Supplementary Material
Robust Inference Using Inverse Probability Weighting

Xinwei Ma* Jingshen Wang†

August 16, 2019

Contents

*Department of Economics, University of California, San Diego.
†Division of Biostatistics, University of California, Berkeley.

I Preliminary Lemmas

I.1 Regular Variation

With finite second moments, weak convergence is not sensitive to delicate tail features. This is captured by the central limit theorem. However, weak convergence of sums of random variables without finite variance relies on additional tail properties. The appropriate notion is regular variation. In this subsection, we take X and Y as some generic univariate random variables, not necessarily the same as in the main paper.

Definition S.1 *A random variable X has regularly varying tail at ∞ with index $-\gamma < 0$, if for all $x > 0$, $\mathbb{P}[X > tx]/\mathbb{P}[X > t] \to x^{-\gamma}$ as $t \to \infty$. Similarly, X has regularly varying tail at $-\infty$ if for all $x > 0$, $\mathbb{P}[X < tx]/\mathbb{P}[X < t] \to x^{-\gamma}$ as $t \to -\infty$. Assume $\mathbb{P}[X > 0] = 1$, then it has regularly varying tail at 0 with index γ if $1/X$ has regularly varying tail at ∞ with index $-\gamma$.*

One special example of regular variation is "approximately polynomial tail": assume $\mathbb{P}[X > x] = c(x)x^{-\gamma}$ with $\gamma > 0$ and $c(x)$ tending to a strictly positive constant, then X has regularly varying tail at ∞ with index $-\gamma$. Following is a complete characterization of regular variation.

Lemma S.1 *Assume X has regularly varying tail at ∞ with index $-\gamma$, then for all x large enough,*

$$\mathbb{P}[X > x] = x^{-\gamma}c(x), \qquad \text{with } c(x) = L(x)\exp\left\{\int_s^x \frac{R(t)}{t}\mathrm{d}t\right\},$$

where $L(x)$ tends to a strictly positive constant, $\lim_{x\to\infty} R(x) = 0$, and s is some strictly positive constant.

If X has regularly varying right tail with index $-\gamma$, then it is clear that $\mathbb{E}[X^\alpha \mathbb{1}_{X>0}]$ exists and is finite for any $\alpha < \gamma$. However, the expectation will be infinite for all $\alpha > \gamma$. For the purpose of studying distributional convergence of sums of heavy-tailed random variables, a more thorough characterization of the truncated moment $\mathbb{E}[X^\alpha \mathbb{1}_{0<X<x}]$ is necessary.

Lemma S.2 *Assume X has a regularly varying right tail at ∞ with index $-\gamma$, then for any $\alpha > \gamma$,*

$$\frac{\mathbb{E}[X^\alpha \mathbb{1}_{0<X<x}]}{x^\alpha \mathbb{P}[X > x]} \to \frac{\gamma}{\alpha - \gamma}, \qquad \text{as } x \to \infty.$$

In the main paper, we take X to be the inverse probability weight multiplied by the binary indicator. However, the primary quantity of interest involves the outcome variable, and it is unclear how multiplication affects the tail behavior. The following lemma gives sufficient conditions under which the product XY has the same tail index as X. Despite being intuitive, it doesn't seem to be available in the literature.

Lemma S.3 *Assume X is nonnegative and has regularly varying tail with index $-\gamma$. Further assume (i) $\mathbb{E}[|Y|^\alpha | X = x]$ is uniformly bounded for some $\alpha > \gamma$, and (ii) there exists a distribution F, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y)|X = x] \to \int \ell(y)F(\mathrm{d}y)$ as $x \to \infty$. Then*

$$\lim_{x\to\infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \lim_{x\to\infty} \mathbb{E}[|Y|^\gamma \mathbb{1}_{Y>0}|X = x], \quad \lim_{x\to\infty} \frac{\mathbb{P}[XY < -x]}{\mathbb{P}[X > x]} = \lim_{x\to\infty} \mathbb{E}[|Y|^\gamma \mathbb{1}_{Y<0}|X = x].$$

Therefore the product XY has regularly varying right (resp. left) tail with index $-\gamma$, if $\lim_{x\to\infty} \mathbb{P}[Y > 0|X = x] > 0$ (resp. $\lim_{x\to\infty} \mathbb{P}[Y < 0|X = x] > 0$).

The first condition that $\mathbb{E}[|Y|^\alpha|X = x]$ is uniformly bounded is intuitive. To ensure the product that XY has the same tail behavior as X, one needs to assume that the tail of Y is thin enough.

In general, it is not possible to drop the second requirement that $Y|X=x$ converges in distribution, unless one is willing to impose additional structures on the conditional distribution. Following is a example, which shows that when the conditional distribution of Y "oscillates" as X tends to infinity, the product XY does not have a regularly varying tail even when Y is bounded.

Example S.1 Assume $Y=1$ for $X\in(2^j,2^{j+1}]$ for $j=1,3,5,\cdots$, and equals 0 otherwise, then on the grid $(2^j)_{j\geq 1}$, XY has right tail:

$$\mathbb{P}[XY>2^j]=\sum_{k=j,\ k\text{ odd}}^{\infty}F_X(2^{k+1})-F(2^k).$$

Now we take limit $j\to\infty$ along the sequence of odd numbers,

$$\lim_{j\to\infty,\ j\text{ odd}}\frac{\mathbb{P}[XY>2^j]}{\mathbb{P}[X>2^j]}=\lim_{j\to\infty,\ j\text{ odd}}\sum_{k=j,k\text{ odd}}^{\infty}\frac{F_X(2^{k+1})-F(2^k)}{\mathbb{P}[X>2^j]}=\left(1-2^{-\gamma}\right)\sum_{k=0}^{\infty}2^{-2k\gamma}=\frac{1-2^{-\gamma}}{1-2^{-2\gamma}}.$$

If we take the limit along the sequence of even numbers,

$$\lim_{j\to\infty,\ j\text{ even}}\frac{\mathbb{P}[XY>2^j]}{\mathbb{P}[X>2^j]}=\left(1-2^{-\gamma}\right)\sum_{k=1}^{\infty}2^{-2k\gamma}=2^{-2\gamma}\frac{1-2^{-\gamma}}{1-2^{-2\gamma}}.$$

Since X has regularly varying tail and the ratio $\mathbb{P}[XY>x]/\mathbb{P}[X>x]$ oscillates between two numbers, we conclude XY does not have regularly varying tail. ‖

I.2 Distributional Convergence

Assume $(X_{i,n})_{1\leq i\leq n,n\geq 1}$ is a triangular array, such that for each n, $(X_{i,n})_{1\leq i\leq n}$ are independently and identically distributed. The following lemma characterizes the asymptotic distribution of the sum $\sum_{i=1}^n X_{i,n}$ (assuming it exists).

Lemma S.4 *Assume $\mathbb{E}[X_{i,n}]=0$ for all n, and that the sum $\sum_{i=1}^n X_{i,n}$ converges in distribution. Then the limiting distribution has a characteristic function given by the canonical form:*

$$\psi(\zeta)=\exp\int_{\mathbb{R}}\frac{e^{i\zeta x}-1-i\zeta x}{x^2}M(\mathrm{d}x),$$

where M is a nonnegative measure satisfying (i) $M(I)<\infty$ for all bounded intervals I, and (ii) the integrals $\int_c^{\infty}x^{-1}M(\mathrm{d}x)$ and $\int_{-\infty}^{-c}x^{-1}M(\mathrm{d}x)$ are finite for all $c>0$.

The next lemma gives conditions under which the distributional convergence of the partial sum, $\sum_{i=1}^n X_{i,n}$, happens.

Lemma S.5 *Assume $\mathbb{E}[X_{i,n}]=0$ for all n, and let F_n be the distribution function of $X_{i,n}$. Then the sum $\sum_{i=1}^n X_{i,n}$ converges in distribution if and only if, for some measure M,*

$$n\mathbb{E}\left[X_{i,n}^2\mathbb{1}_{X_{i,n}\in I}\right]\to M(I)$$

for all compact intervals with $M(\partial I)=0$; and

$$n(1-F_n(c))\to\int_c^{\infty}x^{-2}M(\mathrm{d}x),\qquad nF_n(-c)\to\int_{-\infty}^{-c}x^{-2}M(\mathrm{d}x),$$

for all $c > 0$ with $M(\{c\}) = 0$. In this case, the limiting distribution is infinitely divisible, and its characteristic function is given by the form in Lemma S.4.

II Additional Results

In this section we provide additional results on (i) large sample properties of our local polynomial bias estimator, and (ii) generalizations of our IPW framework to provide robust inference for treatment effect estimands and parameters defined by nonlinear estimating equations.

II.1 Local Polynomial Regression

In the main paper, local polynomial regression is employed for estimating the trimming bias. To be more specific, the outcome variable is regressed on the probability weight in a region local to the origin. That is,

$$\hat{\beta} = \Big[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\Big]' = \underset{\beta_0,\beta_1,\cdots,\beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} D_i \Big[Y_i - \sum_{j=0}^{p} \beta_j e(X_i)^j\Big]^2 \mathbb{1}_{e(X_i)\le h_n},$$

where for ease of exposition we assume that the true probability weights are used. The following lemma characterizes the properties of the local polynomial estimates.

Lemma S.6 *Assume Assumption 1 and 2 hold. In addition, assume (i) $\mu_1(\cdot)$ is $p+1$ times continuously differentiable; (ii) $\mu_2(0) - \mu_1(0)^2 > 0$; and (iii) the bandwidth sequence satisfies $nh_n\mathbb{P}[e(X) \le h_n] \to \infty$ and $nh_n^{2p+3}\mathbb{P}[e(X) \le h_n] = O(1)$. Let $\beta = \Big[\mu_1(0), \mu_1^{(1)}(0), \cdots, \frac{1}{p!}\mu_1^{(p)}(0)\Big]'$ and $\hat{\beta}$ be defined in the above, then*

$$\sqrt{nh_n\mathbb{P}[e(X) \le h_n]} H_n \left(\hat{\beta} - \beta - h_n^{p+1} H_n^{-1} \frac{\mu_1^{(p+1)}(0)}{(p+1)!} S^{-1} R\right) \rightsquigarrow \mathcal{N}\Big(0,\ (\mu_2(0) - \mu_1(0)^2) S^{-1}\Big),$$

where $H_n = \text{diagonal}(1, h_n, h_n^2, \cdots, h_n^p)$, $S = (s_{ij})_{1\le i,j\le p}$ with $s_{ij} = (\gamma_0 - 1)/(\gamma_0 + i + j - 2)$, and $R = (r_i)_{1\le i\le p}$ with $r_i = (\gamma_0 - 1)/(\gamma_0 + i + p)$.

II.2 Treatment Effect Estimation

In this subsection, we extend the IPW framework to provide robust inference for treatment effect estimands when the probability weights can be close to zero and one. Let the binary indicator denote a treatment status: $D = 1$ for the treatment group and 0 for the control group. The corresponding potential outcomes are denoted by $Y(1)$ and $Y(0)$, respectively. The observed outcome is $Y = DY(1) + (1 - D)Y(0)$. We assume that, conditional on the covariates X, potential outcomes $(Y(1), Y(0))$ and the treatment status D are independent. Following the convention in the literature, we use the terminology "propensity score" rather than probability weight. We ignore the issue of using estimated propensity scores for ease of exposition (see Section 2.3 for discussions).

Treatment Effect on the Treated (ATT)

We first consider the treatment effect on the treated estimand: $\tau_0^{\texttt{ATT}} = \mathbb{E}[Y(1) - Y(0)|D = 1]$. Both Assumption 1 and 2 can be modified in a straightforward way.

Assumption ATT *(i) For some $\gamma_0 > 1$, the propensity score has regularly varying tail with index $\gamma_0 - 1$ at 1:*

$$\lim_{t\downarrow 0} \frac{\mathbb{P}[1 - e(X) \le tx]}{\mathbb{P}[1 - e(X) \le t]} = x^{\gamma_0 - 1}, \qquad \text{for all } x > 0.$$

(ii) For some $\varepsilon > 0$, $\mathbb{E}\big[|Y(0) + Y(1)|^{(\gamma_0 \vee 2)+\varepsilon} \big| e(X) = x\big]$ is uniformly bounded. There exists a probability distribution $F_{(0)}$, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y(0))|e(X) = x] \to \int_{\mathbb{R}} \ell(y) F_{(0)}(\mathrm{d}y)$ as $x \uparrow 1$.

Using inverse probability weighting, a natural estimator of $\tau_0^{\mathtt{ATT}}$ is

$$\hat{\tau}_{n,b_n}^{\mathtt{ATT}} = \frac{1}{n_1} \sum_{i=1}^{n} \left[D_i Y_i - \frac{e(X_i)}{1 - e(X_i)} (1 - D_i) Y_i \mathbb{1}_{1-e(X_i) \ge b_n} \right] = \frac{1}{n} \sum_{i=1}^{n} \frac{(D_i - e(X_i)) Y_i}{\hat{\mathbb{P}}[D = 1](1 - e(X_i))} \mathbb{1}_{1-e(X_i) \ge (1-D_i) b_n},$$

where $n_1 = \sum_{i=1}^{n} D_i$ is size of the treatment group, and $\hat{\mathbb{P}}[D = 1] = n_1/n$. It should be clear that propensity scores that are close to 1 will pose a challenge to both estimation and inference. The following proposition characterizes the large sample properties of $\hat{\tau}_{n,b_n}^{\mathtt{ATT}}$.

Proposition S.1 (Asymptotic Distribution of the ATT Estimator) *Assume Assumption ATT holds, $b_n \to 0$, and $\alpha_{(0),+}(0) + \alpha_{(0),-}(0) > 0$, where*

$$\alpha_{(0),+}(x) = \lim_{t \to 1} \mathbb{E}\Big[|Y(0)|^{\gamma_0} \mathbb{1}_{Y(0) > x} \Big| e(X) = t \Big], \ \alpha_{(0),-}(x) = \lim_{t \to 1} \mathbb{E}\Big[|Y(0)|^{\gamma_0} \mathbb{1}_{Y(0) < x} \Big| e(X) = t \Big].$$

Let a_n be defined from

$$\frac{n}{a_n^2} \mathbb{E}\left[\left| \frac{(D - e(X))Y}{\mathbb{P}[D = 0](1 - e(X))} - \tau_0^{\mathtt{ATT}} \right|^2 \mathbb{1}\left(\left| \frac{(D - e(X))Y}{\mathbb{P}[D = 0](1 - e(X))} \right| \le a_n \right) \right] \to 1.$$

(i) If $\gamma_0 \ge 2$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}} (\hat{\tau}_{n,b_n}^{\mathtt{ATT}} - \tau_0^{\mathtt{ATT}} - \mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution. (ii.1) No trimming, light trimming and moderate trimming: if $\gamma_0 < 2$ and $b_n a_n \to t \in [0, \infty)$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}} (\hat{\tau}_{n,b_n}^{\mathtt{ATT}} - \tau_0^{\mathtt{ATT}} - \mathsf{B}_{n,b_n})$ converges in distribution, with the asymptotic characteristic function given by

$$\psi(\zeta) = \exp\left\{ \int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x) \right\},$$
$$\text{where } M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2 - \gamma_0}{\alpha_{(0),+}(0) + \alpha_{(0),-}(0)} |x|^{1-\gamma_0} \Big(\alpha_{(0),+}(-tx) \mathbb{1}_{x<0} + \alpha_{(0),-}(-tx) \mathbb{1}_{x \ge 0} \Big) \right].$$

(ii.2) Heavy trimming: if $\gamma_0 < 2$ and $b_n a_n \to \infty$, let $a_{n,b_n} = \sqrt{n \mathbb{V}[\frac{(D-e(X))Y}{\mathbb{P}[D=1](1-e(X))} \mathbb{1}_{1-e(X) \ge (1-D) b_n}]}$, then $\frac{n}{a_{n,b_n}} (\hat{\tau}_{n,b_n}^{\mathtt{ATT}} - \tau_0^{\mathtt{ATT}} - \mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution.

For the trimmed ATT estimator (i.e., $b_n > 0$), observations from the control group with propensity scores above $1 - b_n$ are discarded, and it can be shown that the trimming bias is

$$\mathsf{B}_{n,b_n} = \frac{1}{\mathbb{P}[D = 1]} \mathbb{E}\Big[e(X) \mathbb{E}[Y(0)|e(X)] \mathbb{1}_{e(X) \ge 1 - b_n} \Big].$$

To implement bias correction, one first regresses the outcome variable on a p-th order polynomial of the propensity score, using only observations from the control group:

$$\Big[\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p \Big]' = \underset{\beta_0, \beta_1, \cdots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^{n} (1 - D_i) \Big[Y_i - \sum_{j=0}^{p} \beta_j e(X_i)^j \Big]^2 \mathbb{1}_{e(X_i) \ge 1 - h_n}.$$

Then the bias is estimated by

$$\hat{\mathsf{B}}_{n,b_n} = \frac{1}{n_1}\sum_{i=1}^{n}\sum_{j=0}^{p}\hat{\beta}_j e(X_i)^{j+1}\mathbb{1}_{e(X_i)\geq 1-b_n}.$$

Average Treatment Effect (ATE)

The average treatment effect, $\tau_0^{\mathtt{ATE}} = \mathbb{E}[Y(1)-Y(0)]$, is another commonly employed treatment effect estimand. Because both small and large propensity scores can lead to "small denominators," Assumptions 1 and 2 have to be properly modified. To be specific, we require

Assumption ATE *(i) For some $\gamma_0 > 1$ and $\omega \in [0,1]$,*

$$\lim_{t\downarrow 0}\frac{\mathbb{P}[e(X)\leq t]}{\mathbb{P}[e(X)\leq t]+\mathbb{P}[1-e(X)\leq t]} = \omega,$$
$$\text{and}\quad \lim_{t\downarrow 0}\frac{\mathbb{P}[e(X)\leq tx]+\mathbb{P}[1-e(X)\leq tx]}{\mathbb{P}[e(X)\leq t]+\mathbb{P}[1-e(X)\leq t]} = x^{\gamma_0-1}, \qquad \text{for all } x>0.$$

(ii) For some $\varepsilon > 0$, $\mathbb{E}[|Y(1)+Y(0)|^{(\gamma_0\vee 2)+\varepsilon}|e(X)=x]$ is uniformly bounded. Further, there exist probability distributions, $F_{(1)}$ and $F_{(0)}$, such that for all bounded and continuous $\ell(\cdot)$, $\mathbb{E}[\ell(Y(1))|e(X)=x]\to\int\ell(y)F_{(1)}(\mathrm{d}y)$ and $\mathbb{E}[\ell(Y(0))|e(X)=1-x]\to\int\ell(y)F_{(0)}(\mathrm{d}y)$ as $x\downarrow 0$.

Note that in part (i), we do not require the two tails of the propensity score having the same index, since it is possible to have $\omega = 0$ or 1. Asymptotically, the heavier tail "wins." Part (ii) takes into account that both potential outcomes can affect the tail behavior of the estimator. The following is a natural estimator of ATE using inverse probability weighting:

$$\begin{aligned}\hat{\tau}_{n,b_n}^{\mathtt{ATE}} &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_iY_i}{e(X_i)}\mathbb{1}_{e(X_i)\geq b_n} - \frac{(1-D_i)Y_i}{1-e(X_i)}\mathbb{1}_{e(X_i)\leq 1-b_n}\right]\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{(2D_i-1)Y_i}{1-D_i+(2D_i-1)e(X_i)}\mathbb{1}_{1-D_i+(2D_i-1)e(X_i)\geq b_n}. \qquad \text{(S.1)}\end{aligned}$$

For ATE estimation, trimming can lead to further complications beyond affecting the limiting distribution and introducing a bias: different trimming thresholds can be applied to the treatment and control groups. For the treatment group ($D=1$), it is natural to discard observations with small propensity scores, while for the control group ($D=0$) observations with large propensity scores will be dropped. To see how having two trimming thresholds can complicate the asymptotic analysis, assume $\omega = 1$ so that the propensity score has a heavier left tail, and Proposition S.2 essentially reduces to Theorem 1. When different trimming thresholds are applied to small and large propensity scores in the treatment and control groups, however, the relative magnitude of the two tails can be overturned. To see this, consider the extreme scenario where fixed trimming is applied to the treatment group but no trimming (or light trimming) for the control group. Then the trimmed ATE estimator will be greatly influenced by the relatively heavier right tail of the propensity score (i.e., "small denominators" in the $D=0$ subsample).

To avoid cumbersome notation and lengthy discussions on each possible scenarios, we instead focus on the "symmetric trimming" in (S.1), which is easy to analyze and implement, but employing different trimming thresholds is also justified in practice. As discussed, trimming introduces a bias which is generally non-negligible. For estimating the ATE, however, it is possible to achieve "small bias" by choosing the two trimming thresholds appropriately. To see this, the trimming bias in (S.1) is $\mathsf{B}_{n,b_n} = \mathbb{E}[\mathbb{E}[Y(0)|e(X)]\mathbb{1}_{e(X)\geq 1-b_n} -$

$\mathbb{E}[Y(1)|e(X)]\mathbb{1}_{e(X)\leq b_n}] \approx \mathbb{E}[Y(0)|e(X)=1]\mathbb{P}[e(X)\geq 1-b_n] - \mathbb{E}[Y(1)|e(X)=0]\mathbb{P}[e(X)\leq b_n]$. Assuming that the propensity score has similar tails at the two ends and that the two conditional expectations have the same sign and magnitude, then it is possible to use different trimming thresholds so that the two components in the bias formula cancel each other. However, this strategy is not always feasible, especially when the two tails behave very differently.

Assumption ATE suffices to characterize the tail of $\frac{(2D-1)Y}{(1-D+(2D-1)e(X))}$. For next result, let

$$\alpha_{(1),+}(x) = \lim_{t\to 0}\mathbb{E}\Big[|Y(1)|^{\gamma_0}\mathbb{1}_{Y(1)>x}\Big|e(X)=t\Big], \quad \alpha_{(1),-}(x) = \lim_{t\to 0}\mathbb{E}\Big[|Y(1)|^{\gamma_0}\mathbb{1}_{Y(1)<x}\Big|e(X)=t\Big],$$

and re-define $\alpha_+(x)$ and $\alpha_-(x)$ as

$$\alpha_+(x) = \omega\alpha_{(1),+}(x) + (1-\omega)\alpha_{(0),-}(-x), \quad \alpha_-(x) = \omega\alpha_{(1),-}(x) + (1-\omega)\alpha_{(0),+}(-x).$$

Proposition S.2 (Asymptotic Distribution of the ATE Estimator) *Assume Assumption ATE holds, $b_n \to 0$, and $\alpha_{(0),+}(0) + \alpha_{(0),-}(0) > 0$. Let a_n be defined from*

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{(2D-1)Y}{1-D+(2D-1)e(X)} - \theta_0\right|^2 \mathbb{1}\left(\left|\frac{(2D-1)Y}{1-D+(2D-1)e(X)}\right| \leq a_n\right)\right] \to 1.$$

(i) If $\gamma_0 \geq 2$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATE}}_{n,b_n} - \tau^{\mathtt{ATE}}_0 - \mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution.
(ii.1) No trimming, light trimming and moderate trimming: if $\gamma_0 < 2$ and $b_n a_n \to t \in [0,\infty)$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATE}}_{n,b_n} - \tau^{\mathtt{ATE}}_0 - \mathsf{B}_{n,b_n})$ converges in distribution, with the asymptotic characteristic function given by

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\zeta x} - 1 - i\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\text{where } M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(tx)\mathbb{1}_{x\geq 0} + \alpha_-(tx)\mathbb{1}_{x<0}\Big)\right].$$

(ii.2) Heavy trimming: if $\gamma_0 < 2$ and $b_n a_n \to \infty$, let $a_{n,b_n} = \sqrt{n\mathbb{V}[\frac{(2D-1)Y}{(1-D+(2D-1)e(X))}\mathbb{1}_{1-D+(2D-1)e(X)\geq b_n}]}$, then $\frac{n}{a_{n,b_n}}(\hat{\tau}^{\mathtt{ATE}}_{n,b_n} - \tau^{\mathtt{ATE}}_0 - \mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution.

Bias correction can be implemented according to Algorithm 1 with a straightforward modification: one first runs two local polynomial regressions, one for the treatment group and the other for the control group:

$$\Big[\hat{\beta}^{\mathtt{l}}_0, \hat{\beta}^{\mathtt{l}}_1, \cdots, \hat{\beta}^{\mathtt{l}}_p\Big]' = \operatorname*{argmin}_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^n D_i\Big[Y_i - \sum_{j=0}^p \beta_j e(X_i)^j\Big]^2 \mathbb{1}_{e(X_i)\leq h_n}$$
$$\Big[\hat{\beta}^{\mathtt{r}}_0, \hat{\beta}^{\mathtt{r}}_1, \cdots, \hat{\beta}^{\mathtt{r}}_p\Big]' = \operatorname*{argmin}_{\beta_0,\beta_1,\cdots,\beta_p} \sum_{i=1}^n (1-D_i)\Big[Y_i - \sum_{j=0}^p \beta_j e(X_i)^j\Big]^2 \mathbb{1}_{e(X_i)\geq 1-h_n}.$$

Then the bias is estimated by

$$\hat{\mathsf{B}}_{n,b_n} = \frac{1}{n}\sum_{i=1}^n\sum_{j=0}^p \Big(\hat{\beta}^{\mathtt{r}}_j \mathbb{1}_{e(X_i)\geq 1-b_n} - \hat{\beta}^{\mathtt{l}}_j \mathbb{1}_{e(X_i)\leq b_n}\Big) e(X_i)^j.$$

We assume the same bandwidth h_n is used for the two local polynomial regressions for simplicity, although in practice different bandwidths can be employed.

Finally, we compare the trimmed ATE estimator $\hat{\tau}^{\mathtt{ATE}}_{n,b_n}$ in (S.1) to another commonly used trimming strategy.

Trimming in (S.1) can be understood as "discarding observations with small denominators." It is different, however, from

$$\check{\tau}^{\mathtt{ATE}}_{n,b_n} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_iY_i}{e(X_i)} - \frac{(1-D_i)Y_i}{1-e(X_i)}\right]\mathbb{1}_{b_n\leq e(X_i)\leq 1-b_n},$$

which "discards observations with small or large propensity scores." To see how bias correction can be conducted for $\check{\tau}^{\mathtt{ATE}}_{n,b_n}$, we note that its bias has four terms:

$$\begin{aligned}
&\mathsf{B}_{n,b_n,1}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \mathbb{E}[Y_i|D_i=0, e(X_i)\leq b_n], && \mathsf{B}_{n,b_n,2}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \mathbb{E}[Y_i|D_i=1, e(X_i)\leq b_n]\\
&\mathsf{B}_{n,b_n,3}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \mathbb{E}[Y_i|D_i=0, e(X_i)\geq 1-b_n], && \mathsf{B}_{n,b_n,4}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \mathbb{E}[Y_i|D_i=1, e(X_i)\geq 1-b_n].
\end{aligned}$$

Term $\mathsf{B}_{n,b_n,1}(\check{\tau}^{\mathtt{ATE}}_{n,b_n})$ and $\mathsf{B}_{n,b_n,4}(\check{\tau}^{\mathtt{ATE}}_{n,b_n})$ can be easily estimated by a sample average without introducing small denominators, as

$$\hat{\mathsf{B}}_{n,b_n,1}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(1-D_i)Y_i}{1-e(X_i)}\mathbb{1}_{e(X_i)\leq b_n}, \qquad \hat{\mathsf{B}}_{n,b_n,4}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \frac{1}{n}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i)}\mathbb{1}_{e(X_i)\geq 1-b_n},$$

and indeed, $\check{\tau}^{\mathtt{ATE}}_{n,b_n} + \hat{\mathsf{B}}_{n,b_n,1}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) + \hat{\mathsf{B}}_{n,b_n,4}(\check{\tau}^{\mathtt{ATE}}_{n,b_n}) = \hat{\tau}^{\mathtt{ATE}}_{n,b_n}$. To correct for the other bias terms, $\mathsf{B}_{n,b_n,2}(\check{\tau}^{\mathtt{ATE}}_{n,b_n})$ and $\mathsf{B}_{n,b_n,3}(\check{\tau}^{\mathtt{ATE}}_{n,b_n})$, it will require the use of our local polynomial bias correction technique.

Whether the researcher should choose $\check{\tau}^{\mathtt{ATE}}_{n,b_n}$ without bias correction (and hence reinterpret the parameter) or employ $\hat{\tau}^{\mathtt{ATE}}_{n,b_n}$ with bias correction depends on the specific dataset and the distribution of the propensity score (or, equivalently, the covariates distributions for the control and treatment group). For example, there might be a spike very close to zero (or one) in the propensity score distribution, or that the propensity score can actually be zero (or one). In either case, bias correction requires extrapolating a local polynomial regression, and hence may not be very reliable.

II.3 General Estimating Equation

We employ the same notation used in Section 1 and 2 of the main paper. Instead of focusing on a population mean, the parameter θ_0 is defined by a possibly nonlinear moment condition $\mathbb{E}[\mu_1(e(X),\theta_0)] = 0$, where $\mu_1(e(X),\theta) = \mathbb{E}[g(Y,X,\theta)|e(X), D=1]$ and g is a known function. Alternatively, we have $\mathbb{E}[Dg(Y_i,X_i,\theta_0)/e(X)] = 0$. For ease of exposition, we assume that both the parameter and the moment condition are univariate. To estimate θ_0, one can solve the following sample analogue:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{D_ig(Y_i,X_i,\hat{\theta}_{n,b_n})}{e(X_i)}\mathbb{1}_{e(X_i)\geq b_n}.$$

As long as the trimming threshold b_n shrinks to zero as the sample size increases, the trimmed estimator $\hat{\theta}_{n,b_n}$ will be consistent for θ_0 under mild regularity conditions (for example, by employing a uniform law of large numbers argument, such as Newey and McFadden 1994). Assuming this is the case, we can employ a Taylor expansion and linearize the estimator:

$$\begin{aligned}
&\frac{n}{a_{n,b_n}}(\hat{\theta}_{n,b_n} - \theta_0 - \Sigma_0\mathsf{B}_{n,b_n}) = \frac{\Sigma_0}{a_{n,b_n}}\sum_{i=1}^{n}\left[\frac{D_iG_i}{e(X_i)}\mathbb{1}_{e(X_i)\geq b_n} - \mathsf{B}_{n,b_n}\right] + o_{\mathrm{p}}(1),\\
&\Sigma_0 = \left(-\mathbb{E}\left[\frac{\partial}{\partial\theta}\mu_1(e(X),\theta_0)\right]\right)^{-1}, \qquad \mathsf{B}_{n,b_n} = -\mathbb{E}[\mu_1(e(X),\theta_0)\mathbb{1}_{e(X)\leq b_n}].
\end{aligned} \tag{S.2}$$

The bias term B_{n,b_n} only represents the leading bias in an asymptotic linear expansion, with higher order bias absorbed into the $o_{\mathrm{p}}(1)$ term. The bias arises because after trimming the estimating equation may not have a zero mean in finite samples. Assuming $\mu_1(\cdot)$ is continuous in its first argument, the bias can be further simplified as $\mathsf{B}_{n,b_n} = -\mu_1(0,\theta_0)\mathbb{P}[e(X) \leq b_n]$, which gives its precise order. From this, one can immediately see that if $\mu_1(x,\theta_0) = 0$ for all x small enough, trimming does not induce any bias, and at the same time can improve the performance of the IPW estimator. Such "small bias" scenario, however, is difficult to justify in practice because it requires that the information provided by observations with small probability weights does not feature in the estimating equation.

Once the estimator has been linearized as above, we can prove a result similar to Theorem 1. To economize notation, define the random variables $G_i(\theta) = g(Y_i, X_i, \theta)$ and $G_i = G_i(\theta_0)$. We make the following assumption.

Assumption GEE *(i) θ_0 is the unique root of $\mathbb{E}[\mu_1(e(X),\theta)] = 0$ in the interior of a compact parameter space Θ. (ii) $g(Y,X,\theta)$ is continuously differentiable in θ, and $\mathbb{E}[\sup_{\theta\in\Theta}|g(Y_i,X_i,\theta)| \vee |\frac{\partial}{\partial\theta}g(Y_i,X_i,\theta)|] < \infty$. (iii) For some $\varepsilon > 0$, $\mathbb{E}[|G|^{(\gamma_0\vee 2)+\varepsilon}|e(X) = x, D = 1]$ is uniformly bounded. There exists a probability distribution F, such that for any bounded and continuous function ℓ, $\mathbb{E}[\ell(G)|e(X) = x, D = 1] \to \int_{\mathbb{R}} \ell(y)F(\mathrm{d}y)$ as $x \downarrow 0$.*

The following proposition characterizes the large-sample properties of the IPW-based GEE estimator.

Proposition S.3 (Asymptotic Distribution of the GEE Estimator) *Assume Assumptions 1 and GEE hold, $b_n \to 0$, and $\alpha_{G,+}(0) + \alpha_{G,-}(0) > 0$, where*

$$\alpha_{G,+}(x) = \lim_{t\to 0}\mathbb{E}\Big[|G|^{\gamma_0}\mathbb{1}_{G>x}\Big|e(X) = t, D = 1\Big], \quad \alpha_{G,-}(x) = \lim_{t\to 0}\mathbb{E}\Big[|G|^{\gamma_0}\mathbb{1}_{G<x}\Big|e(X) = t, D = 1\Big].$$

Let a_n be such that

$$\frac{n}{a_n^2}\mathbb{E}\left[\left|\frac{DG}{e(X)}\right|^2 \mathbb{1}_{|DG/e(X)|\leq a_n}\right] \to 1.$$

(i) If $\gamma_0 \geq 2$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat\theta_{n,b_n} - \theta_0 - \Sigma_0\mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution. (ii.1) No trimming, light trimming and moderate trimming: if $\gamma_0 < 2$ and $b_n a_n \to t \in [0,\infty)$, let $a_{n,b_n} = a_n$, then $\frac{n}{a_{n,b_n}}(\hat\theta_{n,b_n} - \theta_0 - \Sigma_0\mathsf{B}_{n,b_n})$ converges in distribution, with the asymptotic characteristic function given by

$$\psi(\zeta) = \exp\left\{\int_{\mathbb{R}} \frac{e^{i\Sigma_0\zeta x} - 1 - i\Sigma_0\zeta x}{x^2} M(\mathrm{d}x)\right\},$$
$$\text{where } M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_{G,+}(0)+\alpha_{G,-}(0)}|x|^{1-\gamma_0}\Big(\alpha_{G,+}(tx)\mathbb{1}_{x\geq 0} + \alpha_{G,-}(tx)\mathbb{1}_{x<0}\Big)\right].$$

(ii.2) Heavy trimming: if $\gamma_0 < 2$ and $b_n a_n \to \infty$, let $a_{n,b_n} = \sqrt{n\mathbb{V}[DG/e(X)\mathbb{1}_{e(X)\geq b_n}]}$, then $\frac{n}{a_{n,b_n}}(\hat\theta_{n,b_n} - \theta_0 - \Sigma_0\mathsf{B}_{n,b_n})$ converges to the standard Gaussian distribution.

Proposition S.3 can be further generalized to a vector-valued parameter. As long as the moment condition permits identification (and consistent estimation), one can employ the Cramér-Wold device to characterize the limiting distribution.

Selecting the trimming threshold is more complicated, since now the conditional first and second moment cannot be estimated directly. It is possible to employ a three-step procedure. In the first step, one constructs a pilot point estimate. Next, one can estimate the conditional moments applying local polynomial regression, with $G_i(\hat\theta_{n,b_n})$ being the dependent variable. In the final step, the trimming threshold is chosen by plugging the second-step estimated conditional moments into the procedure of Theorem 2.

As a final remark, bias correction is still feasible in this setting by exploiting the asymptotic linear representation in (S.2). To form the bias estimate, one can employ the local polynomial regression technique and regress $G_i(\hat{\theta}_{n,b_n})$ on the probability weights to form an estimate of the bias B_{n,b_n} (Algorithm 1). Then a bias estimate can be constructed as $\hat{\Sigma}_n \mathsf{B}_{n,b_n}$, where $\hat{\Sigma}_n$ estimates Σ_0 as by a sample average.

III Simulation Evidence

In our main simulation design, the probability weight is distributed according to $\mathbb{P}[e(X) \leq x] = x^{\gamma_0 - 1}$ with $\gamma_0 \in \{1.3, 1.5, 1.9\}$. A typical realization with $\gamma_0 = 1.5$ is given in Figure S.1. With $\gamma_0 = 1.5$, the convergence rate of the IPW estimator is $n^{1/3}$. Conditional on the weight and $D = 1$, the outcome variable is generated as $\mu_1(e(X)) + \eta$, where the mean equation is either $\cos(2\pi e(X))$ or $1 - e(X)$, and the error η follows a chi-square distribution with four degrees of freedom, and is centered and scaled to have zero mean and unit variance. In the first specification, the conditional mean function is nonlinear, and a typical realization of the outcome variable is given in Section 3, Figure 1. In the second specification, the leading bias remains the same, but the conditional mean function is linear in the probability weight. Our bias correction technique is therefore expected to perform well.

Throughout, we use 5,000 Monte Carlo repetitions, and for each repetition, 1,000 subsampling iterations are used with subsample size $m = \lfloor n/\log(n) \rfloor$, and the full sample size is $n \in \{2,000,\ 5,000,\ 10,000\}$. We follow Theorem 2 to set the trimming threshold, by solving $\hat{b}_n^s \hat{\mathbb{P}}[e(X_i) \leq \hat{b}_n] = (2n)^{-1}$ with $s \in \{1,\ 1.5,\ 2,\ 3\}$. For $s = 1$, the trimming threshold is rate optimal (in terms of the leading mean squared error) and corresponds to moderate trimming. The other cases fall into the heavy trimming category. Bias correction is based on Algorithm 1 with local linear regression.

The first set of simulation results are collected in Table S.1 and S.2. Under "Conventional" we report bias, standard deviation and root mean squared error of the IPW estimator, both with and without trimming. Note that they have been scaled by $n^{1-1/\gamma_0} = n^{1/3}$. We also report empirical coverage of the conventional Gaussian-based confidence interval under "cov," $[\hat{\theta}_{n,b_n} \pm 1.96 \cdot S_{n,b_n}/\sqrt{n}]$. Average confidence interval length is reported under "|ci|," scaled by $n^{1-1/\gamma_0} = n^{1/3}$. We also include the trimming strategy proposed by Crump *et al.* (2009), and report the performance of the point estimate and the associated Gaussian-based confidence interval. Under "Robust" we report bias, standard deviation, and root mean squared error of the bias-corrected IPW estimator, $\hat{\theta}^{\mathsf{bc}}_{n,b_n}$ (Algorithm 1). Note that without trimming (the first row of each table), the bias correction term is exactly zero, which is why the bias, standard deviation, and root mean squared error remain the same. Under "cov" we report empirical coverage of the subsampling-based confidence interval (Algorithm 2). Also reported is the average length of the subsampling-based confidence interval under "|ci|." In the following, we highlight several observations from Table S.1.

First, inference based on the Gaussian approximation performs poorly, as predicted by our theoretical results. Without trimming, the limiting distribution of the IPW estimator is heavy-tailed (Theorem 1(ii.1)), and hence using critical values computed from Gaussian quantiles lead to confidence intervals that are overly optimistic/narrow. Although heavy trimming can help restore asymptotic Gaussianity (Theorem 1(ii.2)), it is unclear how well distributional approximation based on this result performs in samples of moderate size. In addition, trimming introduces a bias that can significantly shift the limiting distribution away from the target parameter. Indeed, in a sample of size $2,000$, using 0.1 as the trimming threshold will lead to a bias that is so severe that a nominal 95% confidence interval will have almost zero coverage: the researcher essentially changes the target estimand.

Second, it is not surprising that employing a larger trimming threshold can help stabilize the estimator,

leading to a smaller empirical standard deviation. However, the mean squared error increases due to the trimming bias. In addition, by comparing the scaled bias across the three panels in Table S.1, it is clear that the bias is explosive when heavy trimming is used. We also employ the trimming strategy of Crump *et al.* (2009). Since their method is based on minimizing an asymptotic variance term, it is not surprising that it can lead to a relatively large trimming threshold, which in turn implies a large trimming bias.

Third, despite the fact that the conditional mean function is highly nonlinear, our bias correction procedure successfully removes most of the bias, making the subsampling-based confidence interval having an empirical coverage very close to the 95% nominal level. The performance of our inference procedure is quite robust across a range of trimming threshold choices. For the very heavy trimming case, under-coverage remains to be an issue even with bias correction, because it is quite difficult to estimate a nonlinear function local to a point where observations are scarce. In addition, bias correction may introduce extra variability in samples of moderate size. This is again confirmed by our simulation results, and is why we recommend conduct bias correction not only for the main estimator but also in each subsampling iteration.

For the untrimmed IPW estimator (the first row of each table), coverage of the subsampling-based confidence interval is still not very close to the nominal 95% level, and it tends to be wide. Having a wide confidence interval in this case is unavoidable: with small denominators entering the IPW estimator and no trimming, the asymptotic distribution is Lévy stable, which is heavy-tailed. As for the unsatisfactory coverage, it is recognized in the literature that conducting inference for the mean of heavy-tailed random variables is difficult, and coverage of subsampling-based confidence intervals may not be very close to the nominal level. Compared to the untrimmed IPW estimator, the subsampling-based confidence interval using the bias-corrected and trimmed IPW estimator performs much better. Therefore, with small denominators entering the IPW estimator, we recommend employ some degree of trimming according to Theorem 2, and conduct bias correction.

Now we consider how the form of the conditional mean function affects the performance of our procedure. In Table S.2, the conditional mean is a linear function of the probability weight. If this is known a priori, a better estimation strategy is to fit a global linear regression and extrapolate to observations with small probability weights. Such regression-based estimator will converge at the $\sqrt{n}$-rate and will be asymptotically Gaussian. In practice, however, the shape of the conditional mean function is rarely known, so the setting in Table S.2 is best understood as a favorable situation in which our bias correction and inference procedure are expected to perform well. Indeed, the remaining bias is almost zero.

We also provide finite sample comparisons in simulation studies through Table S.3-S.4 for $\gamma_0 = 1.3$ and in Table S.5-S.6 for $\gamma_0 = 1.9$. Encouragingly, we can reach similar conclusions. Meanwhile, we observe that the trimming threshold $\hat{b}_n$ increases with the tail index γ_0, which is in line with our analysis in Theorem 1. Additionally, we see that the effective number of trimmed observations decreases with γ_0. Indeed, with γ_0 closer to 2, the probability weights have a lighter tail at zero, and thus it is sensible that only a smaller fraction of observations needs to be trimmed to achieve desired properties.

IV Proofs

IV.1 Proof of Lemma S.1

See Theorem VIII.9.1 and the corresponding corollary in Feller (1991). ■

IV.2 Proof of Lemma S.2

See Theorem VIII.9.2 in Feller (1991). ■

IV.3 Proof of Lemma S.3

We split the proof into three parts.

Part 1

We first assume X and Y are independent. For simplicity, we denote by F_X and F_Y the distribution functions of X and Y, and $\varepsilon = \alpha - \gamma > 0$. Define $a(y,x)$ be

$$a(y,x) = \frac{1 - F_X(x/y)}{1 - F_X(x)}.$$

Then from the definition of regularly varying functions, one has $\lim_{x\to\infty} a(x,y) = y^\gamma$ for all $y > 0$. Consider the following limit:

$$\lim_{x\to\infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \lim_{x\to\infty} \int_0^\infty a(y,x) F_Y(\mathrm{d}y) = \underbrace{\lim_{x\to\infty} \int_0^{b(x)^{1/(\gamma+\varepsilon)}} a(y,x) F_Y(\mathrm{d}y)}_{\text{(I)}} + \underbrace{\lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty a(y,x) F_Y(\mathrm{d}y)}_{\text{(II)}},$$

where $b(x)$ satisfies $\lim_{x\to\infty} b(x)(1 - F_X(x)) = \infty$ and $\lim_{x\to\infty} b(x)/x^{\gamma+\varepsilon} = 0$. We first show that the second limit is zero:

$$\begin{aligned}
\text{(II)} &= \lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty \frac{1 - F_X(x/y)}{1 - F_X(x)} F_Y(\mathrm{d}y) \le \lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty \frac{1}{1 - F_X(x)} F_Y(\mathrm{d}y) \\
&\le \lim_{x\to\infty} \int_{b(x)^{1/(\gamma+\varepsilon)}}^\infty \frac{y^{\gamma+\varepsilon}}{(1 - F_X(x))b(x)} F_Y(\mathrm{d}y) \le \lim_{x\to\infty} \frac{1}{(1 - F_X(x))b(x)} \mathbb{E}[|Y|^{\gamma+\varepsilon}] = 0.
\end{aligned}$$

Now we consider (I), and show that for all x large enough, the integrand is bounded by an integrable function (of y), hence dominated convergence can be applied. First, we note that for $y \in (0,1)$, $a(y,x) \le 1$ for all x. Therefore we only need to consider $y \in [1, b(x)^{1/(\gamma+\varepsilon)}]$. Since $y \le b(x)^{1/(\gamma+\varepsilon)}$, we have

$$\frac{x}{y} \ge \left(\frac{x^{\gamma+\varepsilon}}{b(x)}\right)^{\frac{1}{\gamma+\varepsilon}},$$

which can be made arbitrarily large for all x large enough. Also note that (where the functions $L(\cdot)$ and $R(\cdot)$ are defined in Lemma S.1)

$$a(y,x) = y^\gamma \frac{L(x/y)}{L(x)} \exp\left\{\int_x^{x/y} \frac{R(t)}{t} \mathrm{d}t\right\},$$

where the ratio $|L(x/y)/L(x)|$ is bounded by a constant for all x large enough, uniformly in y. Similarly, $|R(t)|$ can be chosen to be arbitrarily small, which means the exponential term is bounded by y^ε. Hence, for $y \in [1, b(x)^{1/(\gamma+\varepsilon)}]$,

$$a(y,x) \le C y^{\gamma+\varepsilon},$$

which is integrable with respect to the distribution F_Y. Applying the dominated convergence, one concludes that

$$\lim_{x\to\infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \int_0^\infty y^\gamma F_Y(\mathrm{d}y) = \mathbb{E}[Y^\gamma \mathbb{1}_{Y>0}],$$

so that the product XY also has regularly varying tail with index γ, provided that $\mathbb{P}[Y > 0] > 0$. Similar argument can be applied to characterize the left tail of XY.

Part 2

Now we drop the independence assumption, and assume instead that Y is bounded by a constant C. For simplicity, we use F to denote the limit of the conditional distribution $F_{Y|X=x}$ as $x \to \infty$. Same as before, let $\varepsilon = \alpha - \gamma > 0$. First,

$$\frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \int_0^\infty \frac{\mathbb{P}[Y > x/y | X = y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y) = \int_{x/C}^\infty \frac{\mathbb{P}[Y > x/y | X = y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y).$$

Further, let $U \perp\!\!\!\perp (X, Y)$ be distributed according to F. Since the conditional distribution $Y|X = x$ converges weakly to that of U as $x \to \infty$, one has, for all x large enough,

$$\Big| \mathbb{P}[Y > x | X = y] - \mathbb{P}[U > x] \Big| \leq \eta + \mathbb{1}_{x \in A(y)},$$

where $\eta > 0$ is arbitrary, and the set $A(y)$ takes the form

$$A(y) = \bigcup_{j=1}^J \Big(a_j - \delta(y), a_j + \delta(y) \Big),$$

with $\delta(y)$ monotonically decreasing to zero as $y \to \infty$. (Note that if F is a continuous distribution, then one can simply use $|\mathbb{P}[Y > x|X = y] - \mathbb{P}[U > x]| \leq \eta$. The purpose of introducing $A(y)$ is to handle a discontinuous F.) Then we have

$$\begin{aligned} &\int_{x/C}^\infty \left| \frac{\mathbb{P}[Y > x/y | X = y] - \mathbb{P}[U > x/y]}{\mathbb{P}[X > x]} \right| F_X(\mathrm{d}y) \\ &\leq \eta \frac{\mathbb{P}[X > x/C]}{\mathbb{P}[X > x]} + \sum_{1 \leq j \leq J:\ 0 \leq a_j \leq C} \frac{F_X(x/(a_j - \delta(x/c))) - F_X(x/(a_j + \delta(x/c)))}{\mathbb{P}[X > x]}, \end{aligned}$$

where the right-hand-side has limit ηC^γ. Since η is arbitrary, the left-hand-side tends to zero as $x \to \infty$. As a result, we have

$$\lim_{x \to \infty} \frac{\mathbb{P}[XY > x]}{\mathbb{P}[X > x]} = \lim_{x \to \infty} \int_{x/C}^\infty \frac{\mathbb{P}[U > x/y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y) = \lim_{x \to \infty} \frac{\mathbb{P}[XU > x]}{\mathbb{P}[X > x]}.$$

Since we have $U \perp\!\!\!\perp X$, Part 1 of this proof can be applied to obtain the desired result.

Part 3

Now we drop the boundedness condition on Y. For this purpose, we only need to show that the following

$$\int_0^{x/C} \frac{\mathbb{P}[Y > x/y | X = y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y), \qquad \int_0^{x/C} \frac{\mathbb{P}[U > x/y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y),$$

can be made arbitrarily small by choosing C large enough. We only demonstrate for the first term. By Markov's inequality and the assumption that $\mathbb{E}[|Y|^{\gamma+\varepsilon}|X = x]$ is uniformly bounded, we have

$$\begin{aligned} \int_0^{x/C} \frac{\mathbb{P}[Y > x/y | X = y]}{\mathbb{P}[X > x]} F_X(\mathrm{d}y) &\leq \Big(\sup_x \mathbb{E}[|Y|^{\gamma+\varepsilon}|X = x] \Big) \int_0^{x/C} \frac{y^{\gamma+\varepsilon}}{x^{\gamma+\varepsilon} \mathbb{P}[X > x]} F_X(\mathrm{d}y) \\ &\to \Big(\sup_x \mathbb{E}[|Y|^{\gamma+\varepsilon}|X = x] \Big) C^{-\varepsilon} \frac{\gamma}{\varepsilon}, \end{aligned}$$

where the last line follows from Lemma S.2. ■

IV.4 Proof of Lemma S.4 and S.5

See Section XVII.2 in Feller (1991). ■

IV.5 Proof of Lemma S.6

Define $r(x) = [1, x, \cdots, x^p]'$, then the estimator can be rewritten as

$$\left[\sum_{i=1}^n r(e(X_i))r(e(X_i))'w_i\right]^{-1}\left[\sum_{i=1}^n r(e(X_i))Y_i w_i\right],$$

where $w_i = \mathbb{1}_{e(X_i)\le h_n, D_i=1}$. We use $F_{e(X)}$ to denote the distribution function of the probability weight. We first analyze the "denominator" term. Consider the following:

$$A_n = \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)}\sum_{i=1}^n r(e(X_i)/h_n)r(e(X_i)/h_n)'w_i,$$

whose expectation is given by (we use $\dot r(x)$ to denote the derivative $\mathrm{d}r(x)/\mathrm{d}x$)

$$\begin{aligned}
\mathbb{E}[A_n] &= \frac{1}{F_{e(X)}(h_n)}\int_0^{h_n} r(x/h_n)r(x/h_n)'x/h_n F_{e(X)}(\mathrm{d}x)\\
&= \frac{1}{F_{e(X)}(h_n)}\left[r(1)r(1)'F_{e(X)}(h_n) - \int_0^1 \Big((\dot r(x)r(x)' + r(x)\dot r(x)')x + r(x)r(x)'\Big)F_{e(X)}(xh_n)\mathrm{d}x\right]\\
&\to \left[r(1)r(1)' - \int_0^1 \Big((\dot r(x)r(x)' + r(x)\dot r(x)')x + r(x)r(x)'\Big)x^{\gamma_0-1}\mathrm{d}x\right] = S,
\end{aligned}$$

which is always invertible. Next we show that A_n converges to the expectation computed above. For this purpose, we consider the variance of individual terms in A_n, which is bounded by

$$\begin{aligned}
&\frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)^2}\int_0^{h_n}(x/h_n)^{j+1}F_{e(X)}(\mathrm{d}x) = \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)^2}\left[F_{e(X)}(h_n) - \int_0^1 (j+1)x^j F_{e(X)}(xh_n)\mathrm{d}x\right]\\
\asymp&\frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)}\left[1 - \int_0^1 (j+1)x^j x^{\gamma_0-1}\mathrm{d}x\right] \asymp \frac{1}{n}\frac{1}{h_n F_{e(X)}(h_n)},
\end{aligned}$$

which shrinks to zero under our assumptions.

Now we consider the "numerator" term. Let $\eta_i = Y_i - \mathbb{E}[Y_i|e(X_i), D_i = 1]$ be the residual from conditional expectation projection. Then the following

$$L_n = \sqrt{\frac{1}{nh_n F_{e(X)}(h_n)}}\sum_{i=1}^n r(e(X_i)/h_n)\eta_i w_i$$

has zero mean, and variance given by

$$\begin{aligned}
\mathbb{V}[L_n] &= \frac{1}{F_{e(X)}(h_n)}\int_0^{h_n} x/h_n r(x/h_n)r(x/h_n)'\mathbb{V}[Y|e(X)=x, D=1]F_{e(X)}(\mathrm{d}x)\\
&= (\mu_2(0) - \mu_1(0)^2)\frac{1}{F_{e(X)}(h_n)}\int_0^{h_n} x/h_n r(x/h_n)r(x/h_n)'F_{e(X)}(\mathrm{d}x)(1+o(1))\\
&\to (\mu_2(0) - \mu_1(0)^2)S.
\end{aligned}$$

The Lindeberg condition can be easily verified by calculating higher moments, and L_n will be asymptotically Gaussian provided that $nh_n F_{e(X)}(h_n) \to \infty$. We do not elaborate details here.

Next we consider the bias. Assuming μ_1 is sufficiently smooth, one has

$$\mu_1(x) = \sum_{j=0}^p \frac{1}{j!}\mu_1^{(j)}(0)x^j + \frac{1}{(p+1)!}\mu_1^{(p+1)}(\tilde x)x^{p+1},$$

where $\tilde{x} \in [0, x]$. Now we consider the following

$$\left[\sum_{i=1}^{n} r(e(X_i))r(e(X_i))'w_i\right]^{-1}\left[\sum_{i=1}^{n} r(e(X_i))Y_i w_i\right] - \beta = H_n^{-1}A_n^{-1}\left[\sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}}L_n + h_n^{p+1}R_n\right],$$

where H_n is a diagonal matrix with elements $1, h_n, \cdots, h_n^p$, and R_n is

$$R_n = \frac{1}{(p+1)!}\frac{1}{nh_n^{p+2}F_{e(X)}(h_n)}\sum_{i=1}^{n} r(e(X_i)/h_n)\mu_1^{(p+1)}(\lambda_i e(X_i))e(X_i)^{p+1}w_i,$$

with $\lambda_i \in [0, 1]$. We can show that R_n has expectation

$$\begin{aligned}
\mathbb{E}[R_n] &= \frac{\mu_1^{(p+1)}(0)}{(p+1)!}\frac{1}{F_{e(X)}(h_n)}\left[\int_0^{h_n} r(x/h_n)(x/h_n)^{p+2}F_{e(X)}(\mathrm{d}x)\right](1+o(1)) \\
&= \frac{\mu_1^{(p+1)}(0)}{(p+1)!}\frac{1}{F_{e(X)}(h_n)}\left[r(1)F_{e(X)}(h_n) - \int_0^1 \Big(\dot{r}(x)(x)^{p+2} + (p+1)r(x)(x)^{p+1}\Big)F_{e(X)}(xh_n)\mathrm{d}x\right](1+o(1)) \\
&\to \frac{\mu_1^{(p+1)}(0)}{(p+1)!}\left[r(1) - \int_0^1 \Big(\dot{r}(x)x^{p+2} + (p+1)r(x)x^{p+1}\Big)x^{\gamma_0-1}\mathrm{d}x\right] = \frac{\mu_1^{(p+1)}(0)}{(p+1)!}R,
\end{aligned}$$

and with the same technique used for analyzing A_n,

$$\Big|R_n - \mathbb{E}[R_n]\Big|^2 = o_{\mathrm{p}}(1),$$

which closes the proof. ■

IV.6 Proof of Lemma 1

Let $F_{1/e(X)}$ be the distribution function of the inverse probability weight $1/e(X)$. First consider the tail probability $\mathbb{P}[D/e(X) > x]$:

$$\mathbb{P}[D/e(X) > x] = \mathbb{E}[e(X)\mathbb{1}_{e(X)<x^{-1}}] = \int_0^{x^{-1}} \mathbb{P}[t < e(X) < x^{-1}]\mathrm{d}t = \int_0^1 x^{-1}\mathbb{P}[sx^{-1} < e(X) < x^{-1}]\mathrm{d}s,$$

Therefore,

$$\lim_{x\to\infty}\frac{x\mathbb{P}[D/e(X) > x]}{\mathbb{P}[e(X) < x^{-1}]} = \lim_{x\to\infty}\int_0^1 \frac{\mathbb{P}[sx^{-1} < e(X) < x^{-1}]}{\mathbb{P}[e(X) < x^{-1}]}\mathrm{d}s = \int_0^1 \Big(1 - s^{\gamma_0-1}\Big)\mathrm{d}s = 1 - \frac{\gamma_0 - 1}{\gamma_0},$$

where for the second equality we use the definition of regular variation. As a result, $D/e(X)$ has regularly varying tail with index $-\gamma_0$. The rest follows from Lemma S.3. ■

IV.7 Proof of Theorem 1

Part (i)

We first assume $\gamma_0 > 2$ and there is no trimming ($b_n = 0$). In this case, $DY/e(X)$ has a finite variance, which is also nonzero since $\alpha_+(0) + \alpha_-(0) > 0$. Then we set $a_{n,b_n} = a_n = \sqrt{n\mathbb{V}[DY/e(X)]}$, which satisfies the requirement of the theorem. Then asymptotic Gaussianity follows from the central limit theorem. The case with trimming ($b_n > 0$) follows from the same argument.

Next we consider the $\gamma_0 = 2$ case. Again we demonstrate assuming there is no trimming ($b_n = 0$), and the trimming case can be justified with the same argument. We compute the limits in Lemma S.5 and show that M is a point mass at the origin. Let (recall that we set $a_{n,b_n} = a_n$)

$$W_n = \frac{Z}{a_n}, \qquad Z = \frac{DY}{e(X)} - \theta_0,$$

and F_Z be the distribution function of Z. Without loss of generality, we assume $\alpha_+(0) > 0$, so that $DY/e(X)$ has a regularly varying right tail with index -2. First consider the following,

$$\begin{aligned} n\mathbb{E}[W_n^2 \mathbb{1}_{|W_n|\le c}] &= \frac{n}{a_n^2}\mathbb{E}\left[Z^2 \mathbb{1}_{|Z/a_n|\le c}\right] = \frac{n}{a_n^2}\int_0^{a_n c} x^2 F_{|Z|}(\mathrm{d}x) \\ &= \frac{n}{a_n^2}\int_0^{a_n c} 2t\mathbb{P}[t < |Z| < a_n c]\mathrm{d}t = n\int_0^c 2s\mathbb{P}[a_n s < |Z| < a_n c]\mathrm{d}s \\ &= n\Big(\underbrace{\int_0^1 2s\mathbb{P}[a_n s < |Z| < a_n]\mathrm{d}s}_{\text{(I)}} + \underbrace{\int_0^1 2s\mathbb{P}[a_n < |Z| < a_n c]\mathrm{d}s}_{\text{(II)}} + \underbrace{\int_1^c 2s\mathbb{P}[a_n s < |Z| < a_n c]\mathrm{d}s}_{\text{(III)}}\Big). \end{aligned}$$

Without loss of generality, assume $c > 1$, then

$$0 \le \text{(III)} \le (c^2-1)\Big(F_{|Z|}(a_n c) - F_{|Z|}(a_n)\Big),$$

which implies (using Fatou's lemma)

$$\begin{aligned} \liminf_{n\to\infty}\frac{\text{(I)}}{\text{(III)}} &\ge \liminf_{n\to\infty}\frac{1}{c^2-1}\int_0^1 2s\frac{\mathbb{P}[a_n s < |Z| < a_n]}{\mathbb{P}[a_n < |Z| < a_n c]}\mathrm{d}s \ge \frac{1}{c^2-1}\int_0^1 2s\liminf_{n\to\infty}\frac{\mathbb{P}[a_n s < |Z| < a_n]}{\mathbb{P}[a_n < |Z| < a_n c]}\mathrm{d}s \\ &= \frac{1}{c^2-1}\int_0^1 2s\frac{s^{-2}-1}{1-c^{-2}}\mathrm{d}s = \infty. \end{aligned}$$

With the same reasoning, one has

$$\liminf_{n\to\infty}\frac{\text{(I)}}{\text{(II)}} = \infty.$$

The same conclusion can be shown for any $0 < c < 1$. As a result,

$$n\mathbb{E}[W_n^2 \mathbb{1}_{|W_n|\le c}] = n\cdot\text{(I)}\cdot(1+o(1)) = \frac{n}{a_n^2}\mathbb{E}\left[Z^2\mathbb{1}_{|Z/a_n|\le 1}\right](1+o(1)) \to 1,$$

where the last step follows form the definition of a_n. Applying the same technique, one has

$$\liminf_{n\to\infty}\frac{n\cdot\text{(I)}}{n(1-F_{|Z|}(a_n c))} \to \infty,$$

for any $c > 0$. And since the numerator converges to 1, we have the denominator converges to zero. To summarize, we showed that, for any compact interval I whose interior contains the origin,

$$n\mathbb{E}[W_n^2\mathbb{1}_{|W_n|\in I}] \to 1,$$

and for any $c > 0$,

$$n(1-F_{|Z|}(a_n c)) \to 0.$$

By applying Lemma S.5, it implies that M is a point mass at the origin, and the limiting distribution is Gaussian.

Part (ii.1), no trimming $b_n = 0$

Again we assume, without loss of generality, that $\alpha_+(0) > 0$, so that $DY/e(X)$ has regularly varying right tail with index $-\gamma_0$. For $c > 0$, we compute the following (recall that $a_{n,b_n} = a_n$):

$$n\Big(1-F_Z(a_n c)\Big) = \frac{1-F_Z(a_n c)}{1-F_{|Z|}(a_n)}n\Big(1-F_{|Z|}(a_n)\Big) = \underbrace{\frac{1-F_Z(a_n c)}{1-F_{|Z|}(a_n)}}_{\text{(I)}}\underbrace{\frac{a_n^2(1-F_{|Z|}(a_n))}{\mathbb{E}[|Z|^2\mathbb{1}_{|Z|\le a_n}]}}_{\text{(II)}}\underbrace{\frac{n}{a_n^2}\mathbb{E}[|Z|^2\mathbb{1}_{|Z|\le a_n}]}_{\text{(III)}}.$$

Next, (I) converges to $\frac{\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)}c^{-\gamma_0}$ as Z has regularly varying tails; (II) converges to $\frac{2-\gamma_0}{\gamma_0}$ by Lemma S.2; and (III) converges to 1 due to the definition of a_n. Therefore,

$$n\Big(1-F_Z(a_nc)\Big)\to\frac{\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)}\frac{2-\gamma_0}{\gamma_0}c^{-\gamma_0}=\int_c^\infty\frac{1}{x^2}\Big(\frac{(2-\gamma_0)\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)}x^{1-\gamma_0}\Big)\mathrm{d}x.$$

Similarly, we have, for the left tail,

$$nF_Z(-a_nc)\to\frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\frac{2-\gamma_0}{\gamma_0}c^{-\gamma_0}=\int_c^\infty\frac{1}{x^2}\Big(\frac{(2-\gamma_0)\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}x^{1-\gamma_0}\Big)\mathrm{d}x.$$

Therefore, we conjecture that the measure M in Lemma S.5 takes the following form:

$$M(\mathrm{d}x)=\mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\geq0}+\alpha_-(0)\mathbb{1}_{x<0}\Big)\right].$$

To confirmed, we compute the other condition in Lemma S.5. Fore example, take an interval $I=[c_1,c_2]$ with $c_1>0$,

$$\begin{aligned}n\mathbb{E}[X_n^2\mathbb{1}_{|W_n|\in I}]&=\frac{n}{a_n^2}\int_{a_nc_1}^{a_nc_2}x^2F_Z(\mathrm{d}x)=n\int_0^{a_nc_2}2t\mathbb{P}[(a_nc_1\vee t)<Z<a_nc_2]\mathrm{d}t\\&=\frac{n}{a_n^2}\left(a_n^2c_1^2\mathbb{P}[a_nc_1<Z<a_nc_2]+\int_{a_nc_1}^{a_nc_2}2t\mathbb{P}[t<Z<a_nc_2]\mathrm{d}t\right).\end{aligned}$$

Next, we have

$$nc_1^2\mathbb{P}[a_nc_1<Z<a_nc_2]\to\frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\frac{2-\gamma_0}{\gamma_0}\Big(c_1^{2-\gamma_0}-c_1^2c_2^{-\gamma_0}\Big),$$

and

$$\begin{aligned}\frac{n}{a_n^2}\int_{a_nc_1}^{a_nc_2}2t\mathbb{P}[t<Z<a_nc_2]\mathrm{d}t&=n(\mathbb{P}[Z>a_nc_1])\frac{1}{a_n^2}\int_{a_nc_1}^{a_nc_2}2t\frac{\mathbb{P}[t<Z<a_nc_2]}{\mathbb{P}[Z>a_nc_1]}\mathrm{d}t\\&=n(\mathbb{P}[Z>a_nc_1])\int_{c_1}^{c_2}2s\frac{\mathbb{P}[a_ns<Z<a_nc_2]}{\mathbb{P}[Z>a_nc_1]}\mathrm{d}s\\&\to\frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\frac{2-\gamma_0}{\gamma_0}c_1^{-\gamma_0}\int_{c_1}^{c_2}2s\frac{s^{-\gamma_0}-c_2^{-\gamma_0}}{c_1^{-\gamma_0}}\mathrm{d}s\\&=\frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\frac{2-\gamma_0}{\gamma_0}\left(\frac{2}{2-\gamma_0}(c_2^{2-\gamma_0}-c_1^{2-\gamma_0})-c_2^{-\gamma_0}(c_2^2-c_1^2)\right).\end{aligned}$$

Therefore,

$$\begin{aligned}n\mathbb{E}[X_n^2\mathbb{1}_{|W_n|\in I}]&\to\frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\left(\frac{2-\gamma_0}{\gamma_0}(c_1^{2-\gamma_0}-c_1^2c_2^{-\gamma_0})+\frac{2}{\gamma_0}(c_2^{2-\gamma_0}-c_1^{2-\gamma_0})-\frac{2-\gamma_0}{\gamma_0}c_2^{-\gamma_0}(c_2^2-c_1^2)\right)\\&=\frac{\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\Big(c_2^{2-\gamma_0}-c_1^{2-\gamma_0}\Big)=M(I)\end{aligned}$$

Given the measure M, the characteristic function can be found by evaluating the integral in lemma S.4, which gives

$$\int_{\mathbb{R}}\frac{e^{i\zeta x}-1-i\zeta x}{x^2}M(\mathrm{d}x)=-|\zeta|^{\gamma_0}\frac{\Gamma(3-\gamma_0)}{\gamma_0(\gamma_0-1)}\cos\Big(\frac{\gamma_0\pi}{2}\Big)\left[i\frac{\alpha_+(0)-\alpha_-(0)}{\alpha_+(0)+\alpha_-(0)}\mathrm{sgn}(\zeta)\tan\Big(\frac{\gamma_0\pi}{2}\Big)-1\right].$$

Part (ii.1), light trimming $b_n a_n \to 0$

Take $c > 0$ and first consider the following probability:

$$\int_0^{b_n} x\mathbb{P}[Y > a_n c x | e(X) = x, D = 1] F_{e(X)}(\mathrm{d}x) \leq \int_0^{b_n} x F_{e(X)}(\mathrm{d}x)$$
$$= \mathbb{E}[e(X)\mathbb{1}_{e(X)<b_n}] = \mathbb{P}\left[\frac{D}{e(X)} > b_n^{-1}\right].$$

If $a_n b_n \to 0$, the right-hand-side will be asymptotically negligible compared to $\mathbb{P}[D/e(X) > a_n c]$ for any $c > 0$.

As a result,

$$\frac{\mathbb{P}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n} > a_n c\right]}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]} = \frac{1}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]} \int_{b_n}^1 x\mathbb{P}[Y > a_n c x | e(X) = x, D = 1] F_{e(X)}(\mathrm{d}x)$$
$$= \frac{1}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]} \int_0^1 x\mathbb{P}[Y > a_n c x | e(X) = x, D = 1] F_{e(X)}(\mathrm{d}x) + o(1)$$
$$= \frac{\mathbb{P}\left[\frac{DY}{e(X)} > a_n c\right]}{\mathbb{P}\left[\frac{D}{e(X)} > a_n c\right]} + o(1) \to \alpha_+(0),$$

where the last line follows from Lemma 1. The above shows that, as long as $a_n b_n \to 0$, $\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n}$ and $\frac{DY}{e(X)}$ have the same tail property, meaning that the same Lévy stable limiting distribution emerges under the light trimming regime.

Part (ii.1), moderate trimming $b_n a_n \to t \in (0, \infty)$

As before we ignore the centering, as it is irrelevant for computing tail probabilities (or truncated moments). Let F_U be the limiting distribution of $F_{Y|e(X)=x,D=1}$ as $x \to 0$, $U \perp\!\!\!\perp (X, Y)$ be distributed according to F_U, and $c > 0$. We first compute the following limit (recall that we set $a_{n,b_n} = a_n$ for the moderate trimming scenario):

$$\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq t a_n^{-1}} > a_n c\right] = n\int_0^\infty \mathbb{P}\left[\frac{D}{e(X)}\mathbb{1}_{e(X)\geq t a_n^{-1}} > \frac{a_n c}{x}\right] F_U(\mathrm{d}x)$$
$$= \lim_{n\to\infty} n\int_0^\infty \int_{t/a_n}^{x/(a_n c)} y F_{e(X)}(\mathrm{d}y) F_U(\mathrm{d}x) = \lim_{n\to\infty} n\int_{ct}^\infty \int_{t/a_n}^{x/(a_n c)} y F_{e(X)}(\mathrm{d}y) F_U(\mathrm{d}x).$$

To proceed, note that

$$\int_{t/a_n}^{x/(a_n c)} y F_{e(X)}(\mathrm{d}y) = \int_0^\infty \int_{t/a_n}^{x/(a_n c)} \mathbb{1}_{y>s} F_{e(X)}(\mathrm{d}y)\mathrm{d}s$$
$$= \frac{t}{a_n}\mathbb{P}[(t/a_n) < e(X) < x/(a_n c)] + \int_{t/a_n}^{x/(a_n c)} \mathbb{P}[s < e(X) < x/(a_n c)]\mathrm{d}s$$
$$= \frac{t}{a_n}\left(F_{e(X)}(x/(a_n c)) - F_{e(X)}(t/a_n)\right) + \left(\frac{x}{a_n c} - \frac{t}{a_n}\right) F_{e(X)}(x/(a_n c)) - \int_{t/a_n}^{x/(a_n c)} F_{e(X)}(s)\mathrm{d}s$$
$$= \frac{x}{a_n c} F_{e(X)}(x/(a_n c)) - \frac{t}{a_n} F_{e(X)}(t/a_n) - \int_{t/a_n}^{x/(a_n c)} F_{e(X)}(s)\mathrm{d}s.$$

Therefore,

$$\begin{aligned}
&\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} > a_n c\right] \\
&= \lim_{n\to\infty} n\int_{ct}^{\infty}\left[\frac{x}{a_n c}F_{e(X)}\left(\frac{x}{a_n c}\right) - \frac{t}{a_n}F_{e(X)}\left(\frac{t}{a_n}\right) - \int_{t/a_n}^{x/(a_n c)} F_{e(X)}(y)\mathrm{d}y\right] F_U(\mathrm{d}x) \\
&= \lim_{n\to\infty} n\int_{ct}^{\infty}\left[\frac{x}{a_n c}F_{e(X)}\left(\frac{x}{a_n c}\right) - \frac{t}{a_n}F_{e(X)}\left(\frac{t}{a_n}\right) - \frac{1}{a_n}\int_{t}^{x/c} F_{e(X)}\left(\frac{y}{a_n}\right)\mathrm{d}y\right] F_U(\mathrm{d}x) \\
&= \lim_{n\to\infty}\left[\frac{nF_{e(X)}(1/a_n)}{a_n}\right]\left[\int_{ct}^{\infty}\left[\frac{x}{c}\frac{F_{e(X)}\left(x/(a_n c)\right)}{F_{e(X)}\left(1/a_n\right)} - t\frac{F_{e(X)}\left(t/a_n\right)}{F_{e(X)}\left(1/a_n\right)} - \int_t^{x/c}\frac{F_{e(X)}(y/a_n)}{F_{e(X)}(1/a_n)}\mathrm{d}y\right] F_U(\mathrm{d}x)\right].
\end{aligned}$$

Next, we have

$$\begin{aligned}
\lim_{n\to\infty}\frac{nF_{e(X)}(1/a_n)}{a_n} &= \lim_{n\to\infty} n\mathbb{P}[|DY/e(X)| > a_n]\frac{F_{e(X)}(1/a_n)}{a_n\mathbb{P}[|DY/e(X)| > a_n]} \\
&= \lim_{n\to\infty} \underbrace{\frac{n}{a_n^2}\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq a_n}]}_{\text{(I)}}\underbrace{\frac{a_n^2\mathbb{P}[|DY/e(X)| > a_n]}{\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq a_n}]}}_{\text{(II)}}\underbrace{\frac{F_{e(X)}(1/a_n)}{a_n\mathbb{P}[|DY/e(X)| > a_n]}}_{\text{(III)}}.
\end{aligned}$$

Term (I) converges to 1 due to the definition of a_n; term (II) converges to $(2-\gamma_0)/\gamma_0$ by Lemma S.2; and term (III) converges to $\gamma_0/((\gamma_0-1)(\alpha_+(0)+\alpha_-(0)))$ by Lemma 1. Therefore,

$$\lim_{n\to\infty}\frac{nF_{e(X)}(1/a_n)}{a_n} = \frac{2-\gamma_0}{\gamma_0-1}\frac{1}{\alpha_+(0)+\alpha_-(0)},$$

and

$$\begin{aligned}
&\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} > a_n c\right] \\
=&\frac{2-\gamma_0}{\gamma_0-1}\frac{1}{\alpha_+(0)+\alpha_-(0)}\lim_{n\to\infty}\left[\int_{ct}^{\infty}\left[\frac{x}{c}\frac{F_{e(X)}\left(x/(a_n c)\right)}{F_{e(X)}\left(1/a_n\right)} - t\frac{F_{e(X)}\left(t/a_n\right)}{F_{e(X)}\left(1/a_n\right)} - \int_t^{x/c}\frac{F_{e(X)}(y/a_n)}{F_{e(X)}(1/a_n)}\mathrm{d}y\right] F_U(\mathrm{d}x)\right] \\
=&\frac{2-\gamma_0}{\gamma_0}\frac{1}{\alpha_+(0)+\alpha_-(0)}\left[\int_{ct}^{\infty}\left[\left(\frac{x}{c}\right)^{\gamma_0} - t^{\gamma_0} - \int_t^{x/c} y^{\gamma_0-1}\mathrm{d}y\right] F_U(\mathrm{d}x)\right] \\
=&\frac{2-\gamma_0}{\gamma_0}\frac{1}{\alpha_+(0)+\alpha_-(0)}\left[\int_{ct}^{\infty}\left[\left(\frac{x}{c}\right)^{\gamma_0} - t^{\gamma_0}\right] F_U(\mathrm{d}x)\right] \\
=&\int_c^{\infty}\frac{1}{x^2}\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}x^{1-\gamma_0}\alpha_+(tx)\right]\mathrm{d}x.
\end{aligned}$$

Similarly, we can obtain, for the left tail,

$$\lim_{n\to\infty} n\mathbb{P}\left[\frac{DU}{e(X)}\mathbb{1}_{e(X)\geq ta_n^{-1}} < -a_n c\right] = \int_c^{\infty}\frac{1}{x^2}\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}x^{1-\gamma_0}\alpha_-(tx)\right]\mathrm{d}x,$$

where F_{-U} is the distribution function of $-U$. Define a measure M as

$$M(\mathrm{d}x) = \mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}|x|^{1-\gamma_0}\Big(\alpha_+(tx)\mathbb{1}_{x\geq 0} + \alpha_-(tx)\mathbb{1}_{x<0}\Big)\right],$$

and we verify the other condition in Lemma S.5.

For simplicity, take $I = [c_1, c_2]$ with $0 < c_1 < c_2$ and $t = 1$. Then the truncated second moment is

$$
\begin{aligned}
&\frac{n}{a_n^2}\mathbb{E}\left[\frac{DU^2}{e(X)^2}\mathbb{1}_{e(X)\geq a_n^{-1}}\mathbb{1}\left(a_n c_1 < \frac{DU}{e(X)}\mathbb{1}_{e(X)\geq a_n^{-1}} < a_n c_2\right)\right] \\
=&\frac{n}{a_n^2}\int_{-\infty}^{\infty}\int_0^1 \mathbb{1}_{x\geq a_n^{-1}}\mathbb{1}_{x\in[u/(a_n c_2),u/(a_n c_1)]}\frac{u^2}{x}F_{e(X)}(\mathrm{d}x)F_U(\mathrm{d}u) \\
=&\frac{n}{a_n^2}\int_{c_1}^{\infty}\int_{((u/c_2)\vee 1)/a_n}^{u/(a_n c_1)}\frac{u^2}{x}F_{e(X)}(\mathrm{d}x)F_U(\mathrm{d}u) \\
=&\frac{n}{a_n^2}\int_{c_1}^{\infty} u^2\left[\frac{F_{e(X)}(u/(a_n c_1))}{u/(a_n c_1)} - \frac{F_{e(X)}(((u/c_2)\vee 1)/a_n)}{((u/c_2)\vee 1)/a_n} + \int_{((u/c_2)\vee 1)/a_n}^{u/(a_n c_1)}\frac{1}{x^2}F_{e(X)}(x)\mathrm{d}x\right]F_U(\mathrm{d}u) \\
=&n\int_{c_1}^{\infty} u^2\left[\frac{F_{e(X)}(u/(a_n c_1))}{a_n u/c_1} - \frac{F_{e(X)}(((u/c_2)\vee 1)/a_n)}{a_n((u/c_2)\vee 1)} + \int_{((u/c_2)\vee 1)}^{u/c_1}\frac{1}{a_n x^2}F_{e(X)}(x/a_n)\mathrm{d}x\right]F_U(\mathrm{d}u) \\
=&(1+o(1))\frac{2-\gamma_0}{\gamma_0-1}\frac{1}{\alpha_+ + \alpha_-}\int_{c_1}^{\infty} u^2 \\
&\quad\left[\frac{1}{u/c_1}\frac{F_{e(X)}(u/(a_n c_1))}{F_{e(X)}(1/a_n)} - \frac{1}{(u/c_2)\vee 1}\frac{F_{e(X)}(((u/c_2)\vee 1)/a_n)}{F_{e(X)}(1/a_n)} + \int_{((u/c_2)\vee 1)}^{u/c_1}\frac{1}{x^2}\frac{F_{e(X)}(x/a_n)}{F_{e(X)}(1/a_n)}\mathrm{d}x\right]F_U(\mathrm{d}u) \\
\to&\frac{2-\gamma_0}{\gamma_0-1}\frac{1}{\alpha_+(0)+\alpha_-(0)}\int_{c_1}^{\infty} u^2\left[(u/c_1)^{\gamma_0-2} - ((u/c_2)\vee 1)^{\gamma_0-2} + \int_{((u/c_2)\vee 1)}^{u/c_1} x^{\gamma_0-3}\mathrm{d}x\right]F_U(\mathrm{d}u) \\
=&-\frac{1}{\alpha_+(0)+\alpha_-(0)}\int_{c_1}^{\infty} u^2\left[(u/c_1)^{\gamma_0-2} - ((u/c_2)\vee 1)^{\gamma_0-2}\right]F_U(\mathrm{d}u) \\
=&-\frac{1}{\alpha_+(0)+\alpha_-(0)}\left[\int_{c_1}^{c_2}\frac{u^{\gamma_0}}{c_1^{\gamma_0-2}} - u^2 F_U(\mathrm{d}u) + \int_{c_2}^{\infty}\frac{u^{\gamma_0}}{c_1^{\gamma_0-2}} - \frac{u^{\gamma_0}}{c_2^{\gamma_0-2}}F_U(\mathrm{d}u)\right],
\end{aligned}
$$

which, by simple algebra, can be shown to be the same as $M([c_1, c_2])$. The next step is to replace $DU/e(X)$ by $DY/e(X)$. The same argument used to proved Lemma S.3 applies here, which we do not repeat.

Part (ii.2), heavy trimming $b_n a_n \to \infty$

We verify a Lyapunov condition. Let $0 < \eta < \varepsilon$, where ε is defined in Assumption 2. Consider the following

$$
\begin{aligned}
&\frac{n}{a_{n,b_n}^{2+\eta}}\mathbb{E}\left[\left|\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n} - \theta_0 - \mathsf{B}_{n,b_n}\right|^{2+\eta}\right] \asymp \frac{n}{a_{n,b_n}^{2+\eta}}\mathbb{E}\left[\frac{D|Y|^{2+\eta}}{e(X)^{2+\eta}}\mathbb{1}_{e(X)\geq b_n}\right] \\
\precsim&\frac{n}{a_{n,b_n}^{2+\eta}}\mathbb{E}\left[\frac{1}{e(X)^{1+\eta}}\mathbb{1}_{e(X)\geq b_n}\right] = \frac{n}{a_{n,b_n}^{2+\eta}}\int_1^{1/b_n} x^{1+\eta}F_{1/e(X)}(\mathrm{d}x).
\end{aligned}
$$

Next,

$$
\begin{aligned}
a_{n,b_n} = \sqrt{n\mathbb{V}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n}\right]} &\asymp \sqrt{n\mathbb{E}\left[\frac{D|Y|^2}{e(X)^2}\mathbb{1}_{e(X)\geq b_n}\right]} \asymp \sqrt{n\mathbb{E}\left[\frac{1}{e(X)}\mathbb{1}_{e(X)\geq b_n}\right]} \\
&= \sqrt{n\int_1^{1/b_n} xF_{1/e(X)}(\mathrm{d}x)}.
\end{aligned}
$$

Therefore,

$$\frac{n}{a_{n,b_n}^{2+\eta}}\mathbb{E}\left[\left|\frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n}-\theta_0-\mathsf{B}_{n,b_n}\right|^{2+\eta}\right]\asymp n^{-\eta/2}\left[\int_1^{1/b_n}x^{1+\eta}F_{1/e(X)}(\mathrm{d}x)\right]\left[\int_1^{1/b_n}xF_{1/e(X)}(\mathrm{d}x)\right]^{-1-\eta/2}$$
$$=\underbrace{\frac{\int_1^{1/b_n}x^{1+\eta}F_{1/e(X)}(\mathrm{d}x)}{b_n^{-1-\eta}\mathbb{P}[e(X)\leq b_n]}}_{\text{(I)}}\underbrace{\left[\frac{\int_1^{1/b_n}xF_{1/e(X)}(\mathrm{d}x)}{b_n^{-1}\mathbb{P}[e(X)\leq b_n]}\right]^{-1-\eta/2}}_{\text{(II)}}\underbrace{\left[\frac{1}{nb_n\mathbb{P}[e(X)\leq b_n]}\right]^{\eta/2}}_{\text{(III)}}.$$

By Lemma S.2, both (I) and (II) converges to finite constant. For (III),

$$\frac{1}{nb_n\mathbb{P}[e(X)\leq b_n]}=\frac{\mathbb{P}[|DY/e(X)|\geq b_n^{-1}]}{b_n\mathbb{P}[e(X)\leq b_n]}\frac{1}{n\mathbb{P}[|DY/e(X)|\geq b_n^{-1}]}$$
$$=\underbrace{\frac{\mathbb{P}[|DY/e(X)|\geq b_n^{-1}]}{b_n\mathbb{P}[e(X)\leq b_n]}}_{\text{(III.1)}}\underbrace{\frac{\mathbb{P}[|DY/e(X)|\geq a_n]}{\mathbb{P}[|DY/e(X)|\geq b_n^{-1}]}}_{\text{(III.2)}}\underbrace{\frac{\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq a_n^{-1}}]}{a_n^{-2}\mathbb{P}[|DY/e(X)|\geq a_n^{-1}]}}_{\text{(III.3)}}\underbrace{\frac{1}{na_n^2\mathbb{E}[|DY/e(X)|^2\mathbb{1}_{|DY/e(X)|\leq a_n^{-1}}]}}_{\text{(III.4)}}.$$

Term (III.1) converges to a finite constant by Lemma 1, term (III.3) converges to a finite constant by Lemma S.2, and term (III.4) converges to 1 due to the definition of a_n. Finally, since $a_nb_n\to\infty$, term (III.2) vanishes, which closes the proof. ■

IV.8 Proof of Proposition 1

Part 1: no trimming $b_n=0$

$$\frac{n}{a_{n,b_n}}\left(\hat{\theta}_{n,b_n}-\theta_0\right)=\frac{1}{a_{n,b_n}}\sum_{i=1}^n\left(\frac{D_iY_i}{e(X_i)}-\theta_0\right)+\frac{1}{a_{n,b_n}}\sum_{i=1}^n\frac{D_iY_i}{e(X_i)}\left(\frac{e(X_i)}{\hat{e}(X_i)}-1\right)$$
$$=\frac{1}{a_{n,b_n}}\sum_{i=1}^n\left(\frac{D_iY_i}{e(X_i)}-\theta_0\right)+\left(-\frac{1}{n}\sum_{i=1}^n\frac{D_iY_i}{e(X_i,\tilde{\pi}_n)^2}\frac{\partial e(X_i,\tilde{\pi}_n)}{\partial\pi}\right)\frac{n}{a_{n,b_n}}\left(\hat{\pi}_n-\pi_0\right),$$

where $\tilde{\pi}_n$ is some convex combination of $\hat{\pi}_n$ and π_0, hence $|\tilde{\pi}_n-\pi_0|=O_{\mathrm{p}}(1/\sqrt{n})$. By Assumption 3, the class

$$\left\{\frac{D_iY_i}{e(X_i,\pi)^2}\frac{\partial e(X_i,\pi)}{\partial\pi}:\ |\pi-\pi_0|\leq\varepsilon\right\}$$

is Glivenko-Cantelli, which implies

$$\frac{1}{n}\sum_{i=1}^n\frac{D_iY_i}{e(X_i,\tilde{\pi}_n)^2}\frac{\partial e(X_i,\tilde{\pi}_n)}{\partial\pi}\xrightarrow{\mathrm{p}}\mathbb{E}\left[\frac{DY}{e(X_i)^2}\frac{\partial e(X_i,\pi_0)}{\partial\pi}\right].$$

Therefore, we have

$$\frac{n}{a_{n,b_n}}\left(\hat{\theta}_{n,b_n}-\theta_0\right)=\frac{1}{a_{n,b_n}}\sum_{i=1}^n\left(\frac{D_iY_i}{e(X_i)}-\theta_0-\mathbb{E}\left[\frac{\mu_1(e(X_i))}{e(X_i)}\frac{\partial e(X_i,\pi_0)}{\partial\pi}\right]h(D_i,X_i)\right)+o_{\mathrm{p}}(1).$$

For $\gamma_0>2$, we have $n/a_{n,b_n}\asymp\sqrt{n}$, and the above is asymptotically Gaussian. For the other case, the additional term in the summand is asymptotically negligible.

Part 2: trimming $b_n > 0$

To start,

$$\begin{aligned}
&\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\hat{\pi}_n)}\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}-\theta_0-\mathsf{B}_{n,b_n}\right)\\
=&\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\pi_0)}\mathbb{1}_{e(X_i,\pi_0)\geq b_n}-\theta_0-\mathsf{B}_{n,b_n}\right) &\text{(I)}\\
&+\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\hat{\pi}_n)}\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}-\frac{D_iY_i}{e(X_i,\pi_0)}\mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right), &\text{(II)}
\end{aligned}$$

where asymptotic properties of (I) has been discussed in Theorem 1. For (II), we further expand it as

$$\text{(II)}=\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\left(\frac{D_iY_i}{e(X_i,\hat{\pi}_n)}-\frac{D_iY_i}{e(X_i,\pi_0)}\right)\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}}_{\text{(II.1)}}+\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\frac{D_iY_i}{e(X_i,\pi_0)}\left(\mathbb{1}_{e(X_i,\hat{\pi}_n)\geq b_n}-\mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right)}_{\text{(II.2)}}.$$

By the same argument used for the no trimming case, it satisfies

$$\text{(II.1)}=-\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}A_0h(D_i,X_i)+o_{\mathrm{p}}(1).$$

For (II.2), we first make some auxiliary calculation. Take π be a generic element in the parameter space Π,

$$\frac{e(X_i,\pi)}{e(X_i,\pi_0)}-1=\frac{1}{e(X_i,\pi_0)}\frac{\partial e(X_i,\tilde{\pi})}{\partial\pi}(\pi-\pi_0),$$

where $\tilde{\pi}$ is some convex combination of π and π_0. Next define

$$Z_i(\varepsilon)=\sup_{|\pi-\pi_0|\leq\varepsilon}\left|\frac{1}{e(X_i,\pi_0)}\frac{\partial e(X_i,\pi)}{\partial\pi}\right|.$$

Then we have

$$\left|\mathbb{1}_{e(X_i,\pi)\geq b_n}-\mathbb{1}_{e(X_i,\pi_0)\geq b_n}\right|\leq\mathbb{1}\left(\frac{b_n}{1+Z_i(\varepsilon)\varepsilon}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-Z_i(\varepsilon)\varepsilon}\right)+\mathbb{1}\left(|\pi-\pi_0|>\varepsilon\right).$$

Now fix some $K>0$ and let $\varepsilon=K/\sqrt{n}$ in the above, we have

$$\begin{aligned}
&|\text{(II.2)}|\\
\leq&\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\frac{D_i|Y_i|}{e(X_i,\pi_0)}\mathbb{1}\left(\frac{b_n}{1+Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}}\right)}_{\text{(II.2.1)}}+\underbrace{\text{(II.2)}\cdot\mathbb{1}\left(|\hat{\pi}_n-\pi_0|>\frac{K}{\sqrt{n}}\right)}_{\text{(II.2.2)}}.
\end{aligned}$$

Now take a sequence c_n, we expand (II.2.1) as

$$|\text{(II.2.1)}|\leq\underbrace{\frac{1}{a_{n,b_n}}\sum_{i=1}^{n}\frac{D_i|Y_i|}{e(X_i,\pi_0)}\mathbb{1}\left(\frac{b_n}{1+\frac{K}{\sqrt{n}}c_n}\leq e(X_i,\pi_0)\leq\frac{b_n}{1-\frac{K}{\sqrt{n}}c_n}\right)}_{\text{(II.2.1.1)}}+\underbrace{\text{(II.2.1)}\cdot\mathbb{1}\left(\max_{1\leq i\leq n}Z_i(\frac{K}{\sqrt{n}})>c_n\right)}_{\text{(II.2.1.2)}}.$$

Further,

$$\mathbb{E}[|(\text{II.2.1.1})|] \precsim \frac{n}{a_{n,b_n}}\left[F_{e(X)}\left(\frac{b_n}{1-\frac{K}{\sqrt{n}}c_n}\right) - F_{e(X)}\left(\frac{b_n}{1+\frac{K}{\sqrt{n}}c_n}\right)\right]$$
$$\precsim \frac{n}{a_{n,b_n}}F_{e(X)}(b_n)\left[\left(1+2\frac{\frac{K}{\sqrt{n}}c_n}{1-\frac{K}{\sqrt{n}}c_n}\right)^{\gamma_0-1} - 1\right] \asymp \frac{n}{a_{n,b_n}}F_{e(X)}(b_n)\frac{K}{\sqrt{n}}c_n.$$

In the above, $nF_{e(X)}(b_n)/a_{n,b_n}$ is the rate of the asymptotic bias, and hence to bound this term, it suffices to consider the heavy trimming scenario with $\gamma_0 < 2$. From Lemma 2, we have

$$\frac{n}{a_{n,b_n}}F_{e(X)}(b_n) \precsim \sqrt{nb_n\mathbb{P}[e(X) \le b_n]},$$

which means that

$$(\text{II.2.1.1}) \precsim_{\text{p}} \sqrt{b_n\mathbb{P}[e(X) \le b_n]}c_n \to 0,$$

by Markov's inequality and Assumption 4. In addition, term (II.2.1.2) is asymptotically negligible by Assumption 4. Therefore, for any $\varrho > 0$,

$$\limsup_n \mathbb{P}\left[|(\text{II.2})| > \varrho\right] = \limsup_n \mathbb{P}\left[|\hat{\pi}_n - \pi_0| > \frac{K}{2\sqrt{n}}\right].$$

The left-hand-side is independent of K and the right-hand-side decreases to 0 as $K \uparrow \infty$, we have that (II.2) converges in probability to zero. ■

IV.9 Omitted Details of Remark 2

Assumption 3(ii) and Assumption 4 in Logit models

We first consider Assumption 3(ii).

$$\left|\frac{\mathfrak{L}(X'\pi_0)}{\mathfrak{L}(X'\pi)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X'\pi)\right| = \left|\frac{\mathfrak{L}(X'\pi_0)}{\mathfrak{L}(X'\pi)}\Big(1-\mathfrak{L}(X'\pi)\Big)X\right| = \frac{e^{X'\pi_0}}{e^{X'\pi_0}+1}\frac{e^{X'\pi}+1}{e^{X'\pi}}\frac{1}{e^{X'\pi}+1}|X| \le e^{X'(\pi_0-\pi)}|X|.$$

Then

$$\mathbb{E}\left[\sup_{|\pi-\pi_0|\le\varepsilon}\left|\frac{\mathfrak{L}(X'\pi_0)}{\mathfrak{L}(X'\pi)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X'\pi)\right|\right] \le \mathbb{E}\left[e^{\varepsilon|X|}|X|\right] \le \sqrt{\mathbb{E}[e^{2\varepsilon|X|}]\mathbb{E}[|X|^2]},$$

which will be finite if we have $\mathbb{E}[e^{2\varepsilon|X|}] < \infty$ for some small $\varepsilon > 0$.

Now consider Assumption 4.

$$\left|\frac{1}{\mathfrak{L}(X'\pi_0)}\frac{\partial}{\partial\pi}\mathfrak{L}(X'\pi)\right| = \left|\frac{\mathfrak{L}(X'\pi)}{\mathfrak{L}(X'\pi_0)}\Big(1-\mathfrak{L}(X'\pi)\Big)X\right| = \frac{e^{X'\pi}}{e^{X'\pi}+1}\frac{e^{X'\pi_0}+1}{e^{X'\pi_0}}\frac{1}{e^{X'\pi}+1}|X|$$
$$\le e^{X'(\pi-\pi_0)}\frac{e^{X'\pi_0}+1}{e^{X'\pi}+1}|X| \le e^{X'(\pi-\pi_0)}\Big(e^{X'(\pi_0-\pi)}+1\Big)|X|$$
$$\le \left(e^{X'(\pi-\pi_0)}+1\right)|X|.$$

Therefore,

$$\max_{1\le i\le n}\sup_{\pi:|\pi-\pi_0|\le\varepsilon/\sqrt{n}}\left|\frac{1}{\mathfrak{L}(X_i'\pi_0)}\frac{\partial}{\partial\pi}\mathfrak{L}(X_i'\pi)\right| \le \max_{1\le i\le n}\left(e^{\varepsilon|X_i|/\sqrt{n}}+1\right)\left(\max_{1\le i\le n}|X_i|\right)$$
$$= O_{\text{p}}\left((n^{1/\sqrt{n}}+1)\log(n)\right) = O_{\text{p}}(\log(n)).$$

Assumption 3(ii) and Assumption 4 in Probit models

We first consider Assumption 3(ii).

$$\begin{aligned}\left|\frac{\mathfrak{L}(X'\pi_0)}{\mathfrak{L}(X'\pi)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X'\pi)\right| &= \left|\frac{\Phi(X'\pi_0)\phi(X'\pi)}{\Phi(X'\pi)^2}X\right| = \left|\mathbb{1}_{X'\pi\geq -2}\frac{\Phi(X'\pi_0)\phi(X'\pi)}{\Phi(X'\pi)^2}X + \mathbb{1}_{X'\pi\leq -2}\frac{\Phi(X'\pi_0)\phi(X'\pi)}{\Phi(X'\pi)^2}X\right| \\ &\leq \Phi(-2)^{-2}\Phi(X'\pi_0)\phi(X'\pi)|X| + \mathbb{1}_{X'\pi\leq -2}\frac{\Phi(X'\pi_0)\phi(X'\pi)}{\Phi(X'\pi)^2}|X| \\ &\leq \underbrace{\Phi(-2)^{-2}\Phi(X'\pi_0)\phi(X'\pi)|X|}_{\text{(I)}} + \underbrace{\mathbb{1}_{X'\pi\leq -2}\frac{\phi(X'\pi_0)}{\phi(X'\pi)}\left(\frac{|X'\pi|^3}{|X'\pi|^2-1}\right)^2|X|}_{\text{(II)}},\end{aligned}$$

where for the last line we use Proposition 2.1.2 of Vershynin (2018). Term (I) is easily bounded by

$$\mathbb{E}\left[\sup_{|\pi-\pi_0|\leq\varepsilon}|(\text{I})|\right] \leq \Phi(-2)^{-2}\phi(0)\mathbb{E}[|X|].$$

We can bound (II) by

$$(\text{II}) \leq 4\mathbb{1}_{X'\pi\leq -2}\exp\left\{\frac{1}{2}|X|^2|\pi+\pi_0||\pi-\pi_0|\right\}|X'\pi|^2|X|,$$

Hence

$$\mathbb{E}\left[\sup_{|\pi-\pi_0|\leq\varepsilon}|(\text{II})|\right] \leq 4(|\pi_0|+\varepsilon)^2\mathbb{E}\left[\exp\left\{\frac{1}{2}|X|^2\varepsilon(2|\pi_0|+\varepsilon)\right\}|X|^3\right],$$

which is finite if $\mathbb{E}[e^{\varepsilon(2|\pi_0|+\varepsilon)|X|^2}] < \infty$ for some small $\varepsilon > 0$.

Now consider Assumption 4.

$$\left|\frac{1}{\mathfrak{L}(X'\pi_0)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X'\pi)\right| = \left|\frac{\phi(X'\pi)}{\Phi(X'\pi_0)}X\right| = \frac{\phi(X'\pi)}{\phi(X'\pi_0)}\frac{\phi(X'\pi_0)}{\Phi(X'\pi_0)}|X|,$$

where we can further bound each terms in the above as

$$\frac{\phi(X'\pi)}{\phi(X'\pi_0)} = e^{\frac{1}{2}(|X'\pi_0|^2-|X'\pi|^2)} = e^{\frac{1}{2}(X'(\pi+\pi_0))(X'(\pi-\pi_0))} \leq e^{\frac{1}{2}|X|^2|\pi+\pi_0||\pi-\pi_0|},$$

and

$$\frac{\phi(X'\pi_0)}{\Phi(X'\pi_0)} \leq \frac{1}{\sqrt{2\pi}\Phi(-2)} + \mathbb{1}_{X'\pi_0\leq -2}\frac{|X'\pi_0|^3}{|X'\pi_0|^2-1} \leq \frac{1}{\sqrt{2\pi}\Phi(-2)} + \frac{|X|^3|\pi_0|^3}{|X|^2|\pi_0|^2-1},$$

where in the above, the first inequality is obtained by splitting into two event, $X'\pi_0 > -2$ and $X'\pi_0 \leq -2$, and then applying Proposition 2.1.2 of Vershynin (2018). As a result, we have

$$\begin{aligned}&\max_{1\leq i\leq n}\sup_{\pi:|\pi-\pi_0|\leq\varepsilon/\sqrt{n}}\left|\frac{1}{\mathfrak{L}(X_i'\pi_0)^2}\frac{\partial}{\partial\pi}\mathfrak{L}(X_i'\pi)\right| \\ \leq &\left(\max_{1\leq i\leq n}e^{\frac{1}{2}\varepsilon|X_i|^2|\pi_0|(1+\varepsilon/\sqrt{n})/\sqrt{n}}\right)\left(\max_{1\leq i\leq n}\frac{1}{\sqrt{2\pi}\Phi(-2)}|X_i| + \max_{1\leq i\leq n}\frac{|X_i|^4|\pi_0|^3}{|X_i|^2|\pi_0|^2-1}\right) \\ = &O_{\text{p}}\left(n^{1/\sqrt{n}}\left(\log(n)^{1/2}+\log(n)\right)\right) = O_{\text{p}}(\log(n)).\end{aligned}$$

■

IV.10 Proof of Lemma 2

The bias of $\hat{\theta}_{n,b_n}$ is quite easy to derive. Note that the IPW estimator $\hat{\theta}_n$ is unbiased for θ_0, hence the bias can be written as the following expectation:

$$\begin{aligned}\mathsf{B}_{n,b_n} = \mathbb{E}[\hat{\theta}_{n,b_n}] - \theta_0 &= -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{D_i Y_i}{e(X_i)}\mathbb{1}_{e(X_i)\le b_n}\right] \\ &= -\mathbb{E}\Big[\mathbb{E}[Y|e(X), D=1]\mathbb{1}_{e(X)<b_n}\Big] \approx -\mu_1(0)\cdot\mathbb{P}[e(X)\le b_n],\end{aligned}$$

so that the leading bias vanishes at the rate $\mathbb{P}[e(X) \le b_n]$, unless the data generating process is that the conditional mean shrinks as the probability weight approaches zero.

For the variance of $DY/e(X)\mathbb{1}_{e(X)\ge b_n}$, we note that when $\gamma_0 \in (1,2)$ and $b_n \to 0$, it diverges to infinity. As a result,

$$\mathsf{V}_{n,b_n} = \frac{1}{n}\mathbb{V}\left[\frac{DY}{e(X)}\mathbb{1}_{e(X)\ge b_n}\right] \approx \frac{1}{n}\mathbb{E}\left[\frac{DY^2}{e(X)^2}\mathbb{1}_{e(X)\ge b_n}\right] = \frac{1}{n}\int_{b_n}^{1}\frac{\mathbb{E}[Y^2|e(X)=x, D=1]}{x}\mathrm{d}\mathbb{P}[e(X)\le x].$$

Recall that $\mu_2(0) = \lim_{x\to 0}\mathbb{E}[Y^2|e(X)=x, D=1]$. Choose $c>0$ small enough so that

$$\sup_{x\le c}\Big|\mathbb{E}[Y^2|e(X)=x, D=1] - \mu_2(0)\Big| \le \eta.$$

Then

$$\frac{\int_{b_n}^{1}\mu_2(0)x^{-1}F_{e(X)}(\mathrm{d}x)}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} = 1 + \frac{A+B-C}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)},$$

where

$$\begin{aligned}A &= \int_{c}^{1}\mu_2(0)x^{-1}F_{e(X)}(\mathrm{d}x) \\ B &= \int_{b_n}^{c}\Big(\mu_2(0) - \mathbb{E}[Y^2|e(X)=x, D=1]\Big)x^{-1}F_{e(X)}(\mathrm{d}x) \\ C &= \int_{c}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x).\end{aligned}$$

Note that both A and C are bounded, which means

$$\frac{A}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \to 0, \quad \frac{C}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \to 0.$$

For B, we have

$$\frac{B}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \le \frac{\eta}{\inf_{x\in[0,c]}\mathbb{E}[Y^2|e(X)=x, D=1]},$$

which can be made arbitrarily small (for η close to zero, we can choose c close to zero, which means the denominator will be close to $\mu_2(0) > 0$). Therefore,

$$\frac{\int_{b_n}^{1}\mu_2(0)x^{-1}F_{e(X)}(\mathrm{d}x)}{\int_{b_n}^{1}\mathbb{E}[Y^2|e(X)=x, D=1]x^{-1}F_{e(X)}(\mathrm{d}x)} \to 1.$$

For the final claim, we first note, by a slight modification of Lemma S.2,

$$\frac{b_n^{-1}\mathbb{P}[e(X) \leq b_n]}{\mathbb{E}[e(X)^{-1}\mathbb{1}_{e(X)\geq b_n}]} \to \frac{2-\gamma_0}{\gamma_0 - 1},$$

as $b_n \to 0$. ■

IV.11 Proof of Theorem 2

Part 1: using true probability weights

Let $\hat{F}_{e(X)}(x) = \sum_{i=1}^n \mathbb{1}_{e(X)\leq x}/n$, $c_0 = \mu_2(0)/2\mu_1(0)^2$, and $\hat{c}_n$ be an estimator of c_0. We first consider the behavior of $b^s\hat{F}_{e(X)}(b)$ at b_n (defined in the theorem), which is given by the following probability bound (Markov's inequality):

$$\begin{aligned}
\mathbb{P}\left[n\left|b_n^s\hat{F}_{e(X)}(b_n) - b_n^s F_{e(X)}(b_n)\right| > \delta\right] &\leq n^2\left(\frac{b_n^s}{\delta}\right)^2 \mathbb{E}\left|\hat{F}_{e(X)}(b_n) - F_{e(X)}(b_n)\right|^2 \\
&= n\left(\frac{b_n^s}{\delta}\right)^2 \mathbb{V}\left[\mathbb{1}_{e(X)\leq b_n}\right] \\
&= n\left(\frac{b_n^s}{\delta}\right)^2 F_{e(X)}(b_n)(1 - F_{e(X)}(b_n)) \\
&= \frac{c_0}{\delta^2} b_n^s\Big(1 + o(1)\Big),
\end{aligned}$$

which implies

$$n\left|b_n^s\hat{F}_{e(X)}(b_n) - b_n^s F_{e(X)}(b_n)\right| \xrightarrow{\text{p}} 0.$$

To complete the proof, take some constant $a \in (0,1)$, and define $b_{l,n}$ and $b_{r,n}$ as:

$$b_{l,n}^s F_{e(X)}(b_{l,n}) = \frac{ac_0}{n}, \quad b_{r,n}^s F_{e(X)}(b_{r,n}) = \frac{c_0}{an}.$$

Then it is easy to see that

$$\begin{aligned}
\mathbb{P}\left[\hat{b}_n \leq b_{l,n}\right] &\leq \mathbb{P}\left[b_{l,n}^s\hat{F}_{e(X)}(b_{l,n}) \geq \hat{b}_n^s\hat{F}_{e(X)}(\hat{b}_n)\right] = \mathbb{P}\left[b_{l,n}^s\hat{F}_{e(X)}(b_{l,n}) \geq \frac{\hat{c}_n}{n}\right] \\
&= \mathbb{P}\left[b_{l,n}^s\hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n}) \geq \frac{(1-a)c_0 + (\hat{c}_n - c_0)}{n}\right] \\
&= \mathbb{P}\Big[n\left(b_{l,n}^s\hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n})\right) \geq \underbrace{(1-a)c_0 + (\hat{c}_n - c_0)}_{\xrightarrow{\text{p}}(1-a)c_0 > 0}\Big] \to 0,
\end{aligned}$$

as the first term $n\left(b_{l,n}^s\hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n})\right)$ is $o_{\text{p}}(1)$. Using a similar technique, we can show that $\mathbb{P}[\hat{b}_n \geq b_{r,n}] \to 0$. Therefore,

$$\mathbb{P}\left[b_{l,n} \leq \hat{b}_n \leq b_{r,n}\right] = \mathbb{P}\left[\frac{b_{l,n}}{b_n} \leq \frac{\hat{b}_n}{b_n} \leq \frac{b_{r,n}}{b_n}\right] \to 1.$$

As the choice of a is arbitrary, we only need to show that both $b_{l,n}/b_n$ and $b_{r,n}/b_n$ are arbitrarily close to 1 for all a close to 1. To see this, note that since $b_n \to 0$, one has

$$a = \frac{b_{l,n}^s F_{e(X)}(b_{l,n})}{b_n^s F_{e(X)}(b_n)} = \underbrace{\frac{b_{l,n}^s}{b_n^s} \frac{F_{e(X)}((b_{l,n}/b_n)b_n)}{F_{e(X)}(b_n)}}_{\to (b_{l,n}/b_n)^{\gamma_0 - 1}} = \left(\frac{b_{l,n}}{b_n}\right)^{\gamma_0 - 1 + s} (1 + o(1)).$$

and the same argument applies to $b_{r,n}$.

Part 2: using estimated probability weights

To show that estimated probability weights can be employed, we only need to show that for all $\delta > 0$,

$$\mathbb{P}\left[n \left| b_n^s \hat{F}_{\hat{e}(X)}(b_n) - b_n^s \hat{F}_{e(X)}(b_n) \right| > \delta\right] \to 0,$$

where again b_n is defined in the theorem. From the proof of Proposition 1, we have, for any $|\pi - \pi_0| \le \varepsilon$,

$$\left| \mathbb{1}_{e(X_i,\pi) \ge b_n} - \mathbb{1}_{e(X_i,\pi_0) \ge b_n} \right| \le \mathbb{1}\left(\frac{b_n}{1 + Z_i(\varepsilon)\varepsilon} \le e(X_i, \pi_0) \le \frac{b_n}{1 - Z_i(\varepsilon)\varepsilon} \right),$$

where

$$Z_i(\varepsilon) = \sup_{|\pi - \pi_0| \le \varepsilon} \left| \frac{1}{e(X_i, \pi_0)} \frac{\partial e(X_i, \pi)}{\partial \pi} \right|.$$

Therefore, for any $K > 0$,

$$\begin{aligned}
&\mathbb{P}\left[n \left| b_n^s \hat{F}_{\hat{e}(X)}(b_n) - b_n^s \hat{F}_{e(X)}(b_n) \right| > \delta\right] \\
&\le \mathbb{P}\left[b_n^s \sum_{i=1}^n \mathbb{1}\left(\frac{b_n}{1 + Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}} \le e(X_i, \pi_0) \le \frac{b_n}{1 - Z_i(\frac{K}{\sqrt{n}})\frac{K}{\sqrt{n}}} \right) > \delta \right] + \mathbb{P}\left[|\hat{\pi}_n - \pi_0| \ge \frac{K}{\sqrt{n}} \right] \\
&\le \mathbb{P}\left[b_n^s \sum_{i=1}^n \mathbb{1}\left(\frac{b_n}{1 + \frac{K}{\sqrt{n}} c_n} \le e(X_i, \pi_0) \le \frac{b_n}{1 - \frac{K}{\sqrt{n}} c_n} \right) > \delta \right] + \mathbb{P}\left[\max_{1 \le i \le n} Z_i(\frac{K}{\sqrt{n}}) > c_n \right] + \mathbb{P}\left[|\hat{\pi}_n - \pi_0| \ge \frac{K}{\sqrt{n}} \right],
\end{aligned}$$

where c_n is to be specified. For the first term, one has

$$\begin{aligned}
&\mathbb{E}\left[b_n^s \sum_{i=1}^n \mathbb{1}\left(\frac{b_n}{1 + \frac{K}{\sqrt{n}} c_n} \le e(X_i, \pi_0) \le \frac{b_n}{1 - \frac{K}{\sqrt{n}} c_n} \right) \right] = n b_n^s \left[F_{e(X)}\left(\frac{b_n}{1 - \frac{K}{\sqrt{n}} c_n} \right) - F_{e(X)}\left(\frac{b_n}{1 + \frac{K}{\sqrt{n}} c_n} \right) \right] \\
&\precsim n b_n^s F_{e(X)}(b_n) \left[\left(1 + 2 \frac{\frac{K}{\sqrt{n}} c_n}{1 - \frac{K}{\sqrt{n}} c_n} \right)^{\gamma_0 - 1} - 1 \right] \asymp n b_n^s F_{e(X)}(b_n) \frac{K}{\sqrt{n}} c_n \asymp \frac{K}{\sqrt{n}} c_n \to 0,
\end{aligned}$$

which holds if $c_n = \sqrt{n/\log(n)}$. By our assumption,

$$\mathbb{P}\left[\max_{1 \le i \le n} Z_i\left(\frac{K}{\sqrt{n}} \right) > c_n \right] \to 0.$$

Finally,

$$\mathbb{P}\left[|\hat{\pi}_n - \pi_0| \ge \frac{K}{n} \right]$$

can be made arbitrarily small by taking K large. As

$$\mathbb{P}\left[n \left| b_n^s \hat{F}_{\hat{e}(X)}(b_n) - b_n^s \hat{F}_{e(X)}(b_n) \right| > \delta\right]$$

does not depend on K, this probability converges to 0 for all $\delta > 0$. ■

IV.12 Proof of Proposition 2

We only demonstrate part (ii). To show that Proposition 1 holds with data-driven trimming threshold, first let b_n be defined from

$$b_n^s \mathbb{P}[e(X) \leq b_n] = \frac{1}{2n} \frac{\mu_2(0)}{\mu_1(0)^2},$$

for some $s > 0$. Recall that $s < 1$, $s = 1$ and $s > 1$ correspond to light, moderate and heavy trimming, respectively. Let $\hat{b}_n$ be the estimated trimming threshold from

$$\hat{b}_n^s \left(\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\hat{e}(X) \leq \hat{b}_n} \right) = \frac{1}{2n} \frac{\hat{\mu}_2(0)}{\hat{\mu}_1(0)^2}.$$

Then we consider the following:

$$\begin{aligned}
&\left| \frac{1}{a_{n,b_n}} \sum_{i=1}^n \frac{D_i Y_i}{\hat{e}(X_i)} \left(\mathbb{1}_{\hat{e}(X_i) \geq \hat{b}_n} - \mathbb{1}_{\hat{e}(X_i) \geq b_n} \right) \right| \leq \frac{1}{a_{n,b_n}} \sum_{i=1}^n \frac{D_i |Y_i|}{\hat{e}(X_i)} \left| \mathbb{1}_{\hat{e}(X_i) \geq \hat{b}_n} - \mathbb{1}_{\hat{e}(X_i) \geq b_n} \right| \\
\leq & \underbrace{\left(\frac{1}{a_{n,b_n}} \sum_{i=1}^n \frac{D_i |Y_i|}{\hat{e}(X_i)} \right) \mathbb{1}_{|\hat{b}_n / b_n - 1| \geq \varepsilon_n}}_{\text{(I)}} + \underbrace{\frac{1}{a_{n,b_n}} \sum_{i=1}^n \frac{D_i |Y_i|}{\hat{e}(X_i)} \mathbb{1}_{(1-\varepsilon_n) b_n \leq e(X_i) \leq (1+\varepsilon_n) b_n}}_{\text{(II)}} \\
& + \underbrace{\frac{1}{a_{n,b_n}} \sum_{i=1}^n \frac{D_i |Y_i|}{\hat{e}(X_i)} \left| \mathbb{1}_{(1-\varepsilon_n) b_n \leq e(X_i) \leq (1+\varepsilon_n) b_n} - \mathbb{1}_{(1-\varepsilon_n) b_n \leq \hat{e}(X_i) \leq (1+\varepsilon_n) b_n} \right|}_{\text{(III)}},
\end{aligned}$$

where $0 < \varepsilon_n < 1$ will be specified later. Employing the same technique used to prove Proposition 1, one can show that (III) $= o_{\mathrm{p}}(1)$. Now we consider (I).

$$\text{(I)} \precsim_{\mathrm{p}} \mathbb{P}\left[|\hat{b}_n / b_n - 1| \geq \varepsilon_n \right] = \underbrace{\mathbb{P}\left[\hat{b}_n \leq (1-\varepsilon_n) b_n \right]}_{\text{(I.1)}} + \underbrace{\mathbb{P}\left[\hat{b}_n \geq (1+\varepsilon_n) b_n \right]}_{\text{(I.2)}}.$$

Next, let $b_{l,n} = (1-\varepsilon_n) b_n$, $c_0 = \mu_2(0)/2\mu_1(0)^2$ and $\hat{c}_n = \hat{\mu}_2(0)/2\hat{\mu}_1(0)^2$, then

$$\begin{aligned}
\text{(I.1)} &\leq \mathbb{P}\left[b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) \geq \hat{b}_n^s \hat{F}_{e(X)}(\hat{b}_n) \right] = \mathbb{P}\left[b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) \geq \frac{\hat{c}_n}{n} \right] \\
&\asymp \mathbb{P}\left[b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n}) \geq \frac{\hat{c}_n}{n} - (1-\varepsilon_n)^{\gamma_0+s-1} \frac{c_0}{n} \right] \\
&= \mathbb{P}\left[n \left| b_{l,n}^s \hat{F}_{e(X)}(b_{l,n}) - b_{l,n}^s F_{e(X)}(b_{l,n}) \right| \geq \hat{c}_n - c_0 + \left(1 - (1-\varepsilon_n)^{\gamma_0+s-1} \right) c_0 \right] \\
&\asymp \left[\hat{c}_n - c_0 + \left(1 - (1-\varepsilon_n)^{\gamma_0+s-1} \right) c_0 \right]^{-2} b_n^s,
\end{aligned}$$

which tends to zero provided that $b_n^s / \varepsilon_n^2 \to 0$. It can be shown that (I.2) is also negligible under the same condition. We take $\varepsilon_n = b_n^{s/2-1/4}$. For (II),

$$\begin{aligned}
\text{(II)} &\precsim_{\mathrm{p}} \frac{n}{a_{n,b_n}} \mathbb{P}\left[(1-\varepsilon_n) b_n \leq e(X_i) \leq (1+\varepsilon_n) b_n \right] \asymp \frac{n}{a_{n,b_n}} F_{e(X)}(b_n) \left[\left(1 + \frac{2\varepsilon_n}{1-\varepsilon_n} \right)^{\gamma_0 - 1} - 1 \right] \\
&\asymp \frac{n}{a_{n,b_n}} F_{e(X)}(b_n) \varepsilon_n \precsim \sqrt{n b_n \mathbb{P}[e(X) \leq b_n]} \varepsilon_n \asymp \sqrt{b_n^{1-s} \varepsilon_n^2} = \sqrt{b_n^{1-s} b_n^{s-1/2}} = b_n^{1/4} \to 0,
\end{aligned}$$

where the second line in the above follows from Lemma 2. ■

IV.13 Proof of Theorem 3

For ease of presentation, we assume the true probability weights are used in the local polynomial regression. We split the proof into two parts, according to the behavior of $nF_{e(X)}(b_n)$.

Part 1: $nF_{e(X)}(b_n) \to 0$

With $nF_{e(X)}(b_n) \to 0$, it is clear that this falls into the light trimming scenario. To show that our bias correction does not contribute to the limiting distribution, note that

$$\frac{n}{a_{n,b_n}}|\hat{\mathsf{B}}_{n,b_n} - \mathsf{B}_{n,b_n}| \leq \underbrace{\frac{n}{a_{n,b_n}}|\mathsf{B}_{n,b_n}|}_{o_{\mathrm{p}}(1),\text{ due to light trimming}} + \frac{n}{a_{n,b_n}}\left|\hat{\mathsf{B}}_{n,b_n}\right|.$$

The second term has expansion

$$\frac{n}{a_{n,b_n}}\left|\hat{\mathsf{B}}_{n,b_n}\right| \leq \underbrace{\left|\sum_{j=0}^{p}\hat{\beta}_j\right|}_{O_{\mathrm{p}}(1),\text{ Lemma S.6}} \frac{n}{a_{n,b_n}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{e(X_i)\leq b_n},$$

where by Markov's inequality,

$$\frac{n}{a_{n,b_n}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{e(X_i)\leq b_n} = O_{\mathrm{p}}\left(\frac{n}{a_{n,b_n}}\mathbb{E}[\mathbb{1}_{e(X_i)\leq b_n}]\right) = O_{\mathrm{p}}\left(\frac{n}{a_{n,b_n}}F_{e(X)}(b_n)\right) = o_{\mathrm{p}}(1).$$

Part 1: $nF_{e(X)}(b_n) \gtrsim 1$

We continue our proof assuming $nF_{e(X)}(b_n) \gtrsim 1$. Note that the true bias B_{n,b_n} has order $F_{e(X)}(b_n)$, hence we consider the relative accuracy:

$$\frac{n}{a_{n,b_n}}|\hat{\mathsf{B}}_{n,b_n} - \mathsf{B}_{n,b_n}| \sim \left(\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}\right)\frac{|\hat{\mathsf{B}}_{n,b_n} - \mathsf{B}_{n,b_n}|}{F_{e(X)}(b_n)} \leq \left(\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}\right)\Big(\text{(I)} + \text{(II)} + \text{(III)}\Big),$$

where, by a $(p+1)$-th order Taylor expansion,

$$\begin{aligned}
\text{(I)} &= \sum_{j=0}^{p}(\text{I})_j = \sum_{j=0}^{p}\left(\frac{|\hat{\mu}_1^{(j)}(0) - \mu_1^{(j)}(0)|}{j!F_{e(X)}(b_n)}\frac{1}{n}\sum_{i=1}^{n}e(X_i)^j\mathbb{1}_{e(X_i)\leq b_n}\right),\\
\text{(II)} &= \frac{1}{(p+1)!F_{e(X)}(b_n)}\frac{1}{n}\sum_{i=1}^{n}\mu_1^{(p+1)}(\lambda_i e(X_i))e(X_i)^{p+1}\mathbb{1}_{e(X_i)\leq b_n},\\
\text{(III)} &= \frac{1}{nF_{e(X)}(b_n)}\sum_{i=1}^{n}\Big(\mu_1(e(X_i))\mathbb{1}_{e(X_i)\leq b_n} - \mathbb{E}[\mu_1(e(X_i))\mathbb{1}_{e(X_i)\leq b_n}]\Big),
\end{aligned}$$

with $\lambda_i \in [0,1]$.

Term (III) has zero mean and variance:

$$\mathbb{V}[\text{(III)}] = \frac{1}{nF_{e(X)}(b_n)^2}\mathbb{V}\left[\mu_1(e(X_i))\mathbb{1}_{e(X_i)\leq b_n}\right] \precsim \frac{1}{nF_{e(X)}(b_n)}.$$

Therefore,

$$\left(\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}\right)(\text{III}) \precsim_{\text{p}} \sqrt{\frac{nF_{e(X)}(b_n)}{a_{n,b_n}^2}} \asymp \sqrt{\frac{\sqrt{n}F_{e(X)}(b_n)}{a_{n,b_n}}} \asymp \sqrt{\frac{1}{\sqrt{n}}\sqrt{nb_nF_{e(X)}(b_n)}} = o(1),$$

where we use $nF_{e(X)}(b_n)/a_{n,b_n} \asymp \sqrt{nb_nF_{e(X)}(b_n)}$ from Lemma 2.

Next, for $0 \leq j \leq p$,

$$\mathbb{E}[e(X)^j \mathbb{1}_{e(X)\leq b_n}] = \int_0^{b_n} x^j F_{e(X)}(\mathrm{d}x) = b_n^j F_{e(X)}(b_n) - \int_0^{b_n} jx^{j-1}F_{e(X)}(x)\mathrm{d}x \asymp b_n^j F_{e(X)}(b_n),$$

and its variance has order:

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n e(X_i)^j \mathbb{1}_{e(X)\leq b_n}\right] \leq \frac{1}{n}\mathbb{E}[e(X_i)^{2j}\mathbb{1}_{e(X)\leq b_n}] \asymp \frac{1}{n}F_{e(X)}(b_n)b_n^{2j}.$$

Therefore

$$\frac{1}{F_{e(X)}(b_n)}\frac{1}{n}\sum_{i=1}^n e(X_i)^j \mathbb{1}_{e(X)\leq b_n} = O_{\text{p}}\left(b_n^j\right),$$

which implies that (II) has order:

$$(\text{II}) = O_{\text{p}}\left(b_n^{p+1}\right).$$

By Lemma S.6, term (I) has order:

$$(\text{I}) = O_{\text{p}}\left(\sum_{j=0}^p \sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}}\left(\frac{b_n}{h_n}\right)^j\right) = O_{\text{p}}\left(\sqrt{\frac{1}{nh_nF_{e(X)}(h_n)}}\right).$$

Now we apply Lemma 2 again, which leads to

$$\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}(\text{II}) \precsim_{\text{p}} \sqrt{nb_n^{2p+3}F_{e(X)}(b_n)} \to 0,$$

and

$$\frac{n}{a_{n,b_n}}\mathsf{B}_{n,b_n}(\text{I}) \precsim_{\text{p}} \sqrt{\frac{nb_nF_{e(X)}(b_n)}{nh_nF_{e(X)}(h_n)}} \to 0.$$

■

IV.14 Proof of Theorem 4

Part 1: no trimming ($b_n = 0$)

Define

$$Z = \frac{DY}{e(X)} - \theta_0, \qquad U_n = \frac{1}{a_n}\sum_{i=1}^n Z_i, \qquad V_n = \sqrt{\frac{1}{a_n^2}\sum_{i=1}^n Z_i^2},$$

and recall that we set $a_{n,b_n} = a_n$ if there is no trimming. We first establish the joint limiting distribution of (U_n, V_n^2) under $\gamma_0 < 2$, which is the only interesting case. (Otherwise the self-normalized statistic is asymptotically Gaussian). The argument relies on a modification of the method in Chapter XVII of Feller (1991). To

start, consider the characteristic function

$$\mathbb{E}\left[e^{i(\zeta_1 U_n+\zeta_2 V_n^2)}\right]=\left(\mathbb{E}\left[e^{i(\zeta_1 W_n+\zeta_2 W_n^2)}\right]\right)^n=\left(1+\frac{1}{n}\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}nx^2F_{W_n}(\mathrm{d}x)\right)^n,$$

where

$$W_n=\frac{Z}{a_n}.$$

to proceed, let $K:\mathbb{R}\to(0,\infty)$ be an auxiliary function which is smooth, symmetric, and satisfies $\lim_{x\to\infty}xK(x)=1$. Take, for example, $I=[c_1,c_2]$ to be a compact interval with $0\le c_1<c_2$, following the same argument used to prove Theorem 1,

$$\begin{aligned}
&\int_I K(x)nx^2F_{W_n}(\mathrm{d}x)=\frac{n}{a_n^2}\int_{a_nc_1}^{a_nc_2}K\left(\frac{x}{a_n}\right)x^2F_Z(\mathrm{d}x)\\
=&n\left[K(c_2)c_2^2F_Z(a_nc_2)-K(c_1)c_1^2F_Z(a_nc_1)-\int_{c_1}^{c_2}\left(2xK\left(x\right)+x^2K^{(1)}\left(x\right)\right)F_Z(a_nx)\mathrm{d}x\right]\\
=&n\left[-K(c_2)c_2^2\Big(1-F_Z(a_nc_2)\Big)+K(c_1)c_1^2\Big(1-F_Z(a_nc_1)\Big)+\int_{c_1}^{c_2}\left(2xK\left(x\right)+x^2K^{(1)}\left(x\right)\right)\Big(1-F_Z(a_nx)\Big)\mathrm{d}x\right]\\
\to&\frac{2-\gamma_0}{\gamma_0}\frac{\alpha_+(0)}{\alpha_+(0)+\alpha_-(0)}\left[-K(c_2)c_2^{2-\gamma_0}+K(c_1)c_1^{2-\gamma_0}+\int_{c_1}^{c_2}\left(2xK\left(x\right)+x^2K^{(1)}\left(x\right)\right)x^{-\gamma_0}\mathrm{d}x\right]\\
=&M_K(I),
\end{aligned}$$

where the measure $M_K(\mathrm{d}x)$ is defined as

$$M_K(\mathrm{d}x)=\mathrm{d}x\left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)}K(x)|x|^{1-\gamma_0}\Big(\alpha_+(0)\mathbb{1}_{x\ge0}+\alpha_-(0)\mathbb{1}_{x<0}\Big)\right].$$

The same convergence holds for compact intervals $[c_1,c_2]$ with $c_2\le0$. Finally, we note that

$$\int_{\mathbb{R}}K(x)nx^2F_{W_n}(\mathrm{d}x)\to M_K(\mathbb{R})\in(0,\infty).$$

Therefore, we have the following distributional convergence:

$$\frac{K(x)nx^2F_{W_n}(\mathrm{d}x)}{\int_{\mathbb{R}}K(x)nx^2F_{W_n}(\mathrm{d}x)}\xrightarrow{\mathrm{d}}\frac{M_K(\mathrm{d}x)}{M_K(\mathbb{R})}.$$

As the following is bounded and is continuous in x

$$\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2K(x)}$$

for any $\zeta_1,\zeta_2\in\mathbb{R}$, we have

$$\begin{aligned}
&\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}nx^2F_{W_n}(\mathrm{d}x)=\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2K(x)}K(x)nx^2F_{W_n}(\mathrm{d}x)\\
\to&\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2K(x)}M_K(\mathrm{d}x)=\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}M(\mathrm{d}x),
\end{aligned}$$

where $M(\mathrm{d}x)$ is defined in Theorem 1(ii.1) with $t=0$. To summarize, we showed:

$$\mathbb{E}\left[e^{i(\zeta_1 U_n+\zeta_2 V_n^2)}\right]\to\exp\left\{\int_{\mathbb{R}}\frac{e^{i(\zeta_1 x+\zeta_2 x^2)}-1-i\zeta_1 x}{x^2}M(\mathrm{d}x)\right\}.$$

A similar result was derived in Logan *et al.* (1973). However, our argument only relies on the fact that Z has a regularly varying tail, while they employ the stronger assumption that Z follows a Lévy stable distribution. Given the joint limiting characteristic function, Logan *et al.* (1973) showed that the limiting distribution does not have positive mass on $\mathbb{R} \times \{0\}$, implying that U_n/V_n has a well-defined limiting distribution. Further, the limiting distribution has a smooth density function.

For the self-normalized statistic T_n in Theorem 4, we rely on Proposition 1, which claims that estimating the probability weights in a first step does not contribute to the limiting distribution when $\gamma_0 < 2$. Then with simple algebra,

$$T_n = \frac{U_n}{V_n}\sqrt{\frac{n-1}{n-V_n^2}}.$$

As a result, T_n has the same limiting distribution as U_n/V_n. Therefore, subsampling is valid by standard arguments in Politis and Romano (1994) or Romano and Wolf 1999.

Part 2: trimming ($b_n > 0$)

Define:

$$Z = \frac{DY}{e(X)}\mathbb{1}_{e(X)\geq b_n} - \theta_0 - \mathsf{B}_{n,b_n}, \qquad U_n = \frac{1}{a_{n,b_n}}\sum_{i=1}^n Z_i, \qquad V_n = \sqrt{\frac{1}{a_{n,b_n}^2}\sum_{i=1}^n Z_i^2}.$$

Similar as in the previous part, we first establish the joint limiting distribution of (U_n, V_n^2). Consider the characteristic function:

$$\begin{aligned}\mathbb{E}\left[e^{i(\zeta_1 U_n + \zeta_2 V_n^2)}\right] &= \left(\mathbb{E}\left[e^{i(\zeta_1 W_n + \zeta_2 W_n^2)}\right]\right)^n \\ &= \left(1 + \frac{1}{n}\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} n x^2 F_{W_n}(\mathrm{d}x)\right)^n,\end{aligned}$$

where $W_n = Z/a_{n,b_n}$. Again, let $K : \mathbb{R} \to (0,\infty)$ be an auxiliary function that is smooth, symmetric, and satisfies $\lim_{x\to\infty} xK(x) = 1$.

We split the rest of proof into three cases, the light, moderate and heavy trimming. For all three cases, we show U_n/V_n has a well-defined limiting distribution. And since we focus on $\gamma_0 < 2$, the impact of estimating the probability weights can be ignored. Therefore, the self-normalized statistic T_{n,b_n} has the same limiting distribution as U_n/V_n, and subsampling is valid by standard arguments in Politis and Romano (1994), or Romano and Wolf 1999.

Part 2-1: light trimming ($b_n a_n \to 0$)

The proof is essentially the same as that of the no trimming case. Take, for example, $I = [c_1, c_2]$ to be a compact interval with $0 \leq c_1 < c_2$, then

$$\begin{aligned}&\int_I K(x) n x^2 F_{W_n}(\mathrm{d}x) = \frac{n}{a_{n,b_n}^2}\int_{a_{n,b_n}c_1}^{a_{n,b_n}c_2} K\left(\frac{x}{a_{n,b_n}}\right) x^2 \mathrm{d}F_Z(x) \\ =&n\Big[-K(c_2)c_2^2\Big(1 - F_Z(a_{n,b_n}c_2)\Big) + K(c_1)c_1^2\Big(1 - F_Z(a_{n,b_n}c_1)\Big) + \int_{c_1}^{c_2}\Big(2xK(x) + x^2K^{(1)}(x)\Big)\Big(1 - F_Z(a_{n,b_n}x)\Big)\mathrm{d}x\Big].\end{aligned}$$

The tail probabilities can be calculated as in the proof of Theorem 1(ii.1) with light trimming:

$$\int_I K(x) n x^2 F_{W_n}(\mathrm{d}x) \to M_K(I),$$

where the measure $M_K(\mathrm{d}x)$ is

$$M^{\dagger}(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)} K(x)|x|^{1-\gamma_0} \Big(\alpha_+(0)\mathbb{1}_{x\geq 0} + \alpha_-(0)\mathbb{1}_{x<0} \Big) \right].$$

The same convergence holds for compact intervals $[c_1, c_2]$ with $c_2 \leq 0$. Finally, we note that

$$\int_{\mathbb{R}} K(x) n x^2 F_{W_n}(\mathrm{d}x) \to M_K(\mathbb{R}) \in (0, \infty).$$

Therefore, we have the following distributional convergence:

$$\frac{K(x) n x^2 F_{W_n}(\mathrm{d}x)}{\int_{\mathbb{R}} K(x) n x^2 F_{W_n}(\mathrm{d}x)} \xrightarrow{\mathrm{d}} \frac{M_K(\mathrm{d}x)}{M_K(\mathbb{R})}.$$

Since the following is bounded and continuous,

$$\frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2 K(x)},$$

we have

$$\int_{\mathbb{R}} \frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} n x^2 F_{W_n}(\mathrm{d}x) \to \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} M(\mathrm{d}x),$$

where

$$M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)} |x|^{1-\gamma_0} \Big(\alpha_+(0)\mathbb{1}_{x\geq 0} + \alpha_-(0)\mathbb{1}_{x<0} \Big) \right],$$

as defined in Theorem 1(ii) for the light trimming scenario. To summarize, we showed:

$$\mathbb{E}\left[e^{i(\zeta_1 U_n + \zeta_2 V_n^2)} \right] \to \exp\left\{ \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} M(\mathrm{d}x) \right\},$$

which defines the joint limiting distribution of (U_n, V_n^2).

Part 2-2: moderate trimming ($b_n a_n \to t \in (0, \infty)$)

We do not repeat the lengthy argument. With the tail probability calculations used to prove the moderate trimming scenario of Theorem 1(ii.1), one has

$$\mathbb{E}\left[e^{i(\zeta_1 U_n + \zeta_2 V_n^2)} \right] \to \exp\left\{ \int_{\mathbb{R}} \frac{e^{i(\zeta_1 x + \zeta_2 x^2)} - 1 - i\zeta_1 x}{x^2} M(\mathrm{d}x) \right\},$$

where

$$M(\mathrm{d}x) = \mathrm{d}x \left[\frac{2-\gamma_0}{\alpha_+(0)+\alpha_-(0)} |x|^{1-\gamma_0} \Big(\alpha_+(tx)\mathbb{1}_{x\geq 0} + \alpha_-(tx)\mathbb{1}_{x<0} \Big) \right].$$

Part 2-3: heavy trimming ($b_n a_n \to \infty$)

This case is much easier, and one can directly show that U_n/V_n converges to the standard Gaussian distribution. ■

IV.15 Proof of Proposition S.1

Rewrite the estimator as

$$\hat{\tau}^{\mathtt{ATT}}_{n,b_n} = \frac{c_0}{\hat{c}_n}\frac{1}{n}\sum_{i=1}^{n}\frac{(D_i - e(X_i))Y_i}{c_0(1-e(X_i))}\mathbb{1}_{1-e(X_i)\geq(1-D_i)b_n},$$

where $c_0 = \mathbb{P}[D=1]$, and $\hat{c}_n = n^{-1}\sum_{i=1}^{n} D_i$. We first consider the tail behavior of $(D-e(X))Y/(c_0(1-e(X))$. Note that

$$\mathbb{P}\left[\frac{(D-e(X))Y}{c_0(1-e(X))} > x\right] = \mathbb{P}[D=1]\mathbb{P}\left[\frac{Y(1)}{c_0} > x\middle| D=1\right] + \mathbb{P}[D=0]\mathbb{P}\left[\frac{e(X)Y(0)}{c_0(1-e(X))} < -x\middle| D=0\right],$$

where we take $x>0$. To proceed, let $F_{1-e(X)}$ be the distribution function of $1-e(X)$, then

$$\begin{aligned}
&\lim_{x\downarrow 0}\frac{\mathbb{P}[1-e(X)\leq x|D=0]}{x\mathbb{P}[1-e(X)\leq x]} = \lim_{x\downarrow 0}\frac{\mathbb{P}[D=0|1-e(X)\leq x]}{x\mathbb{P}[D=0]}\\
&= \lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[1-e(X)\leq x]\mathbb{P}[D=0]}\int_0^x y F_{1-e(X)}(\mathrm{d}y)\\
&= \lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[1-e(X)\leq x]\mathbb{P}[D=0]}\left(xF_{1-e(X)}(x) - \int_0^x F_{1-e(X)}(y)\mathrm{d}y\right)\\
&= \lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[1-e(X)\leq x]\mathbb{P}[D=0]}\left(xF_{1-e(X)}(x) - \int_0^1 xF_{1-e(X)}(xy)\mathrm{d}y\right)\\
&= \lim_{x\downarrow 0}\frac{1}{\mathbb{P}[D=0]}\left(1 - \int_0^1 \frac{F_{1-e(X)}(xy)}{F_{1-e(X)}(x)}\mathrm{d}y\right)\\
&= \frac{1}{\mathbb{P}[D=0]}\left(1 - \int_0^1 y^{\gamma_0-1}\mathrm{d}y\right)\\
&= \frac{\gamma_0-1}{\gamma_0}\frac{1}{\mathbb{P}[D=0]}.
\end{aligned}$$

Applying the same argument used to prove Lemma 1, one has

$$\begin{aligned}
&\lim_{x\to\infty}\frac{\mathbb{P}[D=0]\mathbb{P}\left[\frac{e(X)Y(0)}{c_0(1-e(X))} < -x\middle| D=0\right]}{x^{-1}\mathbb{P}[1-e(X)<x^{-1}]}\\
&= \lim_{x\to\infty}\frac{\mathbb{P}[D=0]\mathbb{P}[1-e(X)<x^{-1}|D=0]}{x^{-1}\mathbb{P}[1-e(X)<x^{-1}]}\frac{\mathbb{P}\left[\frac{e(X)Y(0)}{c_0(1-e(X))} < -x\middle| D=0\right]}{\mathbb{P}[1-e(X)<x^{-1}|D=0]}\\
&= \frac{\gamma_0-1}{\gamma_0}c_0^{-\gamma_0}\alpha_{(0),-}(0),
\end{aligned}$$

where

$$\alpha_{(0),-}(x) = \lim_{t\to 1}\mathbb{E}\Big[|Y(0)|^{\gamma_0}\mathbb{1}_{Y(0)<x}\Big|e(X)=t\Big].$$

Therefore,

$$\lim_{x\to\infty}\frac{\mathbb{P}\left[\frac{(D-e(X))Y}{c_0(1-e(X))} > x\right]}{x^{-1}\mathbb{P}[1-e(X)<x^{-1}]} = \frac{\gamma_0-1}{\gamma_0}c_0^{-\gamma_0}\alpha_{(0),-}(0).$$

Similarly, we have

$$\lim_{x\to\infty}\frac{\mathbb{P}\left[\frac{(D-e(X))Y}{c_0(1-e(X))}<-x\right]}{x^{-1}\mathbb{P}[1-e(X)<x^{-1}]}=\frac{\gamma_0-1}{\gamma_0}c_0^{-\gamma_0}\alpha_{(0),+}(0).$$

As a result, $(D-e(X))Y/(c_0(1-e(X))$ has regularly varying tails with index $-\gamma_0$ if $\alpha_{(0),+}(0)+\alpha_{(0),-}(0)>0$. The rest of the proof employs the same argument used to prove Theorem 1. ■

IV.16 Proof of Proposition S.2

We first consider the tail behavior of $(2D-1)Y/(1-D+(2D-1)e(X))$. For this, we note that

$$\mathbb{P}\left[\frac{(2D-1)Y}{1-D+(2D-1)e(X)}>x\right]=\mathbb{P}[D=1]\mathbb{P}\left[\frac{Y(1)}{e(X)}>x\middle|D=1\right]+\mathbb{P}[D=0]\mathbb{P}\left[\frac{Y(0)}{1-e(X)}<-x\middle|D=0\right],$$

where we take $x>0$. Then if $\omega>0$,

$$\begin{aligned}
\lim_{x\downarrow 0}\frac{\mathbb{P}[e(X)\le x|D=1]}{x\mathbb{P}[e(X)\le x]}&=\lim_{x\downarrow 0}\frac{\mathbb{P}[D=1|e(X)\le x]}{x\mathbb{P}[D=1]}\\
&=\lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[e(X)\le x]\mathbb{P}[D=1]}\int_0^x yF_{e(X)}(\mathrm{d}y)\\
&=\lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[e(X)\le x]\mathbb{P}[D=1]}\left(x\mathbb{P}[e(X)\le x]-\int_0^x F_{e(X)}(y)\mathrm{d}y\right)\\
&=\lim_{x\downarrow 0}\frac{1}{x\mathbb{P}[e(X)\le x]\mathbb{P}[D=1]}\left(x\mathbb{P}[e(X)\le x]-\int_0^1 xF_{e(X)}(xy)\mathrm{d}y\right)\\
&=\lim_{x\downarrow 0}\frac{1}{\mathbb{P}[D=1]}\left(1-\int_0^1\frac{F_{e(X)}(xy)}{F_{e(X)}(x)}\mathrm{d}y\right)\\
&=\frac{1}{\mathbb{P}[D=1]}\left(1-\int_0^1 y^{\gamma_0-1}\mathrm{d}y\right)\\
&=\frac{\gamma_0-1}{\gamma_0}\frac{1}{\mathbb{P}[D=1]}.
\end{aligned}$$

Therefore, conditional on $D=1$, the probability weight has regularly varying left tail with index γ_0. Applying the same argument used to prove Lemma 1, one has

$$\begin{aligned}
\lim_{x\to\infty}\frac{\mathbb{P}[D=1]\mathbb{P}\left[\frac{Y(1)}{e(X)}>x\middle|D=1\right]}{x^{-1}\mathbb{P}[e(X)<x^{-1}]}&=\lim_{x\to\infty}\frac{\mathbb{P}[D=1]\mathbb{P}[e(X)<x^{-1}|D=1]}{x^{-1}\mathbb{P}[e(X)<x^{-1}]}\frac{\mathbb{P}\left[\frac{Y(1)}{e(X)}>x\middle|D=1\right]}{\mathbb{P}[e(X)<x^{-1}|D=1]}\\
&=\frac{\gamma_0-1}{\gamma_0}\alpha_{(1),+}(0).
\end{aligned}$$

Similarly, we can show that if $\omega<1$,

$$\lim_{x\downarrow 0}\frac{\mathbb{P}[D=0]\mathbb{P}\left[\frac{Y(0)}{1-e(X)}<-x\middle|D=0\right]}{x^{-1}\mathbb{P}[1-e(X)<x^{-1}]}=\frac{\gamma_0-1}{\gamma_0}\alpha_{(0),-}(0).$$

Together, they imply

$$\lim_{x\to\infty}\frac{x\mathbb{P}\left[\frac{(2D-1)Y}{1-D+(2D-1)e(X)}>x\right]}{\mathbb{P}[e(X)<x^{-1}]+\mathbb{P}[1-e(X)<x^{-1}]}=\frac{\gamma_0-1}{\gamma_0}\Big(\omega\alpha_{(1),+}(0)+(1-\omega)\alpha_{(0),-}(0)\Big).$$

By the same argument,

$$\lim_{x\to\infty}\frac{x\mathbb{P}\left[\frac{(2D-1)Y}{1-D+(2D-1)e(X)}<-x\right]}{\mathbb{P}[e(X)<x^{-1}]+\mathbb{P}[1-e(X)<x^{-1}]}=\frac{\gamma_0-1}{\gamma_0}\Big(\omega\alpha_{(1),-}(0)+(1-\omega)\alpha_{(0),+}(0)\Big).$$

As a result, $(2D-1)Y/(1-D+(2D-1)e(X))$ has regularly varying tail with index $-\gamma_0$ if

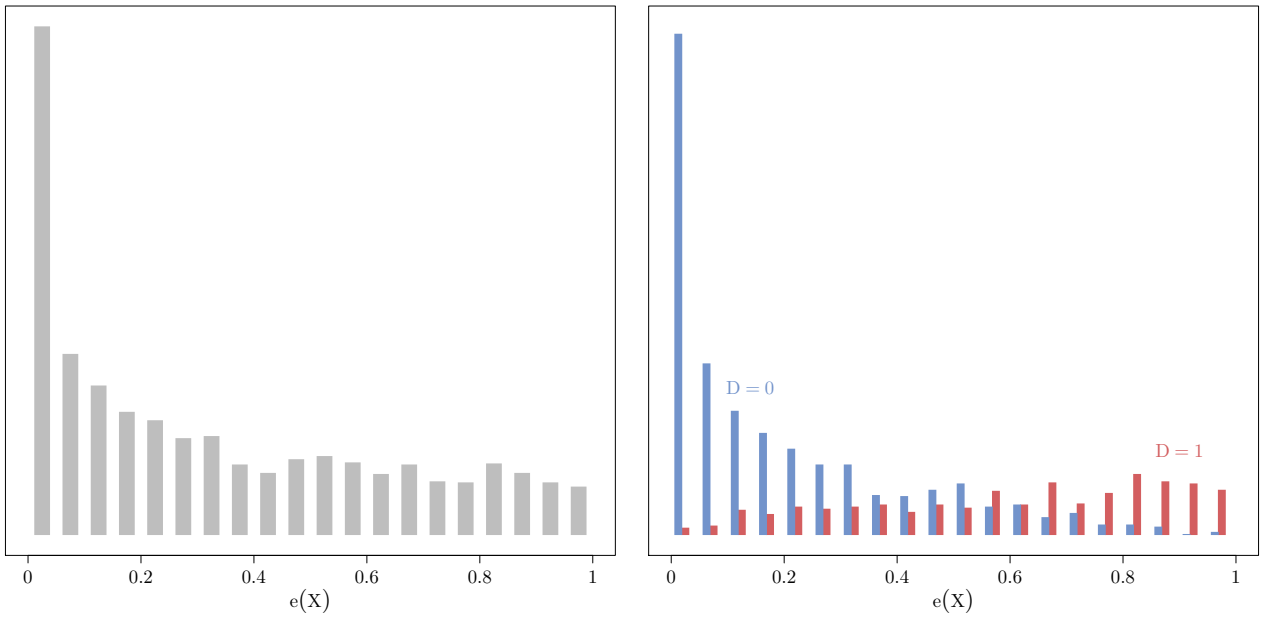
$$\omega\Big(\alpha_{(1),+}(0)+\alpha_{(1),-}(0)\Big)+(1-\omega)\Big(\alpha_{(0),+}(0)+\alpha_{(0),-}(0)\Big)>0.$$

The rest of the proof employs the same argument used to prove Theorem 1. ■

IV.17 Proof of Proposition S.3

This employs the same argument used to prove Theorem 1. ■

V Figures and Tables



(a) (b)

Figure S.1. Illustration of γ_0. Sample size: $n = 2,000$. $\mathbb{P}[e(X) \leq x] = x^{\gamma_0 - 1}$ with $\gamma_0 = 1.5$. (a) Distribution of the Probability Weights. (b) Distribution of the Probability Weights, Separately for Subgroups $D = 1$ (Red) and $D = 0$ (Blue).

Table S.1. Simulation Results. $\gamma_0 = 1.5$, $\mathbb{E}[Y|e(X), D = 1] = \cos(2\pi e(X))$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.377$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.131	3.773	3.776	0.775	7.308	0.131	3.773	3.776	0.844	21.235
0.004	0.170	0.800	1.493	1.694	0.740	5.116	0.238	1.565	1.583	0.924	7.387
0.016	1.338	1.576	0.979	1.855	0.541	3.713	0.465	1.169	1.258	0.926	5.757
0.036	4.606	2.373	0.741	2.486	0.158	2.849	0.628	1.064	1.236	0.913	4.973
0.094	19.225	3.718	0.503	3.752	0.000	1.956	0.711	0.999	1.226	0.906	4.219
		Crump *et al.* (2009)									
0.096	19.996	3.760	0.507	3.794	0.000	1.934	n.a.	n.a.	n.a.	n.a.	n.a.

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.319$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.025	5.681	5.681	0.786	7.948	0.025	5.681	5.681	0.869	37.240
0.002	0.173	0.764	1.546	1.724	0.755	5.336	0.259	1.592	1.613	0.928	7.196
0.010	1.689	1.697	0.966	1.953	0.514	3.717	0.485	1.103	1.205	0.916	5.233
0.025	6.653	2.692	0.714	2.785	0.077	2.805	0.696	0.961	1.187	0.891	4.457
0.072	32.182	4.484	0.478	4.510	0.000	1.885	0.883	0.894	1.257	0.846	3.780
		Crump *et al.* (2009)									
0.096	49.586	5.100	0.429	5.118	0.000	1.666	n.a.	n.a.	n.a.	n.a.	n.a.

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.281$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.053	7.909	7.909	0.787	7.761	0.053	7.909	7.909	0.862	59.629
0.001	0.168	0.781	1.575	1.758	0.757	5.404	0.213	1.609	1.623	0.922	6.944
0.007	1.994	1.812	0.975	2.058	0.477	3.698	0.441	1.086	1.172	0.910	4.870
0.019	8.752	2.971	0.708	3.054	0.037	2.756	0.668	0.916	1.134	0.877	4.097
0.059	47.837	5.175	0.466	5.196	0.000	1.824	0.895	0.831	1.221	0.817	3.490
		Crump *et al.* (2009)									
0.096	99.019	6.442	0.388	6.454	0.000	1.486	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n} = \sum_{i=1}^{n} \mathbb{1}(e(X_i) \leq b_n, D_i = 1)$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95, calculated for the Gaussian-based confidence interval under "Conventional", and calculated for the subsampling-based confidence interval under "Robust"). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. (viii) $\hat{h}_n$: bandwidth for local polynomial bias correction. (ix) Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

Table S.2. Simulation Results. $\gamma_0 = 1.5$, $\mathbb{E}[Y|e(X), D = 1] = 1 - e(X)$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.377$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.132	3.771	3.773	0.774	7.295	0.132	3.771	3.773	0.864	22.017
0.004	0.170	0.800	1.490	1.691	0.742	5.105	0.012	1.569	1.569	0.939	7.755
0.016	1.338	1.569	0.977	1.849	0.543	3.716	0.003	1.172	1.172	0.957	6.029
0.036	4.606	2.357	0.747	2.472	0.165	2.875	0.001	1.063	1.063	0.964	5.228
0.094	19.225	3.730	0.510	3.764	0.000	2.005	0.017	0.984	0.984	0.967	4.530
		Crump *et al.* (2009)									
0.096	19.996	3.775	0.511	3.810	0.000	1.983	n.a.	n.a.	n.a.	n.a.	n.a.

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.319$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.025	5.678	5.678	0.784	7.933	0.025	5.678	5.678	0.873	37.233
0.002	0.173	0.763	1.549	1.726	0.754	5.323	0.031	1.600	1.601	0.935	7.334
0.010	1.689	1.692	0.967	1.949	0.514	3.712	0.015	1.112	1.112	0.956	5.346
0.025	6.653	2.676	0.719	2.771	0.081	2.817	0.015	0.967	0.967	0.963	4.559
0.072	32.182	4.467	0.491	4.494	0.000	1.927	0.019	0.890	0.890	0.964	3.958
		Crump *et al.* (2009)									
0.096	49.586	5.120	0.440	5.139	0.000	1.710	n.a.	n.a.	n.a.	n.a.	n.a.

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.281$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.045	7.909	7.909	0.790	7.747	0.045	7.909	7.909	0.863	59.692
0.001	0.168	0.773	1.571	1.751	0.760	5.391	0.019	1.609	1.609	0.928	7.017
0.007	1.994	1.801	0.973	2.047	0.477	3.689	0.005	1.092	1.092	0.952	4.943
0.019	8.752	2.949	0.710	3.033	0.040	2.760	0.003	0.923	0.923	0.958	4.152
0.059	47.837	5.136	0.474	5.158	0.000	1.856	0.006	0.829	0.829	0.964	3.588
		Crump *et al.* (2009)									
0.096	99.019	6.460	0.395	6.472	0.000	1.524	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n} = \sum_{i=1}^{n} \mathbb{1}(e(X_i) \leq b_n, D_i = 1)$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95, calculated for the Gaussian-based confidence interval under "Conventional", and calculated for the subsampling-based confidence interval under "Robust"). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. (viii) $\hat{h}_n$: bandwidth for local polynomial bias correction. (ix) Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

Table S.3. Simulation Results. $\gamma_0 = 1.3$, $\mathbb{E}[Y|e(X), D = 1] = \cos(2\pi e(X))$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.358$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.299	5.004	5.013	0.611	5.413	0.299	5.004	5.013	0.806	63.074
0.002	0.121	0.874	1.054	1.369	0.567	3.373	0.218	1.123	1.144	0.937	6.832
0.010	1.151	1.452	0.561	1.556	0.284	2.090	0.393	0.750	0.847	0.927	4.661
0.027	4.247	1.955	0.376	1.991	0.016	1.431	0.497	0.674	0.838	0.913	3.715
0.081	17.630	2.673	0.222	2.682	0.000	0.861	0.555	0.646	0.852	0.882	2.886
		Crump *et al.* (2009)									
0.092	20.735	2.761	0.214	2.769	0.000	0.808	n.a.	n.a.	n.a.	n.a.	n.a.

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.301$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.226	5.860	5.864	0.619	5.665	0.226	5.860	5.864	0.829	90.294
0.001	0.105	0.832	1.090	1.372	0.584	3.494	0.230	1.133	1.156	0.933	6.215
0.006	1.456	1.523	0.540	1.616	0.238	2.027	0.388	0.676	0.780	0.912	3.818
0.018	6.323	2.141	0.348	2.169	0.002	1.338	0.498	0.571	0.758	0.871	3.001
0.061	30.642	3.058	0.199	3.064	0.000	0.780	0.591	0.536	0.797	0.824	2.379
		Crump *et al.* (2009)									
0.092	51.712	3.407	0.163	3.411	0.000	0.637	n.a.	n.a.	n.a.	n.a.	n.a.

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.264$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.028	13.288	13.288	0.612	6.605	0.028	13.288	13.288	0.804	332.821
0.000	0.118	0.861	1.126	1.417	0.573	3.498	0.172	1.154	1.167	0.926	5.737
0.004	1.833	1.621	0.514	1.700	0.178	1.931	0.319	0.620	0.697	0.913	3.249
0.013	8.517	2.301	0.323	2.324	0.000	1.261	0.453	0.507	0.679	0.858	2.555
0.050	46.665	3.388	0.182	3.393	0.000	0.718	0.556	0.470	0.728	0.792	2.062
		Crump *et al.* (2009)									
0.092	103.456	4.007	0.138	4.010	0.000	0.528	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n} = \sum_{i=1}^{n} \mathbb{1}(e(X_i) \leq b_n, D_i = 1)$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95, calculated for the Gaussian-based confidence interval under "Conventional", and calculated for the subsampling-based confidence interval under "Robust"). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. (viii) $\hat{h}_n$: bandwidth for local polynomial bias correction. (ix) Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

Table S.4. Simulation Results. $\gamma_0 = 1.3$, $\mathbb{E}[Y|e(X), D = 1] = 1 - e(X)$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.358$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.300	4.999	5.008	0.610	5.407	0.300	4.999	5.008	0.825	64.859
0.002	0.121	0.875	1.048	1.365	0.565	3.368	0.027	1.126	1.126	0.951	7.155
0.010	1.151	1.449	0.559	1.553	0.282	2.092	0.009	0.757	0.757	0.971	4.882
0.027	4.247	1.947	0.375	1.983	0.016	1.446	0.008	0.674	0.674	0.979	3.935
0.081	17.630	2.668	0.229	2.678	0.000	0.895	0.010	0.639	0.639	0.975	3.182
		Crump *et al.* (2009)									
0.092	20.735	2.763	0.219	2.771	0.000	0.842	n.a.	n.a.	n.a.	n.a.	n.a.

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.301$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.226	5.858	5.862	0.617	5.659	0.226	5.858	5.862	0.826	87.993
0.001	0.105	0.832	1.090	1.371	0.582	3.488	0.019	1.139	1.139	0.946	6.347
0.006	1.456	1.521	0.537	1.613	0.234	2.024	0.016	0.682	0.682	0.967	3.911
0.018	6.323	2.134	0.347	2.162	0.002	1.344	0.007	0.574	0.574	0.973	3.075
0.061	30.642	3.043	0.206	3.050	0.000	0.803	0.006	0.533	0.534	0.971	2.524
		Crump *et al.* (2009)									
0.092	51.712	3.408	0.169	3.412	0.000	0.664	n.a.	n.a.	n.a.	n.a.	n.a.

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.264$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.027	13.287	13.287	0.610	6.600	0.027	13.287	13.287	0.804	325.798
0.000	0.118	0.860	1.124	1.415	0.570	3.493	0.008	1.155	1.155	0.936	5.821
0.004	1.833	1.619	0.513	1.698	0.178	1.927	0.014	0.624	0.625	0.967	3.312
0.013	8.517	2.295	0.323	2.317	0.000	1.263	0.003	0.509	0.509	0.973	2.591
0.050	46.665	3.370	0.186	3.375	0.000	0.735	0.008	0.466	0.467	0.971	2.133
		Crump *et al.* (2009)									
0.092	103.456	4.007	0.141	4.010	0.000	0.550	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n} = \sum_{i=1}^{n} \mathbb{1}(e(X_i) \leq b_n, D_i = 1)$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95, calculated for the Gaussian-based confidence interval under "Conventional", and calculated for the subsampling-based confidence interval under "Robust"). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. (viii) $\hat{h}_n$: bandwidth for local polynomial bias correction. (ix) Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

Table S.5. Simulation Results. $\gamma_0 = 1.9$, $\mathbb{E}[Y|e(X), D = 1] = \cos(2\pi e(X))$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.414$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.006	3.612	3.612	0.902	10.703	0.006	3.612	3.612	0.876	13.413
0.013	0.241	0.711	2.325	2.431	0.885	8.840	0.238	2.401	2.413	0.930	9.776
0.032	1.340	1.594	1.966	2.531	0.816	7.654	0.477	2.155	2.207	0.936	8.993
0.057	4.164	2.715	1.702	3.204	0.613	6.691	0.638	2.026	2.124	0.944	8.497
0.119	16.767	4.926	1.386	5.117	0.073	5.484	0.702	1.940	2.063	0.947	8.012
		Crump *et al.* (2009)									
0.101	12.230	4.347	1.457	4.584	0.176	5.756	n.a.	n.a.	n.a.	n.a.	n.a.

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.354$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.030	5.205	5.205	0.918	11.313	0.030	5.205	5.205	0.892	17.278
0.008	0.250	0.732	2.411	2.520	0.903	9.296	0.234	2.464	2.475	0.928	10.000
0.022	1.599	1.742	2.066	2.702	0.821	8.072	0.592	2.213	2.291	0.929	9.076
0.042	5.596	3.141	1.802	3.621	0.566	7.087	0.920	2.080	2.275	0.919	8.548
0.094	26.638	6.330	1.454	6.495	0.015	5.745	1.200	1.965	2.303	0.915	7.933
		Crump *et al.* (2009)									
0.101	30.355	6.680	1.440	6.833	0.005	5.634	n.a.	n.a.	n.a.	n.a.	n.a.

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.315$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.106	3.456	3.458	0.913	11.353	0.106	3.456	3.458	0.890	13.015
0.005	0.229	0.767	2.547	2.660	0.899	9.728	0.186	2.589	2.596	0.934	10.285
0.016	1.836	1.932	2.173	2.908	0.804	8.395	0.548	2.292	2.357	0.926	9.172
0.033	7.162	3.620	1.910	4.093	0.507	7.347	0.934	2.145	2.340	0.913	8.572
0.079	38.011	7.701	1.543	7.854	0.002	5.935	1.383	2.024	2.451	0.888	7.943
		Crump *et al.* (2009)									
0.101	60.651	9.369	1.446	9.480	0.000	5.532	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n} = \sum_{i=1}^{n} \mathbb{1}(e(X_i) \leq b_n, D_i = 1)$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95, calculated for the Gaussian-based confidence interval under "Conventional", and calculated for the subsampling-based confidence interval under "Robust"). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. (viii) $\hat{h}_n$: bandwidth for local polynomial bias correction. (ix) Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

Table S.6. Simulation Results. $\gamma_0 = 1.9$, $\mathbb{E}[Y|e(X), D = 1] = 1 - e(X)$.

(a) $n = 2,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.414$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.021	3.573	3.573	0.910	10.626	0.021	3.573	3.573	0.899	14.614
0.013	0.241	0.693	2.278	2.381	0.893	8.770	0.022	2.359	2.359	0.939	10.343
0.032	1.340	1.565	1.928	2.483	0.822	7.603	0.039	2.119	2.119	0.946	9.383
0.057	4.164	2.680	1.676	3.161	0.626	6.660	0.027	1.991	1.992	0.951	8.798
0.119	16.767	5.065	1.359	5.244	0.062	5.425	0.018	1.892	1.892	0.959	8.198
		Crump *et al.* (2009)									
0.101	12.230	4.395	1.431	4.622	0.171	5.718	n.a.	n.a.	n.a.	n.a.	n.a.

(b) $n = 5,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.354$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.036	5.190	5.191	0.913	11.239	0.036	5.190	5.191	0.901	17.806
0.008	0.250	0.723	2.425	2.531	0.898	9.224	0.005	2.481	2.481	0.936	10.269
0.022	1.599	1.722	2.073	2.695	0.815	8.013	0.048	2.224	2.224	0.943	9.265
0.042	5.596	3.105	1.810	3.594	0.567	7.049	0.075	2.084	2.085	0.950	8.693
0.094	26.638	6.384	1.463	6.549	0.015	5.717	0.066	1.952	1.953	0.955	8.078
		Crump *et al.* (2009)									
0.101	30.355	6.771	1.442	6.923	0.007	5.601	n.a.	n.a.	n.a.	n.a.	n.a.

(c) $n = 10,000$

Trimming		**Conventional**					**Robust** ($\hat{h}_n = 0.315$)				
$\hat{b}_n$	$n_{\leq \hat{b}_n}$	bias	sd	rmse	cov	|ci|	bias	sd	rmse	cov	|ci|
–	–	0.090	3.458	3.460	0.911	11.277	0.090	3.458	3.460	0.900	13.243
0.005	0.229	0.749	2.539	2.647	0.894	9.654	0.030	2.583	2.583	0.936	10.438
0.016	1.836	1.905	2.151	2.873	0.810	8.329	0.004	2.277	2.277	0.943	9.281
0.033	7.162	3.571	1.881	4.036	0.506	7.299	0.008	2.121	2.121	0.949	8.652
0.079	38.011	7.687	1.509	7.834	0.002	5.910	0.010	1.981	1.981	0.955	8.041
		Crump *et al.* (2009)									
0.101	60.651	9.486	1.399	9.589	0.000	5.498	n.a.	n.a.	n.a.	n.a.	n.a.

Notes: (i) $\hat{b}_n$: trimming threshold. (ii) $n_{\leq \hat{b}_n} = \sum_{i=1}^{n} \mathbb{1}(e(X_i) \leq b_n, D_i = 1)$: effective number of trimmed observations. (iii) bias: empirical bias, scaled by n^{1-1/γ_0}. (iv) sd: empirical standard deviation, scaled by n^{1-1/γ_0}. (v) rmse: empirical root mean squared error, scaled by n^{1-1/γ_0}. (vi) cov: coverage probability (nominal level 0.95, calculated for the Gaussian-based confidence interval under "Conventional", and calculated for the subsampling-based confidence interval under "Robust"). (vii) |ci|: average confidence interval length, scaled by n^{1-1/γ_0}. (viii) $\hat{h}_n$: bandwidth for local polynomial bias correction. (ix) Number of Monte Carlo repetitions: 5000. Number of subsampling iterations: 1000. Subsample size: $\lfloor n/\log(n) \rfloor$.

References

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, *96*(1), 187–199.

Feller, W. (1991). *An Introduction to Probability Theory and Its Applications, Volume II*, New York: John Wiley, 2nd edition.

Logan, B. F., Mallows, C. L., Rice, S. O., and Shepp, L. A. (1973). "Limit Distributions of Self-normalized Sums," *Annals of Probability*, *1*(5), 788–809.

Newey, W. K. and McFadden, D. L. (1994). "Large Sample Estimation and Hypothesis Testing," In R. F. Engle and D. L. McFadden (eds.) *Handbook of Econometrics, Volume IV*: Elsevier Science B.V. 2111–2245.

Politis, D. N. and Romano, J. P. (1994). "Large Sample Confidence Regions Based on Subsamples Under Minimal Assumptions," *Annals of Statistics*, *22*(4), 2031–2050.

Romano, J. P. and Wolf, M. (1999). "Subsampling Inference for the Mean in the Heavy-tailed Case," *Metrika*, *50*(1), 55–69.

Vershynin, R. (2018). *High-Dimensional Probability*, Cambridge, U.K.: Cambridge University Press.