Abstract

A popular method for variance reduction in observational causal inference is propensity-based trimming, the practice of removing units with extreme propensities from the sample. This practice has theoretical grounding when the data are homoscedastic and the propensity model is parametric (Yang and Ding, 2018; Crump et al., 2009), but in modern settings where heteroscedastic data are analyzed with non-parametric models, existing theory fails to support current practice. In this work, we address this challenge by developing new methods and theory for sample trimming. Our contributions are three-fold: first, we describe novel procedures for selecting which units to trim. Our procedures differ from previous work in that we trim not only units with small propensities, but also units with extreme conditional variances. Second, we give new theoretical guarantees for inference after trimming. In particular, we show how to perform inference on the trimmed subpopulation without requiring that our regressions converge at parametric rates. Instead, we make only fourth-root rate assumptions like those in the double machine learning literature. This result applies to conventional propensity-based trimming as well and thus may be of independent interest. Finally, we propose a bootstrap-based method for constructing simultaneously valid confidence intervals for multiple trimmed sub-populations, which are valuable for navigating the trade-off between sample size and variance reduction inherent in trimming. We validate our methods in simulation, on the 2007-2008 National Health and Nutrition Examination Survey, and on a semi-synthetic Medicare dataset and find promising results in all settings.

1 Introduction

Traditional methods for estimating causal effects from observational data typically rely on two standard assumptions: unconfoundedness and overlap (Rosenbaum and Rubin, 1984). In practice, observational data often have limited overlap, leading to extreme propensity scores and thus high-variance estimates of the treatment effect. A large body of literature addresses this challenge by modifying the estimand to either exclude or down-weight units with extreme propensity scores, and these methods have been widely adopted in practice (Yang and Ding, 2018; Li et al., 2018; Crump et al., 2009). However, modern data can pose an additional challenge in the form of heavy-tails and heteroscedasticity (Burke et al., 2019; Tripuraneni et al., 2021). In this paper, we address this challenge by exploring sample trimming methods that reduce variance by trimming not only units with extreme propensities, but also units with extreme conditional variances.

Motivation and interpretation To motivate this approach, consider a single unit \((X,Y,Z)\) drawn from a super-population distribution, where \(X\) is a covariate vector, \(Y\) is a response, \(Z\) is
a treatment indicator, and \( e(X) \) is the probability of treatment. An inverse-propensity weighted estimator for \( E[Y] \) is \( YZ/e(X) \), which is unbiased, but is well-known to suffer from extremely high-variance when \( e(X) \) takes small values \(^\text{(Khan and Ugander 2021)}\). As such, the goal of existing sample trimming methods is to preclude this possibility by removing units for which \( e(X) \) takes small values. On the other hand, the inverse-propensity weighted estimate \( YZ/e(X) \) will also have high-variance if \( \text{var}(Y | X) \) is large, an issue which is not addressed by existing methods, but may be a major obstacle when \( \text{var}(Y | X) \) is extremely large for some values of \( X \), as may be the case for, e.g., social network data. Put simply: if we do not believe that we can accurately estimate treatment effects on units with propensities of, say, 0.01, then we must also acknowledge that we cannot accurately estimate treatment effects on units with conditional variances of say, 100, and so we propose to trim these latter units as well.

An important difference between this proposal and existing propensity-based trimming methods is that the sub-population found by propensity-based methods can be interpreted as a population that is likely to receive either treatment or control, and thus may be a natural population of interest. This interpretation does not extend to variance-based trimming methods—instead, variance-based methods can be interpreted as identifying a small population of outliers in the data, whose behavior and response to treatment is very different from that of other units, and trimming these units to focus on an “inlier” population on which treatment effects can be estimated more accurately. In many cases, this inlier population is also of natural interest, since treatment effects on the full population may be dominated by the outliers, and the treatment effect on the inlier population may be more representative of how treatment will affect the majority of units. We demonstrate this phenomenon, along with further interpretive issues, as part of a data example in Section 6.3.

**Inference after trimming** After applying any sample trimming procedure, another challenge immediately arises: how to perform valid inference on the trimmed sub-population. Thus our second contribution in the present work is to provide new theoretical results on inference after sample trimming. Unlike existing work that makes strong rate assumptions or parametric assumptions on the estimation of nuisance components \(^\text{(Crump et al. 2009; Yang and Ding 2018)}\), we use doubly-robust estimators and a carefully chosen estimand to show how valid inference can be performed under weaker conditions on the estimation of nuisance components. These results apply both to our variance-based trimming and to classical propensity-based trimming methods, thus connecting the recent literature on double machine learning with the long-standing practice of sample trimming.

Our third contribution addresses a more subtle, previously unconsidered, aspect of inference after trimming. Roughly speaking, there are several features of a sample trimming method we may be interested in: the amount of variance reduction offered by the trimming, the size of the resulting sub-population, the point estimate on that sub-population, and perhaps even covariate distributions with the sub-population. However, there is no way to smoothly navigate the trade-offs between these considerations. If we trim the sample one way, perform inference, and find the results unfavorable for some reason, we cannot then trim the sample another way and perform valid inference without conditioning on the results of the first sample trimming; this is the problem of selective inference \(^\text{(Taylor and Tibshirani 2015)}\). As a remedy, we introduce a bootstrap-based method that allows an analyst to pre-commit to a small number of trimming methods, and then constructs simultaneously-valid confidence intervals for the sub-populations found by each trimming method. An analyst can then choose freely between the different sub-populations based on problem specific considerations while retaining statistical validity.

To summarize, our work both proposes a new criteria for sample trimming based on conditional...
We adopt a potential outcomes framework with \( n \) units where \( \tau \) with extreme response values. However, Chaudhuri and Hill (2014) are considering a largely different whose contribution to the IPW estimator is extremely large, which also amounts to removing units works in Section 4, but at a high-level, we differ from these previous works in directly modeling variances and propensity scores rather than on propensity scores alone, and develops new theoretical of the full sample by removing possible outliers, and lead to statistically significant conclusions on some of these sub-populations even when no such conclusion was possible on the full population.

1.1 Related work

Our work directly builds on the extensive sample trimming literature, and is especially connected to Crump et al. (2009) and Yang and Ding (2018). We offer a more detailed comparison with these works in Section 4, but at a high-level, we differ from these previous works in directly modeling heteroscedasticity and in assuming weaker conditions on the modeling of nuisance components. One prior work with a similar idea to ours is Chaudhuri and Hill (2014), which proposes to remove units whose contribution to the IPW estimator is extremely large, which also amounts to removing units with extreme response values. However, Chaudhuri and Hill (2014) are considering a largely different problem than us: they are not concerned with variance minimization, consider only classical IPW estimators, and do not modify the estimand as is done in the sample trimming literature.

Our current proposal is also conceptually related to methods in robust statistics and outlier removal. For example, removing units with large residuals from an ordinary least-squares analysis is similar in spirit to the methods we propose here, as are other methods that identify and remove extreme units from the data. We differ from these methods in that our motivation for dropping units is based on variance reduction, not on a contamination model for the data, and in that we emphasize the problem of inference after dropping these units.

2 Model and notation

We adopt a potential outcomes framework with \( n \) units where the tuples \((Y_i(1), Y_i(0), X_i, Z_i)\) are i.i.d. from a super-population distribution \( \mathbb{P} \) over \( \mathbb{R}^2 \times \mathcal{X} \times \{0,1\} \). We assume that \( Y_i(1) \) and \( Y_i(0) \) both have finite variance and that we observe \( Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0) \). We write \( e(x) = \mathbb{P}(Z_i = 1 \mid X_i = x) \) for the propensity score, \( \mu_w(x) = \mathbb{E}[Y_i(w) \mid X_i = x] \), for \( w \in \{0,1\} \), for the conditional means, and \( \sigma_w^2(x) = \text{var}(Y_i(w) \mid X_i = x) \) for the conditional variances. We make the standard unconfoundedness and overlap assumptions that \( Z_i \perp (Y_i(1), Y_i(0)) \mid X_i \) and \( \eta \leq e(x) \leq 1 - \eta \) (Rosenbaum and Rubin [1984]).

Our target of inference is the sample average treatment effect (SATE) and its trimmed analogs,

\[
\tau = \frac{1}{n} \sum_{i=1}^{n} \tau(X_i), \quad \tau_A = \frac{1}{n_A} \sum_{i=1}^{n} \tau(X_i) \mathbb{1}\{X_i \in A\},
\]

where \( \tau(x) = \mu_1(x) - \mu_0(x) \) is a conditional average treatment effect (CATE), \( A \subseteq \mathcal{X} \) is the subset of covariate space we are restricting the covariates to, and \( n_A = \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\} \) is the number of sample units whose covariates lie in \( A \).

In subsequent sections, we employ empirical process notation. We let \( W_i = (X_i, Y_i, Z_i) \) be the entire triplet we observe for unit \( i \), and we write \( \mathbb{P}_n f = \frac{1}{n} \sum_i f(W_i) \) and \( \mathbb{P} f = \int f(w) d\mathbb{P}(w) \). Note that for a random function \( \hat{f} \), \( \mathbb{P} \hat{f} \) is a random variable, since we do not integrate over the randomness in \( \hat{f} \). In contrast, \( \mathbb{E}[\hat{f}] \) is a deterministic quantity that integrates out the randomness in a new sample and in \( \hat{f} \). We also define the norm \( \|f\|_{L^q(\mathbb{P})} = (\mathbb{E}[|f|^q])^{1/q} \).
3 Trimming methods

In this section, we present a framework for sample trimming methods and use this framework to propose a trimming method that accounts for conditional variances. As a starting point, recall the result of Hirano et al. (2003) that the variance of an efficient estimator of $\tau_A$ is given by

$$V_{A}^{\text{eff}} = \frac{1}{P(X \in A)^2} \mathbb{E} \left[ 1\{X \in A\} \left( \frac{\sigma^2(X)}{e(X)} + \frac{\sigma_0^2(X)}{1-e(X)} \right) \right].$$  

(2)

Based on (2), we can see that the key quantity that determines a unit’s contribution to the asymptotic variance is

$$k(x) = \frac{\sigma^2_1(x)}{e(x)} + \frac{\sigma^2_0(x)}{1-e(x)}.$$  

(3)

That is, if many units have large values of $k(X_i)$, then the variance of our estimate of $\tau_A$ will be large, and vice-versa. This idea was made precise by Crump et al. (2009), who showed that (2) is minimized for the set $A$ that thresholds $k(x)$ at a cut-off $\gamma$, that is,

$$\arg\min_A V_{A}^{\text{eff}} = \{ x : k(x) \leq \gamma \},$$  

(4)

for some cut-off $\gamma \in \mathbb{R}$. This result motivates us to consider trimming sets $A$ that have this form i.e. that threshold the function $k(x)$.

Of course, in practice, we do not have direct access to the function $k$ or the choice of $\gamma$ for which the minimum in (1) is attained. Instead, both must be learned from the data, giving us an estimated function $\hat{k}(x)$, an estimated cut-off $\hat{\gamma}$, and a corresponding trimming set $\hat{A} = \{ x : k(x) \leq \hat{\gamma} \}$. The difference between $\hat{A}$ and $A$ is subtle, but will play a crucial role in what follows, particularly in our discussion of inferential issues in Section 4. We now discuss several choices for $\hat{k}$ and $\hat{\gamma}$.

3.1 Choices of $\hat{k}$

How we estimate $k(x)$ depends on what assumptions we are willing to make on $\sigma^2_1(x)$ and $\sigma^2_0(x)$. In particular, we distinguish between two possibilities:

**Homoscedasticity assumed:** if we assume that $\sigma^2_1(x), \sigma^2_0(x)$ are constant in $x$ and equal to each other, then we have that $k(x) \propto 1/(e(x)(1-e(x)))$, and so we can estimate $k$ by first estimating the propensity score by $\hat{e}(x)$, and then setting $\hat{k}(x) = 1/(\hat{e}(x)(1-\hat{e}(x)))$, Note that thresholding this choice of $\hat{k}$ is equivalent to thresholding on $\hat{e}(x)$ itself, and so recovers standard propensity trimming (Crump et al. 2009).

**Heteroscedasticity allowed:** if we are not willing to make the homoscedasticity assumption, then we must also estimate the conditional variances by $\hat{\sigma}^2_1(x), \hat{\sigma}^2_0(x)$, and then use the estimate

$$\hat{k}(x) = \frac{\hat{\sigma}^2_1(x)/\hat{e}(x) + \hat{\sigma}^2_0(x)/(1-\hat{e}(x))}. $$  

(5)

Thus, the usual propensity-based trimming corresponds to choosing $\hat{k}$ based on a homoscedasticity assumption that may or may not be satisfied. In some cases, such as when $Y_i$ is binary and so $\sigma^2_w(x)$ is bounded by 1/4, deviations from this assumption may be negligible. However, in other cases, such as when $Y_i$ is real-valued and has potentially unbounded variance, deviations from this assumption may be significant and worth capturing. In such situations, we propose instead trimming based on the “heteroscedasticity-aware” $\hat{k}$ defined in (5). This is in contrast to propensity-based trimming, which we refer to as “homoscedastic trimming” in light of the underlying homoscedasticity assumption. Going forward, we state all of our results for general $\hat{k}$, making them relevant to both existing (homoscedastic, propensity-based) practice and our new procedures.
3.2 Choices of $\hat{\gamma}$

Next we consider the estimation of $\gamma$. There is a unique choice of $\gamma$ that achieves the minimum in (4), but we may or may not want to target this choice of $\gamma$ depending on how we prioritize other considerations such as simplicity and sub-population size (which can aid in interpretability). Thus we consider three possible choices of $\hat{\gamma}$:

**Constant.** The simplest choice is to pre-commit to a specific value of $\hat{\gamma}$; for example, [Crump et al. (2009)] suggest using $\hat{\gamma} = 1 / 0.1 + 1 / 0.9 \approx 11.1$ for propensity-based trimming.

**Variance minimizing.** If our goal is to actually attain the minimum in (4), then we should choose $\gamma$ to minimize a sample estimate of $V^A_{\text{eff}}$ for $A = \{ x : \hat{k}(x) \leq \gamma \}$,

$$\hat{\gamma} = \arg\min_{\min \hat{k}(X_i) \leq \gamma \leq \max \hat{k}(X_i)} \frac{\frac{1}{n} \sum_i \hat{k}(X_i) 1\{\hat{k}(X_i) \leq \gamma\}}{\left(\frac{1}{n} \sum_i 1\{\hat{k}(X_i) \leq \gamma\}\right)^2}. \quad (6)$$

**Fixed-fraction.** Finally, if our goal is to trim as little of the sample as possible, so that the remaining study population is as close/relevant as possible to the full population, we may take

$$\hat{\gamma} = (1 - \delta) \text{quantile of } \hat{k}(X_1), \cdots, \hat{k}(X_n), \quad (7)$$

so that we only trim a $\delta$-fraction of the data, for some (presumably modest) constant $\delta$.

Each proposed choice of $\hat{k}$ and $\hat{\gamma}$ represents a particular sample trimming method with different properties. For example, homoscedastic trimming with a constant threshold is very common in practice, but may trim a large fraction of the sample. On the other hand, heteroscedastically trimming a fixed-fraction of the sample lends itself to the outcome-outlier removal interpretation discussed in Section [1] and may substantially reduce variance while trimming only a small number of units. Particular choices will depend on problem-specific considerations, but in this work, we especially highlight the value of heteroscedastic trimming a fixed-fraction of units, and focus mainly on this choice in our experiments.

4 Inference after trimming

After we have chosen a trimming method (i.e., after we have estimated a function $\hat{k}$ and a threshold $\hat{\gamma}$ as in the previous section), we obtain a subset $\hat{A}$ of $\mathcal{X}$ and would like to perform inference around $\tau_{\hat{A}}$. Actually performing such inference turns out to be a fairly subtle task. Although it is not uncommon to simply ignore the trimming step and proceed with standard inference, the resulting confidence intervals are not guaranteed to achieve the desired coverage. On the other hand, existing methods that do account for trimming typically require strong assumptions on the modeling process: for example, [Crump et al. (2009)] require that the outcome regression and propensity both be estimated at least an $o(n^{-1/3})$ rate, while [Yang and Ding (2018)] require that the propensity model be parametric and the trimming cut-off $\hat{\gamma}$ be constant. In this paper, we go beyond this prior work by developing a method for valid inference after trimming even when $e(x), \mu_w(x),$ and $\sigma^2_w(x)$ are estimated at the slower $o(n^{-1/4})$ rate. By doing so, we connect the literature on causal inference with machine learning methods to the literature on trimming, allowing practitioners to use doubly-robust estimators on a trimmed sample, and still produce asymptotically valid confidence intervals.
The key to our approach is a careful choice of estimand. In fact, there are two estimands around which we might want to perform inference—the distinction between these has been briefly noted in the literature (Crump et al., 2009; Yang and Ding, 2018), but we hope to provide a more detailed discussion. The first possible estimand is the treatment effect on the sample trimming set $\hat{A}$, that is,

$$
\tau_{\hat{A}} = \frac{1}{n_{\hat{A}}} \sum_{i=1}^{n} \tau(X_i)1\{\hat{k}(X_i) \leq \hat{\gamma}\}, \quad n_{\hat{A}} = \sum_{i} 1\{\hat{k}(X_i) \leq \hat{\gamma}\}, \quad (8)
$$

where $n_{\hat{A}}$ is the trimmed sample size. However, with this estimand, the specific sub-population on which we are performing inference is dependent on the sample, because $\hat{A}$ is a function of the sample. To obtain a sub-population that is meaningful independent of the realized sample, we suppose that $\hat{k}$ and $\hat{\gamma}$ converge to $\hat{k}$ and $\hat{\gamma}$ respectively (see Assumption 2 for a precise statement), and define the limiting sub-population $\bar{A} = \{x : \hat{k}(x) \leq \hat{\gamma}\}$. The estimand corresponding to $\bar{A}$ is

$$
\tau_{\bar{A}} = \frac{1}{\bar{n}_{\bar{A}}} \sum_{i=1}^{n} \tau(X_i)1\{\hat{k}(X_i) \leq \hat{\gamma}\}, \quad \bar{n}_{\bar{A}} = \sum_{i=1}^{n} 1\{\hat{k}(X_i) \leq \hat{\gamma}\}. \quad (9)
$$

To better understand the difference between (8) and (9), it is helpful to consider the special case of homoscedastic trimming with a fixed value of $\hat{\gamma}$. Then, $\tau_{\hat{A}}$ corresponds to the sub-population trimmed by the estimated propensity $\hat{e}(x)$, while, assuming that $\hat{e}(x)$ is consistent for the true propensity, $\tau_{\hat{A}}$ corresponds to the sub-population trimmed by the true propensity $e(x)$. This special case also highlights the pros and cons of each estimand: the advantage of $\tau_{\bar{A}}$ is that we know exactly which units in our sample are part of the target trimming set $\hat{A}$, while the advantage of $\tau_{\bar{A}}$ is that we can interpret it in terms of the true propensity.

Another important difference between these two estimands is that, when using an AIPW-style estimator (Kang and Schafer, 2007; Robins et al., 1994), performing inference around $\tau_{\hat{A}}$ requires much weaker conditions than performing inference around $\tau_{\bar{A}}$: inference at the usual $\sqrt{n}$-rate around $\tau_{\hat{A}}$ requires only a consistency assumption on $\hat{k}$, while inference around $\tau_{\bar{A}}$ requires stronger assumptions on the asymptotics of $\hat{k}$ as well as potentially replacing indicator functions with smooth approximations as in (Yang and Ding, 2018). (We discuss the need for this smoothing in more detail following Theorem 2.) Intuitively, when estimating $\tau_{\bar{A}}$, we do not need to control error terms of the form $\hat{k}(x) - k(x)$ since both the estimator and estimand depend on $\hat{k}$, while we do need to control such terms when estimating $\tau_{\bar{A}}$. This subtle distinction is easy to overlook, but is of central importance to correctly using and interpreting sample trimming methods.

We now present our main result: that an AIPW-style estimator of $\tau_{\bar{A}}$ is first-order equivalent to an i.i.d. sum for any consistent $\hat{k}$, allowing for inference based on normal theory or the bootstrap.

Formally, suppose that we estimate the nuisance parameters $\mu(w)$ and $e$ by $\hat{\mu}(w)$ and $\hat{e}$ so that the following assumption is satisfied:

**Assumption 1.** We assume that $\hat{\mu}(w)$ and $\hat{e}$ are consistent for $\mu(w)$ and $e$ in the $L^2(\mathbb{P})$ norm and satisfy the rate condition

$$
\| (\hat{\mu}(w) - \mu(w))1\{\hat{k}(X_i) \leq \hat{\gamma}\} \|_{L^2(\mathbb{P})} \cdot \| (\hat{e} - e)1\{\hat{k}(X_i) \leq \hat{\gamma}\} \|_{L^2(\mathbb{P})} = o_P(n^{-1/2}). \quad (10)
$$

In Assumption 1, we only require convergence rates on the subset of non-trimmed units. Thus, even if it is quite difficult to estimate the mean and propensity in some extreme parts of the covariate space, this will not be a problem as long as we trim those parts of the covariate space off.

Next, we make the following assumption on $\hat{k}$ and $\hat{\gamma}$.
Assumption 2. We assume \( \hat{k} \) and \( \hat{\gamma} \) are convergent, so that \( \| \hat{k} - \bar{k} \|_{L^2(\mathcal{P})} = o_P(1) \) and \( \hat{\gamma} \stackrel{p}{\to} \gamma \) for some function \( \bar{k} : \mathcal{X} \to \mathbb{R} \) and \( \gamma \in \mathbb{R} \).

Note that we do not assume well-specification of \( \hat{k} \), i.e., that \( \hat{k} \) as in (9) actually converges to \( \sigma_1^2(x)/\epsilon(x) + \sigma_0^2(x)/(1 - \epsilon(x)) \). If this is not true, then we are no longer estimating the optimal trimming, but we are still estimating a well-defined estimand.

Finally, we make the following assumption on the estimation of \( \hat{\mu}(w) \), \( \hat{e} \), and \( \hat{k} \).

Assumption 3. We assume that either (a) \( \hat{\mu}(w) \), \( \hat{e} \), and \( \hat{k} \) are \( K \)-fold cross-fitted, \( K > 1 \) or that (b) \( \hat{\mu}(w) \) and \( \hat{e} \) are restricted to lie in a Donsker class and \( \hat{k} \) is restricted to lie in a VC-subgraph class.

Assumption 3 warrants further discussion. In alternative (a), we mean by cross-fitting that the data are split into \( K \) folds, a model is trained on \( K - 1 \) of them, predictions are made on the remaining fold, and this procedure is repeated for each fold. In alternative (b), we instead restrict the complexity of the fitted functions \( \hat{\mu}(w) \), \( \hat{e} \), and \( \hat{k} \). Both of these are standard conditions in the analysis of the AIPW estimator (Chernozhukov et al. 2018; Kennedy 2016; Andrews 1994), although the condition that \( \hat{k} \) lie in a VC-subgraph class is stronger than typical conditions; we require this stronger condition to handle the fact that \( \hat{k} \) is wrapped in indicator functions for our purposes. Also, both of these assumptions might be preferred for different reasons. Cross-fitting decreases the interpretability of the analysis, since each fold has a different fitted function \( \hat{k} \) and cut-off \( \hat{\gamma} \), while the complexity conditions may be unverifiable in practice. We further compare these two assumptions in Section 3 where we find that cross-fitting is necessary for our confidence intervals to achieve the target coverage, although they continue to achieve reasonable coverage even without cross-fitting, but also that not cross-fitting increases the amount of variance reduced by trimming procedures. In this section, our results apply in either case, although we do not indicate the cross-fitting in our notation for the sake of clarity.

With this preparation, we have the following result, whose proof appears in Appendix A.

**Theorem 1.** Assume that \( \hat{\mu}(w) \), \( \hat{e} \), and \( \hat{\gamma} \) satisfy Assumptions 1, 2, and 3. Consider the doubly-robust scores

\[
\hat{\tau}(X_i) = \hat{\mu}_1(X_i) + \frac{Y_i(1) - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} Z_i - \hat{\mu}_0(X_i) - \frac{Y_i(0) - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)} (1 - Z_i),
\]

and define the doubly-robust estimator

\[
\hat{\tau}_A = \frac{1}{n_A} \sum_{i=1}^{n} \hat{\tau}(X_i) 1\{\hat{k}(X_i) \leq \hat{\gamma}\}.
\]

Then, we have the asymptotically linear expansion

\[
\hat{\tau}_A - \tau_A = \frac{1}{\mathbb{P}(k(X_i) \leq \bar{\gamma})} \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Y_i - \mu_1(X_i)}{\epsilon(X_i)} Z_i + \frac{Y_i(0) - \mu_0(X_i)}{1 - \epsilon(X_i)} (1 - Z_i) \right] 1\{\hat{k}(X_i) \leq \bar{\gamma}\} + o_P(n^{-1/2}).
\]

The value of Theorem 1 is that the right-hand side of (13) is an i.i.d. sum. This means, for instance, that we have the central limit theorem,

\[
\sqrt{n}(\hat{\tau}_A - \tau_A) \xrightarrow{d} N(0, V), \quad V = \frac{1}{\mathbb{P}(k(X_i) \leq \bar{\gamma})^2} \mathbb{E} \left[ 1\{\hat{k}(X_i) \leq \bar{\gamma}\} k(X_i) \right],
\]
and so if we estimate $V$ using the sample variance of the terms of the sum on the RHS of (12), we can form a confidence interval for $\bar{\tau}_A$ using asymptotic normal theory. Similarly, we can bootstrap the terms of the sum on the RHS of (12) to obtain bootstrap confidence intervals.

Our next result is a negative one, and shows that the situation for $\bar{\tau}_A$ is less simple. To state it, we let

$$\hat{S}_A = \sum_{i=1}^n \hat{\tau}(X_i) \mathbf{1}\{\bar{k}(X_i) \leq \bar{\gamma}\} , \quad S_A = \sum_{i=1}^n \tau(X_i) \mathbf{1}\{k(X_i) \leq \bar{\gamma}\},$$

be the numerators of $\hat{\tau}_A$ and $\bar{\tau}_A$ respectively, so that $\hat{\tau}_A = \hat{S}_A/n_A$ and $\bar{\tau}_A = S_A/n_A$. The usual approach to establishing a central limit theorem for $\hat{\tau}_A$ would be to first establish a joint central limit theorem for $\hat{S}_A$ and $n_A$, and then use the delta method obtain a central limit theorem for $\hat{\tau}_A - \bar{\tau}_A$. To illustrate the challenges of this approach, the following theorem assumes that such a procedure could be carried out, and shows what the result would be.

**Theorem 2.** Suppose Assumptions [1][3] hold, and let $\hat{\tau}(X_i)$ be as in (11). Then, if $n^{-1/2}([\hat{S}_A - n_A] - [S_A - n_A])$ converges in distribution, we have that

$$\hat{\tau}_A - \bar{\tau}_A = \frac{1}{\mathbb{P}(k(X_i) \leq \bar{\gamma})} \left( \frac{1}{n} \sum_{i=1}^n \left( \frac{Y_i - \mu_1(X_i)}{e(X_i)} Z_i + \frac{Y_i(0) - \mu_0(X_i)}{1 - e(X_i)} (1 - Z_i) \right) \mathbf{1}\{\bar{k}(X_i) \leq \bar{\gamma}\} + \Delta_n \right),$$

for

$$\Delta_n = \mathbb{P}\left( (\tau(X_i) - \mathbb{E}[\tau(X_i) \mid \bar{k}(X_i) \leq \bar{\gamma}]) \left( \mathbf{1}\{\bar{k}(X_i) \leq \bar{\gamma}\} - \mathbf{1}\{\bar{k}(X_i) \leq \bar{\gamma}\} \right) \right) + o_P(n^{-1/2}). \quad (16)$$

The convergence in distribution assumption above essentially amounts to ensuring that the error term in the delta method/Taylor expansion can be controlled; the point of the theorem is that even if this is true, we do not obtain an asymptotically linear expansion without further assumptions.

To elaborate, the expansion of $\hat{\tau}_A - \bar{\tau}_A$ contains not only an i.i.d. sum, but also an error term $\Delta_n$. This $\Delta_n$ depends on $\mathbf{1}\{\bar{k}(X_i) \leq \bar{\gamma}\} - \mathbf{1}\{\bar{k}(X_i) \leq \bar{\gamma}\}$, and so the asymptotics of $\Delta_n$ depend on the asymptotics of $\bar{k}$ itself, preventing the use of black-box machine learning algorithms to fit $\bar{k}$. Furthermore, even if $\bar{k}$ is fit from a well-specified parametric model, in which case its asymptotics are well understood, analyzing $\Delta_n$ is still challenging because $\bar{k}(X_i)$ is wrapped in non-smooth indicators, preventing the use of standard asymptotic tools like the delta method. One possible path forward is to replace $\hat{\tau}_A$ and $\bar{\tau}_A$ with smoothed equivalents, as Yang and Ding (2018) do. With this smoothing, the usual asymptotic theory is once again applicable, and they then track the contribution of $\Delta_n$ to obtain a central limit theorem for $\hat{\tau}_A - \bar{\tau}_A$ in the case of homoscedastic trimming with a fixed cut-off.

In principle, it is possible to construct other estimators for $\bar{\tau}_A$ that can be analyzed without making parametric assumptions on $\bar{k}$, e.g., by finding an efficient influence function and constructing a one-step estimator (Kennedy, 2022), but this approach may require estimating further moments beyond the conditional mean and variance, complicating the problem significantly.

Taken together, Theorem [1] and Theorem [2] lead us to recommend interpreting the results of a sample-trimmed analysis in terms of $\hat{\tau}_A$, since this approach allows for flexible non-parametric modeling, and is thus the approach we take throughout our experiments.
5 Simultaneous trimming

The results of Section 4 allow us to perform inference around a single trimmed sub-population; in this section, we develop methods for studying treatment effects around multiple trimmed sub-populations.

To illustrate the problem, consider an analyst who decides to use one of the sample trimming methods from Section 3; for example, heteroscedastic trimming with a fixed cut-off \( \hat{\gamma} \) as in (7). Suppose that, after performing inference with the chosen cut-off \( \hat{\gamma} \), the analyst finds that the width of the confidence interval for the treatment effect on the sub-population is essentially the same as the width of the confidence interval for the treatment effect on the full population, i.e., the trimming had almost no effect. In this case, the analyst would likely want to try another, larger, cut-off \( \hat{\gamma} \), and see if the corresponding sub-population has more favorable variance properties.

However, it is now extremely challenging to produce valid confidence intervals for other cut-offs \( \hat{\gamma} \) because of the selection bias introduced by the first step of the analysis. Essentially, any analysis conducted after finding the first trimming to be undesirable would have to be carried out conditional on this fact. There is a growing body of literature that develops selective inference methods to address problems of this kind (Taylor and Tibshirani, 2015), but it is somewhat challenging to develop such methods in this context while still allowing for flexible modeling of \( \hat{k} \).

As an alternative, we propose a method that allows an analyst to pre-commit to several different trimming methods in advance, and then produces a confidence intervals for each method such that the confidence intervals are simultaneously valid. That is, if we construct 95% confidence intervals, we ensure that the probability that any interval fails to cover is at most 5%. We focus our discussion on heteroscedastic trimming with a fixed-fraction cut-off \( \hat{\gamma} \), since this is a case where simultaneous intervals need not be much wider than marginal intervals, for reasons discussed below, and also since this trimming method contains a natural hyperparameter \( \delta \), the fraction of the sample trimmed, that an analyst may wish to explore. The same procedure could also be used in other ways, for example, to explore homoscedastic trimming with a range of several propensity cut-offs.

The idea is as follows: if we specify a grid of trimming fractions \( \{\delta_1, \ldots, \delta_m\} \) in advance, learn the thresholding function \( \hat{k} \), and then choose cut-offs \( \hat{\gamma}_1, \ldots, \hat{\gamma}_m \) according to (7), this procedure gives rise to a set of sub-populations \( \hat{A}_1, \ldots, \hat{A}_m \). Then, the problem of building confidence intervals that are simultaneously valid for all of these sub-populations is equivalent to building an \( \ell^\infty \) confidence region for the vector

\[
\begin{bmatrix}
\hat{\tau}_\Phi^{A_1} & \hat{\tau}_\Phi^{A_2} & \cdots & \hat{\tau}_\Phi^{A_m}
\end{bmatrix}.
\]

But it follows from Theorem 1 that this vector is asymptotically normal, and so we can construct such a confidence region using the bootstrap (Lehmann et al., 2005). The key step is that, rather than taking the quantile of the individual bootstrap replicates, we take the quantiles of the maximum of the bootstrap replicates across all sub-populations. This procedure is described in detail in Algorithm 1. Note that, in each bootstrap iteration, we do not re-fit the propensity, mean, and conditional variance estimates; this is justified by the arguments in Appendix C.10.1 of Dorn et al. (2021).

Of course, the intervals produced by Algorithm 1 will be wider than the intervals produced for a single estimand, since this is the price of the simultaneous validity. However, there are two things we can do to ensure they are much wider: choosing a small value of \( m \), so that we are not attempting to cover many estimands simultaneously, and choosing values of \( \delta_j \) that are close together so that the estimates are strongly positively correlated. In our experiments, we find that the grid \( \{1, 0.9, 0.8, 0.7\} \) gives reasonable results. This approach is informed by the interpretation of heteroscedastic trimming as outlier detection, and allows an analyst to determine whether a large amount of variance in the estimator is being introduced by a handful of units. In these cases, our
Algorithm 1 Simultaneous trimming

**Input:** data \((X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n)\), trimming fractions \(\delta_1, \ldots, \delta_m\), and confidence level \(\alpha\)

**Output:** simultaneously valid confidence intervals for trimmed sub-populations for \(1 \leq j \leq m\)

1: Estimate \(e(x) = \mathbb{P}(Z = 1 \mid X = x)\) by \(\hat{e}\), \(\sigma_w^2(x) = \text{var}(Y(w) \mid X = x)\) for \(w = 0, 1\) by \(\hat{\sigma}_w^2\) and set \(\hat{k}(x)\) as in (5)

2: Set \(\hat{\gamma}_j\) to be the \(1 - \delta_j\) quantile of \(\hat{k}(X_1), \ldots, \hat{k}(X_n)\) and let \(\hat{A}_j = \{x : \hat{k}(x) \leq \hat{\gamma}_m\}\)

3: Compute \(\hat{\tau}_{\hat{A}_j}\) as in Theorem 1 as well as standard errors \(\hat{V}_j\) as in (14)

4: for \(b = 1, \ldots, B\) do

5: Draw a bootstrapped dataset \((\tilde{X}_1, \tilde{Z}_1, \tilde{Y}_1), \ldots, (\tilde{X}_n, \tilde{Z}_n, \tilde{Y}_n)\)

6: Compute point estimates \(\hat{\tau}_{\hat{A}_j}^b\), and standard errors \(\hat{V}_j^b\) on the bootstrapped sample (without refitting \(\hat{k}\))

7: Compute

\[
T_b = \max_{1 \leq j \leq m} \left| \frac{\hat{\tau}_{\hat{A}_j}^b - \hat{\tau}_{\hat{A}_j}}{\sqrt{\hat{V}_j^b}} \right|
\]

8: end for

9: Let \(q\) be the \(1 - \alpha\) quantile of \(T_1, \ldots, T_B\)

10: return \(\hat{\tau}_{\hat{A}_j} \pm q\sqrt{\hat{V}_j}\) for \(1 \leq j \leq m\)

method produces much lower variance estimates on large sub-populations, which may be of interest.

6 Experiments

We now present a series of experiments to evaluate our proposed trimming methods.

Our first set of experiments is a coverage study on simulated data to confirm that intervals produced using the methods of Sections 4 and 5 obtain the target coverage level in large samples. Our second set of experiments is conducted on the National Health and Nutrition Examination Survey (NHANES) [CDC et al., 2015], which has been studied in previous sample trimming work as well (Yang and Ding, 2018; Hsu and Small, 2013), and provides a comparison between our methods and traditional (propensity-based) trimming methods.

Our third set of experiments involves data from the 2022 American Causal Inference Conference (ACIC) data challenge [ACIC, 2022], which is a semi-synthetic dataset designed to mimic a real dataset of Medicare interventions. We use this data to demonstrate the value of the simultaneous trimming method of Section 5 in particular, constructing confidence intervals around several sub-populations simultaneously. We see that the confidence intervals for some of the sub-populations are much narrower than those around the full population. Most importantly, having set aside concerns of selective inference, we can freely choose which of these populations to report the effect on based on our preferences between confidence interval width and sample size.

6.1 Coverage experiments

We begin by studying the coverage rates of confidence intervals for the estimators in Section 4. To do so, we generate data from the following model:

\[
\begin{align*}
X_1, X_2 &\sim N(0, 1) \\
\epsilon &\sim N(0, 1 + (X_2)_+) \\
Y(0) &= 2X_1 - X_2 + \epsilon \\
Y(1) &= Y(0) + \tau \\
e(X) &= \frac{1}{1 + 2 \exp(X_2 - X_1)}.
\end{align*}
\]
Table 1: Coverage of 95% confidence intervals based on Theorem 1 for treatment effects when heteroscedastically trimming none, 5%, and 10% of the sample on data generated from (17) for different sample sizes, as well as coverage of simultaneous confidence intervals constructed using Algorithm 1. All results are based on 1000 trials. We see that, for large n, we achieve approximately the target 95% coverage.

|               | Target | 500  | 1000 | 2000 | 4000 | 8000 | 16000 | 32000 |
|---------------|--------|------|------|------|------|------|-------|-------|
| Cross-fitted  | δ = 0.1| 0.878| 0.915| 0.928| 0.947| 0.952| 0.952 | 0.942 |
|               | δ = 0.05| 0.871| 0.903| 0.927| 0.945| 0.952| 0.94 | 0.937 |
|               | δ = 0  | 0.855| 0.879| 0.915| 0.936| 0.955| 0.937 | 0.94  |
|               | Simultaneous | 0.866| 0.892| 0.922| 0.938| 0.955| 0.948 | 0.942 |
| Not cross-fitted | δ = 0.9 | 0.839| 0.844| 0.884| 0.885| 0.902| 0.901 | 0.893 |
|               | δ = 0.95 | 0.836| 0.843| 0.877| 0.883| 0.898| 0.894 | 0.897 |
|               | δ = 1   | 0.858| 0.861| 0.892| 0.904| 0.911| 0.923 | 0.922 |
|               | Simultaneous | 0.839| 0.842| 0.874| 0.886| 0.903| 0.9  | 0.895 |

and set a constant treatment effect $\tau = 1$. This is a straightforward linear model with two important features: first, the errors $\epsilon$ are heteroscedastic, meaning that there is in fact a non-trivial conditional variance to model, and the propensity $e(X)$ does not come from a standard logistic model, necessitating the use of non-parametric methods. We threshold the propensities to lie in $[0.05, 0.95]$—this is to obtain convergence to asymptotic coverage rates in reasonable sample sizes. With more extreme propensities, we expect to find similar results, but in possibly larger sample sizes than those considered here. Then, we estimate the propensities and conditional means using the \texttt{regression_forest} function from the R package \texttt{grf} \cite{Tibshirani2021}; we fit conditional variances by subtracting estimates of $E[Y_i \mid X_i]^2$ and $E[Y_i^2 \mid X_i]$ (this procedure may theoretically produce negative values, but does not in our case, suggesting that the regressions are fairly accurate). We run this procedure both with and without cross-fitting, so as to compare the two options in Assumption 3.

Finally, based on these estimates, we heteroscedastically trim a $1 - \delta$ fraction of the data for $\delta = 0, 0.05, 0.1$. For each of these sub-populations, we construct marginal 95% confidence intervals using Theorem 1, as well as simultaneous 95% confidence intervals for all three sub-populations using Algorithm 1. Because the true treatment effect is constant, the treatment effect is still $\tau$ for each sub-population, and so we can estimate the coverage rates of these confidence intervals across multiple trials. Results are shown in Table 1.

These results confirm that our intervals generally achieve the desired coverage levels; we make two further comments. First, comparing the cross-fitted and non-cross-fitted results, we see that the cross-fitted intervals are closer to the target 95% coverage level, but that the non-cross-fitted still achieve reasonable coverage of about 90% in large samples. This suggests that, if exact coverage is extremely important in an application, it is advisable to cross-fit. There are, however, some drawbacks to cross-fitting, which we discuss in Section 6.2. Second, the sample sizes in Table 1 at which coverage is achieved might seem relatively large, but these are consistent with previous studies of the coverage of random forest based methods \cite{Athey2019}.
6.2 NHANES experiments

Following prior work on sample trimming methods, we analyze the 2007-2008 U.S. National Health and Nutrition Examination Survey data (CDC et al., 2015), attempt to estimate the effect of smoking on blood lead levels as in Yang and Ding (2018); Hsu and Small (2013), and compare our trimming methods to other common trimming methods. The dataset contains 3,340 subjects, 679 of whom are smokers, and 2,661 of whom are not. The response, blood lead level, ranges from 0.18 μg/dl to 33.10 μg/dl, in the population, suggesting that there may be heteroscedasticity in the data. As before, we fit the propensities and conditional means using the regression_forest function from the R package grf (Tibshirani et al., 2021); we fit conditional variances by subtracting estimates of \(E[Y_i | X_i] \) and \(E[Y_i^2 | X_i] \) (this may again theoretically produce negative values, but does not in our case, suggesting that the regressions are fairly accurate). All regressions use gender, education level, income, and race as covariates, (these are the same covariates as in Yang and Ding (2018), except that we do not use age as a covariate—the reason for this is that later, we will examine the covariate distribution of our trimmed sub-populations, and the results are more interpretable with fewer covariates.)

Additionally, in this example, we do not cross-fit our regressions. This is because we find empirically that cross-fitting often significantly decreases the efficiency of our procedure i.e., that the sub-populations found by cross-fitted regressions have much higher variance than those found by non-cross-fitted regressions. To understand why this happens, consider the unit \(Y_i \) with the highest conditional-variance. This \(Y_i \) may be significantly larger than any of the other responses in the sample, and depending on the quality of our covariates, we may not be able to correctly model how large this \(Y_i \) is without actually observing it in the sample. Thus, when we cross-fit, our model is less accurate on exactly those units that are most important for us to model accurately. When we do not cross-fit, we slightly over-fit to the observed responses, but this ensures that we trim the correct units. Furthermore, we saw in Section 6.1 that we can still obtain reasonable coverage without cross-fitting, and so we do not cross-fit in this example.

We compare three different sample trimming methods on this data to the baseline of no sample trimming. In the language of Section 3, these are homoscedastic trimming with the constant cut-off of 11.1 (recall that \(1/0.1 + 1/0.9 \approx 11.1 \)) proposed by Crump et al. (2009), which we refer to from here on as the 0/1/0.9 rule, homoscedastic trimming with a variance-minimizing cut-off, and heteroscedastic trimming with variance-minimizing cut-off. For all of these methods, we perform inference using the estimator of Theorem 1. Point estimates and 95% confidence intervals constructed from asymptotic normal theory are shown in Table 2 and visualized in Figure 1. We note that, for the optimal homoscedastic trimming, the implied propensity cut-offs found are 0.04 and 0.96.

We see from these results that, in this problem, sample trimming based only on propensities offers little variance reduction—the variance-minimizing homoscedastic trimming and the 0.1/0.9 rule both actually slightly increase the variance of our point estimate. On the other hand, the variance-minimizing heteroscedastic trimming offers a 17% reduction in variance and visibly smaller confidence intervals, as seen in Figure 1.

We can also see from Figure 1 and Table 2 that the estimated treatment effect for the sub-population found by the heteroscedastic trimming is quite different from the estimated treatment effects for the sub-populations found by other trimming methods. This is true even though the sub-population found by optimal heteroscedastic trimming is only about 15% smaller than the sub-population found by the 0.1/0.9 rule. It is also to be expected, since units with high conditional variances are intuitively likely to have high conditional means as well, and thus dropping them from the sample may substantially change the treatment effect. This change in treatment effect also means that it is important to understand the sub-population we are trimming down to and how it
Figure 1: Normal theory 95% confidence intervals for estimated treatment effects using three different sample trimming methods (optimal variance-aware trimming, optimal propensity trimming, the 0.1/0.9 rule) and also no trimming. We see that the homoscedastic trimming based on propensities alone offers no variance reduction relative to not trimming at all, but that heteroscedastic trimming reduces variance more significantly, albeit around a noticeably different point estimate.

| Method              | \( \hat{\tau} \) | \( \hat{\sigma} \) | Conf. interval       | Sample fraction |
|---------------------|-------------------|---------------------|---------------------|-----------------|
| Optimal var-aware   | 0.438             | 0.057               | (0.327, 0.55)       | 0.511           |
| Optimal propensity  | 0.707             | 0.095               | (0.522, 0.893)      | 0.697           |
| 0.1/0.9 rule        | 0.714             | 0.102               | (0.514, 0.913)      | 0.642           |
| None                | 0.734             | 0.073               | (0.591, 0.877)      | 1               |

Table 2: Point estimates and standard errors for sample trimming with various methods. We see that the propensity based trimming methods offer little to no variance reduction, while the homoscedasticity-aware method appreciably reduces variance. However, the sub-population found by the heteroscedasticity-aware method is smaller than the others, containing only about half of the sample, and has a noticeably different point estimate.

differs from the full population. We explore these differences by analyzing the distribution of the education covariate in the full sample, the variance-minimizing heteroscedastic sub-population, and the variance-minimizing homoscedastic sub-population, shown in Table [3].

We see in Table [3] that, broadly speaking, the variance-minimizing heteroscedastic trimming more aggressively trims units with low education levels, while the variance-minimizing homoscedastic trimming more aggressively trims units with high education levels. The difference can be understood by looking at the average values of our fitted propensities and variances within each education level, which are also shown in Table [3]. These reveal that less educated units have higher propensities and higher conditional variances, while more educated units have lower propensities and lower conditional variances. Thus the homoscedastic trimming removes many units with high levels of education because of their low propensities, while the heteroscedastic trimming allows for the low variance of those units to compensate for their small propensities; similarly, the heteroscedastic trimming removes many units with low education levels because of their high variance, while the homoscedastic trimming allows those units to remain because of their relatively high propensities.
Table 3: Distribution of education level in the full sample of the NHANES data, in the optimal variance-aware trimming set $\hat{A}^{\text{het}}$, and in the optimal propensity trimming set $\hat{A}^{\text{hom}}$, along with average values of fitted models for $\hat{e}$, $\hat{\sigma}_0$, and $\hat{\sigma}_1$ within each education level. We see, for instance, that $\hat{A}^{\text{het}}$ includes fewer people with some high-school than $\hat{A}^{\text{hom}}$ does, because even though their estimated propensities are not very small, their estimated variances are very large.

| Education level | Sample | $\hat{A}^{\text{het}}$ | $\hat{A}^{\text{hom}}$ | $\hat{e}$ | $\hat{\sigma}_0$ | $\hat{\sigma}_1$ |
|-----------------|--------|------------------------|------------------------|--------|----------------|----------------|
| No HS           | 444    | 130                    | 273                    | 0.174  | 1.918          | 7.337          |
| Some HS         | 550    | 334                    | 530                    | 0.297  | 2.487          | 4.869          |
| HS              | 838    | 568                    | 711                    | 0.255  | 1.345          | 3.563          |
| Some college    | 850    | 404                    | 688                    | 0.207  | 1.42           | 3.822          |
| College         | 656    | 272                    | 125                    | 0.072  | 1.41           | 3.158          |

Table 4: Pointwise and simultaneous confidence intervals on subsets of different sizes of the ACIC data. The treatment has a small positive effect on the full population, but this is only due to a few outliers. In fact, the effect of the treatment is negative on large sub-populations of the data, and the simultaneous confidence intervals show that this negative effect is significant on the sub-populations of 80% and 70% of the data, even though the effect of the treatment is not significant on the full population. Because our confidence intervals and uniformly valid, we may choose which population to report the effect on, or simply report all four.

| Sample frac. $\hat{\tau}$ | $\hat{\tau}$ | $\hat{\sigma}$ | Pointwise CI | Simultaneous CI |
|-----------------------------|--------------|----------------|-------------|-----------------|
| 1                           | 6.551        | 19.813         | $(-31.941, 45.043)$ | $(-41.966, 55.068)$ |
| 0.9                         | -29.952      | 11.4           | $(-51.951, -7.953)$ | $(-57.867, -2.038)$ |
| 0.8                         | -43.379      | 10.129         | $(-62.591, -24.166)$ | $(-68.182, -18.575)$ |
| 0.7                         | -36.382      | 9.255          | $(-54.459, -18.305)$ | $(-59.044, -13.72)$ |

Thus, in this application, trimming the sample based on propensities alone is not able to provide any meaningful variance reduction, but heteroscedasticity-aware sample trimming is able to identify a sub-population on which the variance of the estimated treatment effect is smaller than the variance of the estimated treatment effect on the full population. Further analysis of this sub-population shows that it is roughly a sub-population of highly educated individuals. Of course, whether or not this sub-population is actually of interest will depend on analyst-specific considerations, and our goal here is not to suggest that this sub-population is necessarily preferable to the full population. Indeed, the question of whether inference around a sub-population is relevant to a given domain problem is central to all trimming methods. Even the propensity-trimmed sub-population, with its down-weighting of more educated individuals, represents a sub-population that may or may not be preferable to the full population. Instead, we hope this experiment shed lights on how our proposed methods compares to existing methods, and clarifies what they are able to offer.

6.3 ACIC experiments

Our final set of experiments involves data from the 2022 American Causal Inference Conference’s data challenge [ACIC 2022]. This dataset is a semi-synthetic dataset constructed to mimic data from evaluations of Medicare interventions on the U.S. health-care system. This dataset is of
Figure 2: Left: estimated trimming function $\hat{k}$ versus AIPW scores (the terms of the sum in (12)); right: pointwise and simultaneous confidence intervals on subsets of different sizes of the ACIC data. The right plot shows that even though the treatment effect is not significant on the full population, we are able to identify large sub-populations on which the treatment effect is significant, and give confidence intervals that are simultaneously valid across all four sub-populations. The left plot explains why the treatment effects on the sub-populations are much smaller and of different sign than the treatment effect on the full population: the units with the largest values of $\hat{k}$ have large positive treatment effects, and trimming them reduces the estimated average treatment effect.

particular interest to us because the response, Medicare spending, is extremely heavy-tailed, with some patients spending a few hundred dollars and others spending tens of thousands of dollars. In such a setting, trimming the sample to exclude some outlier patients is likely to offer significant variance reduction.

The data consists of observations of patients over four years, with each patient assigned to a medical practice, and treatment assigned at the practice level. We make several simplifications to the problem: we restrict to a single year of data and work with a subset of 20% of the data for computational reasons. We treat the practice to which a patient is assigned as a covariate, and set the probability that a patient is treated to be equal to the probability that their practice is treated. Finally, we do not account for the cluster structure of the data in our analysis (Abadie et al., 2017); sample trimming in the presence of cluster structure is an exciting direction for future work. After these simplifications, there are 65,609 patients in the sample, whose propensities range from 0.11 to 0.85, suggesting that there are not significant overlap issues in the data. Meanwhile, responses range from $-7.20$ (indicating a patient paid Medicare, rather than the other way around) to $67,847$, suggesting there may potentially be significant heteroscedasticity.

For our analysis, we use the simultaneous trimming method of Section 5. Specifically, we use heteroscedastic fixed-fraction trimming with fractions $\{0, 0.1, 0.2, 0.3\}$; these methodological choices are based on our prior belief that there is a small fraction of patients with extreme responses, and that trimming these patients may significantly reduce variance. (For reference, the optimal heteroscedastic trimming trims nearly 90% of the data, and is thus not an appealing option in this case.) We fit propensity scores using practice-level covariate and conditional means and variances using patient-level covariates. As before, all regressions are fit using the \texttt{regression} function from the R package \texttt{grf} (Tibshirani et al., 2021). For the same reasons discussed in Section 6.2, we do not cross-fit these regressions.

The results of this analysis are reported in Table 4 and visualized in Figure 2. These results paint a clear picture of the effects of the intervention: we see from the left panel of Figure 2 that there is
significant heterogeneity with the treatment increasing costs for some patients, and decreasing costs for others, and that there are a handful of outliers with large AIPW scores that significantly affect our estimate. In the right panel of the same figure, we find that the treatment effect is slightly positive on the full population, but with a wide confidence interval due to the variance introduced by units with large responses. However, after trimming a small fraction of the data, the treatment effect is negative and the confidence interval around it is much narrower, because we have trimmed off the extreme units from the left panel. Furthermore, for all three sub-populations, we can conclude that this negative effect is statistically significant, in part because of the much lower variance of these populations. The change in point estimate also contributes to significance, but we note that if the sub-populations had the same standard error as the full population, their confidence intervals would contain zero. Thus the combination of simultaneous valid confidence intervals and heteroscedastic trimming effectively served to remove the handful of units distorting our estimate of the treatment effect and allowed us to identify a large sub-population of the data on which the treatment has a statistically significant negative effect.

Crucially, we would also not be able to draw this conclusion using classical sample trimming methods, since we would not know in advance which size of sub-population would be of interest. If we chose this sub-population adaptively from the data, the resulting inference would not be valid. By using our domain knowledge of the heavy-tailed behavior of the response and pre-committing to a set of trimming fractions, we are able to elucidate the structure of the response and draw statistically valid inferences.

The structure of this example—where the treatment effect on the full population is not significant but the treatment effect on a sub-population is—is illustrative, but not necessarily representative of all settings. In particular, it may well be the case that the treatment effect is not significant for any of the sub-populations considered, or significant for all of them. However, the goal of our method is not to find a sub-population on which there is a significant effect, but rather to find sub-populations with smaller confidence intervals than the full population, and allow for an analyst to choose freely between these and the full population. In this way, regardless of issues of statistical significance, the method can provide insight on the structure of treatment effects in the presence of outliers, and provide an analyst with statistically valid statements about potentially interesting sub-populations.

7 Conclusion

In this paper we have proposed new methods for sample trimming that account for not only extreme propensities but also extreme conditional variances, and developed new theoretical results on inference after sample trimming. Our experiments show that our proposed estimators and confidence intervals achieve good coverage, reduce variance even when propensity-based trimming alone cannot, and enable analysts to choose between several different sub-populations based on any criteria they choose.

There are several interesting directions for future work, the most prominent of which would be to develop further results for inference around the limiting estimand $\tau_A$ of Section 4. These would likely require stronger assumptions than the ones we make, but would be of value when an estimand that is independent of the sample realization is needed. Other open problems include better methods for trading off sample size and variance reduction, extending our results to settings with dependence between units, and considering trimming methods that do not rely on one-dimensional summary functions.

Finally, there is also a broader set of questions around trimming methods more generally that are worth further exploration. The framework we have taken here, of performing inference around a
modified estimand, is standard in the literature, but may not be ideal from an analyst’s perspective, who may be unsure of which estimand is actually relevant to them. This is especially pertinent to our simultaneous trimming methods, since the path from a set of simultaneous confidence intervals around multiple estimands to a binary decision on implementing a policy is unclear. These issues are closely related to those that arise in problems with heterogeneous treatment effects, where a treatment may help some parts of the population but harm others, making the decision to treat or not fraught with complications. For these reasons, methods that do not shift the estimand but instead bound the impact of the trimmed units may be especially useful, and are worthy of further study.

Acknowledgements

We thank Kevin Guo for helpful discussions and feedback on early versions of this work.

References

Alberto Abadie, Susan Athey, Guido W Imbens, and Jeffrey Wooldridge. When should you adjust standard errors for clustering? Technical report, National Bureau of Economic Research, 2017.

ACIC. American causal inference conference 2022 data challenge, 2022. URL https://acic2022.mathematica.org/

Donald WK Andrews. Empirical process methods in econometrics. Handbook of econometrics, 4:2247–2294, 1994.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. The Annals of Statistics, 47(2):1148–1178, 2019.

Pauline Burke et al. Measuring average treatment effect from heavy-tailed data. arXiv preprint arXiv:1905.09252, 2019.

CDC et al. National health and nutrition examination survey. nhanes 2007-2008 public data general release file documentation, 2009, 2015.

Saraswata Chaudhuri and Jonathan B Hill. Heavy tail robust estimation and inference for average treatment effects. Technical report, Working paper, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1):187–199, 2009.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. arXiv preprint arXiv:2112.11449, 2021.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71(4):1161–1189, 2003.

Jesse Y Hsu and Dylan S Small. Calibrating sensitivity analyses to observed covariates in observational studies. Biometrics, 69(4):803–811, 2013.
Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Edward H Kennedy, Sivaraman Balakrishnan, and Max G’Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics*, 48(4):2008–2030, 2020.

Samir Khan and Johan Ugander. Adaptive normalization for ipw estimation. *arXiv preprint arXiv:2106.07695*, 2021.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

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

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Julie Tibshirani, Susan Athey, Erik Sverdrup, and Stefan Wager. *grf: Generalized Random Forests*, 2021. URL https://CRAN.R-project.org/package=grf R package version 2.0.2.

Nilesh Tripuraneni, Dhruv Madeka, Dean Foster, Dominique Perrault-Joncas, and Michael I. Jordan. A framework for the meta-analysis of randomized experiments with applications to heavy-tailed response data, 2021.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Jon Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

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

In our proofs, we employ empirical process notation. We let \( W_i = (X_i, Y_i, Z_i) \) be the entire triplet we observe for unit \( i \), and we write \( \mathbb{P}_n f = \frac{1}{n} \sum f(W_i) \) and \( \mathbb{P} f = \int f(w) d\mathbb{P}(w) \). Note that for a random function \( \hat{f} \), \( \mathbb{P} \hat{f} \) is a random variable, since we do not integrate over the randomness in \( \hat{f} \). In contrast, \( \mathbb{E}[\hat{f}] \) is a deterministic quantity that integrates out the randomness in a new sample and in \( \hat{f} \). We also define the norm \( ||f||_{L^q(P)} = (\mathbb{P}|f|^q)^{1/q} \).

A.1 Proof of Theorem 1

Before proceeding with the main proof, we introduce a helpful lemma.

**Lemma 1.** Suppose \( \hat{k} \) and \( \hat{\gamma} \) satisfy Assumption 2. Then

\[
\|1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\}\|_{L^2(P)} \xrightarrow{P} 0. \tag{18}
\]

**Proof.** We begin by writing

\[
\|1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\}\|_{L^2(P)} \leq \|1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\}\|_{L^2(P)} + \|1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\}\|_{L^2(P)}, \tag{19}
\]

by the triangle inequality, and then bound each of the two terms on the RHS of \( \text{(19)} \) separately. For the first term on the RHS of \( \text{(19)} \), we have

\[
\|1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\}\|_{L^2(P)}^2 = \int \left( 1\{\hat{k}(x) \leq \hat{\gamma}\} - 1\{\hat{k}(x) \leq \hat{\gamma}\} \right)^2 d\mathbb{P}(x), \tag{20}
\]

\[
= \int \left( \min(\hat{\gamma}, \hat{\gamma}) - \hat{k}(x) \leq \min(\hat{\gamma}, \hat{\gamma}) - \hat{k}(x) \right) d\mathbb{P}(x), \tag{21}
\]

\[
\leq B|\hat{\gamma} - \gamma|, \tag{22}
\]

\[
= o_P(1), \tag{23}
\]

where \( B \) is an upper bound on the density of \( \hat{k}(x) \), and the final equality uses the fact that \( \hat{\gamma} \) converges to \( \gamma \) in probability.

For the second term on the RHS of \( \text{(19)} \), we have

\[
\|1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\}\|_{L^2(P)} = \int \left( 1\{\hat{k}(x) \leq \hat{\gamma}\} - 1\{\hat{k}(x) \leq \hat{\gamma}\} \right)^2 d\mathbb{P}(x), \tag{24}
\]

\[
= \int 1\{\min(\hat{k}(x), \hat{k}(x)) \leq \hat{\gamma} \leq \max(\hat{k}(x), \hat{k}(x)) \} d\mathbb{P}(x). \tag{25}
\]

We assume for convenience now that \( \hat{k}(x) \leq \hat{k}(x) \) for all \( x \); the case where this assumption does not hold is analogous. Then, to show that \( \text{(25)} \) is \( o_P(1) \), take any \( \epsilon > 0 \) and consider

\[
\int 1\{\hat{k}(x) \leq \hat{\gamma} - \epsilon, \hat{k}(x) \geq \hat{\gamma} + \epsilon \} d\mathbb{P}(x) \leq \int 1\{|\hat{k}(x) - \hat{k}(x)| \geq 2\epsilon \} d\mathbb{P}(x). \tag{26}
\]

The RHS is of \( \text{(26)} \) is \( o_P(1) \) by Assumption 2 and the LHS converges to \( \text{(25)} \) as \( \epsilon \to 0 \) by dominated convergence. Thus we can send \( \epsilon \to 0 \) and conclude that \( \text{(25)} \) is \( o_P(1) \) as well. Finally, combining \( \text{(23)} \), the arguments above, and \( \text{(19)} \), we conclude that \( \text{(18)} \) holds. \( \square \)
With this lemma in hand, we can now prove Theorem 1.

\textbf{Proof of Theorem 1.} We begin with some notation. Define

\[
\psi_1(W; \hat{\mu}_0, \hat{k}, \hat{\gamma}) = (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) - \tau(X_i)) \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\}
\]

and

\[
\psi_2(W; \hat{\mu}_0, \hat{\epsilon}, \hat{k}, \hat{\gamma}) = \left( \frac{Y_i(1) - \hat{\mu}_1(X_i)}{\hat{\epsilon}(X_i)} Z_i + \frac{Y_i(0) - \hat{\mu}_0(X_i)}{1 - \hat{\epsilon}(X_i)} \right) \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\}.
\]

Essentially, \(\psi_1\) contains terms that will vanish asymptotically, and \(\psi_2\) contains the terms that will contribute to the asymptotically linear expansion.

With these definitions we have.

\[
\hat{\tau}_A - \tau_A = \frac{\mathbb{P}_n \left( \psi_1(W; \hat{\mu}_0, \hat{k}, \hat{\gamma}) + \psi_2(\hat{\mu}_0, \hat{\epsilon}, \hat{k}, \hat{\gamma}) \right)}{n A/n}
\]

(27)

We now proceed in two steps. The first step is to show that

\[
\frac{n A}{n} \mathbb{P} (\hat{k}(X_i) \leq \hat{\gamma})
\]

(28)

and the second is to show that

\[
\mathbb{P}_n \left( \psi_1(W; \hat{\mu}_0, \hat{k}, \hat{\gamma}) + \psi_2(\hat{\mu}_0, \hat{\epsilon}, \hat{k}, \hat{\gamma}) \right) = \mathbb{P}_n \psi_2(W; \mu_0, \epsilon, \hat{k}, \hat{\gamma}) + o_P(n^{-1/2}).
\]

(29)

Once we establish (28) and (29), the desired result (13) follows from Slutsky’s lemma.

For the first step, we begin by noting that \(n A/n = \mathbb{P}_n \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\}\) and then computing

\[
\mathbb{P}_n \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} = (\mathbb{P}_n - \mathbb{P}) \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\},
\]

(30)

\[
= (\mathbb{P}_n - \mathbb{P}) \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + o_P(n^{-1/2}),
\]

(31)

\[
= \mathbb{P}_n \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + \mathbb{P}\{\hat{k}(X_i) \leq \hat{\gamma}\} - \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + o_P(n^{-1/2}),
\]

(32)

\[
\leq \mathbb{P}_n \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + \mathbb{P}\{\hat{k}(X_i) \leq \hat{\gamma}\} - \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + o_P(n^{-1/2}),
\]

(33)

\[
= \mathbb{P}_n \mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\} + o_P(1),
\]

(34)

\[
= \mathbb{P}\{\hat{k}(X_i) \leq \hat{\gamma}\} + o_P(1).
\]

(35)

where the second equality uses either Lemma 2 of [Kennedy et al. (2020)] and Assumption 3(a) or Lemma 19.24 of [Van der Vaart (2000)], Assumption 3(b), and the fact that \(\mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\}\) is VC-subgraph and thus Donsker because \(\hat{k}\) lies in a VC-subgraph class and the indicator is a monotone function ([Wellner et al. (2013)]. The fifth equality uses Lemma 1 and the fact that \(L^2\)-convergence implies \(L^1\)-convergence, and the sixth equality uses the law of large numbers. Since (35) implies (28), we are done with the first step.

We now proceed to the second step of showing (29). It is straightforward to check that, given our assumptions, \(\psi_1(W; \hat{\mu}_0, \hat{k}, \hat{\gamma})\) converges in \(L^2(\mathbb{P})\) to \(\psi_1(W; \mu_0, \hat{k}, \hat{\gamma})\), and similarly for \(\psi_2\). Also, under Assumption 3(b), \(\psi_1\) and \(\psi_2\) are both Donsker. This is because, since \(\hat{k}\) lies in a VC-subgraph class, \(\mathbb{1}\{\hat{k}(X_i) \leq \hat{\gamma}\}\) also lies in a VC-subgraph class and thus a Donsker class as well. Since Donsker classes are closed under addition and multiplication ([Wellner et al. (2013)], it follows that \(\psi_1\) and \(\psi_2\) lie in a Donsker class.
With this in mind, we first consider \( \mathbb{P}_n \psi_1 \) and find that
\[
\mathbb{P}_n \psi_1(W; \hat{\mu}(w), \hat{k}, \hat{\gamma}) = (\mathbb{P}_n - \mathbb{P}) \psi_1(W; \hat{\mu}(w), \hat{k}, \hat{\gamma}) + \mathbb{P} \psi_1(W; \hat{\mu}(w), \hat{k}, \hat{\gamma}),
\]
(36)
\[
= (\mathbb{P}_n - \mathbb{P}) \psi_1(W; \mu(w), \hat{k}, \hat{\gamma}) + \mathbb{P} \psi_1(W; \hat{\mu}(w), \hat{\gamma}) + o_P(n^{-1/2}),
\]
(37)
\[
= \mathbb{P} \psi_1(W; \hat{\mu}(w), \hat{k}, \hat{\gamma}) + o_P(n^{-1/2}),
\]
(38)
where the second equality uses either Assumption 3(a) and Lemma 2 of Kennedy et al. (2020) or Assumption 3(b), Lemma 19.24 of Van der Vaart (2000), and the arguments of the preceding paragraph. The third equality follows from the fact that \( \psi_1(W; \hat{\mu}(w), \hat{k}, \hat{\gamma}) = 0 \) identically.

Similarly,
\[
\mathbb{P}_n \psi_2(W; \hat{\mu}(w), \hat{\gamma}) = (\mathbb{P}_n - \mathbb{P}) \psi_2(W; \hat{\mu}(w), \hat{\gamma}) + \mathbb{P} \psi_2(W; \hat{\mu}(w), \hat{\gamma}),
\]
(39)
\[
= (\mathbb{P}_n - \mathbb{P}) \psi_2(W; \mu(w), \hat{\gamma}) + \mathbb{P} \psi_2(W; \hat{\mu}(w), \hat{\gamma}),
\]
(40)
\[
= \mathbb{P} \psi_2(W; \hat{\mu}(w), \hat{\gamma}) + o_P(n^{-1/2}),
\]
(41)
again using either Assumption 3(a) and Lemma 2 of Kennedy et al. (2020) or Assumption 3(b) and Lemma 19.24 of Van der Vaart (2000) for the second equality, and the fact that \( \mathbb{P} \psi_2(W; \mu(w), \hat{\gamma}) = 0 \) identically for the third equality.

Taken together, (38) and (41) imply that
\[
\mathbb{P}_n (\psi_1 + \psi_2) = \mathbb{P}_n \psi_2(W; \mu(w), \hat{\gamma}) + \mathbb{P} \left( \psi_1(W; \hat{\mu}(w), \hat{\gamma}) + \psi_2(W; \hat{\mu}(w), \hat{\gamma}) \right).
\]
(42)

Now, the second term of (42) is
\[
= \mathbb{P} \left[ \left( \frac{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)}{\hat{e}(X_i)} + \frac{Y_i - \hat{\mu}_0(X_i)}{e(X_i)} \right) \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \right]
\]
(43)
\[
= \sum_{w \in \{0, 1\}} \mathbb{P} \left[ \left( \frac{\mu(w) - \mu(X_i)}{\hat{e}(X_i)} + \frac{Y_i - \mu(X_i)}{e(X_i)} \right) \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \right].
\]
(44)

We bound the \( w = 1 \) term; the \( w = 0 \) term is analogous. The \( w = 1 \) term is, after applying the tower rule conditional on \( X_i \) and simplifying,
\[
= \mathbb{P} \left[ \left( \frac{(\hat{\mu}_1(X_i) - \mu_1(X_i)) \hat{e}(X_i) + (\mu_1(X_i) - \hat{\mu}_1(X_i)) e(X_i)}{\hat{e}(X_i)} \right) \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \right]
\]
(45)
\[
\leq \mathbb{P} \left[ (\hat{\mu}_1(X_i) - \mu_1(X_i)) (\hat{e}(X_i) - e(X_i)) \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \right]
\]
(46)
\[
\leq \sqrt{\mathbb{P} \left( (\hat{\mu}_1(X_i) - \mu_1(X_i))^2 \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \right) \cdot \mathbb{P} \left( (\hat{e}(X_i) - e(X_i))^2 \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \right)}
\]
(47)
\[
= \| (\hat{\mu}_1(X_i) - \mu_1(X_i)) \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \|_{L^2(\mathbb{P})} \times \| (\hat{e}(X_i) - e(X_i)) \mathbb{1} \{ \hat{k}(X_i) \leq \hat{\gamma} \} \|_{L^2(\mathbb{P})},
\]
(48)
where the first inequality uses the fact that \( \hat{e}(X_i) \) is bounded away from 0 and 1 for sufficiently large \( n \) by consistency and the second uses Cauchy-Schwarz. Then, (48) is \( o_P(n^{-1/2}) \) by Assumption 1, which establishes (29) and completes the proof.

A.2 Proof of Theorem 2

In this section, we analyze \( \hat{\tau}^{\hat{A}} - \tau^{\hat{A}} \). This is slightly more involved than the results of the previous section because both the numerators and the denominators of \( \hat{\tau}^{\hat{A}} \) and \( \tau^{\hat{A}} \) differ. Thus we first
analyze the numerators, then the denominators, and then combine our results to obtain the theorem. To simplify notation, let
\[ \hat{S}_A = \sum_{i=1}^{n} \hat{\tau}(X_i) 1\{ \hat{k}(X_i) \leq \hat{\gamma} \}, \quad S_A = \sum_{i=1}^{n} \tau(X_i) 1\{ k(X_i) \leq \gamma \}, \] (49)
be the numerators of \( \hat{S}_A \) and \( S_A \), so that \( \hat{S}_A = \hat{S}_A/n_A \) and \( S_A = S_A/n_A \).

The following lemma analyzes the difference between \( \hat{S}_A \) and \( S_A \).

**Lemma 2.** We have
\[
\frac{\hat{S}_A - S_A}{n} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i - \mu_1(X_i)}{e(X_i)} Z_i + \frac{Y_i(0) - \mu_0(X_i)}{1 - e(X_i)} (1 - Z_i) \right) 1\{ \hat{k}(X_i) \leq \hat{\gamma} \}
+ \mathbb{P}[\tau(X_i)(1\{ \hat{k}(X_i) \leq \hat{\gamma} \} - 1\{ k(X_i) \leq \gamma \})] + o_P(n^{-1/2}),
\] (50)

**Proof.** As before, we begin by defining
\[ \psi(W; \hat{\mu}(w), \hat{\gamma}) = \left( \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{Y_i(1) - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} Z_i - \frac{Y_i(0) - \hat{\mu}_0(X_i)}{1 - \hat{e}(X_i)} (1 - Z_i) \right) 1\{ \hat{k}(X_i) \leq \hat{\gamma} \}, \]
so that \( \hat{S}_A/n = \mathbb{P}_n \psi(W; \hat{\mu}(w), \hat{\gamma}) \). We can check that \( \psi(W; \hat{\mu}(w), \hat{\gamma}) \) converges in \( L^2 \) to \( \psi(W; \mu(w), e, \hat{k}, \gamma) \), and so we can then write
\[ \hat{S}_A/n = \mathbb{P}_n \psi(W; \hat{\mu}(w), \hat{\gamma}) = (\mathbb{P}_n - \mathbb{P}) \psi(W; \mu(w), \hat{\gamma}) + \mathbb{P} \psi(W; \hat{\mu}(w), \hat{\gamma}), \]
\[ = (\mathbb{P}_n - \mathbb{P}) \psi(W; \mu(w), e, \hat{k}, \gamma) + \mathbb{P} \psi(W; \hat{\mu}(w), \hat{\gamma}) + o_P(n^{-1/2}), \]
\[ = S_A/n + \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i(1) - \mu_1(X_i)}{e(X_i)} - \frac{Y_i(0) - \mu_0(X_i)}{1 - e(X_i)} \right) 1\{ \hat{k}(X_i) \leq \hat{\gamma} \} \]
\[ + \mathbb{P}(\psi(W; \hat{\mu}(w), \hat{\gamma}) - \psi(W; \mu(w), e, \hat{k}, \gamma)) + o_P(n^{-1/2}), \]
where the third equality uses Assumption 3(a) and Lemma 2 of Kennedy et al. (2020) or Assumption 3(b), Lemma 19.24 of Van der Vaart (2000), and arguments similar to those in the proof of Theorem 1. Now, rearranging (53) gives
\[ \frac{\hat{S}_A - S_A}{n} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i(1) - \mu_1(X_i)}{e(X_i)} Z_i - \frac{Y_i(0) - \mu_0(X_i)}{1 - e(X_i)} (1 - Z_i) \right) \]
\[ + \mathbb{P}(\psi(W; \hat{\mu}(w), \hat{\gamma}) - \psi(W; \mu(w), e, \hat{k}, \gamma)) + o_P(n^{-1/2}), \]
\[ = \mathbb{P} \left[ \left( \hat{\mu}_1(X_i) + \frac{Y_i(1) - \hat{\mu}_1(X_i)}{\hat{e}(X_i)} \right) 1\{ \hat{k}(X_i) \leq \hat{\gamma} \} - \left( \mu_1(X_i) + \frac{Y_i(1) - \mu_1(X_i)}{e(X_i)} \right) 1\{ \bar{k}(X_i) \leq \gamma \} \right], \]
(58)
\[ = \mathbb{P} \left[ \left( \hat{\mu}_1(X_i) \hat{e}(X_i) + (Y_i(1) - \hat{\mu}_1(X_i)) e(X_i) \right) 1\{ \hat{k}(X_i) \leq \hat{\gamma} \} - \mu_1(X_i) 1\{ \bar{k}(X_i) \leq \gamma \} \right], \]
(59)
\[ = \mathbb{P} \left[ \hat{\mu}_1 \hat{e} 1\{ \hat{k} \leq \hat{\gamma} \} + \mu_1 e 1\{ \hat{k} \leq \hat{\gamma} \} - \hat{\mu}_1 e 1\{ \hat{k} \leq \hat{\gamma} \} - \mu_1 \hat{e} 1\{ \hat{k} \leq \hat{\gamma} \} \right]. \]
(60)
where we omit the covariates \(X_i\) in (60) for brevity.

Then, we can add and subtract \(\mu_1 e 1\{\hat{k} \leq \hat{\gamma}\}\) from the numerator of (60) and factor it as

\[
\hat{\mu}_1 \hat{e} 1\{\hat{k} \leq \hat{\gamma}\} + \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} - \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} + \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} - \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\},
\]

(61)

\[
= (\hat{\mu}_1 - \mu_1)(\hat{e} - e) 1\{\hat{k} \leq \hat{\gamma}\} + \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} - 1\{\hat{k} \leq \hat{\gamma}\}.
\]

(62)

Finally, substituting this back into (60), we find that (60) is

\[
= \mathbb{P} \left[ (\hat{\mu}_1 - \mu_1)(\hat{e} - e) 1\{\hat{k} \leq \hat{\gamma}\} + \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} - 1\{\hat{k} \leq \hat{\gamma}\} \right],
\]

(63)

\[
= \mathbb{P} \left[ \frac{(\hat{\mu}_1 - \mu_1)(\hat{e} - e)}{\hat{e}} 1\{\hat{k} \leq \hat{\gamma}\} \right] + \mathbb{P} \left[ \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} - 1\{\hat{k} \leq \hat{\gamma}\} \right],
\]

(64)

\[
= \mathbb{P} \left[ \mu_1 e 1\{\hat{k} \leq \hat{\gamma}\} - 1\{\hat{k} \leq \hat{\gamma}\} \right] + o_p(n^{-1/2}),
\]

(65)

where the third equality repeats the bounding of (48).

Now, (65) tracks the contribution of treatment terms to (57); the contribution of the control terms is analogous, so we conclude that (57) is, up to an \(o_p(n^{-1/2})\) error term,

\[
\frac{1}{n} \sum_{i=1}^{n} \left( \frac{Y_i(1) - \mu_1(X_i)}{e(X_i)} + \frac{Y_i(0) - \mu_0(X_i)}{1 - e(X_i)} \right) 1\{\hat{k}(X_i) \leq \hat{\gamma}\} + \mathbb{P} \left[ \tau(X_i) 1\{\hat{k}(X_i) \leq \hat{\gamma}\} - 1\{\hat{k}(X_i) \leq \hat{\gamma}\} \right],
\]

completing the proof.

For the denominators \(n_{\hat{\gamma}}\) and \(n_{\hat{\beta}}\), we recall (32), which shows that

\[
\frac{n_{\hat{\beta}} - n_{\hat{\gamma}}}{n} = \mathbb{P}_n 1\{\hat{k}(X_i) \leq \hat{\gamma}\} - \mathbb{P}_n 1\{\hat{k}(X_i) \leq \hat{\beta}\},
\]

(66)

\[
= \mathbb{P}(1\{\hat{k}(X_i) \leq \hat{\beta}\} - 1\{\hat{k}(X_i) \leq \hat{\beta}\}) + o_p(n^{-1/2}).
\]

(67)

Finally, combining Lemma 2 and (67) gives the theorem.

\textbf{Proof of Theorem 3}. A first-order Taylor expansion of the function \(f(x, y) = x/y\) shows that

\[
\frac{\hat{S}_{\hat{\beta}}}{n_{\hat{\beta}}} = \frac{\hat{S}_{\hat{\beta}}}{n_{\hat{\beta}}} + \frac{(\hat{S}_{\hat{\beta}} - S_{\hat{\beta}})/n}{n_{\hat{\beta}}/n} - \frac{S_{\hat{\beta}}/n}{(n_{\hat{\beta}}/n)^2} \cdot \frac{n_{\hat{\beta}} - n_{\hat{\beta}}}{n} + o_p \left( \left\| \frac{[\hat{S}_{\hat{\beta}}/n - S_{\hat{\beta}}/n]}{n} \right\|_2 \right),
\]

(68)

\[
= \frac{S_{\hat{\beta}}}{n_{\hat{\beta}}} + \frac{(\hat{S}_{\hat{\beta}} - S_{\hat{\beta}})/n}{n_{\hat{\beta}}/n} - \frac{S_{\hat{\beta}}/n}{(n_{\hat{\beta}}/n)^2} \cdot \frac{n_{\hat{\beta}} - n_{\hat{\beta}}}{n} + o_p(n^{-1/2}),
\]

(69)

because \([\hat{S}_{\hat{\beta}}/n - S_{\hat{\beta}}/n]/n_{\hat{\beta}}/n] = O_p(n^{-1/2})\) by assumption. Rearranging, we find

\[
\hat{\tau}_{\hat{\beta}} - \tau_{\hat{\beta}} = \frac{(\hat{S}_{\hat{\beta}} - S_{\hat{\beta}})/n}{n_{\hat{\beta}}/n} - \frac{S_{\hat{\beta}}/n}{(n_{\hat{\beta}}/n)^2} \cdot \frac{n_{\hat{\beta}} - n_{\hat{\beta}}}{n} + o_p(n^{-1/2}),
\]

(70)

\[
= \frac{1}{\mathbb{P}(k(X_i) \leq \hat{\gamma}) + o_p(1)} \cdot \frac{\hat{S}_{\hat{\beta}} - S_{\hat{\beta}}}{n} - \frac{\mathbb{E}[\tau(X_i) 1\{\hat{k}(X_i) \leq \hat{\gamma}\}] + o_p(1)}{(\mathbb{P}(k(X_i) \leq \hat{\gamma}) + o_p(1))^2} \cdot \frac{n_{\hat{\beta}} - n_{\hat{\beta}}}{n} + o_p(n^{-1/2}),
\]

(71)

\[
= \frac{1}{\mathbb{P}(k(X_i) \leq \hat{\gamma})} \cdot \frac{\hat{S}_{\hat{\beta}} - S_{\hat{\beta}}}{n} - \frac{\mathbb{E}[\tau(X_i) 1\{k(X_i) \leq \hat{\gamma}\}] + o_p(1)}{\mathbb{P}(k(X_i) \leq \hat{\gamma})} \cdot \frac{n_{\hat{\beta}} - n_{\hat{\beta}}}{n} + o_p(n^{-1/2}),
\]

(72)

\[
= \frac{1}{\mathbb{P}(k(X_i) \leq \hat{\gamma})} \cdot \frac{\hat{S}_{\hat{\beta}} - S_{\hat{\beta}}}{n} - \frac{\mathbb{E}[\tau(X_i) 1\{k(X_i) \leq \hat{\gamma}\}] + o_p(1)}{\mathbb{P}(k(X_i) \leq \hat{\gamma})} \cdot \frac{n_{\hat{\beta}} - n_{\hat{\beta}}}{n} + o_p(n^{-1/2}),
\]

(73)
where the second equality is the law of large numbers, the third follows from bounding the contribution of the $o_P(1)$ terms, and the fourth is the definition of conditional expectation. Then, substituting Lemma 2 and (67) into (73) and simplifying gives the result.