5941	-.0005	.0002	.1630	.6294
TT(X,$k_n^{(x)}$)	43	.0028	.0033	.1636	.4708	.0018	.0002	.1986	.9803	36	.0009	.0003	.1130	.4695	-.008	.0005	.1365	.4900
TT(Y)	43	-.0131	-.0020	.3534	1.872	.0248	.0161	.4280	2.847	36	-.0065	-.0105	.2319	.2074	-.0153	-.0098	.2755	1.752
		$\beta = 2$ ($\kappa = 1.25$)				$\beta = 2$ ($\kappa = 1.5$)					$\beta = 2$ ($\kappa = 1.25$)				$\beta = 2$ ($\kappa = 1.5$)			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	.0048	.0001	.9474	21.81	.0101	.0044	.7679	15.48	0	-.0052	-.0019	.7042	20.82	-.0058	.0029	.7880	20.01
TT-BC(Z)	1	.0012	.0006	.2582	2.182	.0008	.0024	.2727	1.793	.4	-.0002	-.0001	.2202	1.786	.0026	.0011	.2731	.9982
TT(X,$k_n^{(x)}$)	43	.0004	.0009	.2161	2.603	.0014	.0029	.2428	1.778	36	-.0005	-.0012	.1602	1.862	.0020	.0010	.1731	1.055
TT(Y)	43	-.0546	-.0268	.9156	7.104	.2761	-.0071	5.578	12.20	36	.0203	.0062	1.932	10.01	-.0574	-.0152	.5498	4.984

		$(Y_0, Y_1, X) \sim$ **Norm,** $U \sim$ **Lap**				$(Y_0, Y_1, X) \sim$ **Lap,** $U \sim$ **Norm**					$(Y_0, Y_1, X) \sim$ **Norm,** $U \sim$ **Lap**				$(Y_0, Y_1, X) \sim$ **Lap,** $U \sim$ **Norm**			
		$\beta = .25$ ($\kappa = 17$)				$\beta = .25$ ($\kappa = 5$)					$\beta = .25$ ($\kappa = 17$)				$\beta = .25$ ($\kappa = 5$)			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0018	-.0016	.2076	.3912	.0023	.0020	.2157	.6367	0	-.0009	-.0021	.1287	.4928	-.0011	-.0011	.1365	.7204
TT-BC(Z)	1	-.0029	-.0044	.2089	.5164	.0033	.0043	.2082	.5288	.4	-.0008	-.0004	.1292	.4356	-.0013	-.0021	.1306	.9142
TT(X,$k_n^{(x)}$)	43	-.0013	-.0005	.1553	.5728	.0001	.0013	.1769	.7379	36	-.0004	-.0015	.1017	.6008	-.0006	-.0001	.1174	.6657
TT(Y)	43	.0125	.0138	.2000	.5619	-.0029	-.0103	.2379	.6255	36	.0096	.0118	.1300	.5567	-.0024	-.0120	.1413	.9829
		$\beta = 1$ ($\kappa = 2$)				$\beta = 1$ ($\kappa = 2$)					$\beta = 1$ ($\kappa = 2$)				$\beta = 1$ ($\kappa = 2$)			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0034	-.0042	.2661	1.131	.0014	-.0008	.7626	17.69	0	-.0029	-.0050	.1696	.9470	-.0089	-.0032	.7094	.2153
TT-BC(Z)	1	-.0026	-.0032	.2171	.4071	.0014	.0007	.2418	.7800	.4	-.0037	-.0036	.1465	.6245	-.0009	-.0003	.1744	.9857
TT(X,$k_n^{(x)}$)	43	-.0005	-.0009	.1610	.6371	-.0006	-.0006	.1976	.4325	36	-.0017	-.0026	.1095	.5110	-.0011	-.0030	.1328	.6897
TT(Y)	43	-.0148	-.0146	.2602	.5327	-.0663	.0052	1.474	9.239	36	-.0117	-.0134	.1626	.4112	-.0455	-.0027	1.456	10.22
		$\beta = 2$ ($\kappa = 1.25$)				$\beta = 2$ ($\kappa = 1.5$)					$\beta = 2$ ($\kappa = 1.25$)				$\beta = 2$ ($\kappa = 1.5$)			
Estimator	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$	Tr%	Mean	Med	MSE	KS$_{.05}$	Mean	Med	MSE	KS$_{.05}$
No Trim	0	-.0012	-.0046	.5675	11.22	-.0028	.0001	.8648	19.23	0	.0007	-.0003	.3290	7.072	.0117	.0019	1.088	25.13
TT-BC(Z)	1	.0008	.0001	.2734	1.231	-.0013	.0019	.2795	1.872	.4	.0005	.0030	.1735	.7289	-.0003	.0038	.2448	2.154
TT(X,$k_n^{(x)}$)	43	.0001	.0002	.1890	1.085	.0003	-.0010	.2759	2.754	36	-.0003	-.0005	.1317	.7510	.0013	.0049	.2168	3.345
TT(Y)	43	-.0098	.0218	.7151	5.099	-.0719	.0103	1.948	11.22	36	.0093	.0091	.3311	2.774	-.0505	-.0121	1.523	12.21

The treatment assignment is $D = I(.25 + \beta X > U)$, hence Z has an asymmetric distribution. The true propensity score $p(X)$ is used to compute Z. "No Trim" is the untrimmed estimator $\tilde{\theta}_n$; "TT(Z)" is the tail-trimmed estimator $\hat{\theta}_n^{(tz)}$ and "TT–BC(Z)" is the bias-corrected tail-trimmed $\hat{\theta}_n^{(tz:o)}$: both use *sample mean-centering* for trimming. "TT(X)" is $\theta_n^{(tx)}$; and "TT(X,k)" is the adaptive version $\hat{\theta}_n^{(tx)}$ of $\theta_n^{(tx)}$. "TT(Y)" is $\hat{\theta}_n^{(ty)}$. KS$_{.05}$ is the Kolmogorov-Smirnov test statistic divided by its 5% critical value: values above 1 indicate rejection of standard normality at the 5% level. Tr% is the percent of observations Z_i trimmed. κ is the tail index of $Z = h(X)Y$. Other than KS$_{.05}$, all values are averages over the randomly drawn 10,000 samples.