Sharp Sensitivity Analysis for Inverse Propensity Weighting via Quantile Balancing *

Jacob Dorn
Department of Economics
Princeton University

Kevin Guo
Department of Statistics
Stanford University

Abstract

Inverse propensity weighting (IPW) is a popular method for estimating treatment effects from observational data. However, its correctness relies on the untestable (and frequently implausible) assumption that all confounders have been measured. This paper introduces a robust sensitivity analysis for IPW that estimates the range of treatment effects compatible with a given amount of unobserved confounding. The estimated range converges to the narrowest possible interval (under the given assumptions) that must contain the true treatment effect. Our proposal is a refinement of the influential sensitivity analysis by Zhao, Small, and Bhattacharya (2019), which we show gives bounds that are too wide even asymptotically. This analysis is based on new partial identification results for Tan (2006)'s marginal sensitivity model.

Keywords: unobserved confounding, partial identification, quantile regression

1 Introduction

Estimating treatment effects from observational data is difficult because "treated" and "control" samples typically differ on many characteristics besides treatment status. For example, consumers of nutritional supplements may be wealthier or more health-conscious than those not taking supplements. One popular tool for adjusting for such baseline imbalances is Inverse Propensity Weighting (IPW) [4, 19]. This technique re-weights treated and untreated samples to be similar along all observed characteristics and then compares outcomes in the weighted samples. The crucial assumption underlying this approach is that the weighted samples do not systematically differ along important *unobserved* characteristics. This "unconfoundedness" assumption is untestable, and often implausible.

This paper studies how much can be learned when unconfoundedness does not hold, but one can bound the plausible degree of unobserved confounding. In particular, given a "sensitivity assumption" controlling the degree of selection, we aim to answer two questions:

(1) *Sensitivity analysis.* Can we bound how much the IPW point estimate from our "primary analysis" might change if unobserved confounding were properly accounted for?
(2) *Partial identification.* Can we characterize the most informative bounds that could possibly be obtained from the sensitivity assumption with even an infinite amount of observational data?

*This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2039656. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We are grateful to the associate editor and several anonymous referees for careful feedback. We are also grateful for comments from Guillaume Basse, Bo Honoré, Nathan Kallus, Michal Kolesár, David Lee, Xinran Li, Ulrich Müller, Karl Schulze, Dylan Small, Angela Zhou, Qingyuan Zhao, and seminar participants at Berkeley, Cambridge, and Princeton.

The specific sensitivity assumption used in this paper is the "marginal sensitivity model" of [56], which is a variant of Rosenbaum's famous "Γ sensitivity model" [48, 45, 46] that is better suited for IPW analyses. This sensitivity assumption is quite popular in causal inference; see [56, 27, 28, 29, 26, 60, 34, 49, 50, 54] for an incomplete list of references. As we will see, it lends itself to computationally-efficient sensitivity analyses which are simple enough to explain to any practitioner comfortable with IPW.

Recently, Zhao, Small, and Bhattacharya [60] (hereafter ZSB) introduced an interpretable IPW sensitivity analysis for the marginal sensitivity model that has been largely responsible for the recent resurgence of interest in this sensitivity assumption. However, they did not answer the partial identification question, leaving open the possibility that more informative bounds could be obtained from the same data and assumptions. Indeed, there are no existing partial identification results for the marginal sensitivity model that can be used to benchmark a sensitivity analysis.

The first main contribution of this paper is to provide a complete answer to the partial identification question (2). We derive closed-form expressions for the largest and smallest values of the "usual" estimands (e.g. average treatment effect) compatible with the marginal sensitivity assumption. These expressions show that the ZSB bounds are essentially always conservative because they ignore an infinite collection of constraints implied by the distribution of observed characteristics. [56] also identified these constraints, but deemed it intractable to incorporate them all in a sensitivity analysis. In contrast, our partial identification results show that this collection can actually be reduced to a *single* constraint which is easy to incorporate.

Our second main contribution is to introduce a new IPW sensitivity analysis, which we call the *quantile balancing* method. The method is a simple refinement of the ZSB sensitivity analysis, and has several desireable features:

(i) The quantile balancing sensitivity interval is always a subset of the ZSB interval. Outside of knife-edge cases, it is a strict subset.
(ii) When the outcome's conditional quantiles can be estimated consistently, the bounds converge to the sharp partial identification region for the average treatment effect (the best possible bounds that can be obtained under the marginal sensitivity model). With some abuse of terminology, we say that quantile balancing is "sharp."
(iii) Under standard assumptions for IPW inference, the bounds can be converted into confidence intervals using the same percentile bootstrap scheme proposed by ZSB.
(iv) When the estimated quantiles are inconsistent, the sensitivity interval is too wide rather than too narrow and the confidence intervals over-cover rather than under-cover. In other words, our intervals are guaranteed to be valid, regardless of the quality of the additional input we demand.

We apply the quantile balancing method in several simulated examples and one real-data application, and find that it can substantially tighten the ZSB bounds when the covariates are good predictors of the outcome. We also extend our analysis to Augmented IPW (AIPW) estimators. That analysis shows that a slight refinement of the ZSB method is sharp under "additive-noise" data generating processes, though the refinement makes little difference in practice. One shortcoming we will mention up-front is that our statistical guarantees assume the outcome is continuously-distributed in order to enable quantile regression. Since our partial identification results also apply to discrete outcomes, we conjecture that the quantile balancing procedure could be modified to give sharp bounds in that setting too.

1.1 Setting and background

We consider the Neyman-Rubin potential outcomes model with a binary treatment [42, 51]. We observe i.i.d. samples (X_i, Y_i, Z_i) from a distribution P, where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ is a vector of covariates, $Z_i \in \{0, 1\}$ is a binary treatment assignment indicator, and $Y_i \in \mathbb{R}$ is a real-valued outcome.

We assume that each sample (X_i, Y_i, Z_i) is obtained by coarsening a "full data" sample $(X_i, Y_i(0), Y_i(1), Z_i, U_i)$. Here, $Y_i(0)$ and $Y_i(1)$ are potential outcomes and U_i is a vector of unobserved confounders of unspecified dimension. The observed outcome is related to the potential outcomes through the consistency relation $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$.

The goal is to use the observed data to draw inferences about a causal estimand ψ_0. For the purposes of exposition, we initially focus on the counterfactual means $\psi_{\mathrm{T}} = \mathbb{E}[Y(1)]$ and $\psi_{\mathrm{C}} = \mathbb{E}[Y(0)]$, although the examples of most practical interest are the average treatment effect (ATE) and the average treatment effect on the treated (ATT).

$$\psi_{\mathrm{ATE}} = \mathbb{E}[Y(1) - Y(0)]$$
$$\psi_{\mathrm{ATT}} = \mathbb{E}[Y(1) - Y(0)|Z = 1].$$

With minor modification, our identification results can also be applied to more complex estimands, including policy values [3, 27] and weighted average treatment effects. However, we do not present those extensions in this paper.

Under the unconfoundedness assumption $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid X$, all of the above quantities can be consistently estimated from the observed data using inverse propensity weighting. IPW estimators work by reweighting the observed sample by some function of the propensity score $e(x) := P(Z = 1|X = x)$. For example, if the estimand of interest is ψ_{T}, the (stabilized) IPW estimator is given by (1):

$$\hat{\psi}_{\mathrm{T}} = \frac{\mathbb{E}_n[YZ/\hat{e}(X)]}{\mathbb{E}_n[Z/\hat{e}(X)]} \tag{1}$$

Here, $\hat{e}(\cdot)$ is an estimate of the propensity score $e(\cdot)$ and $\mathbb{E}_n[\cdot]$ is shorthand for $\frac{1}{n}\sum_{i=1}^n [\cdot]_i$. An unstabilized version of $\hat{\psi}_{\mathrm{T}}$ which uses only the numerator of (1) is also common. Related estimators for the other estimands considered will be denoted by $\hat{\psi}_{\mathrm{C}}, \hat{\psi}_{\mathrm{ATE}}$, and $\hat{\psi}_{\mathrm{ATT}}$. See the articles by [4] or [19] for their exact formulas.

We will assume some conditions which are required for identification and estimation under unconfoundedness: overlap ($0 < e(X) < 1$ almost surely) and one outcome moment ($\mathbb{E}_P[|Y|] < \infty$). However, we will not assume unconfoundedness.

2 The marginal sensitivity model

The marginal sensitivity model introduced by [56] is a relaxation of unconfoundedness which has been applied in many causal inference problems. This one-parameter sensitivity assumption allows for the existence of unobserved confounders U, but limits the degree of selection bias that can be attributed to these confounders.

Assumption Λ. (Marginal sensitivity model)
There exists a vector of unmeasured confounders U that, if measured, would lead to unconfoundedness: $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid (X, U)$. However, within each stratum of the observed covariates, measuring U can only change the odds of treatment by at most a factor of Λ, i.e. if we set $e_0(x, u) := P(Z = 1|X = x, U = u)$, then (2) holds with probability one.

$$\Lambda^{-1} \leq \frac{e_0(X, U)/[1 - e_0(X, U)]}{e(X)/[1 - e(X)]} \leq \Lambda \tag{2}$$

The statement of the marginal sensitivity model presented in [56] and [60] uses the potential outcomes $(Y(0), Y(1))$ in place of the unobserved variable U. However, as pointed out by a referee, these assumptions are equivalent.

To avoid confusion between e_0 and e, we will follow [29] and refer to e_0 as the "true propensity score" and e as the "nominal propensity score."

Like Rosenbaum's famous "Γ sensitivity model", Assumption Λ controls the degree of unobserved confounding with a single parameter. When $\Lambda = 1$, measuring additional confounders cannot change the odds of treatment at all, i.e. treatment assignment is unconfounded. As Λ increases, stronger forms of confounding are allowed. For advice on how to choose this parameter, see [21]. For more on the relationship between this and Rosenbaum's model, see [60] Section 7.1. The marginal sensitivity assumption is "nonparametric"

in the sense that no assumptions are needed about how e_0 depends on u. Even the dimension of the vector U does not need to be specified.

To see how Assumption Λ can be used for sensitivity analysis, begin by considering how an oracle statistician who observed the confounders U_i might estimate ψ_{T}. One strategy would be to use the IPW estimator (3), which is consistent under weak assumptions.

$$\hat{\psi}^*_{\mathrm{T}} = \frac{\sum_{i=1}^n Y_i Z_i / e_0(X_i, U_i)}{\sum_{i=1}^n Z_i / e_0(X_i, U_i)}. \tag{3}$$

In reality, $\{U_i\}_{i \leq n}$ are not observed, but under Assumption Λ, it is possible to *bound* the true propensity scores $e_0(X_i, U_i)$. In particular, the vector $(e_0(X_1, U_1), \cdots, e_0(X_n, U_n))$ must belong to the ZSB constraint set $\mathcal{E}_n(\Lambda)$ defined in (4).

$$\mathcal{E}_n(\Lambda) = \left\{ \bar{e} \in \mathbb{R}^n \,:\, \Lambda^{-1} \leq \frac{\bar{e}_i/(1-\bar{e}_i)}{e(X_i)/[1-e(X_i)]} \leq \Lambda \right\} \tag{4}$$

ZSB proposed bounding the oracle statistician's IPW estimator (3) with the largest and smallest IPW estimates that can be obtained using putative propensities in $\mathcal{E}_n(\Lambda)$.

$$[\hat{\psi}^-_{\mathrm{T,ZSB}}, \hat{\psi}^+_{\mathrm{T,ZSB}}] = \left[\min_{\bar{e} \in \mathcal{E}_n(\Lambda)} \frac{\sum_{i=1}^n Y_i Z_i / \bar{e}_i}{\sum_{i=1}^n Z_i / \bar{e}_i}, \max_{\bar{e} \in \mathcal{E}_n(\Lambda)} \frac{\sum_{i=1}^n Y_i Z_i / \bar{e}_i}{\sum_{i=1}^n Z_i / \bar{e}_i} \right]. \tag{5}$$

Since the interval (5) contains the consistent estimator $\hat{\psi}^*_{\mathrm{T}}$, the distance between the true estimand ψ_{T} and the nearest point in the sensitivity interval tends to zero. ZSB show that this conclusion holds even if the nominal propensity score e is replaced by a suitably consistent estimate $\hat{e}$ in the definition of $\mathcal{E}_n(\Lambda)$, which is important for practical applications as e is typically not known in observational studies.

This simple idea is intuitive enough to explain to any practitioner who is comfortable with IPW and has been extended to estimands other than ψ_{T}. ZSB also consider ψ_{ATE} and ψ_{ATT}. Related work by [27, 28, 26, 34] takes the idea substantially further. [56] applied a similar idea to a different propensity-score-based estimator and [1, 39, 57] used similar approaches in survey sampling problems.

2.1 Sharpness and data-compatibility

The aforementioned works do not address the asymptotic optimality of the interval $[\hat{\psi}^-_{\mathrm{T,ZSB}}, \hat{\psi}^+_{\mathrm{T,ZSB}}]$. Does it converge to a limiting set containing all values of ψ_{T} compatible with Assumption Λ and no others? Sensitivity analyses with this asymptotic optimality property are called "sharp" in the partial identification literature.

Sharpness is important for interpreting the results of a sensitivity analysis. If the primary analysis finds a positive treatment effect but the bounds associated with a very small value of Λ include zero, one might be tempted to conclude that the primary analysis is sensitive to unobserved confounding. However, unless the bounds are known to be sharp, this inference is not warranted even in large samples. Perhaps the bounds were just too conservative.

Despite its attractive features, the ZSB sensitivity analysis is not sharp. It can be arbitrarily conservative. To illustrate this, consider a simple joint distribution of observables:

$$\begin{aligned} X &\sim \mathcal{N}(0, \sigma^2) \\ Z \mid X &\sim \text{Bernoulli}(\tfrac{1}{2}) \\ Y \mid X, Z &\sim \mathcal{N}(X, 1). \end{aligned} \tag{6}$$

Suppose that a data analyst receives i.i.d. samples (X_i, Y_i, Z_i) from this distribution and is willing to posit that Assumption Λ is satisfied with $\Lambda = 2$. Let $\phi(\cdot)$ and z_τ denote the density and τ-th quantile of the standard normal distribution, respectively. The following result, which follows from Theorem 2 in Section 3.1, writes the set of values of ψ_{T} compatible with Assumption Λ explicitly in terms of these quantities and shows that this "partially identified" set is smaller than the limiting ZSB interval.

Corollary 1. (ZSB is asymptotically conservative)
Let (X_i, Y_i, Z_i) be i.i.d. samples from the joint distribution (6).

(i) The set of values of ψ_{T} compatible with the bound $\Lambda = 2$ and the distribution (6) is the interval $[\pm\frac{3}{4}\phi(z_{2/3})] \approx [\pm 0.27]$.

(ii) However, with probability one, $[\pm 0.27\sqrt{\sigma^2+1}] \subseteq [\hat{\psi}^-_{\mathrm{T,ZSB}}, \hat{\psi}^+_{\mathrm{T,ZSB}}]$ for all large n.

The precise meaning of (i) is the following: for any $\psi_{\mathrm{T}} \in [\pm\frac{3}{4}\phi(z_{2/3})]$, it is possible to construct a distribution Q for the full data $(X, Y(0), Y(1), Z, U)$ which marginalizes to (6), satisfies Assumption Λ with $\Lambda = 2$, and has $\mathbb{E}_Q[Y(1)] = \psi_{\mathrm{T}}$. On the other hand, for any ψ_{T} not in this interval, it is impossible to construct such a distribution.

Corollary 1 implies that the ZSB interval typically includes many values of ψ which cannot possibly be reconciled with the data. The explanation for this conservatism is that the odds-ratio bound (2) does not capture all of the restrictions on the true propensity score e_0. Additional information can be found in the marginal distribution of the *observed* characteristics. For example, in the context of Corollary 1, consider the putative propensity score (7).

$$\bar{e}(x,u) = \begin{cases} 1/3 & \text{if } x < 0 \\ 2/3 & \text{if } x \geq 0 \end{cases} \tag{7}$$

This certainly satisfies the odds-ratio bound (2) — and is therefore a possible value of $\bar{e}$ in the ZSB optimization problem (5) — but it could not possibly be the true propensity score e_0. If it were, we would observe $P(Z = 1|X \geq 0) = \frac{2}{3}$, while the observed data distribution P demands that $P(Z = 1|X \geq 0) = \frac{1}{2}$. Another way of saying this is that $\bar{e}$ does not *marginalize* to the nominal propensity score:

$$\begin{aligned} 1/2 &= P(Z = 1|X = x) \\ &= \int P(Z = 1|X = x, U = u)\, \mathrm{d}P(u|X = x) \\ &\neq \int \bar{e}(x,u)\, \mathrm{d}P(u|X = x) \\ &= \begin{cases} 1/3 & \text{if } x < 0 \\ 2/3 & \text{if } x \geq 0 \end{cases}. \end{aligned}$$

In short, this choice of $\bar{e}$ is allowed in the domain of the ZSB optimization problem but is incompatible with the distribution of observed data.

This example suggests that it should be possible to improve upon the ZSB bounds by only optimizing over the subset of $\mathcal{E}_n(\Lambda)$ which is "data compatible." However, this is easier said than done, because the observed data distribution actually imposes an infinite number of constraints on putative propensity scores $\bar{e}$. For example, the true e_0 "balances" all integrable functions $h : \mathcal{X} \to \mathbb{R}$:

$$\begin{aligned} \mathbb{E}[h(X)Z/e_0(X,U)] &= \mathbb{E}[h(X)\mathbb{E}[Z|X,U]/e_0(X,U)] \\ &= \mathbb{E}[h(X)e_0(X,U)/e_0(X,U)] \\ &= \mathbb{E}[h(X)]. \end{aligned} \tag{8}$$

Every such h gives rise to a testable "balancing constraint" (9) which can be used to rule out incompatible values of $\bar{e}$.

$$\frac{\mathbb{E}_n[h(X)Z/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \approx \mathbb{E}[h(X)] \tag{9}$$

In other words, any sharp sensitivity analysis must contend with an infinite number of constraints, which is typically computationally intractable [6, 14]. Previous works have considered relaxing these constraints by balancing only a finite set of functions [56, 57], but the resulting bounds are generally not sharp.

While this paper proceeds under the "superpopulation" model of causal inference, the idea that observable quantities can constrain unobserved variables can also be applied in the "finite population" model. See [57] for an application of this idea to partial identification in survey sampling problems.

3 Partial identification results

In this section, we show that at the *population* level, it is possible to characterize the sharp bounds for $\psi_0 \in \{\psi_{\mathrm{T}}, \psi_{\mathrm{C}}, \psi_{\mathrm{ATT}}, \psi_{\mathrm{ATE}}\}$ without ignoring or relaxing any of the infinitely many balancing constraints on the true propensity score. We apply these partial identification results to finite-sample sensitivity analysis in Section 4.

To state these results formally, we need a few pieces of additional notation. Recall that Assumption Λ requires the true propensity score $e_0(X, U)$ to satisfy the following odds-ratio bound:

$$\Lambda^{-1} \leq \frac{e_0(X,U)/[1-e_0(X,U)]}{e(X)/[1-e(X)]} \leq \Lambda.$$

Therefore, it is natural to define $\mathcal{E}_\infty(\Lambda)$ to be the set of all random variables $\bar{E}$ which satisfy the same condition:

$$\mathcal{E}_\infty(\Lambda) := \left\{ \bar{E} \,:\, \Lambda^{-1} \leq \frac{\bar{E}/(1-\bar{E})}{e(X)/(1-e(X))} \leq \Lambda \text{ with probability one} \right\}. \tag{10}$$

This can be viewed as the "population" version of the ZSB constraint set $\mathcal{E}_n(\Lambda)$.

Additionally, we define the conditional distribution function $F(y|x,z)$ and quantile function $Q_t(x,z)$ by:

$$F(y|x,z) = P(Y \leq y \mid X = x, Z = z)$$
$$Q_t(x,z) = \inf\{q \in \mathbb{R} \,:\, F(q|x,z) \geq t\}.$$

Since these functions only refer to observed quantities, they are identified from the observed-data distribution.

3.1 Partial identification via quantile balancing

Our first partial identification result shows that to compute optimal bounds for ψ_{T}, the infinitely-many balancing constraints described in Section 2.1 can actually be reduced to a *single* constraint. In particular, it suffices to minimize/maximize the function $\bar{E} \mapsto \mathbb{E}[YZ/\bar{E}]$ over the set of putative propensity scores $\bar{E} \in \mathcal{E}_\infty(\Lambda)$ that "balance" a particular conditional quantile of Y.

Theorem 1. (Optimal bounds for ψ_{T})
For any $\Lambda \geq 1$, the set of values of ψ_{T} compatible with the observed data distribution and Assumption Λ is a closed interval $[\psi_{\mathrm{T}}^-, \psi_{\mathrm{T}}^+]$. Moreover, if we define $\tau = \frac{\Lambda}{\Lambda+1}$, then the interval endpoints solve (11) and (12).

$$\psi_{\mathrm{T}}^- = \min_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}] \quad \textit{subject to} \quad \mathbb{E}[Q_{1-\tau}(X,1)Z/\bar{E}] = \mathbb{E}[Q_{1-\tau}(X,1)] \tag{11}$$

$$\psi_{\mathrm{T}}^+ = \max_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}] \quad \textit{subject to} \quad \mathbb{E}[Q_{\tau}(X,1)Z/\bar{E}] = \mathbb{E}[Q_{\tau}(X,1)]. \tag{12}$$

We will highlight a few important takeaways from this theorem. First, if one adds additional balancing constraints of the form $\mathbb{E}[h(X)Z/\bar{E}] = \mathbb{E}[h(X)]$ in (11) and (12), the value of these problems will not change. Thus, for the purposes of computing population-level bounds, the quantile balancing constraints in Theorem 1 capture all the information in the observed data. Second, the fact that only a single conditional quantile appears in each of the sharp bounds for ψ_{T} reflects a special advantage of the marginal sensitivity model. For alternative sensitivity assumptions, sharp bounds often involve distinct quantiles $Q_{\tau(x)}$ for each covariate level [33, 35], complicating estimation by potentially requiring estimates of the entire conditional quantile process [36, 53]. Third, this result shows that the ZSB sensitivity analysis for IPW can only be sharp when the conditional quantiles of Y do not depend on X at all, and can therefore be refined outside pathological cases. AIPW-based variants of the ZSB sensitivity analysis will generally refine the IPW bounds since some of the variability in the quantiles of Y will be absorbed by the regression function. We discuss AIPW sensitivity analysis in Section 4.2.

We can extend the theorem to other estimands. To bound ψ_{C}, exchange the labels "treated" and "control" and apply Theorem 1. Sharp bounds on ψ_{C} can be translated into sharp bounds on ψ_{ATT} using the relation $\psi_{\mathrm{ATT}} = \frac{\mathbb{E}[Y]-\psi_{\mathrm{C}}}{P(Z=1)}$.

Corollary 2. (Optimal bounds for ψ_{C} and ψ_{ATT})
In the setting of Theorem 1, the partially identified set for ψ_{C} is the interval $[\psi_{\mathrm{C}}^-, \psi_{\mathrm{C}}^+]$, where the interval endpoints solve (13) and (14).

$$\psi_{\mathrm{C}}^- = \min_{\bar{E}\in\mathcal{E}_\infty(\Lambda)} \mathbb{E}[Y\tfrac{1-Z}{1-\bar{E}}] \quad \textit{subject to} \quad \mathbb{E}[Q_{1-\tau}(X,0)\tfrac{1-Z}{1-\bar{E}}] = \mathbb{E}[Q_{1-\tau}(X,0)] \tag{13}$$

$$\psi_{\mathrm{C}}^+ = \max_{\bar{E}\in\mathcal{E}_\infty(\Lambda)} \mathbb{E}[Y\tfrac{1-Z}{1-\bar{E}}] \quad \textit{subject to} \quad \mathbb{E}[Q_{\tau}(X,0)\tfrac{1-Z}{1-\bar{E}}] = \mathbb{E}[Q_{\tau}(X,0)] \tag{14}$$

The partially identified set for ψ_{ATT} is the interval $[\psi_{\mathrm{ATT}}^-, \psi_{\mathrm{ATT}}^+]$, where $\psi_{\mathrm{ATT}}^{\mp} = \frac{\mathbb{E}[Y]-\psi_{\mathrm{C}}^{\pm}}{P(Z=1)}$.

Sharp bounds for ψ_{ATE} can be obtained by subtracting sharp bounds for ψ_{T} and ψ_{C}. Equivalently, these bounds can be obtained by solving optimization problems with two quantile balancing constraints. Although this result is superficially similar to Theorem 1 and Corollary 2, its proof requires a novel construction, which we discuss in Section 3.3.

Theorem 2. (Optimal bounds for ψ_{ATE})
For any $\Lambda \geq 1$, the set of values of ψ_{ATE} compatible with the observed data distribution and Assumption Λ is a closed interval $[\psi_{\mathrm{ATE}}^-, \psi_{\mathrm{ATE}}^+]$ where $\psi_{\mathrm{ATE}}^- = \psi_{\mathrm{T}}^- - \psi_{\mathrm{C}}^+$ and $\psi_{\mathrm{ATE}}^+ = \psi_{\mathrm{T}}^+ - \psi_{\mathrm{C}}^-$.

In certain special cases, the partially identified set for ψ_{ATE} can be computed more explicitly. These explicit bounds are useful for gaining intuition about the main factors that make a causal estimate more or less robust to unobserved confounding. Corollary 3, which is a corollary of our later work, gives such bounds in the Gaussian outcome model (15).

$$\begin{aligned} X &\sim P_X \\ Z \mid X &\sim \text{Bernoulli}(e(X)) \\ Y \mid X, Z &\sim \mathcal{N}(\mu(X,Z), \sigma^2(X)). \end{aligned} \tag{15}$$

Corollary 3. (Simpler bounds for Gaussian data)
Suppose the observed-data distribution has the factorization (15), with $0 < e(X) < 1$ almost surely and $\mathbb{E}[|\mu(X,Z)|] < \infty$. Let $\psi_{\mathrm{ATE}} = \mathbb{E}[\mu(X,1) - \mu(X,0)]$ be the nominal ATE. Then the partially identified set for the ATE under Assumption Λ is:

$$[\psi_{\mathrm{ATE}}^-, \psi_{\mathrm{ATE}}^+] = [\psi_{\mathrm{ATE}} \pm \tfrac{\Lambda^2-1}{\Lambda}\phi(\Phi^{-1}(\tfrac{\Lambda}{\Lambda+1}))\mathbb{E}[\sigma(X)]]. \tag{16}$$

Here, ϕ and Φ are the standard normal density and distribution function, respectively.

For a fixed bound Λ on the degree of unobserved confounding, the formula (16) shows that two key features map the observed data distribution to robustness. The first is the magnitude of the nominal ATE: all else equal, larger nominal effects are more robust. The second is the average noise level $\mathbb{E}[\sigma(X)]$: the better the measured variables predict the outcome, the less unobserved confounding can affect our estimates. In the extreme case where X and Z perfectly predict Y, then the ATE remains point-identified no matter how large Λ is, as long as overlap holds. These insights are not specific to the marginal sensitivity model. In alternative sensitivity models, they have also been observed by [47, 22], [11], and others. [47, 22, 11], and others.

3.2 Data-compatible propensity scores

Although the qualitative implications of Corollary 3 are plausible, we nevertheless find the quantile balancing formulas of Section 3.1 to be counterintuitive. After all, it is certainly not true that every random variable

$\bar{E} \in \mathcal{E}_\infty(\Lambda)$ satisfying $\mathbb{E}[Q_\tau(X,1)Z/\bar{E}] = \mathbb{E}[Q_\tau(X,1)]$ could plausibly be the true propensity score $e_0(X,U)$. Indeed, the constraints of the quantile-balancing optimization problems do not even enforce that $\mathbb{E}[Z/\bar{E}] = 1$. Our intuition for why the ZSB procedure is conservative suggests the quantile balancing formulas should be conservative as well.

To explain how these results are possible, we begin by characterizing which random variables $\bar{E}$ could plausibly be the true propensity score $e_0(X,U)$. The calculation (8) indicates that $\bar{E}$ should at least satisfy $\mathbb{E}[h(X)Z/\bar{E}] = \mathbb{E}[h(X)]$ for all integrable h, or equivalently, $\mathbb{E}[Z/\bar{E}|X] = 1$. Proposition 1 shows that for the purposes of bounding ψ_{T}, this is actually the *only* constraint on $\bar{E}$ implied by the distribution of observables. Similar results appear in [8, 43, 56, 17, 20, 15, 60].

Proposition 1. (Characterizing data-compatible propensity scores)
For any random variable $\bar{E} \in \mathcal{E}_\infty(\Lambda)$ satisfying $\mathbb{E}[Z/\bar{E}|X] = 1$, there is a distribution Q for $(X, Y(0), Y(1), Z, U)$ with the following properties:

(i) The distribution of the observables (X, Y, Z) is the same under P and Q.
(ii) Q satisfies Assumption Λ.
(iii) $\mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\bar{E}]$.

In short, this result says that $\mathbb{E}[YZ/\bar{E}]$ is a plausible value of ψ_{T} as long as $\mathbb{E}[Z/\bar{E}|X] = 1$. It is not hard to show that the converse also holds: if ψ is a plausible value of ψ_{T}, then $\psi = \mathbb{E}[YZ/\bar{E}]$ for some random variable $\bar{E}$ satisfying $\mathbb{E}[Z/\bar{E}|X] = 1$. As a result, the optimal bounds for ψ_{T} can be obtained by solving the variational problems in Corollary 4.

Corollary 4. *The partially identified set for ψ_{T} is an interval whose endpoints solve:*

$$\psi_{\mathrm{T}}^- = \min_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}] \quad \textit{subject to} \quad \mathbb{E}[Z/\bar{E}|X] = 1 \tag{17}$$

$$\psi_{\mathrm{T}}^+ = \max_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}] \quad \textit{subject to} \quad \mathbb{E}[Z/\bar{E}|X] = 1 \tag{18}$$

Even though the variational problems (17) and (18) can be infinite-dimensional optimization problems with infinitely-many constraints, they have several nice features that enable them to be solved explicitly. Some straightforward algebraic manipulation shows that the problem (18) can be written as:

$$\begin{aligned} \text{maximize} \quad & \mathbb{E}[\mathbb{E}[YZ/\bar{E}|X]] \\ \text{subject to} \quad & \mathbb{E}[Z/\bar{E}|X] = 1 \\ \text{and} \quad & 1 + \tfrac{1-e(X)}{e(X)}\Lambda^{-1} \leq 1/\bar{E} \leq 1 + \tfrac{1-e(X)}{e(X)}\Lambda. \end{aligned} \tag{19}$$

Not only is this problem *linear* in the decision "variable" $1/\bar{E}$, it also separates across levels of X. Therefore, it suffices to separately solve (20) for each $x \in \mathcal{X}$.

$$\begin{aligned} \text{maximize} \quad & \mathbb{E}[YZ/\bar{E}|X = x] \\ \text{subject to} \quad & \mathbb{E}[Z/\bar{E}|X = x] = 1 \\ \text{and} \quad & 1 + \tfrac{1-e(x)}{e(x)}\Lambda^{-1} \leq 1/\bar{E} \leq 1 + \tfrac{1-e(x)}{e(x)}\Lambda \end{aligned} \tag{20}$$

The problem (20) requires us to maximize one expectation subject to an equality constraint on another expectation. This resembles the problem solved by the Neyman-Pearson lemma, and in fact is a special case of the generalization due to [12]. The optimization problems posed in Theorem 1 also fall in this class. It turns out that both of these problems have a common solution, given in Proposition 2.

Proposition 2. (Formulas for the worst-case propensity scores)
There exist $\bar{E}_-$, $\bar{E}_+ \in \mathcal{E}_\infty(\Lambda)$ satisfying $\mathbb{E}[Z/\bar{E}_-|X] = \mathbb{E}[Z/\bar{E}_+|X] = 1$ and also (21) and (22).

$$1/\bar{E}_- = \begin{cases} 1 + \tfrac{1-e(X)}{e(X)}\Lambda^{+1} & \textit{if } Y < Q_{1-\tau}(X,1) \\ 1 + \tfrac{1-e(X)}{e(X)}\Lambda^{-1} & \textit{if } Y > Q_{1-\tau}(X,1) \end{cases} \tag{21}$$

$$1/\bar{E}_+ = \begin{cases} 1 + \frac{1-e(X)}{e(X)}\Lambda^{+1} & \text{if } Y > Q_\tau(X,1) \\ 1 + \frac{1-e(X)}{e(X)}\Lambda^{-1} & \text{if } Y < Q_\tau(X,1) \end{cases} \tag{22}$$

Further, $\bar{E}_-$ solves both (11) and (17), and $\bar{E}_+$ solves both (12) and (18).

The form of the propensity score $\bar{E}_+$ gives us insight into the confounding structure which maximizes ψ_{T}: in the worst case, all observations with "high" values of Y are unlikely to be treated and thus receive large propensity weight, while all observations with "low" values of Y are likely to be treated and thus receive small propensity weight. The cutoff between high and low is chosen to satisfy the data-compatibility condition $\mathbb{E}[Z/\bar{E}_+|X] = 1$.

This argument presented in this section extends immediately to ψ_{C} by swapping treatment and control labels, extends to ψ_{ATT} by the argument given in Section 3.1, and can extend to other sensitivity models of the form $e_{\min}(X) \leq e_0(X,U) \leq e_{\max}(X)$ by modifying the constraints of (20).

3.3 Data compatibility for the ATE

To extend the argument from Section 3.2 to the ATE requires additional care. Although $\psi^+_{\mathrm{ATE}} = \psi^+_{\mathrm{T}} - \psi^-_{\mathrm{C}}$ is certainly a *valid* upper bound for the partially identified set for ψ_{ATE}, it is not obviously a sharp one. Proposition 1 only implies that there exists a distribution Q matching the observed-data distribution which has $\mathbb{E}_Q[Y(1)] = \psi^+_{\mathrm{T}}$ and another distribution Q' which has $\mathbb{E}_{Q'}[Y(0)] = \psi^-_{\mathrm{C}}$, but these distributions need not be the same. In other words, the two bounds may not be simultaneously achievable.

Theorem 2 indicates that the worst-case bounds on the counterfactual means are simultaneously achievable in the marginal sensitivity model. This is a surprising result, given that simultaneous achievability is *not* expected to hold in the closely-related Rosenbaum sensitivity model. In that model, [58] derived sharp bounds on ψ_{T} and ψ_{C} but required an extra symmetry assumption on the distribution of potential outcomes to establish sharpness of the resulting ATE bounds.

The key to our bounds on ψ_{ATE} is the following claim, which strengthens Proposition 1.

Proposition 3. (Simultaneous achievability)
For any random variable $\bar{E} \in \mathcal{E}_\infty(\Lambda)$ satisfying $\mathbb{E}[Z/\bar{E}|X] = \mathbb{E}[(1-Z)/(1-\bar{E})|X] = 1$, there is a distribution Q for the full data $(X, Y(0), Y(1), Z, U)$ with the following properties:
(i) The distribution of the observables (X, Y, Z) is the same under P and Q.
(ii) Q satisfies Assumption Λ.
(iii) $\mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\bar{E}]$ and $\mathbb{E}_Q[Y(0)] = \mathbb{E}_P[Y(1-Z)/(1-\bar{E})]$.

Unlike Proposition 1, this result does not follow from the existing data-compatibility characterizations of [8, 43, 56, 60] and instead requires an original construction. Given this result, one can derive Theorem 2 as a consequence of Theorem 1 and Corollary 2.

4 Sensitivity analysis

In this section, we give our proposals for translating the population-level partial identification results of Section 3 into practical sensitivity analyses. Our main proposal, which we call the *quantile balancing* method, conducts a sensitivity analysis for IPW estimators by modifying the ZSB proposal to incorporate the sufficient constraints derived in Section 3.1. We also discuss extensions of our sensitivity analysis to the AIPW estimator of [44] which are simpler to implement but only sharp under homoscedasticity.

Throughout this section, we take $\Lambda \geq 1$ to be fixed and set $\tau = \Lambda/(\Lambda+1)$.

4.1 Sensitivity analysis via quantile balancing

We begin by describing our IPW sensitivity analysis for the average treated potential outcome. Theorem 1 implies that the largest value of ψ_{T} compatible with Assumption Λ solves the optimization problem (23):

$$\psi_{\mathrm{T}}^{+} = \max_{\bar{E} \in \mathcal{E}_{\infty}(\Lambda)} \frac{\mathbb{E}[YZ/\bar{E}]}{\mathbb{E}[Z/\bar{E}]} \quad \text{s.t.} \quad \begin{pmatrix} \mathbb{E}[Q_\tau(X,1)Z/\bar{E}] \\ \mathbb{E}[Z/\bar{E}] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[Q_\tau(X,1)Z/e(X)] \\ \mathbb{E}[Z/e(X)] \end{pmatrix}. \tag{23}$$

In the above display, we have included an additional constraint $\mathbb{E}[Z/\bar{E}] = \mathbb{E}[Z/e(X)]$ which motivates our finite-sample procedure without affecting the optimization problem value.

Our proposal is to estimate ψ_{T}^{+} by replacing all of the unknown quantities in (23) with empirical counterparts. We estimate ψ_{T}^{-} by following the same principle. To translate these estimates into confidence intervals, we employ the same simple percentile bootstrap scheme as ZSB.

We will be concrete about what optimization problem we are proposing to solve. Let $\hat{Q}_\tau(x,z)$ be an estimate of the conditional quantile function of Y obtained by some kind of quantile regression (e.g. [2, 31, 37, 55]). Let $\hat{e}$ be the data analyst's estimate of the nominal propensity score e from their primary analysis. We define $\hat{\psi}_{\mathrm{T}}^{+}$ as the solution to the empirical maximization problem (24).

$$\hat{\psi}_{\mathrm{T}}^{+} = \max_{\bar{e} \in \mathcal{E}_n(\Lambda)} \frac{\mathbb{E}_n[YZ/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \quad \text{s.t.} \quad \begin{pmatrix} \mathbb{E}_n[\hat{Q}_\tau(X,1)Z/\bar{e}] \\ \mathbb{E}_n[Z/\bar{e}] \end{pmatrix} = \begin{pmatrix} \mathbb{E}_n[\hat{Q}_\tau(X,1)Z/\hat{e}(X)] \\ \mathbb{E}_n[Z/\hat{e}(X)] \end{pmatrix} \tag{24}$$

The lower bound $\hat{\psi}_{\mathrm{T}}^{-}$ is defined similarly, but with maximization replaced by minimization and $\hat{Q}_\tau(x,z)$ replaced by another quantile estimate $\hat{Q}_{1-\tau}(x,z)$. We call $\hat{\psi}_{\mathrm{T}}^{+}$ and $\hat{\psi}_{\mathrm{T}}^{-}$ the *quantile balancing bounds* for ψ_{T}.

Two features of this proposal require some explanation. The first feature to explain is the inclusion of the constraint $\mathbb{E}_n[Z/\bar{e}] = \mathbb{E}_n[Z/\hat{e}(X)]$ in (24). At the population level, Theorem 1 shows that only the constraint $\mathbb{E}[Q_\tau(X,1)Z/\bar{E}] = \mathbb{E}[Q_\tau(X,1)Z/e(X)]$ is relevant. However, in finite samples, this additional constraint improves robustness when $\hat{Q}_\tau$ is an inaccurate estimate of Q_τ and also simplifies the associated computation. The second feature to explain is why the right-hand side of the constraints in (24) have an "IPW" form (i.e. $\mathbb{E}_n[\hat{Q}_\tau(X,1)Z/\hat{e}(X)]$) rather than a "sample average" form (i.e. $\mathbb{E}_n[\hat{Q}_\tau(X,1)]$). If $\mathbb{E}_n[\hat{Q}_\tau(X,1)Z/\hat{e}(X)] \neq \mathbb{E}_n[\hat{Q}_\tau(X,1)]$, then a sample average version of (24) may have no feasible propensities. With the IPW form, $\bar{e}_i = \hat{e}(X_i)$ is always feasible.

Now that we have explained our proposed sensitivity analysis, we will collect several immediate properties of the quantile balancing bounds:

(i) When $\Lambda = 1$ (i.e. no confounding is allowed), the quantile balancing bounds collapse to the usual IPW estimate of ψ_{T} under unconfoundedness.
(ii) The quantile balancing bounds are sample bounded, i.e. $\min_i Y_i \leq \hat{\psi}_{\mathrm{T}}^{-} \leq \hat{\psi}_{\mathrm{T}}^{+} \leq \max_i Y_i$.
(iii) The quantile balancing bounds are always a subset of the ZSB bounds and, outside of knife-edge cases, are a strict subset.
(iv) The optimization problem (24) is convex and can be solved efficiently. In fact, it reduces to a standard quantile regression problem. See Appendix Afor implementation details.

The property (i) leads us to call quantile balancing a "sensitivity analysis for IPW." One can also apply quantile balancing to unstabilized IPW estimators at the cost of properties (ii) and (iii). See Appendix B for computational details, including for Augmented IPW estimators.

The quantile balancing idea extends easily to other causal estimands. To compute bounds for ψ_{C}, one only needs to exchange the definitions of "treated" and "control" and solve the same optimization problem. Subtracting the bounds for ψ_{T} and ψ_{C} gives bounds for ψ_{ATE}, and bounds for ψ_{ATT} follow from a similar principle (see Appendix A for the exact formula).

To form confidence intervals based on quantile balancing, we follow [60] and propose using the percentile bootstrap. If $[\hat{\psi}_b^{-}, \hat{\psi}_b^{+}]$ are quantile balancing bounds estimated in the b^{th} of B bootstrap samples, we report the quantile balancing $1-\alpha$ confidence interval as:

$$\mathrm{CI}(\alpha) = [Q_{\alpha/2}(\{\hat{\psi}_b^{-}\}_{b\in[B]}), Q_{1-\alpha/2}(\hat{\psi}_b^{+}\}_{b\in[B]})]. \tag{25}$$

As is standard for bootstrap-based IPW inference, we require re-estimating the nominal propensity score separately in each bootstrap replication. That requirement does not extend to the conditional quantiles. While the conditional quantiles can be re-estimated within bootstraps, our inference results will also apply if they are taken from the main dataset. This helps keep inference computationally tractable.

4.2 Implications for AIPW sensitivity analysis

The quantile balancing sensitivity analysis described above requires the data analyst to perform several quantile regressions. Our partial identification results imply that, in certain "additive-noise" data generating processes, a data analyst whose primary analysis was conducted using the AIPW estimator can perform sharp sensitivity analysis without performing any quantile regressions.

To explain how, we begin by describing the modeling assumption. Suppose the observed outcome Y has the following signal-plus-noise representation:

$$Y = \mu(X, Z) + \epsilon \quad \text{with} \quad \mathbb{E}[\epsilon] = 0, \epsilon \perp\!\!\!\perp (X, Z). \tag{26}$$

Such models frequently arise in the regression applications [see, e.g. 18, Chapter 3] and fit quite well in the real-data example we present in Section 5.2 below.

The additive-noise assumption (26) implies that the conditional quantiles of the residuals ϵ are constant. In particular, the assumption implies $Q_\tau(x, z) = \mu(x, z) + Q_\tau(\epsilon)$, where $Q_\tau(\epsilon)$ is the τ-th quantile of the noise. Therefore, Theorem 1 and some algebra imply that the sharp upper bound for ψ_{T} has the following formula:

$$\psi^+_{\mathrm{T}} = \max_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \left\{ \mathbb{E}[\mu(X, 1) + \frac{\mathbb{E}[(Y - \mu(X, 1))Z/\bar{E}]}{\mathbb{E}[Z/\bar{E}]} \right\} \quad \text{s.t.} \quad \mathbb{E}[Z/\bar{E}] = \mathbb{E}[Z/e(X)]. \tag{27}$$

Similar formulas can be derived for $\psi^-_{\mathrm{T}}, \psi^+_{\mathrm{C}}, \psi^-_{\mathrm{C}}$. This formula is convenient after an AIPW primary analysis, which requires estimates of all the nuisance parameters in this equation.

A natural estimate of ψ^+_{T} is the finite-sample analogue of (15).

$$\hat{\psi}^+_{\mathrm{T,AIPW}} = \max_{\bar{e} \in \mathcal{E}_n(\Lambda)} \left\{ \mathbb{E}_n[\hat{\mu}(X, 1)] + \frac{\mathbb{E}_n[(Y - \hat{\mu}(X, 1))Z/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \right\} \quad \text{s.t.} \quad \mathbb{E}_n[Z/\bar{e}] = \mathbb{E}_n[Z/\hat{e}(X)] \tag{28}$$

The estimated bound $\hat{\psi}^+_{\mathrm{T,AIPW}}$ grows with Λ and recovers the original (stabilized) AIPW estimator when $\Lambda = 1$. One can also not divide by $\mathbb{E}_n[Z/\bar{e}]$ in (28) to recover the unstabilized AIPW estimator at $\Lambda = 1$.

The estimator (28) slightly modifies the proposal in Section 6.2 of [60] to include the balancing constraint $\mathbb{E}_n[Z/\bar{e}] = \mathbb{E}_n[Z/\hat{e}(X)]$. In theory, this constraint is necessary to achieve sharpness in the additive-noise model (26). However, the simulations presented in Section 5 find that when the additive-noise model holds, this constraint scarcely refines the stabilized point estimates while somewhat degrading the coverage of bootstrap confidence intervals.

4.3 Theoretical properties

We now state some theoretical properties of the quantile balancing bounds $[\hat{\psi}^-, \hat{\psi}^+]$ which apply when the outcome Y has a continuous distribution. In short, the bounds are sharp when quantiles are estimated consistently and are valid even when quantiles are estimated inconsistently. Moreover, the percentile bootstrap yields valid confidence intervals if standard IPW inference conditions are satisfied and quantiles are estimated parametrically.

To obtain these results, we need a few conditions. The first condition collects some standard IPW consistency requirements which we expect the data analyst to have already assumed in their primary analysis.

Condition 1. (IPW assumptions)
The nominal propensity score e satisfies $\varepsilon \leq e(X) \leq 1 - \varepsilon$ with probability one for some $\varepsilon > 0$. The estimated nominal propensity score $\hat{e}(\cdot) \equiv \hat{e}(\cdot, \{X_i, Z_i\}_{i \leq n})$ is uniformly consistent, and the variance of Y is finite.

The second condition requires that the outcome Y has a bounded conditional density which is positive near the relevant conditional quantiles. This is a common identification condition for quantile regression [2, 5]. However, it means our theoretical guarantees do not apply when Y is discrete.

Condition 2. (Density)
The conditional distribution of $Y \mid X, Z$ has a uniformly bounded density $f(y|x,z)$. For each $(x,z) \in \mathcal{X} \times \{0,1\}$, the map $y \mapsto f(y|x,z)$ is continuous and positive near $Q_{1-\tau}(x,z)$ and $Q_\tau(x,z)$.

Finally, we make some assumptions about how the quantiles are estimated. For the standard linear quantile regression method of [31], one only needs to check that the regressors in the quantile regression have finite variance. We cover generic (possibly nonlinear) methods by requiring sample splitting to avoid overfitting. The specific form of sample splitting analyzed in our proofs is "cross-fitting" [52, 41, 10], but leave-one-out or out-of-bag quantile estimates perform similarly in simulations. Based on our practical experience, we recommend using some kind of sample splitting even when the quantile model is linear.

Condition 3. (Quantile estimates)
For each $t \in \{1-\tau, \tau\}$, one of the following holds for the estimated quantile function $\hat{Q}_t$:
(i) $\hat{Q}_t(x,z) = \hat{\beta}_t(z)^\top h(x)$ for some fixed "features" $h_j(X)$ with finite variance.
(ii) $\hat{Q}_t(x,z)$ is estimated using cross-fitting and satisfies Condition N in the supplementary materials.

Condition 3 is essentially "algorithmic," and neither (i) nor (ii) impose any accuracy requirements on the estimated conditional quantiles. The appendix conditions in (ii) are technical to state but very mild. Under Conditions 1 and 2, they are satisfied by quantile estimates based on nearest-neighbors [55], kernels [7], and random forests [2, 37].

Under these conditions, we have the following result on the asymptotic sharpness of the quantile balancing bounds.

Theorem 3. (Sharpness and robustness)
For any $\psi_0 \in \{\psi_\mathrm{T}, \psi_\mathrm{C}, \psi_\mathrm{ATT}, \psi_\mathrm{ATE}\}$, let $[\psi^-, \psi^+]$ be its partially identified interval under Assumption Λ and let $[\hat{\psi}^-, \hat{\psi}^+]$ be the corresponding quantile balancing interval. Assume Conditions 1, 2, and 3.
(i) If the quantile regression estimates are consistent, then $\hat{\psi}^- \xrightarrow{p} \psi^-$ and $\hat{\psi}^+ \xrightarrow{p} \psi^+$.
(ii) Even if the quantile models are misspecified, we still have $\hat{\psi}^- \leq \psi^- + a_n$ and $\psi^+ - b_n \leq \hat{\psi}^+$, where $a_n = o_P(1)$ and $b_n = o_P(1)$.

The same conclusions hold for the AIPW-based bounds introduced in Section 4.2 when the outcome regression is estimated by linear regression, i.e. sharpness under an additive-noise model and validity in general. However, while AIPW is doubly-robust under unconfoundedness, the validity of the corresponding AIPW quantile balancing bounds relies on correct specification of the nominal propensity score.

The result (ii) shows that even when quantiles are not estimated consistently, the quantile balancing bounds are still valid; we will offer some intuition on why this novel robustness property holds. At the population level, the worst-case propensity score $\bar{E}_+$ defined in Proposition 2 "balances" all integrable function of X, so intuitively, we should expect that it "nearly" balances the estimated quantile function $\hat{Q}_\tau(\cdot, 1)$ in finite samples even if $\hat{Q}_\tau(\cdot, 1)$ is not particularly close to $Q_\tau(\cdot, 1)$. That suggests a vector of propensities very close to the true worst-case propensity vector will belong to the feasible set $\mathcal{E}_n(\Lambda)$. Since the quantile balancing upper bound $\hat{\psi}^+_\mathrm{T}$ is defined as a maximum over the feasible set, it will be at least as large as a quantity close to ψ^+_T. This roughly explains why validity holds even under misspecification.

The validity of the confidence interval (25) follows under stronger parametric assumptions. We prove an inference result assuming the nominal propensity score is estimated by a correctly-specified parametric model and the conditional quantiles are estimated by a (potentially misspecified) parametric model.

Theorem 4. (Inference)
Let $[\psi^-, \psi^+]$ be as in Theorem 3, and let $\mathrm{CI}(\alpha)$ be as in (25). Suppose Conditions 1, 2, and 3.(i) are

satisfied, and also that the nominal propensity score is estimated by a regular parametric model (e.g. logistic regression). Then we have

$$\liminf_{n\to\infty} \mathbb{P}([\psi^-, \psi^+] \subseteq \mathrm{CI}(\alpha)) \geq 1 - \alpha \tag{29}$$

for any $\alpha \in (0, 1)$.

We have found that these confidence intervals can under-cover the identified set in finite samples when the quantiles are correctly specified. In our simulations, the use of cross-fit conditional quantile estimates largely resolves the issue with minimal effect on point estimates, so we advocate for the use of such estimators in practice. Although we do not have theoretical support for the confidence interval $\mathrm{CI}(\alpha)$ when quantiles are estimated by a nonlinear model, we find that approach performs reasonably well in the simulations of Section 5 as long as cross-fit quantiles are used.

5 Numerical examples

In this section, we illustrate the finite-sample performance of our proposed sensitivity analyses on several simulated datasets and one real-data example.

5.1 Simulated data

We consider two data-generating processes (DGPs) in our simulated examples. The two DGPs differ in the conditional distribution of Y given (X, Z), but otherwise can be described as follows:

$$\begin{aligned} X &\sim \mathrm{Uniform}([-1, 1]^5) \\ Z \mid X &\sim \mathrm{Bernoulli}\left(\tfrac{1}{1+\exp(-\sum_{j=1}^5 X_j/\sqrt{5})}\right) \\ Y \mid X, Z &\sim \mathcal{N}(\mu(X), \sigma^2(X)). \end{aligned} \tag{30}$$

In the first DGP, we use $\mu(x) = x_1 + \cdots + x_5$ and $\sigma(x) = 1$. In the second DGP, we use $\mu(x) = \frac{3}{2}\mathrm{sign}(x_1) + \mathrm{sign}(x_2)$ and $\sigma(x) = 2 + \mathrm{sign}(x_3) + \mathrm{sign}(x_4)$. The estimand of interest is the ATE and we fix $\Lambda = 2$, i.e. unobserved confounders can double or halve the odds of treatment.

We compare five methods for obtaining bounds on the partially identified set:

1. `QB-Linear` applies the quantile balancing method of Section 4 with quantiles estimated using linear quantile regression on $X_1, \ldots, X_5$.
2. `QB-Forest` applies quantile balancing with quantiles estimated using the random forest method from [2].
3. `ZSB` applies the main IPW method from [60], described in Section 2.1.
4. `ZSB-AIPW` applies the AIPW-based method from Section 6.2 of [60], described in Section 4.2. This requires an estimate of the outcome model $\mu(X, Z) = \mathbb{E}[Y|X, Z]$. We use a situationally-appropriate outcome model, linear regression in DGP1 and random forest regression in DGP2.
5. `AIPW+1` applies the AIPW-based method introduced in Section 4.2. We call this `AIPW+1` because it refines `ZSB-AIPW` to incorporate an additional "one-balancing" constraint $\mathbb{E}_n[Z/\bar{e}] = \mathbb{E}_n[Z/\hat{e}(X)]$.

All methods estimate the nominal propensity score by logistic regression. We use 5-fold cross-fitting in all of our quantile regressions. We do not re-estimate quantiles or random forest models within bootstraps.

Figure 1 shows the distribution of upper and lower bound point estimates from each of these five methods, estimated using 2,000 simulations with $n = 1,000$ observations each. Simulations at other sample sizes are presented in Appendix B. Dashed lines indicate the true partially identified region. The results conform to the asymptotic predictions of Section 4: (i) when the quantile models are "correctly specified," the quantile balancing point estimates are nearly unbiased; (ii) under misspecification, the range of `QB` point estimates is

too wide rather than too narrow; (iii) the `ZSB` range of point estimates is too wide in both cases; and (iv) AIPW-based methods give nearly-sharp bounds in the additive-noise DGP1 but conservative bounds in the heteroscedastic DGP2. We also find that the `+1` constraint in `AIPW+1`, which is necessary for sharpness in theory, has minimal practical impact in either DGP.

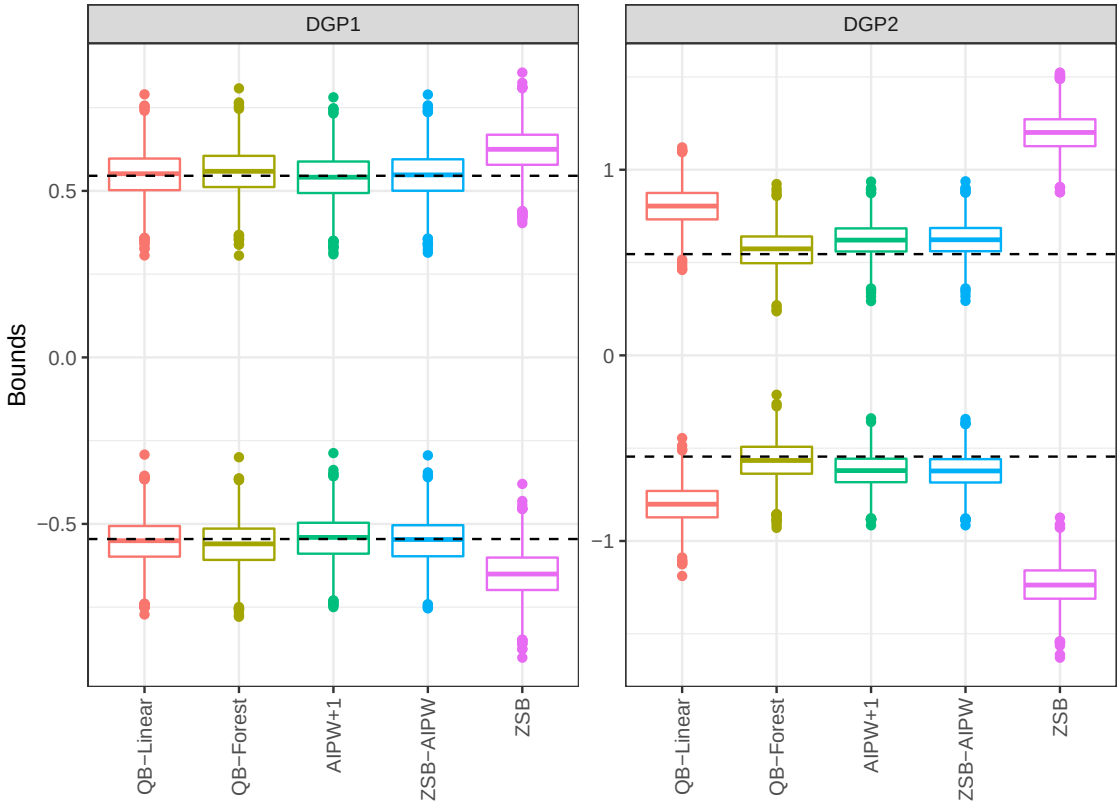

Figure 1: *Boxplots of the ATE upper and lower bound point estimates for both DGPs and all considered methods. The dashed line indicates the boundary of the true partially identified set. In DGP1, all methods but `ZSB` are correctly specified and give reasonably accurate bounds. In DGP2, the `Forest` method is well-suited to the piecewise-constant conditional quantiles and gives the most accurate bounds.*

Figure 2 shows the coverage for 95% bootstrap confidence intervals based on each of the five methods. In DGP1, both quantile balancing methods have nearly nominal coverage, but AIPW-based methods undercover and the `+1` constraint exacerbates the undercoverage. In DGP2 the `QB-Forest` method achieves nearly nominal coverage, while all other methods overcover. The `ZSB` method overcovers for both DGPs.

5.2 Real data

In this section, we apply our proposed sensitivity analysis to a subsample of data from the 1966-1981 National Longitudinal Survey (NLS) of Older and Young Men. We wish to estimate the impact of union membership on wages. Specifically, we consider the ATE of union membership on log wages. For illustrative reasons, we focus on the 1978 cross-section of Young Men and restrict our attention to craftsmen and laborers not enrolled in school. Our estimates are thus based on a sample of 668 respondents with measurements of wages, union membership, and eight covariates.

For our primary analysis, we use IPW to adjust for baseline imbalances in covariates between union and

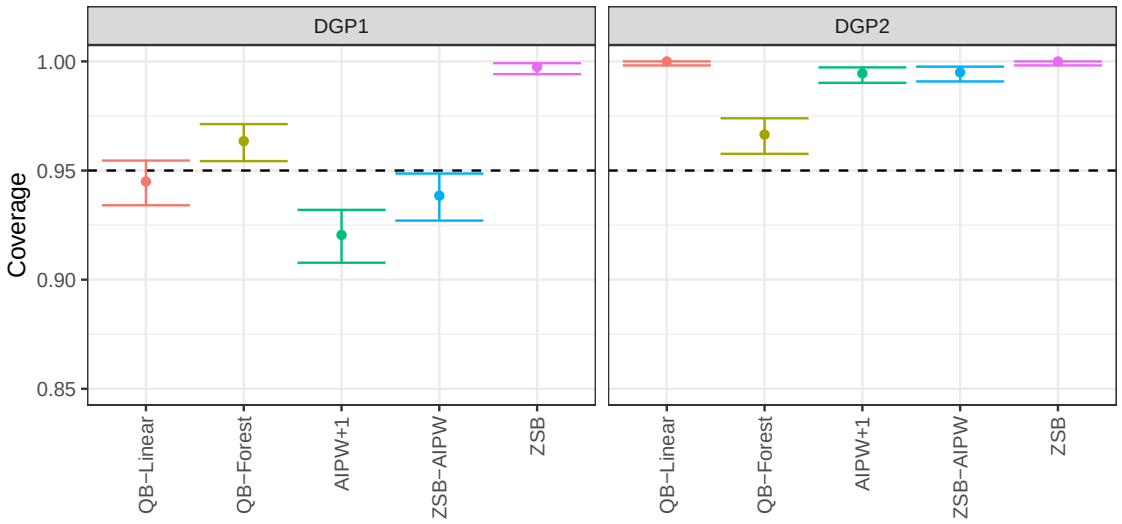

Figure 2: *The coverage of nominal level 95% bootstrap confidence intervals based on five different methods. Estimates are based on* $2,000$ *simulations each with* $n = 1,000$ *observations. The error bars are binomial confidence intervals for the true coverage probability.*

nonunion samples. Table 1 reports the covariate balance between union and nonunion samples before and after weighting by the (estimated) inverse propensity score. On several important characteristics, inverse propensity weighting dramatically improves balance across the two samples.

	Unweighted		Weighted	
Covariate	Union	Nonunion	Union	Nonunion
Age	30.1	30.0	30.0	30.0
Black	24%	24%	23%	24%
Metropolitan	74%	57%	66%	65%
Southern	32%	53%	42%	42%
Married	78%	75%	76%	76%
Manufacturing	42%	32%	37%	38%
Laborer	23%	15%	18%	18%
Education	12.2	11.7	12.1	12.0

Table 1: *Covariate means among the nonunion and union subsamples, along with the means in the weighted samples. In red, we highlight particularly large imbalances. In the weighted samples, propensity weights are estimated using logistic regression.*

The IPW point estimate of the ATE is 0.23 with an associated 90% confidence interval of $[0.18, 0.27]$. Thus, our primary analysis concludes that union membership has a positive effect on wages, at least on average among craftsmen and laborers. Both the point estimate and the confidence interval are in agreement with prior literature studying the same problem using cross-sectional data. See [24, 25] for overviews. An AIPW-based primary analysis gives the same point estimate and confidence interval, up to rounding.

[16], [38], and many other economists have argued that cross-sectional estimates of the union premium overestimate the true causal effect because higher-skill workers are simultaneously more likely to be selected for union jobs and earn higher wages. Here, "skill" refers to an unobserved confounder which is only partially captured by the measured covariates. Is it plausible that the positive effect we find in the IPW analysis could be entirely due to selection on skill? A sensitivity analysis may help address this question.

Figure 3 reports point estimate ranges and 90% bootstrap confidence intervals from quantile balancing,

the ZSB-IPW method, and the ZSB-AIPW method for several values of the sensitivity parameter Λ. For quantile balancing, we estimate conditional quantiles using linear quantile regression with five-fold cross fitting. For AIPW, we use linear regression for the outcome model.

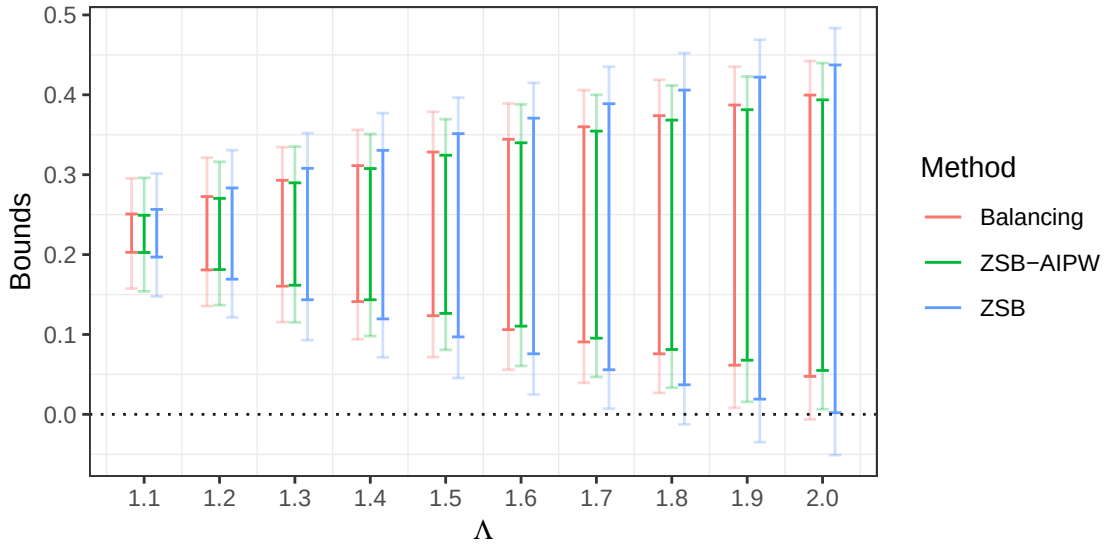

Figure 3: *Point estimate ranges and 90% bootstrap confidence intervals for the ATE in the NLS dataset. For the quantile balancing method, conditional quantiles are estimated using the linear quantile regression method of [31], with five-fold cross-fitting.*

All three sensitivity analyses show that the positive effect found in the primary analysis is fairly robust to unobserved confounding, but quantile balancing and ZSB-AIPW refine the baseline ZSB-IPW interval. Even if the odds of union membership for "skilled" workers were nearly double ($\Lambda = 1.9$) the odds for "typical" workers with the same observed covariates, the quantile balancing and AIPW sensitivity analyses analysis would still find a statistically significant positive treatment effect. Meanwhile, when $\Lambda = 1.8$, the ZSB confidence intervals already include the null. In this application, quantile balancing only slightly refines the ZSB range. Moreover, quantile balancing and ZSB-AIPW yield very similar ranges and confidence intervals. This is to be expected from the discussion in Section 4.2, as an "additive noise" model appears quite plausible in this application.

To put these sensitivities in context, we follow [29] and compute the degree to which the (estimated) odds of union membership could change if *measured* confounders were omitted from the dataset. Caveats to this approach and more sophisticated empirical calibration strategies are discussed in [21, 59, 11]. No measured confounders except `Laborer` and `South` were able to nearly double or halve the odds of union membership for any respondent. We interpret these results as showing that the qualitative conclusions of the primary analysis are fairly robust to unobserved confounding by skill.

Incidentally, longitudinal estimates of union wage effects — which control for individual-specific effects like "skill" — come to similar conclusions as the one suggested by our sensitivity analysis. Although treatment effect estimates from longitudinal studies are generally smaller than those from cross-sectional studies, they still find evidence in favor of the "union premium" [9, 24, 16].

6 Conclusion

We have shown that quantile balancing — a simple modification of the popular ZSB sensitivity analysis — is feasible, robust, and sharp. This new sensitivity analysis for IPW is based on novel partial identification

results for [56]'s marginal sensitivity model.

We will point to several interesting directions for future work. While our partial identification results focus on counterfactual means and a few treatment effects, it should be possible to extend our partial identification results to more complex estimands of the type considered in [26, 28, 27, 29, 34]. Perhaps a similarly compact sensitivity analysis could even apply to dynamic treatment regimes. Future work could also investigate data-compatibility in the finite population model. In addition, while our IPW identification arguments generalize to any sensitivity assumption that only restricts the propensity score in a pointwise fashion (i.e. $e_{\min}(x) \leq e_0(x,u) \leq e_{\max}(x)$), the practicality of our sensitivity analysis and its theoretical properties rely on the marginal sensitivity model quite heavily. It would be interesting to see if a practical and sharp sensitivity analysis could be developed for other sensitivity assumptions in this class.

References

[1] Aronow, P. M. and D. K. K. Lee (2013). Interval estimation of population means under unknown but bounded probabilities of sample selection. *Biometrika 100*(1), 235–240.

[2] Athey, S., J. Tibshirani, and S. Wager (2019, 04). Generalized random forests. *Annals of Statistics 47*(2), 1148–1178.

[3] Athey, S. and S. Wager (2021). Policy learning with observational data. *Econometrica 89*(1), 133–161.

[4] Austin, P. C. and E. A. Stuart (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine 34*(28), 3661–3679.

[5] Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019). Conditional quantile processes based on series or many regressors. *Journal of Econometrics 213*(1), 4–29. Annals: In Honor of Roger Koenker.

[6] Beresteanu, A., I. Molchanov, and F. Molinari (2011). Sharp identification regions in models with convex moment predictions. *Econometrica 79*(6), 1785–1821.

[7] Bhattacharya, P. K. and A. K. Gangopadhyay (1990). Kernel and Nearest-Neighbor Estimation of a Conditional Quantile. *Annals of Statistics 18*(3), 1400 – 1415.

[8] Birmingham, J., A. Rotnitzky, and G. M. Fitzmaurice (2003). Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society: Series B 65*(1), 275–297.

[9] Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics 18*(1), 5 – 46.

[10] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*(1), C1–C68.

[11] Cinelli, C. and C. Hazlett (2020). Making sense of sensitivity: Extending omitted variables bias. *Journal of the Royal Statistical Society: Series B 82*, 39–67.

[12] Dantzig, G. B. and A. Wald (1951). On the fundamental lemma of neyman and pearson. *Annals of Mathematical Statistics 22*(1), 87–93.

[13] Darling, D. A. (1953, 06). On a class of problems related to the random division of an interval. *Annals of Mathematical Statistics 24*(2), 239–253.

[14] Davezies, L. and X. D'Haultfoeuille (2016). A new characterization of identified sets in partially identified models.

[15] Franks, A. M., A. D. Amour, and A. Feller (2020). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association 115*(532), 1730–1746.

[16] Freeman, R. B. (1984). Longitudinal analyses of the effects of trade unions. *Journal of Labor Economics 2*(1), 1–26.

[17] Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica 79*(2), 437–452.

[18] Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

[19] Hirano, K. and G. W. Imbens (2002). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services & Outcomes Research Methodology 2*, 259–278.

[20] Hristache, M. and V. Patilea (2017, 05). Conditional moment models with data missing at random. *Biometrika 104*(3), 735–742.

[21] Hsu, J. Y. and D. S. Small (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics 69*(4), 803–811.

[22] Hsu, J. Y., D. S. Small, and P. R. Rosenbaum (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association 108*(501), 135–148.

[23] Imai, K. and M. Ratkovic (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B 76*(1), 243–263.

[24] Jakubson, G. (1991). Estimation and testing of the union wage effect using panel data. *Review of Economic Studies 58*(5), 971–991.

[25] Johnson, G. (1975). Economic analysis of trade unionism. *American Economic Review 65*(2), 23–28.

[26] Kallus, N., X. Mao, and A. Zhou (2019). Interval estimation of individual-level causal effects under unobserved confounding. In K. Chaudhuri and M. Sugiyama (Eds.), *Proceedings of Machine Learning Research*, Volume 89 of *Proceedings of Machine Learning Research*, pp. 2281–2290. PMLR.

[27] Kallus, N. and A. Zhou (2018). Confounding-robust policy improvement. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31*, pp. 9269–9279. Curran Associates, Inc.

[28] Kallus, N. and A. Zhou (2020a). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 22293–22304. Curran Associates, Inc.

[29] Kallus, N. and A. Zhou (2020b). Minimax-optimal policy learning under unobserved confounding. *Management Science*, 1–20.

[30] Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

[31] Koenker, R. W. and G. Bassett (1978). Regression quantiles. *Econometrica 46*(1), 33–50.

[32] Kosorok, M. (2008). *Introduction to empirical processes and semiparametric inference*. Springer series in statistics. Springer.

[33] Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies 76*, 1071–1102.

[34] Lee, K., F. J. Bargagli-Stoffi, and F. Dominici (2020). Causal rule ensemble: Interpretable inference of heterogeneous treatment effects.

[35] Masten, M. A. and A. Poirier (2018). Identification of treatment effects under conditional partial independence. *Econometrica 86*(1), 317–351.

[36] Masten, M. A., A. Poirier, and L. Zhang (2020). Assessing sensitivity to unconfoundedness: Estimation and inference.

[37] Meinshausen, N. (2006, December). Quantile regression forests. *Journal of Machine Learning Research 7*, 983–999.

[38] Mellow, W. (1981). Unionism and wages: A longitudinal analysis. *Review of Economics and Statistics 63*(1), 43–52.

[39] Miratrix, L. W., S. Wager, and J. R. Zubizarreta (2018). Shape-constrained partial identification of a population mean under unknown probabilities of sample selection. *Biometrika 105*(1), 103–114.

[40] Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In *Econometric Theory*, Volume 4 of *Handbook of Econometrics*, pp. 2111 – 2245. Elsevier.

[41] Newey, W. K. and J. M. Robins (2017). Cross-fitting and fast remainder rates for semiparametric estimation. CeMMAP working papers CWP41/17, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

[42] Neyman, J. (1923, 11). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

[43] Robins, J. M., A. Rotnitzky, and D. O. Scharfstein (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, New York, NY, pp. 1–94. Springer New York.

[44] Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression-coefficients when some regressors are not always observed. *Journal of the American Statistical Association 89*(427), 846–866.

[45] Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika 74*(1), 13–26.

[46] Rosenbaum, P. R. (2002, 08). Covariance adjustment in randomized experiments and observational studies. *Statistical Science 17*(3), 286–327.

[47] Rosenbaum, P. R. (2005). Heterogeneity and causality. *The American Statistician 59*(2), 147–152.

[48] Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer.

[49] Rosenman, E., G. Basse, A. Owen, and M. Baiocchi (2020). Combining observational and experimental datasets using shrinkage estimators.

[50] Rosenman, E. T. R. and A. B. Owen (2021). Designing experiments informed by observational studies. *Journal of Causal Inference 9*(1), 147–171.

[51] Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology 66*(5), 688–701.

[52] Schick, A. (1986, 09). On asymptotically efficient estimation in semiparametric models. *Annals of Statistics 14*(3), 1139–1151.

[53] Semenova, V. (2020). Better Lee bounds.

[54] Soriano, D., E. Ben-Michael, P. J. Bickel, A. Feller, and S. D. Pimentel (2021). Interpretable sensitivity analysis for balancing weights.

[55] Stone, C. J. (1977, 07). Consistent nonparametric regression. *Annals of Statistics 5*(4), 595–620.

[56] Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association 101*(476), 1619–1637.

[57] Tudball, M., Q. Zhao, R. Hughes, K. Tilling, and J. Bowden (2019). An interval estimation approach to sample selection bias.

[58] Yadlowsky, S., H. Namkoong, S. Basu, J. Duchi, and L. Tian (2018). Bounds on the conditional and average treatment effect with unobserved confounding factors.

[59] Zhang, B. and D. S. Small (2020). A calibrated sensitivity analysis for matched observational studies with application to the effect of second-hand smoke exposure on blood lead levels in children. *Journal of the Royal Statistical Society: Series C 69*(5), 1285–1305.

[60] Zhao, Q., D. S. Small, and B. B. Bhattacharya (2019). Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B 81*(4), 735–761.

A Appendix: implementation

This appendix describes how the quantile balancing sensitivity analysis can be implemented using standard solvers for linear quantile regression (A.1), e.g. the `quantreg` package in `R` or the `qreg` function in `Stata`. It also gives the formulas for the ATT bounds (A.2), which were omitted from the main text, and offers a discussion of formulas for AIPW estimators (A.3).

Throughout this appendix, $\Lambda \geq 1$ is fixed and we set $\tau = \Lambda/(\Lambda+1)$. We also use the notation $\mathbb{E}_n[\cdot]$ as shorthand for the average $\frac{1}{n}\sum_{i=1}^n [\cdot]_i$.

A.1 Computing bounds with weighted quantile regression

We begin by considering computation of ψ_{T}^+. We consider the more general optimization problem (31) for some function $g : X \to \mathbb{R}^k$ containing an "intercept." In the main text, we assumed $g(x) = (1, \hat{Q}_\tau(x, 1))$.

$$\max_{\bar{e}\in\mathcal{E}_n(\Lambda)} \frac{\mathbb{E}_n[YZ/\bar{e}]}{\mathbb{E}_n[Z/\hat{e}(X)]} \quad \text{subject to} \quad \mathbb{E}_n[g(X)Z/\bar{e}] = \mathbb{E}_n[g(X)Z/\hat{e}(X)]. \tag{31}$$

Let $\rho_\tau(u) = u(\tau - \mathbb{I}\{u < 0\})$ be the quantile regression "check" function [30, 31] and define the weighted linear quantile regression objective as:

$$\mathcal{L}_n(\gamma) := \mathbb{E}_n[\rho_\tau(Y - \gamma^\top g(X))Z\tfrac{1-\hat{e}(X)}{\hat{e}(X)}]. \tag{32}$$

The following proposition shows that any minimizer of $\mathcal{L}_n$ can be used to compute the solution of (31).

Lemma 1. *Suppose $\hat{e}(X_i) \in (0,1)$ for all i, let $\hat{\gamma}$ minimize the weighted quantile regression objective $\mathcal{L}_n$ and let $\hat{V}_i = \mathrm{sign}(Y_i - \hat{\gamma}^\top g(X))$. Then the optimal objective value in the quantile balancing problem (31) is:*

$$\frac{\mathbb{E}_n[(Y - \hat{\gamma}^\top g(X))Z(1 + \Lambda^{\hat{V}}(1-\hat{e}(X))/\hat{e}(X))] + \mathbb{E}_n[\hat{\gamma}^\top g(X)Z/\hat{e}(X)]}{\mathbb{E}_n[Z/\hat{e}(X)]}.$$

Proof. See the supplementary materials. □

This same approach can be used to compute a lower bound for ψ_{T} by replacing Y with $-Y$, applying Lemma 1, and then negating the answer. Upper and lower bounds for ψ_{C} can then be obtained by replacing Z by $1-Z$ and $\hat{e}(X)$ by $1-\hat{e}(X)$ and then applying the same procedure. Subtracting the upper and lower bounds for ψ_{T} and ψ_{C} as in Theorem 2 gives bounds on ψ_{ATE}.

A.2 Bounds for the ATT

Next, we describe the standard quantile balancing bounds for ψ_{ATT}. Let $\bar{Y}(1)$ be the average value of Y_i among treated observations. We define the quantile balancing upper bound for the ATT as the solution to the optimization problem (33), where $g_+(x) = (1, \hat{Q}_{1-\tau}(x, 0))$.

$$\hat{\psi}^+_{\mathrm{ATT}} = \max_{\bar{e}\in\mathcal{E}_n(\Lambda)} \quad \bar{Y}(1) - \frac{\sum_{Z_i=0} Y_i \frac{\bar{e}_i}{1-\bar{e}_i}}{\sum_{Z_i=0} \frac{\bar{e}_i}{1-\bar{e}_i}} \quad \text{s.t.} \quad \sum_{Z_i=0} g_+(X_i)\frac{\bar{e}_i}{1-\bar{e}_i} = \sum_{Z_i=0} g_+(X_i)\frac{\hat{e}_i}{1-\hat{e}_i} \tag{33}$$

The lower bound $\hat{\psi}^-_{\mathrm{ATT}}$ is defined similarly, but with maximization replaced by minimization and $g_+(x)$ replaced by $g_-(x) := (1, \hat{Q}_\tau(x, 0))$. When $\Lambda = 1$, the two bounds collapse to the ordinary (stabilized) IPW estimate of the ATT under unconfoundedness [4, 23]. These bounds can also be computed using a variant of Lemma 1, but we omit the details.

A.3 AIPW computation

Here, we give formulas for three increasingly sharp AIPW sensitivity analyses. These were briefly discussed in the main text in Sections 4.1 and 4.2. For simplicity, we focus our discussion on the estimand $\psi_{\mathrm{T}} = \mathbb{E}[Y(1)]$.

Recall that, under unconfoundedness, the stabilized and unstabilized AIPW estimators of ψ_{T} have the following formulas:

$$\hat{\psi}^{(stab)}_{\mathrm{T}} = \mathbb{E}_n[\hat{\mu}(X,1)] + \frac{\mathbb{E}_n[Z(Y-\hat{\mu}(X,1))/\hat{e}(X)]}{\mathbb{E}_n[Z/\hat{e}(X)]}$$
$$\hat{\psi}^{(unstab)}_{\mathrm{T}} = \mathbb{E}_n[\hat{\mu}(X,1)] + \mathbb{E}_n[Z(Y-\hat{\mu}(X,1))/\hat{e}(X)]$$

Analysts whose primary analysis was conducted using the stablized AIPW estimator $\hat{\psi}^{(stab)}_{\mathrm{T}}$ may consider using any of the following three estimators for $\hat{\psi}^+_{\mathrm{T}}$:

$$\max_{\bar{e}\in\mathcal{E}_n(\Lambda)} \left\{ \mathbb{E}_n[\hat{\mu}(X,1)] + \frac{\mathbb{E}_n[Z(Y-\hat{\mu}(X,1))/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \right\} \tag{ZSB-AIPW}$$

$$\max_{\bar{e}\in\mathcal{E}_n(\Lambda)} \left\{ \mathbb{E}_n[\hat{\mu}(X,1)] + \frac{\mathbb{E}_n[Z(Y-\hat{\mu}(X,1))/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \right\} \quad \text{s.t.} \quad \mathbb{E}_n[Z/\bar{e}] = \mathbb{E}_n[Z/\hat{e}(X)] \tag{AIPW+1}$$

$$\max_{\bar{e}\in\mathcal{E}_n(\Lambda)} \left\{ \mathbb{E}_n[\hat{\mu}(X,1)] + \frac{\mathbb{E}_n[Z(Y-\hat{\mu}(X,1))/\bar{e}]}{\mathbb{E}_n[Z/\bar{e}]} \right\} \quad \text{s.t.} \quad \begin{pmatrix} \mathbb{E}_n[\hat{Q}^{(\epsilon)}_\tau(X,1)Z/\bar{e}] \\ \mathbb{E}_n[Z/\bar{e}] \end{pmatrix} = \begin{pmatrix} \mathbb{E}_n[\hat{Q}^{(\epsilon)}_\tau(X,1)Z/\hat{e}(X)] \\ \mathbb{E}_n[Z/\hat{e}(X)] \end{pmatrix} \tag{QB-AIPW}$$

`ZSB-AIPW` is the [60] proposal for stabilized AIPW estimators, which is generally not sharp. `AIPW+1` was described in Section 4.2 and adds a "balancing-ones" constraint which is necessary and sufficient for sharpness under homoscedastic additive noise models. `QB-AIPW` additionally balances $\hat{Q}^{(\epsilon)}_\tau(x,z)$, an estimate of the τ-th conditional quantile of the residual $\epsilon = Y - \mu(X,Z)$, which is necessary for sharpness under heteroscedastic models. All three approaches can be extended to $\hat{\psi}^{(unstab)}_{\mathrm{T}}$ by removing the term $\mathbb{E}_n[Z/\bar{e}]$ from the objective, though this change may impose a substantial cost with `ZSB` approach.

The additional constraints as we move from `ZSB-AIPW` to `AIPW+1` to `QB-AIPW` come at a cost. In certain simulations (see Appendix B), the constraints lead to substantial undercoverage of bootstrap confidence

Method	DGP1			DGP2		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
QB-Linear	90.7%	94.2%	94.5%	99.3%	100%	100%
Linear QB-AIPW	79.5%	90.1%	90.8%	95.5%	99.9%	100%
Linear AIPW+1	84.5%	92%	92%	97.4%	99.9%	100%
Linear ZSB-AIPW	87.2%	93%	93.8%	98%	100%	100%
QB-Forest	91.1%	95.6%	96.4%	98.1%	96.7%	96.7%
Forest QB-AIPW	91.4%	96.4%	97%	98%	96.7%	97.5%
Forest AIPW+1	94.6%	97.2%	97.6%	99.7%	99.1%	99.5%
Forest ZSB-AIPW	96.5%	97.8%	98%	99.9%	99.1%	99.5%
ZSB	96.9%	99.5%	99.8%	100%	100%	100%

Table 2: Table of rates at which various methods' 95% bootstrap confidence intervals cover the full identified sets in both DGPs and with increasing sample sizes.

intervals. The added constraint in `AIPW+1` is necessary for sharpness in additive-noise models but typically only refines the `ZSB-AIPW` estimate slightly. The `QB-AIPW` estimator, which requires an additional residual nuisance estimate, can be sharp under more general models. However, the `QB-AIPW`'s under-coverage is particularly extreme.

B Appendix: additional simulation results

This appendix presents additional simulation results beyond those appearing in the main text. For these simulations, we use the same two DGPs as Section 5 but include additional estimators and sample sizes.

In our simulations, we compare the four types of methods described in Section 5 (`QB`, `ZSB-AIPW`, `AIPW+1` and `ZSB`) along with the `QB-AIPW` method described in Section A.3. The `QB-AIPW` method requires an estimate of $Q_\alpha^{(\epsilon)}(x, z)$, the α-th conditional quantile of the residual $\epsilon = Y - \mu(X, Z)$. For this, we use an estimator of the form:

$$\hat{Q}_\alpha^{(\epsilon)}(x, z) = \hat{Q}_\alpha(x, z) - \hat{\mu}(x, z).$$

Here, $\hat{\mu}$ is an estimate of the conditional mean of Y and $\hat{Q}_\alpha$ is an estimate of the α-th conditional quantile of Y.

Figure 4 presents point estimates from all five methods when outcome regressions and conditional quantiles are estimated using linear models. In DGP1, most methods perform well when $n \geq 500$ except `ZSB`, which is noticeably conservative. However, when $n = 100$, `QB-AIPW` is noticeably aggressive. Meanwhile, in DGP2, all methods are conservative at all sample sizes because the quantile models are misspecified.

Figure 5 presents the same results when outcome regressions and conditional quantiles are estimated using random forest models. In DGP1 the results are qualitatively similar to the results in Figure 4 except `QB-AIPW` is no longer aggressive. Meanwhile, in DGP2, the `QB` and `QB-AIPW` methods yield sharper bounds than the other methods at all sample sizes, since the others do not account for heteroscedasticity.

Table 2 present coverage of bootstrap 95% confidence intervals for all methods, DGPs, and sample sizes. In DGP1, all methods but those that eventually over-cover exhibit under-coverage at $n = 100$. This is especially extreme for linear AIPW methods. As the sample size increases, `QB` with linear quantiles achieves near-nominal coverage, while the AIPW-based methods with linear regression estimates continue to under-cover. ZSB-AIPW's asymptotic conservativeness seems to be useful for offsetting this under-coverage. With forest nuisance estimates, `QB` continues to achieve near-nominal coverage in DGP1, while AIPW-based methods eventually over-cover. In DGP2, once we get beyond 100 observations, only `QB`-based methods achieve coverage below 99%. In those settings, `QB`-based methods with forest quantile estimates achieve near-nominal coverage.

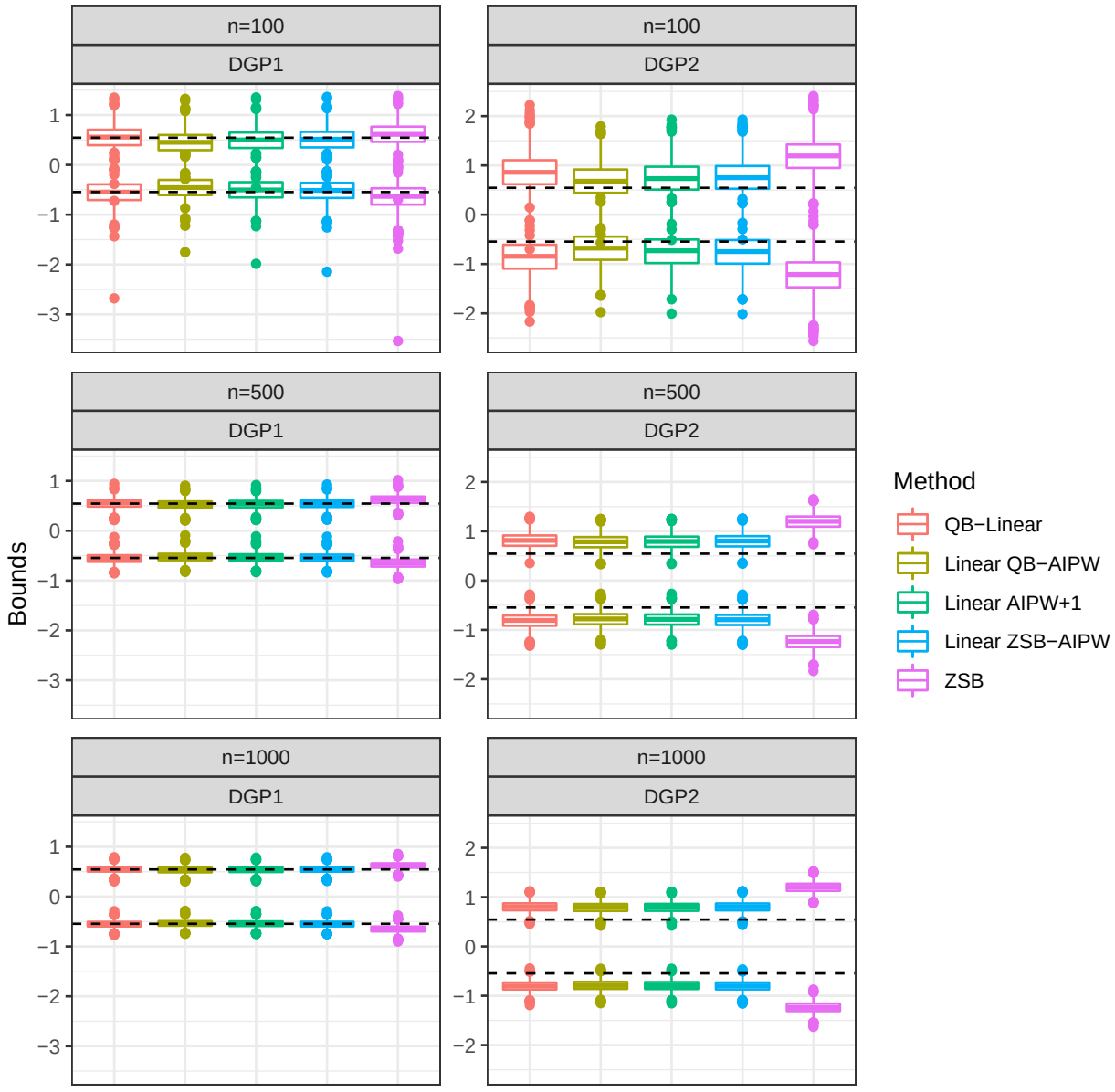

Figure 4: Box plot of point estimates across simulations with linear quantile and outcome regression nuisances.

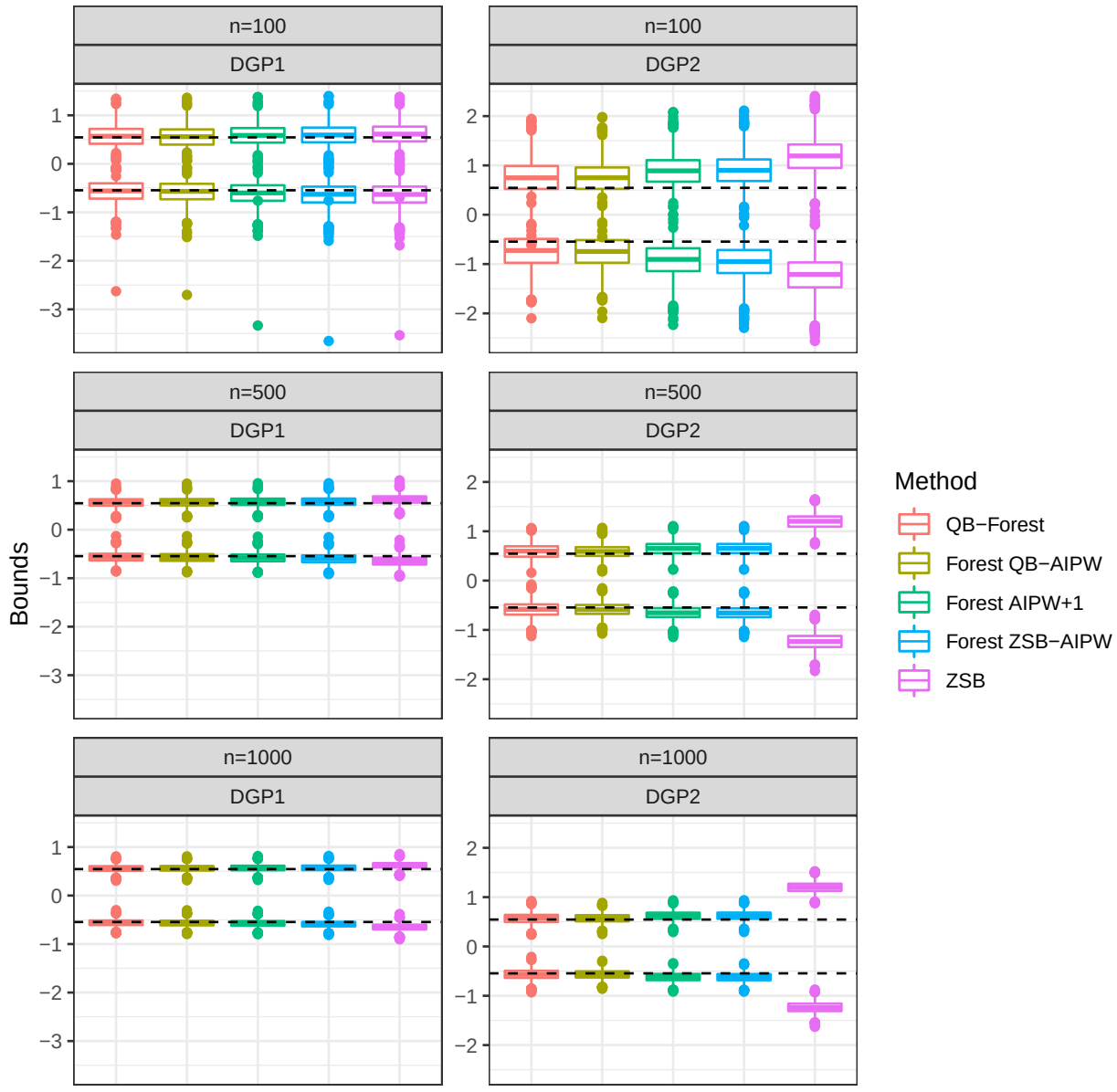

Figure 5: Box plot of point estimates across simulations with random forest-based quantile and outcome regression estimates.

C Appendix: proofs

This appendix collects proofs of the results in the main text along with supporting results. The organization of the appendix is to first prove all the Propositions appearing in the main text from first to last, including immediately-related Theorems, then all of the remaining Theorems in the main text from first to last. Proofs of non-immediate Corollaries are placed immediately after their main source. Proofs of substantial Lemmas are placed immediately before the argument in which they are first used. For readers hoping to follow all of the arguments from beginning to end, we recommend reading the results in the following order:

1. The proof of Proposition 1 in Section C.1 (Page 25).
2. The proof of Corollary 4 in Section C.2 (Page 26).
3. The proofs of Proposition 2 and Theorem 1 in Section C.3 (Page 28).
4. The proofs of Proposition 3 and Theorem 2 in Section C.4 (Page 28).
5. The proof of Corollary 3 in Section C.5 (Page 32)
6. The proof of Corollary 1 in Section C.6 (Page 33).
7. The proof of Lemma 1 in Section C.7 (Page 33).
8. The proof of Theorem 3, which is split across Sections C.8 (Page 34) and C.9 (Page 36) for the linear and non-linear cases, respectively.
9. The proof of Theorem 4 in Section C.10 (Page 43).

Many of the proofs depend on results from earlier in the list, but no proof depends on any results appearing later in the list.

Throughout, we will use the following notation. For an integer $n \geq 1$, $[n]$ denotes the set $\{1, \cdots, n\}$. If $\{a_n\}$ and $\{b_n\}$ are sequences of real numbers, then $a_n \precsim b_n$ means $a_n = \mathcal{O}(b_n)$ and $a_n \sim b_n$ means $a_n/b_n \to 1$. Similarly, if $\{A_n\}$ and $\{B_n\}$ are sequences of random variables, then $A_n \precsim_P B_n$ means $A_n = \mathcal{O}_P(B_n)$ and $A_n \sim_P B_n$ means $A_n/B_n \xrightarrow{p} 1$. We adopt the convention that $a/b = 0$ when a and b are both zero.

We also make use of some standard empirical process notation. For a (possibly random) function $f : \mathcal{X} \times \mathbb{R} \times \{0,1\} \to \mathbb{R}$, we will write $Pf := \int f \mathrm{d}P$ and $\mathbb{E}_n f := \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i, Z_i)$. For any vector $v = (v_1, ..., v_n)$, we take $\mathbb{E}_n v = \frac{1}{n}\sum_{i=1}^n v_i$. For any $p \in [1, \infty)$, we define $||f||_{L^p(P)} = (P|f|^p)^{1/p}$ and $||f||_{L^p(\mathbb{P}_n)} = (\mathbb{E}_n|f|^p)^{1/p}$. When $p = \infty$, we set $||f||_{L^\infty(P)} = \inf\{t : P(|f| \leq t) = 1\}$ and $||f||_{L^\infty(\mathbb{P}_n)} = \max_{i \leq n} |f(X_i, Y_i, Z_i)|$.

C.1 Proof of Proposition 1

We instead show the more general result:

Proposition 1B. *Let $(X, Y, Z) \sim P$, and let $e_{\min}, e_{\max} : \mathcal{X} \to (0, 1]$ be any two functions. For any random variable $\bar{E} \in (0, 1]$ satisfying $\mathbb{E}[Z/\bar{E}|X] = 1$ and $Z/e_{\max}(X) \leq Z/\bar{E} \leq Z/e_{\min}(X)$, we can construct random variables $(Y(0), Y(1), U)$ on the same probability space as $(X, Y, Z, \bar{E})$ and an associated putative propensity score $\bar{e}(X, U) := \mathbb{E}[Z|X, U]$ satisfying the following properties:*

(i) $Y = ZY(1) + (1 - Z)Y(0)$.

(ii) $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid (X, U)$ *and* $e_{\min}(X) \leq \bar{e}(X, U) \leq e_{\max}(X)$.

(iii) $Z/\bar{e}(X, U) = Z/\bar{E}$.

To recover the result of Proposition 1 from Proposition 1B, define $e_{\min}(x) = e(x)/(e(x) + [1 - e(x)]\Lambda)$ and $e_{\max}(x) = e(x)/(e(x) + [1 - e(x)]/\Lambda)$. Then let Q be the joint distribution of $(X, Y(0), Y(1), Z, U)$. Item (i) implies Q is data compatible, item (ii) and $\bar{E} \in \mathcal{E}_\infty(\Lambda)$ imply Q satisfies Assumption Λ, and item (iii) implies $\mathbb{E}_Q[Y(1)] = \mathbb{E}_Q[YZ/\bar{e}(X, U)] = \mathbb{E}_P[YZ/\bar{E}]$.

Proof. We begin by constructing $Y(0), Y(1)$ and U. Let $(X, Y, Z, \bar{E})$ be as in the proposition, and suppose we have access to independent random variables $V_1, V_2 \sim \text{Uniform}[0, 1]$ which are also jointly independent of $(X, Y, Z, \bar{E})$. Define the following collection of conditional distribution functions:

$$F(y|x, z) = P(Y \leq y|X = x, Z = z)$$

$$
\begin{aligned}
G(y|x,z,\bar{e}) &= P(Y \le y | X = x, Z = z, \bar{E} = \bar{e}) \\
H(\bar{e}|x,z) &= P(\bar{E} \le \bar{e} | X = x, Z = z) \\
K(u|x) &= \int_{-\infty}^{u} \frac{e(x)}{1-e(x)} \frac{1-\bar{e}}{\bar{e}} \, \mathrm{d}H(\bar{e}|x,1)
\end{aligned}
$$

One can verify that the conditions $\mathbb{E}[Z/\bar{E}|X] = 1$ and $\bar{E} > 0$ imply $K(u|x)$ is a proper CDF for each x. Using these functions, we define U, $Y(1)$, and $Y(0)$ by:

$$
\begin{aligned}
U &= Z\bar{E} + (1-Z)K^{-1}(V_2|X) \\
Y(1) &= ZY + (1-Z)G^{-1}(V_1|X,1,U) \\
Y(0) &= ZF^{-1}(V_1|X,0) + (1-Z)Y.
\end{aligned}
$$

We adopt the convention that $J^{-1}(s) := \inf\{t : J(t) \ge s\}$ whenever J is a distribution function, so that these quantities are well-defined even when some of these conditional distribution functions are not invertible.

With the construction done, we now verify the properties stated in the Proposition.

(i) This is immediate from the definition of $Y(0)$ and $Y(1)$.

(ii) We compute the distribution of $(Y(0), Y(1))$ given $X, U, Z = 1$ and the distribution of $(Y(0), Y(1))$ given $X, U, Z = 0$.

$$
\begin{aligned}
P(Y(0) \le y_0, Y(1) \le y_1 | X, U, Z = 1) &= P(F^{-1}(V_1|X,0) \le y_0, Y \le y_1 | X, U, Z = 1) \\
&= P(F^{-1}(V_1|X,0) \le y_0 | X, U, Z = 1) G(y_1|X,1,U) \\
&= P(F^{-1}(V_1|X,0) \le y_0 | X) G(y_1|X,1,U) \\
&= F(y_0|X,0) G(y_1|X,1,U) \\
P(Y(0) \le y_0, Y(1) \le y_1 | X, U, Z = 0) &= P(Y \le y_0, G^{-1}(V_1|X,1,U) \le y_1 | X, U, Z = 0) \\
&= P(Y \le y_0 | X, U, Z = 0) P(G^{-1}(V_1|X,1,U) \le y_1 | X, U, Z = 0) \\
&= P(Y \le y_0 | X, Z = 0) G(y_1|X,1,U) \\
&= F(y_0|X,0) G(y_1|X,1,U).
\end{aligned}
$$

Since these are the same, $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid (X, U)$.

A short calculation using Bayes' theorem shows that $\bar{e}(X, U) = U$.

$$
\begin{aligned}
\bar{e}(x,u) &= e(x) \frac{\mathrm{d}P(u|X=x, Z=1)}{\mathrm{d}P(u|X=x)} \\
&= e(x) \frac{\mathrm{d}P(u|x,1)/\mathrm{d}H(u|x,1)}{\mathrm{d}P(u|x)/\mathrm{d}H(u|x,1)} \\
&= \frac{e(x)}{e(x) + (1-e(x)) \frac{e(x)}{1-e(x)} \frac{1-u}{u}} \\
&= u
\end{aligned}
$$

Since the support of $K(\cdot|x)$ is a subset of the support of $H(\cdot|x,1)$, the assumption $Z/e_{\max}(X) \le \bar{E} \le Z/e_{\min}(X)$ implies $e_{\min}(X) \le U \le e_{\max}(X)$ almost surely, so $e_{\min}(X) \le e(X,U) \le e_{\max}(X)$.

(iii) The event $Z = 1$ implies $U = \bar{E}$, so $Z/e(X,U) = Z/U = Z/\bar{E}$.

□

C.2 Proof of Corollary 4

Proof. For this proof, we need a mathematically precise definition of the partially identified set. Let $\mathcal{P}(\Lambda)$ be the set of all probability distributions Q on $\mathcal{X} \times \mathbb{R} \times \mathbb{R} \times \{0,1\} \times \mathbb{R}^k$ (for some $k \ge 1$) satisfying the

following properties:

(i) If $(X, Y(0), Y(1), Z, U) \sim Q$, then $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid (X, U)$.
(ii) If we define $Y = ZY(1) + (1 - Z)Y(0)$, then the law of (X, Y, Z) under Q is the observed-data distribution P.
(iii) The odds ratio between $Q(Z = 1 \mid X, U)$ and $Q(Z = 1 \mid X)$ is bounded between Λ^{-1} and Λ almost surely.

The partially identified set for ψ_{T} is the set $\Psi_{\mathrm{T}} = \{\mathbb{E}_Q[Y(1)] : Q \in \mathcal{P}(\Lambda)\}$. We begin by verifying (18) by bounding ψ_{T}^+ below and then bounding it above.

For any random variable $\bar{E}$ on the same probability space as (X, Y, Z) satisfying $\mathbb{E}_P[Z/\bar{E}|X] = 1$, Proposition 1 implies that we may construct a distribution $Q \in \mathcal{P}(\Lambda)$ for which $\mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\bar{E}]$. Therefore, $\psi_{\mathrm{T}}^+ = \sup \Psi_{\mathrm{T}} \geq \mathbb{E}_P[YZ/\bar{E}]$. Since this inequality holds for every $\bar{E}$ satisfying $\mathbb{E}_P[Z/\bar{E}|X] = 1$, it holds for the supremum over $\bar{E}$. This proves one side of the equality (18).

For the other side, for any distribution $Q \in \mathcal{P}(\Lambda)$, we may write:

$$\begin{aligned}\mathbb{E}_Q[Y(1)] &= \mathbb{E}_Q[YZ/Q(Z = 1 \mid X, U)] \\ &= \mathbb{E}_Q[YZ \times \mathbb{E}[1/Q(Z = 1 \mid X, U) \mid X, Y, Z]]\end{aligned}$$

Since $\mathbb{E}[1/Q(Z = 1 \mid X, U) \mid X, Y, Z]$ is $\sigma(X, Y, Z)$-measurable, there exists a measurable function $\bar{e}_Q(x, y, z)$ such that $e_Q(X, Y, Z) = 1/\mathbb{E}[1/Q(Z = 1 \mid X, U) \mid X, Y, Z]$. Hence, if we define the random variable $\bar{E}$ on the same probability space on which P is defined by $\bar{E} = \bar{e}_Q(X, Y, Z)$, then we have:

$$\begin{aligned}\mathbb{E}_Q[Y(1)] &= \mathbb{E}_Q[YZ/\bar{e}_Q(X, Y, Z)] \\ &= \mathbb{E}_P[YZ/\bar{e}_Q(X, Y, Z)] \\ &= \mathbb{E}_P[YZ/\bar{E}].\end{aligned}$$

Finally, we check that $\bar{E}$ has the required properties. For any integrable function $h : \mathcal{X} \to \mathbb{R}$, we have:

$$\begin{aligned}\mathbb{E}_P[h(X)Z/\bar{E}] &= \mathbb{E}_P[h(X)Z/\bar{e}_Q(X, Y, Z)] \\ &= \mathbb{E}_Q[h(X)Z/\bar{e}_Q(X, Y, Z)] \\ &= \mathbb{E}_Q[h(X)Z\mathbb{E}[1/Q(Z = 1 \mid X, U) \mid X, Y, Z]] \\ &= \mathbb{E}_Q[h(X)] \\ &= \mathbb{E}_P[h(X)]\end{aligned}$$

Since this holds for every h, we may conclude $\mathbb{E}_P[Z/\bar{E} \mid X] = 1$. Finally, since, conditional on X, the support of $\bar{e}_Q(X, Y, Z)$ (under P or Q) is the same as that of $Q(Z = 1 \mid X, U)$, we may conclude that the following holds with probability one:

$$1 + \tfrac{1-e(X)}{e(X)}\Lambda^{-1} \leq 1/\bar{E} \leq 1 + \tfrac{1-e(X)}{e(X)}\Lambda$$

Hence, $\bar{E} \in \mathcal{E}_\infty(\Lambda)$, implying $\mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\bar{E}] \leq \sup_{\bar{E} \in \mathcal{E}_\infty(\Lambda)} \mathbb{E}[YZ/\bar{E}]$ s.t. $\mathbb{E}[Z/\bar{E}|X] = 1$. Since Q is arbitrary, the inequality continues to hold after taking the supremum over $Q \in \mathcal{P}(\Lambda)$ on both sides. This proves the other side of (18).

The equality (17) follows from an identical argument.

We complete the proof by showing that the identified set is an interval. Suppose $\psi = \alpha\psi_{\mathrm{T}}^- + (1-\alpha)\psi_{\mathrm{T}}^-$ for some $\alpha \in [0, 1]$. Suppose $\bar{E}_-$ and $\bar{E}_+$ solve (17) and (18), respectively. Define $\bar{E}^* = 1/[\alpha/\bar{E}_- + (1-\alpha)/\bar{E}_+]$. Then $\bar{E}^* \in \big[\min\{\bar{E}_-, \bar{E}_+\}, \max\{\bar{E}_-, \bar{E}_+\}\big]$, so $\bar{E}^* \in \mathcal{E}_\infty(\Lambda)$. In addition:

$$\mathbb{E}[Z/\bar{E}^*|X] = \alpha\mathbb{E}[Z/\bar{E}_-|X] + (1-\alpha)\mathbb{E}[Z/\bar{E}_+|X] = \alpha + (1-\alpha) = 1$$

Therefore, by Proposition 1, $\alpha\psi_{\mathrm{T}}^- + (1-\alpha)\psi_{\mathrm{T}}^-$ is in the partially identified set. □

C.3 Proof of Proposition 2 and Theorem 1

Proof. We begin by proving Proposition 2, which is sufficient to make Theorem 1 a simple implication of Corollary 4. By symmetry, it suffices to show that $\bar{E}_+$ solves both (12) and (18), where (12) is from Theorem 1 and (18) is from Corollary 4.

First, we show that there exists $\bar{E}_+ \in \mathcal{E}_\infty(\Lambda)$ with the properties stated in Proposition 2. Define $e_{\min}(x) = e(x)/(e(x) + [1-e(x)]/\Lambda)$ and $e_{\max}(x) = e(x)/(e(x) + [1-e(x)]\Lambda)$. For any $\gamma \in [e_{\min}(x), e_{\max}(x)]$, define $e_\gamma(x,y)$ by:

$$\bar{e}_\gamma(x,y) = \begin{cases} e_{\min}(x) & \text{if } y > Q_\tau(x,1) \\ e_{\max}(x) & \text{if } y < Q_\tau(x,1) \\ \gamma & \text{if } y = Q_\tau(x,1) \end{cases}$$

We claim that for all x, there exists $\gamma(x) \in [e_{\min}(x), e_{\max}(x)]$ solving $\mathbb{E}[Z/\bar{e}_{\gamma(x)}(X,Y)|X=x] = 1$. We will prove this by applying the intermediate value theorem to the continuous function $w_x(\gamma) := \mathbb{E}[Z/e_\gamma(X,Y)|X = x]$. If we took $\gamma = e_{\max}(x)$, then we would have:

$$\begin{aligned} w_x(e_{\max}(x)) &= F(Q_\tau(x,1)|x,1)(e(x) + [1-e(x)]/\Lambda) + (1 - F(Q_\tau(x,1)|x,1))(e(x) + [1-e(x)]\Lambda) \\ &\leq e(x) + (1-e(x))(\tau/\Lambda + (1-\tau)\Lambda) \\ &= 1 \end{aligned}$$

and a similar calculation shows $w_x(e_{\min}(x)) \geq 1$. Thus, there is some $\gamma(x) \in [e_{\min}(x), e_{\max}(x)]$ which solves $\mathbb{E}[Z/\bar{e}_{\gamma(x)}(X,Y)|X=x] = 1$. Therefore, $\bar{E}_+ := \bar{e}_{\gamma(X)}(X,Y)$ belongs to $\mathcal{E}_\infty(\Lambda)$ and satisfies $\mathbb{E}[Z/\bar{E}_+|X] = 1$.

Now we show that any random variable $\bar{E}_+$ satisfying the requirements of the proposition solves the quantile balancing problem (12). It is easy to see that $\bar{E}_+$ is feasible in (12), since $\mathbb{E}[Q_\tau(X)Z/\bar{E}_+] = \mathbb{E}[Q_\tau(X)\mathbb{E}[Z/\bar{E}_+|X]] = \mathbb{E}[Q_\tau(X)]$. Moreover, for any other $\bar{E} \in \mathcal{E}_\infty(\Lambda)$ which balances Q_τ, we may write:

$$\begin{aligned} \mathbb{E}[YZ/\bar{E}] &= \mathbb{E}[Q_\tau(X,1)Z/\bar{E} + (Y - Q_\tau(X,1))Z/\bar{E}] \\ &\leq \mathbb{E}[Q_\tau(X,1)] + \mathbb{E}[(Y - Q_\tau(X,1))Z/\bar{E}_+] \\ &= \mathbb{E}[Q_\tau(X,1)Z/\bar{E}_+] + \mathbb{E}[(Y - Q_\tau(X,1)Z/\bar{E}_+] \\ &= \mathbb{E}[YZ/\bar{E}_+]. \end{aligned}$$

The inequality step follows because $1/\bar{E}_+$ takes on the maximum allowable value whenever $(Y - Q_\tau(X,1))Z$ is positive and the minimal allowable value whenever $(Y - Q_\tau(X,1))Z$ is negative, so $(Y - Q_\tau(X,1))Z/\bar{E}_+$ is always larger than $(Y - Q_\tau(X,1))Z/\bar{E}$. Since $\bar{E}$ is arbitrary, this proves $\bar{E}_+$ solves (12).

Finally, $\bar{E}_+^*$ solves the less constrained problem (12) and is feasible in the more constrained problem (18), so it solves (18) as well. This proves Proposition 2.

Now we proceed to Theorem 1. To prove that the partially identified set is an interval, observe that the set

$$\mathcal{W} = \{1/\bar{E} \,:\, \bar{E} \in \mathcal{E}_\infty(\Lambda), \mathbb{E}_P[Z/\bar{E}|X] = 1\}$$

is convex. By Corollary 4, the partially identified set is the image of $\mathcal{W}$ under the linear function $W \mapsto \mathbb{E}[YZW]$. Therefore, the partially identified set is a convex set in $\mathbb{R}$, i.e. an interval.

The formulas for the interval endpoints follow immediately from Corollary 4 and Proposition 2. These results also show that the endpoints are attained, so that the partially identified interval is closed. □

C.4 Proof of Proposition 3 and Theorem 2

Proof. We will divide the proof of Proposition 3, where we begin, into several steps. Rather than explicitly constructing a distribution Q with $\mathbb{E}_Q[Y(1)] = \mathbb{E}_P[YZ/\bar{E}]$ and $\mathbb{E}_Q[Y(0)] = \mathbb{E}_P[Y(1-Z)/(1-\bar{E})]$ for each $\bar{E}$ satisfying the conditions of the Proposition, we will instead construct the extremal distributions $Q_{+,+}$, $Q_{+,-}$, $Q_{-,+}$ and $Q_{-,-}$ that attain the endpoints of the partially identified set for ψ_{T} and ψ_{C}. Then, we will show that we can achieve any mixture. This will establish Proposition 3.

C.4.1 Notation

We begin by recording some notation that will be used throughout the proof. By Theorem 1 and Proposition 2 (and their generalizations to the estimand ψ_{C}), the extremal potential outcomes have the following formulas:

$$\begin{aligned}
\psi_{\mathrm{T}}^{+} &= \mathbb{E}[YZ/\bar{E}_{\mathrm{T}}^{+}] \\
\psi_{\mathrm{T}}^{-} &= \mathbb{E}[YZ/\bar{E}_{\mathrm{T}}^{-}] \\
\psi_{\mathrm{C}}^{+} &= \mathbb{E}[Y(1-Z)/(1-\bar{E}_{\mathrm{C}}^{+})] \\
\psi_{\mathrm{C}}^{-} &= \mathbb{E}[Y(1-Z)/(1-\bar{E}_{\mathrm{C}}^{-})]
\end{aligned}$$

where the worst-case propensity scores $\bar{E}_{\mathrm{T}}^{-}, \bar{E}_{\mathrm{T}}^{+}, \bar{E}_{\mathrm{C}}^{-}, \bar{E}_{\mathrm{C}}^{+}$ are random variables which satisfy the following:

$$\bar{E}_{\mathrm{T}}^{+} = \begin{cases} \frac{e(X)}{e(X)+[1-e(X)]\Lambda} & \text{if } Y > Q_{\tau}(X,1) \\ \frac{e(X)}{e(X)+[1-e(X)]/\Lambda} & \text{if } Y < Q_{\tau}(X,1) \end{cases}$$

$$\bar{E}_{\mathrm{T}}^{-} = \begin{cases} \frac{e(X)}{e(X)+[1-e(X)]/\Lambda} & \text{if } Y > Q_{1-\tau}(X,1) \\ \frac{e(X)}{e(X)+[1-e(X)]\Lambda} & \text{if } Y < Q_{1-\tau}(X,1) \end{cases}$$

$$\bar{E}_{\mathrm{C}}^{+} = \begin{cases} \frac{e(X)}{e(X)+[1-e(X)]/\Lambda} & \text{if } Y > Q_{\tau}(X,0) \\ \frac{e(X)}{e(X)+[1-e(X)]\Lambda} & \text{if } Y < Q_{\tau}(X,0) \end{cases}$$

$$\bar{E}_{\mathrm{C}}^{-} = \begin{cases} \frac{e(X)}{e(X)+[1-e(X)]\Lambda} & \text{if } Y > Q_{1-\tau}(X,0) \\ \frac{e(X)}{e(X)+[1-e(X)]/\Lambda} & \text{if } Y < Q_{1-\tau}(X,0) \end{cases}.$$

The formulas for $\bar{E}_{\mathrm{C}}^{-}$ and $\bar{E}_{\mathrm{C}}^{+}$ can be derived by exchanging the roles of Z and $1-Z$ (and correspondingly the roles of $e(X)$ and $1-e(X)$) and then applying Proposition 2.

C.4.2 Constructing $Q_{+,-}$

We now construct the distribution $Q_{+,-}$ which attains the upper bound on ψ_{T} and the lower bound on ψ_{C}. We will actually construct random variables $Y(0), Y(1), U$ on the same probability space as (X, Y, Z), with associated plausible propensity score $\bar{e}(X,U) := \mathbb{E}[Z|X,U]$, that satisfy the following requirements:

(a) $Y = Y(1)Z + Y(0)(1-Z)$.
(b) $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid (X,U)$.
(c) $\bar{e}(X,U) \in \mathcal{E}_{\infty}(\Lambda)$.
(d) $\mathbb{E}[Y(1)] = \psi_{\mathrm{T}}^{+}$ and $\mathbb{E}[Y(0)] = \psi_{\mathrm{C}}^{-}$.

We then take $Q_{+,-}$ to be the joint distribution of $(X, Y(0), Y(1), Z, U)$.

We start with the construction. Let $(X,Y,Z) \sim P$ and $(V_1, V_2) \sim \mathrm{Uniform}[0,1]^2$ independently of (X,Y,Z). Let $F(y|x,z) = P(Y \le y | X = x, Z = z)$ and $\bar{H}(y|x,z) = P(Y = y | X = x, Z = z)$. Let $T = \tau Z + (1-\tau)(1-Z)$, and define the binary "confounder" U by:

$$U = \mathbb{I}\{Y > Q_T(X,Z)\} + \mathbb{I}\{Y = Q_T(X,Z), V_1 \bar{H}(Y|X,Z) < F(Y|X,Z) - T\}.$$

Define the conditional CDF of Y to sample from by $G(y|x,z,u) = P(Y \le y | X = x, U = u, Z = z)$, and construct $Y(0), Y(1)$ by:

$$\begin{aligned}
Y(1) &= ZY + (1-Z)G^{-1}(V_2|X, Z=1, U) \\
Y(0) &= ZG^{-1}(V_2|X, Z=0, U) + (1-Z)Y.
\end{aligned}$$

This concludes the construction. We now verify that $Y(0), Y(1), U$ satisfy the required properties (a) – (d) from the start of this sub-section.

(a) This is immediate from the definition of $Y(0)$ and $Y(1)$.

(b) We prove (b) by computing the joint distribution of $(Y(0), Y(1))$ given $X, U, Z = 1$ and also the joint distribution of $(Y(0), Y(1))$ given $X, U, Z = 0$.

$$\begin{aligned}
P(Y(0) \le y_0, Y(1) \le y_1 | X, U, Z = 1) &= P(G^{-1}(V_2|X, 0, U) \le y_0, Y \le y_1 | X, U, Z = 1) \\
&= G(y_0|X, 0, U) P(Y \le y_1 | X, U, Z = 1) \\
&= G(y_0|X, 0, U) G(y_1|X, 1, U) \\
P(Y(0) \le y_0, Y(1) \le y_1 | X, U, Z = 0) &= P(Y \le y_0, G^{-1}(V_2|X, 1, U) \le y_1 | X, U, Z = 0) \\
&= P(Y \le y_0 | X, U, Z = 0) G(y_1|X, 1, U) \\
&= G(y_0|X, 0, U) G(y_1|X, 1, U)
\end{aligned}$$

Since these are the same, $(Y(0), Y(1)) \perp\!\!\!\perp Z \mid (X, U)$.

(c) We establish (c) by directly computing $\bar{e}(X, U)$. First, observe that $\mathbb{E}[U|X, Z = 1] = 1 - \tau$.

$$\begin{aligned}
\mathbb{E}[U|X, Z = 1] &= P(Y > Q_\tau(X, Z)|X, Z = 1) + P(Y = Q_\tau(X, Z), V_1 < \tfrac{F(Q_\tau(X,1)|X,1) - \tau}{\bar{H}(Q_\tau(X,1)|X,1)} \mid X, 1) \\
&= 1 - F(Q_\tau(X, 1)|X, 1) + \bar{H}(Q_\tau(X, 1)|X, 1) \tfrac{F(Q_\tau(X,1)|X,1) - \tau}{\bar{H}(Q_\tau(X,1)|X,1)} \\
&= 1 - \tau
\end{aligned}$$

A similar calculation shows $\mathbb{E}[U|X, Z = 0] = \tau$. Therefore, we have:

$$\begin{aligned}
\bar{e}(x, 0) &= P(Z = 1 \mid X = x, U = 0) \\
&= \frac{e(x) P(U = 0|X = x, Z = 1)}{e(x) P(U = 0|X = x, Z = 1) + [1 - e(x)] P(U = 0|X = x, Z = 0)} \\
&= \frac{e(x)\tau}{e(x)\tau + [1 - e(x)](1 - \tau)} \\
&= \frac{e(x)}{e(x) + [1 - e(x)]/\Lambda} \\
\bar{e}(x, 1) &= P(Z = 1 \mid X = x, U = 1) \\
&= \frac{e(x) P(U = 1|X = x, Z = 1)}{e(x) P(U = 1|X = x, Z = 1) + [1 - e(x)] P(U = 1|X = x, Z = 0)} \\
&= \frac{e(x)(1 - \tau)}{e(x)(1 - \tau) + [1 - e(x)]\tau} \\
&= \frac{e(x)}{e(x) + [1 - e(x)]\Lambda}
\end{aligned}$$

Both $\bar{e}(x, 1)$ and $\bar{e}(x, 0)$ satisfy the bounded odds ratio condition, so $\bar{e}(X, U) \in \mathcal{E}_\infty(\Lambda)$.

(d) The explicit formulas for $\bar{e}(X, U)$ obtained in the proof of (c) shows that $\bar{e}(X, U)$ satisfies:

$$\bar{e}(X, U) = \begin{cases} \frac{e(X)}{e(X) + [1 - e(X)]\Lambda} & \text{if } U = 1 \\ \frac{e(X)}{e(X) + [1 - e(X)]/\Lambda} & \text{if } U = 0 \end{cases} \tag{34}$$

If $Z = 1$, then $Y > Q_\tau(X, 1)$ implies $U = 1$ while $Y < Q_\tau(X, 1)$ implies $U = 0$. Therefore, by comparing (34) with the formula for $\bar{E}^+_{\mathrm{T}}$, we may conclude that $Z/\bar{e}(X, U) = Z/\bar{E}^+_{\mathrm{T}}$, except possibly on the event $Y = Q_\tau(X, 1)$. Moreover, we can check that $\mathbb{E}[Z/\bar{e}(X, U)|X] = 1$.

$$\begin{aligned}
\mathbb{E}[Z/\bar{e}(X, U)|X] &= e(X) \mathbb{E}[1/\bar{e}(X, U)|X, Z = 1] \\
&= e(X)(P(U = 0|X, Z = 1)/\bar{e}(X, 0) + P(U = 1|X, Z = 1)/\bar{e}(X, 1)) \\
&= e(X)(\tau(1 + \tfrac{1 - e(X)}{e(X)} \Lambda^{-1}) + (1 - \tau)(1 + \tfrac{1 - e(X)}{e(X)} \Lambda))
\end{aligned}$$

$$= 1$$

Therefore, $\mathbb{E}[YZ/\bar{e}(X,U)] = \psi_{\mathrm{T}}^{+}$ by Proposition 2.

Similarly, when $Z = 0$, then $Y > Q_{1-\tau}(X,0)$ implies $U = 1$ and $Y < Q_{1-\tau}(X,0)$ implies $U = 0$. Therefore, by comparing (34) with the formula for $\bar{E}_{\mathrm{C}}^{-}$, we can conclude $(1-Z)/(1-\bar{e}(X,U)) = (1-Z)/(1-\bar{E}_{\mathrm{C}}^{-})$, except possibly on the event $Y = Q_{1-\tau}(X,0)$. Moreover, we can check that $\mathbb{E}[(1-Z)/(1-\bar{e}(X,U))|X] = 1$.

$$\begin{aligned}
\mathbb{E}[(1-Z)/(1-\bar{e}(X,U))|X] &= (1-e(X))\mathbb{E}[1/(1-\bar{e}(X,U))|X, Z=0] \\
&= (1-e(X))(\tfrac{P(U=0|X,Z=0)}{1-\bar{e}(X,0)} + \tfrac{P(U=1|X,Z=1)}{1-\bar{e}(X,1)}) \\
&= (1-e(X))((1-\tau)\tfrac{1-e(X)+e(X)\Lambda}{1-e(X)} + \tau\tfrac{1-e(X)+e(X)/\Lambda}{1-e(X)}) \\
&= 1
\end{aligned}$$

Therefore, by an argument similar to the proof of Proposition 2, we have $\mathbb{E}[YZ/\bar{e}(X,U)] = \psi_{\mathrm{C}}^{-}$.

C.4.3 Constructing the other extremal distributions

Next, we construct the other extremal distributions. We start with the distribution $Q_{+,+}$ that attains ψ_{T}^{+} and ψ_{C}^{+}.

Define $Y' = ZY + (1-Z)(-Y)$. Applying the construction from Section C.4.2 to the data (X, Y', Z) yields potential outcomes $(Y(0)', Y(1)')$ and a binary confounder U' satisfying the consistency relation $Y' = Y(1)'Z + Y(0)'(1-Z)$ and the unconfoundedness condition $(Y(0)', Y(1)') \perp\!\!\!\perp Z \mid (X, U')$. Moreover, if we define $Q'_t(x,z)$ to be the t-th conditional quantile of Y' given $X = x, Z = z$, then $e'(X,U') := \mathbb{E}[Z|X,U']$ will satisfy:

$$Z/e'(X,U') = \begin{cases} Z\left(1 + \frac{1-e(X)}{e(X)}\Lambda^{+1}\right) & \text{if } Y' > Q'_{\tau}(X,1) \\ Z\left(1 + \frac{1-e(X)}{e(X)}\Lambda^{-1}\right) & \text{if } Y' < Q'_{\tau}(X,1) \end{cases}$$

$$(1-Z)/(1-e'(X,U')) = \begin{cases} (1-Z)\left(1 + \frac{e(X)}{1-e(X)}\Lambda^{-1}\right) & \text{if } Y' > Q'_{1-\tau}(X,0) \\ (1-Z)\left(1 + \frac{e(X)}{1-e(X)}\Lambda^{+1}\right) & \text{if } Y' < Q'_{1-\tau}(X,0) \end{cases}$$

and also $\mathbb{E}[Z/e'(X,U')|X] = \mathbb{E}[(1-Z)/(1-e'(X,U')|X] = 1$.

Observe that when $Z = 1$, $Y' = Y$ and $Q'_{\tau}(X,1) = Q_{\tau}(X,1)$. Therefore, $Z/e'(X,U') = Z/\bar{E}_{\mathrm{T}}^{+}$, except possibly on the event $Y = Q_{\tau}(X,1)$. As a result, Proposition 2 and Theorem 1 imply:

$$\begin{aligned}
\mathbb{E}[Y(1)'] &= \mathbb{E}[Y'Z/e'(X,U')] \\
&= \mathbb{E}[Y'Z/\bar{E}_{\mathrm{T}}^{+}] \\
&= \psi_{\mathrm{T}}^{+}
\end{aligned}$$

On the other hand, when $Z = 0$, we have $Y' = -Y$ and $Q'_{1-\tau}(X,0) = -Q_{\tau}(X,0)$. On this event, $Y' > Q'_{1-\tau}(X,0)$ is equivalent to $Y < Q_{\tau}(X,0)$, so $(1-Z)/(1-e'(X,U') = (1-Z)/(1-\bar{E}_{\mathrm{C}}^{+})$, except possibly on the event $Y = Q_{\tau}(X,0)$. Similarly, Proposition 2 and Corollary 2 imply:

$$\begin{aligned}
\mathbb{E}[Y(0)'] &= \mathbb{E}[Y'(1-Z)/(1-e'(X,U'))] \\
&= -\mathbb{E}[Y(1-Z)/(1-\bar{E}_{\mathrm{C}}^{+})] \\
&= -\psi_{\mathrm{C}}^{+}
\end{aligned}$$

Finally, we define $Y(0) = -Y(0)', Y(1) = Y(1)'$ and $U = U'$. Then the data $(Y(0), Y(1), Z, X, U)$ will satisfy Assumption Λ and also $\mathbb{E}[Y(1)] = \psi_{\mathrm{T}}^{+}$, $\mathbb{E}[Y(0)] = \psi_{\mathrm{C}}^{-}$.

To construct $Q_{-,-}$, apply the preceding construction to $Y'' = -Y'$. To construct $Q_{-,+}$, apply the construction in Section C.4.2 to $Y''' = -Y$.

C.4.4 Creating all convex combinations

Finally, we show that for any ψ_T satisfying $\psi_T^- \le \psi_T \le \psi_T^+$ and any ψ_C satisfying $\psi_C^- \le \psi_C \le \psi_C^+$, there is a data-compatible distribution Q satisfying Assumption Λ with $\mathbb{E}_Q[Y(1)] = \psi_T$ and $\mathbb{E}_Q[Y(0)] = \psi_C$.

Since the vector (ψ_T, ψ_C) lies in the convex hull of the points $(\psi_T^-, \psi_C^-), (\psi_T^-, \psi_C^+), (\psi_T^+, \psi_C^-), (\psi_T^+, \psi_C^+)$, there exists nonnegative weights w_1, w_2, w_3, w_4 summing to one and satisfying:

$$\begin{pmatrix}\psi_T \\ \psi_C\end{pmatrix} = w_1\begin{pmatrix}\psi_T^- \\ \psi_C^-\end{pmatrix} + w_2\begin{pmatrix}\psi_T^- \\ \psi_C^+\end{pmatrix} + w_3\begin{pmatrix}\psi_T^+ \\ \psi_C^-\end{pmatrix} + w_4\begin{pmatrix}\psi_T^+ \\ \psi_C^+\end{pmatrix}$$

Let $M \sim \text{Multinomal}(\{1, \cdots, 4\}, (w_1, \cdots, w_4))$, and sample $(X, Y(0), Y(1), Z, U) \sim Q_{-,-}$ when $M = 1$, $Q_{-,+}$ when $M = 2$, $Q_{+,-}$ when $M = 3$ and $Q_{+,+}$ when $M = 4$. Finally, let Q be the distribution of $(X, Y(0), Y(1), Z, U')$ where $U' = (U, M)$.

It is clear that the distribution Q is data-compatible and satisfies Assumption Λ, since it is the mixture of distributions satisfying these conditions. Moreover, it is easy to check that $\mathbb{E}_Q[Y(1)] = \mathbb{E}_Q[\mathbb{E}[Y(1)|M]] = w_1\psi_T^- + w_2\psi_T^- + w_3\psi_T^+ + w_4\psi_T^+ = \psi_T$. By the same reasoning, $\mathbb{E}_Q[Y(0)] = \psi_C$.

C.4.5 Proof of Theorem 2

We now proceed to prove Theorem 2.

As in the proof of Corollary 4, let $\mathcal{P}(\Lambda)$ be the set of full-data distributions Q compatible with Assumption Λ and the observed-data distribution P. Then we may write:

$$\begin{aligned}\psi_{\text{ATE}}^+ &= \sup_{Q\in\mathcal{P}(\Lambda)} \mathbb{E}_Q[Y(1) - Y(0)] \\ &\le \sup_{Q\in\mathcal{P}(\Lambda)} \mathbb{E}_Q[Y(1)] - \inf_{Q\in\mathcal{P}(\Lambda)} \mathbb{E}_Q[Y(0)] \\ &= \psi_T^+ - \psi_C^-.\end{aligned}$$

In the other direction, Proposition 2 implies that there exists worst-case propensity scores $\bar{E}_T^+$ and $\bar{E}_C^-$ in $\mathcal{E}(\Lambda)$ satisfying $\psi_T^+ = \mathbb{E}_P[YZ/\bar{E}_T^+]$ and $\psi_C^- = \mathbb{E}_P[Y(1-Z)/(1-\bar{E}_C^-)]$ such that if we define $\bar{E} = Z\bar{E}_T^+ + (1-Z)\bar{E}_C^-$, then $\bar{E}$ satisfies the hypotheses of Proposition 3. Therefore, Proposition 3 implies that there exists a distribution $Q \in \mathcal{P}(\Lambda)$ for which $\mathbb{E}_Q[Y(1)-Y(0)] = \psi_T^+ - \psi_C^-$. Therefore $\sup_{Q\in\mathcal{P}(\Lambda)} \mathbb{E}_Q[Y(1)-Y(0)] \ge \psi_T^+ - \psi_C^-$.

The arguments so far imply $\psi_T^+ - \psi_C^- \ge \psi_{\text{ATE}}^+ = \sup_{Q\in\mathcal{P}(\Lambda)} \mathbb{E}_Q[Y(1) - Y(0)] \ge \psi_T^+ - \psi_C^-$. Thus, $\psi_{\text{ATE}}^+ = \psi_T^+ - \psi_C^-$. By exactly the same reasoning, $\psi_{\text{ATE}}^- = \psi_T^- - \psi_C^+$.

Finally, it remains to show that the partially identified set for ψ_{ATE} is a closed interval. By Proposition 3, the partially identified set for the ATE contains the set $\{\psi_T - \psi_C : (\psi_T, \psi_C) \in [\psi_T^-, \psi_T^+] \times [\psi_C^-, \psi_C^+]\}$, which is a closed interval. Moreover, the preceding calculation shows it does not contain any other points. Thus, the partially identified set is a closed interval. □

C.5 Proof of Corollary 3

Proof. First, we will compute the partially identified set for ψ_T. Let z_τ denote the τ-th quantile of the standard normal distribution. Since the conditional distribution of $Y \mid X = x, Z = 1$ is continuous for every x, Proposition 2 implies $\psi_T^+ = \mathbb{E}[YZ/\bar{E}_+]$ where $\bar{E}_+$ satisfies the following:

$$1/\bar{E}_+ = \begin{cases} 1 + \frac{1-e(X)}{e(X)}\Lambda^{+1} & \text{if } Y \ge \mu(X,1) + \sigma(X)z_\tau \\ 1 + \frac{1-e(X)}{e(X)}\Lambda^{-1} & \text{if } Y < \mu(X,1) + \sigma(X)z_\tau \end{cases}$$

Let $C(x) = \mu(x,1) + \sigma(x)z_\tau$. Write $\mathbb{E}[YZ/\bar{E}_+] = \mathbb{E}[e(X)\mathbb{E}[Y/\bar{E}_+|X, Z = 1]]$ and evaluate the inner expectation as follows:

$$\mathbb{E}[Y/\bar{E}_+|X, Z = 1] = \tau\mathbb{E}[Y/\bar{E}_+|X, Z = 1, Y < C(X)] + (1-\tau)\mathbb{E}[Y/\bar{E}_+|X, Z = 1, Y \ge C(X)]$$

$$= \tau \mathbb{E}[Y|X, Z=1, Y<C(X)] + \tau \tfrac{1-e(X)}{e(X)} \Lambda^{-1} \mathbb{E}[Y|X, Z=1, Y<C(X)]$$
$$= \tfrac{\mu(X,1)}{e(X)} + \tfrac{\Lambda-1}{\Lambda} \tfrac{1-e(X)}{e(X)} \sigma(X) \tfrac{\phi(z_\tau)}{1-\tau}$$

In the last step, we used the inverse Mills ratio formula for the expectation of a truncated Gaussian distribution. Simplifying gives $\psi_{\mathrm{T}}^{+} = \mathbb{E}[\mu(X,1)] + \frac{\Lambda^2-1}{\Lambda}\phi(z_\tau)\mathbb{E}[(1-e(X))\sigma(X)]$.

At this point, we can immediately generalize the above calculation to all other potential outcome bounds. By applying the preceding calculation to $-Y$ and negating the answer, we may conclude:

$$\psi_{\mathrm{T}}^{-} = \mathbb{E}[\mu(X,1)] - \tfrac{\Lambda^2-1}{\Lambda}\phi(z_\tau)\mathbb{E}[(1-e(X))\sigma(X)].$$

By exchanging the roles of Z and $1-Z$ (and correspondingly the roles of $e(X)$ and $1-e(X)$), we then obtain the bounds:

$$\psi_{\mathrm{C}}^{+} = \mathbb{E}[\mu(X,0)] + \tfrac{\Lambda^2-1}{\Lambda}\phi(z_\tau)\mathbb{E}[e(X)\sigma(X)]$$
$$\psi_{\mathrm{C}}^{-} = \mathbb{E}[\mu(X,0)] - \tfrac{\Lambda^2-1}{\Lambda}\phi(z_\tau)\mathbb{E}[e(X)\sigma(X)]$$

Finally, subtracting the sharp bounds on ψ_{T} and ψ_{C} as justified by Theorem 2 gives the conclusion of Corollary 3. □

C.6 Proof of Corollary 1

Proof. The partially identified set for ψ_{T} follows from the proof of Corollary 3, so we only need to show that the ZSB interval is asymptotically too wide. Let $\hat{\psi}_{\mathrm{T,ZSB}}^{+}$ be as in (5). Let $\bar{E}^* = \frac{1}{3} + \frac{1}{3}\mathbb{I}\{Y \le 0.27\sqrt{\sigma^2+1}\}$, and notice that $Y \mid Z=1 \sim \mathcal{N}(0, \sigma^2+1)$. Then a straightforward calculation using the Inverse Mills ratio formula gives:

$$\frac{\mathbb{E}[YZ/\bar{E}^*]}{\mathbb{E}[Z/\bar{E}^*]} = \frac{\phi(0.27)\sqrt{\sigma^2+1}}{2-\Phi(0.27)} > 0.276\sqrt{\sigma^2+1}$$

The strong law of large numbers implies $\liminf \hat{\psi}_{\mathrm{T,ZSB}}^{+} \ge \liminf(\mathbb{E}_n YZ/\bar{E}^*)/(\mathbb{E}_n Z/\bar{E}^*) > 0.27\sqrt{\sigma^2+1}$ almost surely. The lower bound follows by symmetry.

Note that the ZSB approach remains conservative even in the case $\sigma^2 = 0$, in which the identified set is $[\pm\frac{3}{4}\phi(z_{2/3})] \subset [\pm 0.276]$ and the ZSB AIPW approach we discuss in Section 4.2 is equivalent to this ZSB IPW approach. □

C.7 Proof of Lemma 1

Proof. If $\Lambda = 1$, the claim holds trivially, so we proceed assuming $\Lambda > 1$.

Let $\hat{W}_i = Z_i(1-\hat{e}(X_i))/\hat{e}(X_i)$. Since $\mathcal{L}_n$ is convex, computing the subdifferential optimality criterion for $\hat{\gamma}$ shows that there exists a vector $\Delta \in [\Lambda^{-1}, \Lambda]^n$ such that $\mathbb{E}_n \hat{W} g(X)(\Delta - 1) = 0$ and $\Delta_i = \Lambda^{\mathrm{sign}(Y_i - \hat{\gamma}^\top h(X_i))}$ whenever $Y_i \neq \hat{\gamma}^\top h(X_i)$.

We will first show that $\bar{e}_i^* := (1 + \Delta_i(1-\hat{e}_i)/\hat{e}_i)^{-1}$ solves (31). It is clear that $\bar{e}_i^*$ belongs to $\mathcal{E}_n(\Lambda)$. Moreover, we have $0 = \mathbb{E}_n \hat{W} g(X)(\Delta - 1) = \mathbb{E}_n g(X) Z/\bar{e}^* - \mathbb{E}_n g(X) Z/\hat{e}(X)$. Therefore $\bar{e}^*$ is a feasible solution to (31).

Optimality of $\bar{e}_i^*$ follows from Theorem 3.1 in [12]. The main technical requirement to apply that result is that $\mathbb{E}_n g(X) Z/\hat{e}(X)$ is in the relative interior of $\{\mathbb{E}_n g(X) Z/\tilde{e} : \tilde{e} \in \mathcal{E}_n(\Lambda)\}$. If $0 < \hat{e}_i < 1$ for all i, then this condition is satisfied by the open mapping theorem and the fact that $1/\hat{e}$ is an interior point of $1/\mathcal{E}_n(\Lambda)$.

Finally, we show the desired equivalence:

$$\frac{\mathbb{E}_n YZ/\bar{e}^*}{\mathbb{E}_n Z/\hat{e}(X)} =_i \frac{\mathbb{E}_n(Y - \hat{\gamma}^\top g(X))Z/\bar{e}^* + \mathbb{E}_n \hat{\gamma}^\top g(X) Z/\bar{e}^*}{\mathbb{E}_n Z/\hat{e}(X)}$$

$$
\begin{aligned}
&=_{ii} \frac{\mathbb{E}_n(Y - \hat{\gamma}^\top g(X))Z/\bar{e}^* + \mathbb{E}_n \hat{\gamma}^\top g(X) Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)} \\
&=_{iii} \frac{\mathbb{E}_n(Y - \hat{\gamma}^\top g(X))Z(1 + \Lambda^{\hat{V}}(1-\hat{e}(X))/\hat{e}(X)) + \mathbb{E}_n \hat{\gamma}^\top g(X) Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)}
\end{aligned}
$$

There, step i adds and subtracts the term $\mathbb{E}_n \hat{\gamma}^\top g(X) Z/\bar{e}^*$ in the numerator, step ii uses the fact that $\bar{e}^*$ "balances" $g(X)$, and step iii restates $\bar{e}^*$ in terms of $\hat{V}$. Since $\frac{\mathbb{E}_n YZ/\bar{e}^*}{\mathbb{E}_n Z/\hat{e}(X)}$ is the objective value from (31), this proves Lemma 1. □

C.8 Proof of Theorem 3 for linear quantiles

In this section, we give the proof of Theorem 3 under the assumption that $\hat{Q}_\tau(x,z) = \hat{\beta}(z)^\top h(x)$ for some "features" $h : \mathcal{X} \to \mathbb{R}^k$ with finite variance. Results for K-fold cross-fit linear estimates hold by viewing the folds as random and interacting the features with the fold identities to produce features in $\mathbb{R}^{k*K}$. We assume throughout that h contains an "intercept", i.e. $h_1(x) \equiv 1$. For simplicity, we only give the arguments for the estimator $\hat{\psi}^+_{\mathrm{T}}$. Results for other quantile balancing bounds follow by essentially the same arguments. Since this estimator only involves a single estimated quantile function, we will lighten the notation by writing $Q(x)$ and $\hat{Q}(x)$ in place of $Q_\tau(x,1)$ and $\hat{Q}_\tau(x,1)$.

C.8.1 Supporting lemmas

The proofs will make use of several easy lemmas.

Lemma 2. *Assume that Conditions 1 and 2 hold, and also that $Q(x) = \beta_0^\top h(x)$ for some $\beta_0 \in \mathbb{R}^d$. Further suppose that $\mathbb{E}[h(X)h(X)^\top]$ is finite and nonsingular. Let $\hat{\gamma}$ minimize the loss function $\mathcal{L}_n(\gamma) = \mathbb{E}_n \rho_\tau(Y - \gamma^\top h(X)) Z \frac{1-\hat{e}(X)}{\hat{e}(X)}$. Then $\hat{\gamma} \xrightarrow{p} \beta_0$.*

Proof. Define the population loss function $\mathcal{L}$ by:

$$
\begin{aligned}
\mathcal{L}(\gamma) &= \mathbb{E}_P[\rho_\tau(Y - \gamma^\top h(X)) Z \tfrac{1-e(X)}{e(X)}] \\
&= \mathbb{E}_P[(1-e(X))\mathbb{E}[\rho_\tau(Y - \gamma^\top h(X))|X, Z=1]]
\end{aligned}
$$

By Condition 2, β_0 is the unique minimizer of $\mathbb{E}[\rho_\tau(Y - \gamma^\top h(X))|X = x, Z = 1]$ for each $x \in \mathcal{X}$, and hence the unique minimizer of $\mathcal{L}$.

We will show that $\mathcal{L}_n$ converges to $\mathcal{L}$ pointwise in probability. For each $\gamma \in \mathbb{R}^d$, we have:

$$
\begin{aligned}
\mathcal{L}_n(\gamma) &= \mathbb{E}_n \rho_\tau(Y - \gamma^\top h(X)) Z \tfrac{1-\hat{e}(X)}{\hat{e}(X)} \\
&= \mathbb{E}_n \rho_\tau(Y - \gamma^\top h(X)) Z \tfrac{1-e(X)}{e(X)} + \mathbb{E}_n \rho_\tau(Y - \gamma^\top h(X)) Z(1/\hat{e}(X) - 1/e(X)) \\
&= \mathbb{E}_n \rho_\tau(Y - \gamma^\top h(X)) Z \tfrac{1-e(X)}{e(X)} + \mathcal{O}(||\rho_\tau(y - \gamma^\top h(x))||_{L^2(\mathbb{P}_n)} ||1/\hat{e} - 1/e||_{L^2(\mathbb{P}_n)}) \\
&= \mathcal{L}(\gamma) + o_P(1)
\end{aligned}
$$

where the last step is by the law of large numbers and Condition 1. The conclusion $\hat{\gamma} \xrightarrow{p} \beta_0$ now follow from general consistency results for convex M-estimators, e.g. Theorem 2.7 in [40]. □

Lemma 3. *Let $\hat{U}_i = \mathrm{sign}(Y_i - \hat{Q}(X_i))$. Then we have the inequality:*

$$
\hat{\psi}^+_{\mathrm{T}} \leq \frac{\mathbb{E}_n(Y - \hat{Q}(X))Z(1 + \Lambda^{\hat{U}}(1-\hat{e}(X))/\hat{e}(X)) + \mathbb{E}_n \hat{Q}(X) Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)} \tag{35}
$$

Proof. By Lemma 1, $\hat{\psi}_{\mathrm{T}}^{+}$ would be exactly equal to the right-hand side of (35) if $\hat{U}_i$ were replaced by $\hat{V}_i = \mathrm{sign}(Y_i - \hat{\gamma}_0 - \hat{\gamma}_1 \hat{Q}(X_i))$ where $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1)$ comes from $\mathcal{L}_n$ in Lemma 3. However, $(Y_i - \hat{Q}(X_i))\Lambda^{\hat{U}_i}$ is (weakly) larger than $(Y_i - \hat{Q}(X_i))\Lambda^{\hat{V}_i}$ for every i, since $\hat{U}_i$ exactly matches the sign of $Y_i - \hat{Q}(X_i)$ while $\hat{V}_i$ might not. Making this replacement index-by-index gives (35). □

Lemma 4. *Let $U_i = \mathrm{sign}(Y_i - Q(X_i))$. Then we have the inequality:*

$$\hat{\psi}_{\mathrm{T}}^{+} \geq \frac{\mathbb{E}_n(Y - \hat{\gamma}^\top h(X))Z(1 + \Lambda^U(1 - \hat{e}(X))/\hat{e}(X)) + \mathbb{E}_n \hat{\gamma}^\top h(X) Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)} \tag{36}$$

where $\hat{\gamma}$ is as in Lemma 2.

Proof. For the purposes of this proof, let $\bar{\psi}_{\mathrm{T}}^{+}$ be the solution to the "feature-balancing" problem:

$$\bar{\psi}_{\mathrm{T}}^{+} = \max_{\bar{e} \in \mathcal{E}_n(\Lambda)} \frac{\sum_{i=1}^n Y_i Z_i/\bar{e}_i}{\sum_{i=1}^n Z_i/\bar{e}_i} \quad \text{s.t.} \quad \mathbb{E}_n h(X) Z/\bar{e} = \mathbb{E}_n h(X) Z/\hat{e}(X).$$

It is clear that $\hat{\psi}_{\mathrm{T}}^{+} \geq \bar{\psi}_{\mathrm{T}}^{+}$, since the feature balancing problem has the same objective as the quantile balancing problem but faces more constraints. Lemma 1 implies the $\bar{\psi}_{\mathrm{T}}^{+}$ would be exactly equal to the right-hand side of (36) if we replaced U_i by $\hat{U}_i = \mathrm{sign}(Y_i - \hat{\gamma}^\top h(X_i))$. However, $(Y_i - \hat{\gamma}^\top h(X_i))\Lambda^{U_i}$ is (weakly) smaller than $(Y_i - \hat{\gamma}^\top h(X_i))\Lambda^{\hat{U}_i}$ for every i, since $\hat{U}_i$ exactly matches the sign of $Y_i - \hat{\gamma}^\top h(X_i)$ while U_i might not. Making this replacement index-by-index gives (36). □

C.8.2 Proof of main result

Now we prove Theorem 3(i), which we restate to make the regularity conditions more precise.

Theorem 3(i). (Sharpness for ψ_{T}^{+})
Assume Conditions 1, 2, and 3.(i). If $Q(x) = \beta_0^\top h(x)$ for some $\beta_0 \in \mathbb{R}^k$ and $\hat{\beta} \xrightarrow{p} \beta_0$, then $\hat{\psi}_{\mathrm{T}}^{+} = \psi_{\mathrm{T}}^{+} - o_P(1)$. However, even if $Q(x) \neq \beta^\top h(x)$ for any β, we still have $\hat{\psi}_{\mathrm{T}}^{+} \geq \psi_{\mathrm{T}}^{+} - o_P(1)$.

Proof. We start by proving the upper bound $\hat{\psi}_{\mathrm{T}}^{+} \leq \psi_{\mathrm{T}}^{+} + o_P(1)$ in the well-specified case. Lemma 3 gives the following upper bound on the quantile balancing estimator:

$$\hat{\psi}_{\mathrm{T}}^{+} \leq \frac{\mathbb{E}_n(Y - \hat{Q}(X))Z(1 + \Lambda^{\hat{U}}(1 - \hat{e}(X))/\hat{e}(X)) + \mathbb{E}_n \hat{Q}(X) Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)}.$$

Condition 1 implies $\mathbb{E}_n Z/\hat{e}(X) \xrightarrow{p} 1$, and the consistency of $\hat{\beta}$ implies $\mathbb{E}_n \hat{Q}(X) Z/\hat{e}(X) \xrightarrow{p} \mathbb{E}[Q(X)]$. To establish the upper bound, it remains to show $\mathbb{E}_n(Y - \hat{Q}(X))Z(1 + \Lambda^{\hat{U}}(1 - \hat{e}(X))/\hat{e}(X))$ converges to $\psi_{\mathrm{T}}^{+} - \mathbb{E}[Q(X)]$.

The first step is to replace the estimated propensity score $\hat{e}$ appearing in this quantity by the true nominal propensity score e. The Cauchy-Schwarz inequality and Condition 1 imply:

$$\begin{aligned}
\mathbb{E}_n(Y - \hat{Q}(X))Z\Lambda^{\hat{U}}\left(\tfrac{1-\hat{e}(X)}{\hat{e}(X)} - \tfrac{1-e(X)}{e(X)}\right) &= \mathcal{O}(||Y - \hat{\beta}^\top h(X)||_{L^2(\mathbb{P}_n)} \times ||1/\hat{e}(X) - 1/e(X)||_{L^2(\mathbb{P}_n)}) \\
&= \mathcal{O}_P((||Y||_{L^2(\mathbb{P}_n)} + ||\hat{\beta}^\top h(X)||_{L^2(\mathbb{P}_n)}) \times \varepsilon^{-2} ||\hat{e}(X) - e(X)||_{\mathcal{L}^\infty(\mathbb{P}_n)}) \\
&= \mathcal{O}_P((||Y||_{L^2(\mathbb{P}_n)} + ||Q(X)||_{L^2(\mathbb{P}_n)}) \times o_P(1)) \\
&= o_P(1)
\end{aligned}$$

Thus, $\mathbb{E}_n(Y - \hat{Q}(X))Z(1 + \Lambda^{\hat{U}} \frac{1-\hat{e}(X)}{\hat{e}(X)}) = \mathbb{E}_n(Y - Q(X))Z(1 + \Lambda^{\hat{U}} \frac{1-e(X)}{e(X)}) + o_P(1)$.

The next step is to replace $\hat{U}$ and $\hat{Q}(X)$ by $U = \mathrm{sign}(Y - Q(X))$ and $Q(X)$, respectively. For this, we employ a uniform convergence argument. For each $\beta \in \mathbb{R}^k$, define the function $f_\beta(x, y, z)$ by:

$$f_\beta(x, y, z) = (y - \beta^\top h(x)) z (1 + \Lambda^{\mathrm{sign}(y - \beta^\top h(x))} \tfrac{1-e(x)}{e(x)}).$$

Standard Glivenko-Cantelli (GC) preservation arguments (c.f. [32]) show that the class $\mathcal{F} = \{f_\beta : ||\beta - \beta_0|| \leq 1\}$ is GC, so we have the uniform convergence $\sup_{f \in \mathcal{F}} |\mathbb{E}_n f - Pf| = o_P(1)$. Moreover, the map $\beta \mapsto Pf_\beta$ is continuous at β_0, which can be seen by noticing that as $\beta \to \beta_0$, $f_\beta(x, y, z) \to f_{\beta_0}(x, y, z)$ for almost every (x, y, z) (exceptions occur when $y = \beta_0^\top x$, but Condition 2 implies this happens with probability zero) and then applying the dominated convergence theorem. Thus, we have:

$$\begin{aligned}
\mathbb{E}_n(Y - \hat{Q}(Z))Z(1 + \Lambda^{\hat{U}} \tfrac{1-e(X)}{e(X)}) &= \mathbb{E}_n f_{\hat{\beta}}(X, Y, Z) \\
&= Pf_{\hat{\beta}}(X, Y, Z) + o_P(1) \\
&= Pf_{\beta_0}(X, Y, Z) + o_P(1) \\
&= \mathbb{E}[(Y - Q(X))Z/\bar{E}_+] + o_P(1) \\
&= \psi_{\mathrm{T}}^+ - \mathbb{E}[Q(X)] + o_P(1)
\end{aligned}$$

Combining these various results gives $\hat{\psi}_{\mathrm{T}}^+ \leq \psi_{\mathrm{T}}^+ + o_P(1)$. This establishes the upper bound in the well-specified case.

Now we turn to the lower bound, $\hat{\psi}_{\mathrm{T}}^+ \geq \psi_{\mathrm{T}}^+ - o_P(1)$, beginning in the correctly-specified case. Lemma 4 lower bounds the quantile balancing estimator by a variant of the "feature balancing" estimator:

$$\hat{\psi}_{\mathrm{T}}^+ \geq \frac{\mathbb{E}_n(Y - \hat{\gamma}^\top h(X))Z(1 + \Lambda^U(1 - \hat{e}(X))/\hat{e}(X)) + \mathbb{E}_n \hat{\gamma}^\top h(X) Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)}.$$

We will show that this lower bound is at least $\psi_{\mathrm{T}}^+ - o_P(1)$. We may assume without loss of generality that $\mathbb{E}[h(X)h(X)^\top]$ is full rank, since excising features that are linear combinations of other ones has no effect on the feature balancing estimator. In the preceding display, the denominator $\mathbb{E}_n Z/\hat{e}(X)$ converges to one, so we can focus on the two terms in the numerator.

Since Lemma 2 implies that $\hat{\gamma}$ is consistent, exactly the same arguments from the upper bound show $\mathbb{E}_n \hat{\gamma}^\top h(X) Z/\hat{e}(X) \xrightarrow{p} \mathbb{E}[Q(X)]$. Moreover, the argument from the upper bound shows that $\hat{e}$ can be replaced by e in the expression $\mathbb{E}_n(Y - \hat{\gamma}^\top h(X))Z(1 + \Lambda^U(1 - \hat{e}(X))/\hat{e}(X))$. Some manipulation shows that $1 + \Lambda^U(1 - e(X))/e(X) = 1/\bar{E}_+$ almost surely, where $\bar{E}_+$ is the worst-case propensity score defined in Proposition 2. Therefore, we may write:

$$\begin{aligned}
\mathbb{E}_n(Y - \hat{\gamma}^\top h(X))Z(1 + \Lambda^U \tfrac{1-\hat{e}(X)}{\hat{e}(X)}) &= \mathbb{E}_n(Y - \hat{\gamma}^\top h(X))Z/\bar{E}_+ + o_P(1) \\
&= \mathbb{E}_n(Y - Q(X))Z/\bar{E}_+ + \mathcal{O}_P(||\hat{\gamma} - \beta_0||) + o_P(1) \\
&= \psi_{\mathrm{T}}^+ - \mathbb{E}[Q(X)] + o_P(1)
\end{aligned}$$

Combining these various results gives $\hat{\psi}_{\mathrm{T}}^+ \geq \psi_{\mathrm{T}}^+ - o_P(1)$. This establishes the lower bound in the well-specified case.

Finally, we extend the lower bound to the misspecified case. If $Q(x) \neq \beta^\top h(x)$ for any β, then we can lower bound $\hat{\psi}_{\mathrm{T}}^+$ by the feature-balancing estimator that balances $h(x)$ *and* the true quantile $Q(x)$. This brings us back to the well-specified case, so the preceding arguments show $\hat{\psi}_{\mathrm{T}}^+ \geq \psi_{\mathrm{T}}^+ - o_P(1)$. □

C.9 Proof of Theorem 3 for nonlinear quantiles

In this section, we prove Theorem 3 when quantiles are estimated by a nonlinear model. As in the case of linear quantiles, we will give the argument for the estimator $\hat{\psi}_{\mathrm{T}}^+$. As such, we will continue to use $\hat{Q}(x)$ and $Q(x)$ as shorthand for $\hat{Q}_\tau(x, 1)$ and $Q_\tau(x, 1)$.

C.9.1 Regularity conditions

As alluded to in Condition 3, we require nonlinear models to be estimated using a form of sample splitting called "cross-fitting" [10, 52, 41]. We briefly describe the procedure, mostly to fix notation.

The sample $\{(X_i, Y_i, Z_i)\}$ is divided into K disjoint "folds" $\mathcal{F}_1, \cdots, \mathcal{F}_K$ of approximately equal size. For each $k \in [K]$, a quantile estimate $\hat{Q}_{-k}$ is obtained using observations not in $\mathcal{F}_k$. Finally, we set $\hat{Q}_i = \sum_{k=1}^{K} \hat{Q}_{-k}(X_i)\mathbb{I}\{i \in \mathcal{F}_k\}$. In this way, no observation is used to obtain its own quantile estimate. In the extreme case where K is equal to the sample size, this is simply "leave-one-out" estimation. However, in cross-fitting, K is taken to be a fixed constant.

We also require the fitted quantiles $\hat{Q}_i$ to satisfy an additional regularity condition.

Condition N. *For some $\alpha, \beta > 0$, we have $\max_{i \le n} |\hat{Q}_i| = o_P(n^\alpha)$ and $\mathbb{P}(0 < |\hat{Q}_i - \hat{Q}_j| < n^{-\beta}$ for some $(i, j)) \to 0$.*

This condition rules out gross "outliers" in $\hat{Q}_i$ which are difficult to balance. The condition $\max_i |\hat{Q}_i| = o_P(n^\alpha)$ alone is not sufficient for this, because it is not an affine-invariant assumption. One can take an arbitrarily poorly-behaved estimate $\hat{Q}$ and scale it to be bounded by one without changing the estimator $\hat{\psi}^+_{\mathrm{T}}$. The separation requirement rules out this trick.

It is not hard to find examples of estimators which satisfy this condition. For example, under Conditions 1 and 2, Condition N is satisfied by any estimator whose fitted values $\{\hat{Q}_i\}$ only take values in the observed outcomes $\{Y_i\}$ (e.g. [37, 55, 2] will satisfy it with $\alpha = \frac{1}{2}$ and any $\beta > 2$).[1]

C.9.2 Supporting lemmas

To simplify the proof, we separate out a preliminary convergence result as a lemma. Throughout this proof and the next, we will use the following notation: for a function f, $\mathbb{E}_n^k f$ denotes the fold-k average $\frac{1}{|\mathcal{F}_k|}\sum_{i \in \mathcal{F}_k} f(X_i, Y_i, Z_i)$.

Lemma 5. *Assume Condition 1. Suppose $||\hat{Q}_{-k} - Q||_{L^2(P)} \xrightarrow{p} 0$ for each $k \in [K]$. Then $||\hat{Q} - Q||_{L^2(\mathbb{P}_n)} = o_P(1)$ and $\mathbb{E}_n \hat{Q} Z/\hat{e}(X) = \mathbb{E}[Q(X)] + o_P(1)$.*

Proof. Start with the first claim. For any $k \in [K]$, applying Markov's inequality conditionally on $\{(X_i, Y_i, Z_i)\}_{i \notin \mathcal{F}_k}$ gives $\mathbb{E}_n^k(\hat{Q}_{-k}(X) - Q(X))^2 = \mathcal{O}_P(||\hat{Q}_{-k} - Q||^2_{L^2(P)}) = o_P(1)$. Averaging over $k \in [K]$ gives the desired result.

For the second claim, write:

$$\begin{aligned}
\mathbb{E}_n \hat{Q} Z/\hat{e}(X) &= \mathbb{E}_n QZ/e(X) + \mathbb{E}_n(\hat{Q} - Q)Z/e(X) + \mathcal{O}(||\hat{Q}||_{L^2(\mathbb{P}_n)}||1/\hat{e} - 1/e||_{L^2(\mathbb{P}_n)}) \\
&= \mathbb{E}[Q(X)] + \mathcal{O}(||\hat{Q} - Q||_{L^2(\mathbb{P}_n)}/\varepsilon) + o_P(1) \\
&= \mathbb{E}[Q(X)] + o_P(1).
\end{aligned}$$

□

C.9.3 Proof of main result

Now we are ready to prove Theorem 3 for nonlinear quantile models in the case of the estimand ψ^+_{T}. We restate the result to make the quantile consistency assumption precise.

Theorem 3(ii). *Assume Conditions 1, 2, 3.(ii), and N. If $||\hat{Q}_{-k} - Q||_{L^2(P)} = o_P(1)$ for each $k \in [K]$, then $\hat{\psi}^+_{\mathrm{T}} = \psi^+_{\mathrm{T}} - o_P(1)$. However, even if $||\hat{Q}_{-k} - Q||_{L^2(P)} \not\to 0$, we still have $\hat{\psi}^+_{\mathrm{T}} \ge \psi^+_{\mathrm{T}} - o_P(1)$.*

[1] The upper bound follows from the well-known fact that the maximum of n i.i.d. observations from a distribution with finite variance has magnitude $o_P(n^{1/2})$. Therefore, $\max_i |\hat{Q}_i| \le \max_j |Y_j| = o_P(n^{1/2})$. For the lower bound, it suffices to show that $\mathbb{P}(\min_{i \ne j} |Y_i - Y_j| < n^{-\beta}) \to 0$ whenever $\beta > 2$. Let $F_Y(y) = P(Y \le y)$, and let $B < \infty$ be a uniform bound on $F'_Y(\cdot)$; this exists since $f(y|x,z)$ is uniformly bounded by Condition 2. Then $\mathbb{P}(\min_{i \ne j} |Y_i - Y_j| < n^{-\beta}) \le \mathbb{P}(\Delta \le Bn^{-\beta})$ where $\Delta = \min_{i \ne j} |F_Y(Y_i) - F_Y(Y_j)|$. Theorem 8.2 in [13] shows that $n^2\Delta \rightsquigarrow$ Exponential(1), so $\mathbb{P}(n^2\Delta \le Bn^{-(\beta-2)}) \to 0$.

Proof. We start by proving $\hat{\psi}_{\mathrm{T}}^{+} \leq \psi_{\mathrm{T}}^{+} + o_P(1)$ when the quantile model is consistent. This part of the proof follows roughly the same template as the corresponding proof in the linear case. Lemma 3 implies:

$$\hat{\psi}_{\mathrm{T}}^{+} \leq \frac{\mathbb{E}_n(Y - \hat{Q})Z(1 + \Lambda^{\hat{U}}(1 - \hat{e}(X))/\hat{e}(X)) + \mathbb{E}_n\hat{Q}Z/\hat{e}(X)}{\mathbb{E}_n Z/\hat{e}(X)}$$

where $\hat{U}_i = \mathrm{sign}(Y_i - \hat{Q}_i)$. Since $\mathbb{E}_n Z/\hat{e}(X) \xrightarrow{p} 1$ by Condition 1 and $\mathbb{E}_n\hat{Q}Z/\hat{e}(X) \xrightarrow{p} \mathbb{E}[Q(X)]$ by Lemma 5, it remains to show that $\mathbb{E}_n(Y - \hat{Q})Z(1 + \Lambda^{\hat{U}}(1 - \hat{e}(X))/\hat{e}(X))$ converges to $\psi_{\mathrm{T}}^{+} + o_P(1)$. By the same reasoning as in the linear case, we may replace $\hat{e}(X)$ by $e(X)$ in this quantity without changing its value much. Thus, we may write:

$$\begin{aligned}
\mathbb{E}_n(Y - \hat{Q})Z(1 + \Lambda^{\hat{U}}\tfrac{1-\hat{e}(X)}{\hat{e}(X)}) &= \mathbb{E}_n(Y - \hat{Q})Z(1 + \Lambda^{\hat{U}}\tfrac{1-e(X)}{e(X)}) + o_P(1) \\
&=_i \mathbb{E}_n(Y - Q(X))Z(1 + \Lambda^{\hat{U}}\tfrac{1-e(X)}{e(X)}) + \mathcal{O}(\varepsilon^{-1}||\hat{Q}(X) - Q(X)||_{L^2(\mathbb{P}_n)}) + o_P(1) \\
&=_{ii} \mathbb{E}_n(Y - Q(X))Z(1 + \Lambda^{\hat{U}}\tfrac{1-e(X)}{e(X)}) + o_P(1) \\
&=_{iii} \mathbb{E}_n(Y - Q(X))Z/\bar{E}_+ + \mathcal{O}(||Y - Q(X)||_{L^2(\mathbb{P}_n)}||Z\Lambda^{\hat{U}} - Z\Lambda^{U}||_{L^2(\mathbb{P}_n)}) + o_P(1) \\
&=_{iv} \psi_{\mathrm{T}}^{+} - \mathbb{E}[Q(X)] + \mathcal{O}_P(||Z\Lambda^{\hat{U}} - Z\Lambda^{U}||_{L^2(\mathbb{P}_n)}) + o_P(1)
\end{aligned}$$

Here, i adds and subtracts a term then applies Cauchy-Schwarz, ii applies Lemma 5 to conclude $||\hat{Q} - Q||_{L^2(\mathbb{P}_n)} = o_P(1)$, iii adds and subtacts $\mathbb{E}_n(Y - Q(X))Z/\bar{E}_+$ and applies Cauchy-Schwarz, and iv holds by Proposition 2 and the law of large numbers.

It remains to prove that $||Z\Lambda^{\hat{U}} - Z\Lambda^{U}||_{L^2(\mathbb{P}_n)} = o_P(1)$, or equivalently (up to constants) that $\mathbb{E}_n Z\mathbb{I}\{\hat{U} \neq U\} = o_P(1)$. For each $k \in [K]$, we may apply Chebyshev's inequality conditional on $\{(X_i, Y_i, Z_i)\}_{i \notin \mathcal{F}_k}$ to conclude:

$$\left|\mathbb{E}_n^k Z\mathbb{I}\{\hat{U} \neq U\} - \int z\mathbb{I}\{\mathrm{sign}(y - \hat{Q}_{-k}(x)) \neq \mathrm{sign}(y - Q(x))\}\,\mathrm{d}P(x, y, z)\right| = o_P(1)$$

The integral in the preceding display tends to zero in probability. To see this, recall that Condition 2 requires the conditional density $f(y|x, z)$ to be uniformly bounded by some $B < \infty$, so we may write:

$$\begin{aligned}
\int z\mathbb{I}\{\mathrm{sign}(y - \hat{Q}_{-k}(x)) \neq \mathrm{sign}(y - Q(x))\}\mathrm{d}P(x, y, z) &= \int_{\mathcal{X}} e(x) \int_{\hat{Q}_{-k}(x) \wedge Q(x)}^{\hat{Q}_{-k}(x) \vee Q(x)} f(y|x, 1)\,\mathrm{d}y\,\mathrm{d}P_X(x) \\
&\leq \int_X (1 - \varepsilon)B|\hat{Q}_{-k}(x) - Q(x)|\,\mathrm{d}P_X(x) \\
&\precsim ||\hat{Q}_{-k} - Q||_{\mathcal{L}^1(P)} \\
&\leq ||\hat{Q}_{-k} - Q||_{L^2(P)} \\
&= o_P(1).
\end{aligned}$$

Thus, $\mathbb{E}_n^k Z\mathbb{I}\{\hat{U} \neq U\} = o_P(1)$. Averaging over k gives $\mathbb{E}_n Z\mathbb{I}\{\hat{U} \neq U\} = o_P(1)$, and so $\hat{\psi}_{\mathrm{T}}^{+} \leq \psi_{\mathrm{T}}^{+} + o_P(1)$.

Now, we turn to the lower bound, which is substantially more difficult. We wish to show $\hat{\psi}_{\mathrm{T}}^{+} \geq \psi_{\mathrm{T}}^{+} - o_P(1)$ whether or not $\hat{Q}_{-k}$ converges to Q. For each $k \in [K]$, define $\hat{\psi}^{+}(k)$ by:

$$\hat{\psi}^{+}(k) = \max_{\bar{e}_k \in \mathcal{E}_{n,k}(\Lambda)} \mathbb{E}_n^k Y Z/\bar{e}_k \quad \text{subject to} \quad \begin{pmatrix} \mathbb{E}_n^k \hat{Q}_{-k} Z/\bar{e}_k \\ \mathbb{E}_n^k Z/\bar{e}_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}_n^k \hat{Q}_{-k} Z/\hat{e} \\ \mathbb{E}_n^k Z/\hat{e}(X) \end{pmatrix} \tag{37}$$

where $\mathcal{E}_{n,k}(\Lambda)$ is the projection of $\mathcal{E}_n(\Lambda)$ onto the coordinates in $\mathcal{F}_k$. Clearly, $\hat{\psi}_{\mathrm{T}}^{+} \times \mathbb{E}_n Z/\hat{e}(X) \geq \sum_k \hat{\psi}_{+}(k)|\mathcal{F}_k|/n$, so it suffices to prove $\hat{\psi}^{+}(k) \geq \psi_{\mathrm{T}}^{+} - o_P(1)$ for each k.

We will make some notational simplifications. The remainder of the proof will focus on showing $\hat{\psi}_{\mathrm{T}}^{+}(1) \geq \psi_{\mathrm{T}}^{+} - o_P(1)$. For convenience, we will assume $\mathcal{F}_1 = [n_1]$ where $n_1 \sim n/K$ almost surely. As an additional

simplification, we will assume that $\varepsilon/2 \leq \hat{e}_i \leq 1-\varepsilon/2$ for all i. Mechanically, this can always be done by "trimming" the estimated propensity score. Condition 1 implies the trimming has no effect in large samples, so it is only used as a theoretical device to simplify calculations. Finally, recall that we have defined $\hat{W}_i = Z_i(1-\hat{e}_i)/\hat{e}_i$.

We will construct an propensity vector $\bar{e}^*$ satisfying the constraints of (37) with the property that $\bar{\psi}^1 := \mathbb{E}_n^1 YZ/\bar{e}^*$ converges to ψ_{T}^+. Since $\hat{\psi}^+(1) \geq \bar{\psi}^1$, this will show $\hat{\psi}^+(1) \geq \psi_{\mathrm{T}}^+ - o_P(1)$. A natural first idea is to take the idealized propensity score $\bar{e}_i^* = (1+\theta_i(1-\hat{e}(X_i))/\hat{e}(X_i))^{-1}$, where $\theta_i = \Lambda^{U_i}$. This mimics the true worst-case propensity score, but uses $\hat{e}(X_i)$ in place of $e(X_i)$ to satisfy the odds-ratio constraint. It is not hard to see that this would result in a sharp estimate of ψ_{T}^+ by classic IPW logic.

$$\begin{aligned}\mathbb{E}_n^1 YZ(1+\theta\tfrac{1-\hat{e}(X)}{\hat{e}(X)}) &= \mathbb{E}_n^1 YZ(1+\theta\tfrac{1-e(X)}{e(X)}) + \mathcal{O}(||YZ||_{\mathcal{L}^1(\mathbb{E}_n)} \times ||1/e - 1/\hat{e}||_\infty) \\ &= \mathbb{E}_n^1 YZ/\bar{E}_+ + o_P(1) \\ &= \psi_{\mathrm{T}}^+ + o_P(1)\end{aligned} \tag{38}$$

However, this choice of $\bar{e}^*$ is not guaranteed to satisfy the "balancing" constraints of (37). Our construction perturbs this "ideal" choice to gain feasibility.

Our construction will be somewhat convoluted, so it is worth taking a moment to explain the high-level idea. First, we discard a small number of gross "outliers" to produce a set of "inliers" $\mathcal{I}_{j^*}$ whose fitted quantiles are relatively easy to balance. We then produce a feasible propensity $\bar{e}^*$ by assigning the outliers the nominal propensity score $\hat{e}(X_i)$ and perturbing the inliers' idealized propensity score by a small amount. We show the resulting lower bound $\bar{\psi}^1 = \mathbb{E}_n^1 YZ/\bar{e}^*$ is a consistent (albeit impractical) estimator of ψ_{T}^+.

We start by extracting a set of inliers $\mathcal{I}_{j^*} \subseteq [n_1]$ in the following fashion: set $\mathcal{I}_1 = [n_1]$, and for $2 \leq j \leq 4\beta+3$, recursively define $\mathcal{I}_j$ by:

$$\mathcal{I}_j = \{i \in \mathcal{I}_{j-1} \,:\, |\hat{Q}_i - \bar{Q}_{j-1}| \leq 2^{(j-1)} n^{-(j-1)/4}\} \tag{39}$$

where $\bar{Q}_{j-1} = (\sum_{i\in\mathcal{I}_{j-1}} \hat{W}_i\hat{Q}_i)/(\sum_{i\in\mathcal{I}_{j-1}} \hat{W}_i)$ is the weighted average value of $\hat{Q}_i$ within $\mathcal{I}_{j-1}$. We set $\mathcal{I}_{4\beta+4} = \emptyset$. Let j^* be the first stage in the above procedure at which an $n_1^{-1/8}$ fraction of the "weight" in $\mathcal{I}_j$ comes from outliers:

$$j^* = \min\left\{j \,:\, \frac{\sum_{i\in\mathcal{I}_j\setminus\mathcal{I}_{j+1}} \hat{W}_i}{\sum_{i\in\mathcal{I}_j} \hat{W}_i} \geq n_1^{-1/8}\right\} \tag{40}$$

It is easy to verify that j^* is well-defined (the set is not empty) whenever $Z_i = 1$ for some index $i \leq n_1$. For completeness, when that does not happen, we arbitrarily set $j^* = 4\beta+3$.

With this definition of j^*, we ensure the total "weight" on discarded outliers is asymptotically negligible. Since $\sum_{i\in\mathcal{I}_j\setminus\mathcal{I}_{j+1}} \hat{W}_i \leq n_1^{-1/8}\sum_{i\in\mathcal{I}_j} \hat{W}_i$ for all $j < j^*$, we have:

$$\sum_{i\notin\mathcal{I}_{j^*}} \hat{W}_i = \sum_{j<j^*}\sum_{i\in\mathcal{I}_j\setminus\mathcal{I}_{j+1}} \hat{W}_i \leq (4\beta+2)n_1^{-1/8}\sum_{i\in\mathcal{I}_1} Z_i(1-\hat{e}_i)/\hat{e}_i = o_P(n_1).$$

Therefore the inliers $\mathcal{I}_{j^*}$ will constitute most of the "weight" in the sample, i.e.

$$\frac{1}{n_1}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i = \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{W}_i - o_P(1) \geq (2/\varepsilon)\frac{1}{n_1}\sum_{i=1}^{n_1} Z_i - o_P(1) \tag{41}$$

We now perturb the idealized propensity for inliers in $\mathcal{I}_{j^*}$. Set $R_i = (\hat{Q}_i - \bar{Q}_{j^*})\mathbb{I}\{j^* \neq 4\beta+3\} + \mathbb{I}\{j^* = 4\beta+3\}$, and define $\lambda_1, \lambda_2, \alpha$ by:

$$\lambda_1 = \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i(1-\theta_i)}{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i|R_i|} \times (1 + \mathbb{I}\{j^* \neq 4\beta+3\})$$

$$\lambda_2 = \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i(1-\theta_i - \lambda_1 \mathbb{I}\{R_i \geq 0\})}{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i}$$
$$\alpha = \min\{1, (|\lambda_1| + |\lambda_2|)/(1-\Lambda^{-1})\}.$$

Finally, construct $\bar{e}^*$ by:

$$1/\bar{e}^* = \begin{cases} 1/\hat{e}_i & \text{if } i \notin \mathcal{I}_{j^*} \\ 1 + \frac{1-\hat{e}_i}{\hat{e}_i}(\alpha + (1-\alpha)(\theta_i + \lambda_1 \mathbb{I}\{R_i \geq 0\} + \lambda_2)) & \text{if } i \in \mathcal{I}_{j^*}. \end{cases}$$

We may verify that, with probability tending to one, we were successful in satisfying the constraints of (37).

The odds-ratio condition is satisfied as follows. If $\alpha = 1$ or $i \notin \mathcal{I}_{j^*}$, the odds ratio condition for i is satisfied trivially, so we proceed assuming $\alpha = \frac{|\lambda_1|+|\lambda_2|}{\Lambda^{-1}-1}$ and $i \in \mathcal{I}_{j^*}$. For the upper portion of the odds-ratio condition:

$$\begin{aligned}
\frac{(1-\bar{e}_i)/\bar{e}_i}{(1-\hat{e}_i)/\hat{e}_i} &= \alpha + (1-\alpha)\left(\theta_i + \lambda_i \mathbb{I}\{R_i \geq 0\} + \lambda_2\right) \\
&\leq \alpha\left(1 - \Lambda + \Lambda\right) + (1-\alpha)\left(\Lambda + |\lambda_1| + |\lambda_2|\right) \\
&= \Lambda + (|\lambda_1| + |\lambda_2|)\left(\frac{1-\Lambda}{1-\Lambda^{-1}} + (1-\alpha)\right) \\
&\leq \Lambda + (|\lambda_1| + |\lambda_2|)\left(-\Lambda + 1\right) \\
&\leq \Lambda
\end{aligned}$$

For the lower portion of the odds-ratio condition:

$$\begin{aligned}
\frac{(1-\bar{e}_i)/\bar{e}_i}{(1-\hat{e}_i)/\hat{e}_i} &= \alpha + (1-\alpha)\left(\theta_i + \lambda_i \mathbb{I}\{R_i \geq 0\} + \lambda_2\right) \\
&\geq \alpha\left(1 - \Lambda^{-1} + \Lambda^{-1}\right) + (1-\alpha)\left(\Lambda^{-1} - |\lambda_1| - |\lambda_2|\right) \\
&= \Lambda^{-1} + (|\lambda_1| + |\lambda_2|)\left(\frac{1-\Lambda^{-1}}{1-\Lambda^{-1}} + \alpha - 1\right) \\
&= \Lambda^{-1} + \alpha\left(|\lambda_1| + |\lambda_2|\right) \\
&\geq \Lambda^{-1}
\end{aligned}$$

We now proceed to balancing. If $\sum_{i=1}^{n_1} \hat{W}_i = 0$, we balance everything vacuously, so we proceed assuming otherwise. Our first substantive calculation verifies that $\bar{e}^*$ balances ones, i.e. $\mathbb{E}_n^1 Z/\bar{e}^* = \mathbb{E}_n^1 Z/\hat{e}(X)$:

$$\begin{aligned}
\mathbb{E}_n^1(Z/\bar{e}^* - Z/\hat{e}(X)) &= \frac{1}{n_1}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i(\alpha + (1-\alpha)(\theta_i + \lambda_1 \mathbb{I}\{R_I \geq 0\} + \lambda_2) - 1) \\
&= (1-\alpha)\frac{1}{n_1}\left(\lambda_2 \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i - \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i(1 - \theta_i - \lambda_1 \mathbb{I}\{R_i \geq 0\})\right) \\
&= 0
\end{aligned}$$

The final equality holds by the definition of λ_2.

To verify that $\bar{e}^*$ also balances $\hat{Q}_{-k}$ with probability tending to one, we use the following decomposition:

$$\begin{aligned}
\mathbb{E}_n^1 \hat{Q} Z(1/\bar{e}^* - 1/\hat{e}(X)) &= \mathbb{I}\{j^* \neq 4\beta + 3\} \times \bar{Q}_{j^*} \mathbb{E}_n^1 Z(1/\bar{e}^* - 1/\hat{e}(X)) && (42) \\
&+ \mathbb{I}\{j^* \neq 4\beta + 3\} \times \mathbb{E}_n^1 (\hat{Q} - \bar{Q}_{j^*}) Z(1/\bar{e}^* - 1/\hat{e}(X)) && (43) \\
&+ \mathbb{I}\{j^* = 4\beta + 3\} \times \mathbb{E}_n^1 \hat{Q} Z(1/\bar{e}^* - 1/\hat{e}(X)) && (44)
\end{aligned}$$

Since $\bar{e}^*$ balances constants, (42) is also zero.

The term (43) requires a lengthier argument. On the event $j^* \neq 4\beta+3$, we have $\mathbb{E}_n^1(\hat{Q}-\bar{Q}_{j^*})Z(1/\bar{e}^* - 1/\hat{e}(X)) = \mathbb{E}_n^1 RZ(1/\bar{e}^* - 1/\hat{e}(X))$, which the following calculation shows is identically zero when $j^* \neq 4\beta+3$:

$$\begin{aligned}
\mathbb{E}_n^1 RZ(1/\bar{e}^* - 1/\hat{e}) &=_i (1-\alpha)\frac{1}{n_1}\Bigg(\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i(1-\theta_i) - \lambda_1 \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i \geq 0\}\Bigg) \\
&=_{ii} (1-\alpha)\frac{1}{n_1}\Bigg(\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i(1-\theta_i) - \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i(1-\theta_i) \times \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i \geq 0\}}{\frac{1}{2}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i |R_i|}\Bigg) \\
&=_{iii} 0
\end{aligned}$$

Step i follows since $\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i = 0$ on the event $\{j^* \neq 4\beta+3\}$, step ii substitutes in the definition of λ_1, and step iii exploits the identity $\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i \geq 0\} = \frac{1}{2}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i |R_i|$:

$$\begin{aligned}
\frac{1}{2}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i |R_i| &= \frac{1}{2}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i \geq 0\} + \frac{1}{2}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i (-R_i)\mathbb{I}\{R_i < 0\} \\
&= \frac{1}{2}\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i \geq 0\} + \frac{1}{2}\Bigg(\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i - \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i < 0\}\Bigg) \\
&= \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i R_i \mathbb{I}\{R_i \geq 0\}
\end{aligned}$$

Thus, (43) $= 0$.

The final term (44) is more subtle. For any $i, j \in \mathcal{I}_{4\beta+3}$, $|\hat{Q}_i - \bar{Q}_{4\beta+2}|, |\hat{Q}_j - \bar{Q}_{4\beta+2}| \leq 2^{(4\beta+2)} n^{-(\beta+1/4)}$, so $|\hat{Q}_i - \hat{Q}_j| \precsim n^{-(\beta+1/4)}$. However, by Condition N, all distinct values of $\hat{Q}_i$ are separated by distance $n^{-\beta}$ with probability approaching one. Thus, with high probability, all values of $\hat{Q}_i$ in $\mathcal{I}_{4\beta+3}$ are identical to a constant $\hat{Q}_0$. In that case (44) $= \hat{Q}_0 \times \mathbb{E}_n^1 z(1/\bar{e}^* - 1/\hat{e}(X)) = \hat{Q}_0 \times 0$.

Combining these various cases yields the conclusion $\mathbb{E}_n^1 \hat{Q} Z(1/\bar{e}^* - 1/\hat{e}(X)) = 0$ with probability tending to one. Thus, $\bar{e}^*$ is (with high probability) feasible in (37).

Next, we check that $\bar{\psi}^1$ converges to ψ_{T}^+.

The first step in this consistency calculation is to prove that $\lambda_1 = o_P(1)$ and $\lambda_2 = o_P(1)$. Conditional on $\{(X_i, Z_i)\}_{i\leq N}$ and $\hat{Q}_{-k}$, the only randomness remaining in λ_1 comes from the θ_i values. For observations i with $Z_i = 1$, θ_i takes on the value Λ^{-1} with probability τ and Λ with probability $1-\tau$. Since $\tau = \Lambda/(\Lambda+1)$, simple algebra gives $\mathbb{E}[(1-\theta_i)|Z_i = 1, X_i] = 0$. Hence, $\mathbb{E}[\lambda_1|\mathcal{G}] = 0$ where $\mathcal{G} = \sigma(\{(X_i, Z_i)\}_{i\leq N}, \hat{Q}_{-k})$. Chebyshev's inequality implies $\lambda_1 = \mathcal{O}_P(\sqrt{\mathrm{Var}(\lambda_1|\mathcal{G})})$, so it suffices to show the conditional variance of λ_1 vanishes. Note that $\mathrm{Var}(\theta_i|\mathcal{G}) \leq c(\Lambda)$ for some constant $c(\Lambda)$, and $1-\theta_i$ is (conditionally) independent of $1-\theta_j$ when $i \neq j$. Therefore, we may write:

$$\mathrm{Var}(\lambda_1|\mathcal{G}) \precsim \frac{\sum_{i\in\mathcal{I}_{j^*}} (\hat{W}_i R_i)^2}{(\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i |R_i|)^2} \mathbb{I}\{j^* \neq 4\beta+3\} + \frac{\sum_{i\in\mathcal{I}_{j^*}} (\hat{W}_i R_i)^2}{(\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i |R_i|)^2} \mathbb{I}\{j^* = 4\beta+3\}. \tag{45}$$

Without loss of generality, assume that the exponent α in Condition N is zero. This can always be achieved by rescaling $\hat{Q}_i$ by $n^{-\alpha}$ and making a corresponding change to the lower bound β. Hence:

$$\begin{aligned}
\frac{\sum_{i\in\mathcal{I}_{j^*}} (\hat{W}_i R_i)^2}{(\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i |R_i|)^2} \mathbb{I}\{j^* \neq 4\beta+3\} &\leq_i \frac{\sum_{i\in\mathcal{I}_{j^*}} (\hat{W}_i R_i)^2}{(\sum_{i\in\mathcal{I}_{j^*}\setminus\mathcal{I}_{j^*=1}} \hat{W}_i |R_i|)^2} \mathbb{I}\{j^* \neq 4\beta+3\} \\
&\leq_{ii} \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i^2 R_i^2}{(\sum_{i\in\mathcal{I}_{j^*}\setminus\mathcal{I}_{j^*+1}} \hat{W}_i 2^{j^*} n^{-j^*/4})^2} \mathbb{I}\{j^* \neq 4\beta+3\}
\end{aligned}$$

$$
\begin{aligned}
&\leq_{iii} \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i^2 (2^{j^*} n^{-(j^*-1)/4})^2}{(\sum_{i\in\mathcal{I}_{j^*}\setminus\mathcal{I}_{j^*+1}} \hat{W}_i 2^{j^*} n^{-j^*/4})^2} \mathbb{I}\{j^* \neq 4\beta+3\} \\
&\leq_{iv} n^{1/2} \times \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i^2}{(n_1^{-1/8} \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i)^2} \mathbb{I}\{j^* \neq 4\beta+3\} \\
&\precsim_{v} \frac{n^{3/4}}{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i} \mathbb{I}\{j^* \neq 4\beta+3\} \\
&=_{vi} \mathcal{O}_P(n^{-1/4})
\end{aligned}
$$

Step i makes the denominator smaller by removing positive terms. Step ii is justified because, on the event $j^* \neq 4\beta+3$, $|R_i| \geq 2^{j^*} n^{-j^*/4}$ for all $i \in \mathcal{I}_{j^*} \setminus \mathcal{I}_{j^*+1}$ by (39). Step iii requires some more justification. If $j^* = 1$, then $R_i = |\hat{Q}_i - \bar{Q}_1| \leq 2\max_i |\hat{Q}_i| \leq 2$. If $1 < j^* \neq 4\beta+3$, then $|R_i| = |\bar{Q}_i - \bar{Q}_{j^*}| \leq |\bar{Q}_i - \bar{Q}_{j^*-1}| + |\bar{Q}_{j^*-1} + \bar{Q}_{j^*}| \leq 2^{j^*} n^{-(j^*-1)/4}$. In either case, $|R_i| \leq 2^{j^*} n^{-(j^*-1)/4}$. Step iv rearranges and invokes the definition of j^*, while step v uses the fact that $n_1 \leq n$ and our trimming assumption on $\hat{e}_i$ ensures the ratio of $\hat{W}_i/\hat{W}_i^2$ is bounded above and below when $Z_i \neq 1$. Step vi holds by (41).

The second term (45) can be controlled by a similar calculation. In fact, it is easier since $R_i = 1$ on the event $j^* = 4\beta+3$. That omitted calculation shows that $\mathrm{Var}(\lambda_1|\mathcal{G}) = o_P(1)$, and hence $\lambda_1 = o_P(1)$.

To show $\lambda_2 = o_P(1)$, start by writing λ_2 as the difference of two terms.

$$
\lambda_2 = \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i(1-\theta_i)}{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i} - \lambda_1 \frac{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i \mathbb{I}\{R_i \geq 0\}}{\sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i}
$$

The first term is $o_P(1)$ by the same argument as the one for λ_1 when $j^* = 4\beta+3$. The second term is the product of λ_1 and a quantity less than one. Since $\lambda_1 = o_P(1)$, this shows the second term is $o_P(1)$ as well.

Finally, we ready to show that $\bar{\psi}^1 = \psi_{\mathrm{T}}^+ - o_P(1)$. By (38), it suffices to show the distance between $\mathbb{E}_n^1 YZ/\bar{e}^*$ and $\mathbb{E}_n^1 YZ(1+\theta_i \frac{1-\hat{e}(X_i)}{\hat{e}(X_i)})$ is vanishing. We expand this difference as the sum of several terms:

$$
\begin{aligned}
\mathbb{E}_n^1 YZ/\bar{e}^* - \mathbb{E}_n^1 YZ(1+\theta_i \tfrac{1-\hat{e}(X)}{\hat{e}(X)}) &= \frac{1}{n_1} \sum_{i\notin\mathcal{I}_{j^*}} \hat{W}_i Y_i (1-\theta_i) && (46) \\
&+ \alpha \frac{1}{n_1} \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i Y_i (1-\theta_i) && (47) \\
&+ (1-\alpha) \frac{1}{n_1} \sum_{i\in\mathcal{I}_{j^*}} \hat{W}_i Y_i (\lambda_1 \mathbb{I}\{R_i \geq 0\} + \lambda_2). && (48)
\end{aligned}
$$

The term (46) can be handled as follows:

$$
\begin{aligned}
\left| \frac{1}{n_1} \sum_{i\notin\mathcal{I}_{j^*}} \hat{W}_i Y_i (\theta_i - 1) \right| &\precsim \left(\frac{1}{n_1} \sum_{i\notin\mathcal{I}_{j^*}} \hat{W}_i^2 \right)^{1/2} \left(\frac{1}{n_1} \sum_{i=1}^{n_1} |Y_i|^2 \right)^{1/2} \\
&\precsim \left(\frac{1}{n_1} \sum_{i\notin\mathcal{I}_{j^*}} \hat{W}_i \right)^{1/2} (\mathbb{E}[Y^2] + o_P(1)) \\
&= o_P(1)
\end{aligned}
$$

where we have used (41) in the final step. To analyze (47), use the fact that $|\lambda_1| \vee |\lambda_2| = o_P(1)$, and hence $\alpha = o_P(1)$. Since $\frac{1}{n_1}\sum_{i=1}^{n_1} \hat{W}_i |Y_i| = \mathcal{O}_P(1)$, the product vanishes. Finally, (48) is smaller than $\mathcal{O}_P(1) \times \frac{1}{n_1}\sum_{i=1}^{n_1} \hat{W}_i |Y_i| (|\lambda_1| + |\lambda_2|) = o_P(1)$.

Putting it all together, we have shown $\mathbb{E}_n^1 YZ/\bar{e}^* - \mathbb{E}_n^1 YZ(1+\theta_i \frac{1-\hat{e}(X)}{\hat{e}(X)}) = o_P(1)$, and hence $\hat{\psi}_+(1) \geq \bar{\psi}^1 = \psi_{\mathrm{T}}^+ - o_P(1)$. □

C.10 Proof of Theorem 4

In this section, we prove Theorem 4. For brevity, we only prove the validity of the bootstrap upper bound for ψ_{T}^{+}, and restrict our attention to the case where the nominal propensity score is estimated by logistic regression. By symmetry, the result extends to ψ_{T}^{-} and the other estimands of interest, and the proof can easily be modified to handle other parametric propensity models like probit regression. As in the proof of Theorem 3(i), we abbreviate $\hat{Q}_\tau(x,1)$ and $Q_\tau(x,1)$ by $\hat{Q}(x)$ and $Q(x)$, respectively, and results for K-fold cross-fit linear quantile estimates hold by viewing the folds as random and interacting the features with the fold identities to produce features in $\mathbb{R}^{k\times K}$.

For convenience, we restate the theorem in this special case to make the regularity conditions more precise.

Theorem 4(i). (Inference for $\psi_{\mathbf{T}}^{+}$)
Assume Conditions 1, 2, and 3.(i). Suppose that the nominal propensity score $\hat{e}$ is consistently estimated by logistic regression, and the covariate space $\mathcal{X}$ is bounded.[2] *Suppose the number of bootstrap samples $B \equiv B_n$ tends to infinity. Then we have:*

$$\liminf_{n\to\infty} \mathbb{P}(\psi_{\mathrm{T}}^{+} \le Q_{1-\alpha}(\{\hat{\psi}_b^{+}\}_{b\in[B]}) \ge 1-\alpha$$

for all $\alpha \in (0,1)$.

Proof. We begin by introducing some notation. For $i \le n$, let $(X_i^*, Y_i^*, Z_i^*) \sim \mathbb{P}_n$ be bootstrap observations, and let $\mathbb{E}_n^* = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i^*,Y_i^*,Z_i^*)}$ denote the bootstrap empirical distribution. Let $\hat{\theta}^*$ be the logistic regression coefficient vector estimated on the bootstrap dataset, and set $\hat{e}^*(x) = 1/[1+\exp(-x^\top\hat{\theta}^*)]$. Further define the bootstrap ZSB constraint set $\mathcal{E}_n^*(\Lambda)$ by:

$$\mathcal{E}_n^*(\Lambda) = \left\{ \bar{e} \in \mathbb{R}^n \,:\, \Lambda^{-1} \le \frac{\bar{e}_i/[1-\bar{e}_i]}{\hat{e}^*(X_i^*)/[1-\hat{e}^*(X_i^*)]} \le \Lambda \text{ for all } i \le n \right\}$$

and the bootstrap quantile balancing estimator $\hat{\psi}_*^{+}$ by:

$$\hat{\psi}_*^{+} = \max_{\bar{e}\in\mathcal{E}_n^*(\Lambda)} \frac{\sum_{i=1}^n Y_i Z_i/\bar{e}_i}{\sum_{i=1}^n Z_i/\bar{e}_i} \quad \text{s.t.} \quad \begin{pmatrix} \mathbb{E}_n^* \hat{Q}(X) Z/\bar{e} \\ \mathbb{E}_n^* Z/\bar{e} \end{pmatrix} = \begin{pmatrix} \mathbb{E}_n^* \hat{Q}(X) Z/\hat{e}^*(X) \\ \mathbb{E}_n^* Z/\hat{e}^*(X) \end{pmatrix}.$$

The estimated quantile $\hat{Q}$ in the definition of $\hat{\psi}_*^{+}$ comes from the original dataset, but the rest of the argument will go through even if it is re-estimated within each bootstrap sample.

The first step of the proof is to reduce our task to that of proving bootstrap consistency for a much simpler estimator under the assumption that $Q(x) = \beta_0^\top h(x)$. Define the bootstrap *feature balancing* estimator $\bar{\psi}_*^{+}$ by:

$$\bar{\psi}_*^{+} = \max_{\bar{e}\in\mathcal{E}_n^*(\Lambda)} \frac{\sum_{i=1}^n Y_i Z_i/\bar{e}_i}{\sum_{i=1}^n Z_i/\bar{e}_i} \quad \text{s.t.} \quad \mathbb{E}_n^* h(X) Z/\bar{e} = \mathbb{E}_n^* h(X) Z/\hat{e}^*(X).$$

Adding constraints to the balancing problem reduces the objective, so $\hat{\psi}_*^{+} \ge \bar{\psi}_*^{+}$ deterministically and the quantiles of the bootstrap distribution of $\hat{\psi}_*^{+}$ are above the quantiles of the bootstrap distribution of $\bar{\psi}_*^{+}$. A further reduction can be obtained by defining the estimator $\mathring{\psi}_*^{+}$ by:

$$\mathring{\psi}_*^{+} = \frac{\mathbb{E}_n^*(Y - \hat{\gamma}^{*\top} h(X)) Z (1 + \Lambda^{\mathrm{sign}(Y-Q(X))}(1-\hat{e}^*(X))/\hat{e}^*(X)) + \mathbb{E}_n^* \hat{\gamma}^{*\top} h(X) Z/\hat{e}^*(X)}{\mathbb{E}_n^* Z/\hat{e}^*(X)}$$

$$\hat{\gamma}^* = \operatorname*{argmin}_{\gamma\in\mathbb{R}^k} \mathbb{E}_n^* \rho_\tau(Y - \gamma^\top h(X)) Z \tfrac{1-\hat{e}^*(X)}{\hat{e}^*(X)}.$$

[2] This is needed for logistic regression to be compatible with the strong overlap requirement of Condition 1, although examining the proof shows it could be relaxed to the existence of certain exponential moments as in [60], Assumption C.1(3).

This estimator is not actually implementable as it depends on the true quantile Q through the term $\text{sign}(Y - Q(X))$. Still, the proof of Lemma 4 implies $\bar{\psi}_*^+ \geq \mathring{\psi}_*^+$, so it suffices to prove the validity of the percentile bootstrap for the estimator $\mathring{\psi}_*^+$.

The rest of this proof will be dedicated to proving the validity of the percentile bootstrap for the estimator $\mathring{\psi}_*^+$. Let θ_0 be the true logistic regression coefficient vector. For any $\theta \in \mathbb{R}^d, \beta \in \mathbb{R}^k, \psi \in \mathbb{R}$, define the estimating equation $m_{\theta,\beta,\psi}(x,y,z)$ by:

$$m_{\theta,\gamma,\psi}(x,y,z) = \begin{bmatrix} x(z - 1/(1+e^{-\theta^\top x})) \\ h(x)(\tau - \mathbb{I}\{\gamma^\top h(x) < 0\})ze^{\theta^\top x} \\ (y - \gamma^\top h(x))z(1 + \Lambda^{\text{sign}(y-Q(x))}e^{\theta^\top x} + \gamma^\top h(x)z(1+e^{-\theta^\top x}) - \psi z(1+e^{-\theta^\top x}) \end{bmatrix}.$$

and define $M(\theta,\gamma,\psi) = Pm_{\theta,\gamma,\psi}(X,Y,Z)$. If the linear quantile model is correctly specified (i.e. $Q(x) = \beta_0^\top h(x)$ for some $\beta_0 \in \mathbb{R}^d$), then $(\theta_0, \beta_0, \psi_{\text{T}}^+)$ solve the estimating equation $M(\theta_0, \gamma_0, \psi_{\text{T}}^+) = 0$. Meanwhile, the estimators $(\hat{\theta}^*, \hat{\gamma}^*, \mathring{\psi}_*^+)$ (approximately) solve the bootstrap estimating equation:

$$M_n^*(\hat{\theta}^*, \hat{\gamma}^*, \mathring{\psi}_*^+) = \mathbb{E}_n^* m_{\hat{\theta}^*,\hat{\gamma}^*,\mathring{\psi}_*^+}(X,Y,Z) = o_P(n^{-1/2}).$$

Therefore, we are in a position to apply the standard theory of bootstrap Z-estimators, at least in the correctly-specified case $Q(x) = \beta_0^\top h(x)$.

Specifically, we will apply Theorem 10.6 in [32], but prove bootstrap consistency by more direct means. Since logistic regression and weighted quantile regression are both convex optimization problems, the consistency $\hat{\theta}^* \xrightarrow{p} \theta_0$ and $\hat{\gamma}^* \xrightarrow{p} \beta_0$ follow from the bootstrap law of large numbers and Theorem 2.7 in [40]. From this, the result $\mathring{\psi}_*^+ \xrightarrow{p} \psi_{\text{T}}^+$ follows from the same argument used in the proof of Theorem 3.(i), with all applications of the law of large numbers replaced by the bootstrap law of large numbers. It remains to check Assumption (C) in Theorem 10.6 of [32]). Exercise 10.5.5 in [32] verifies this for the logistic regression estimating equation $x(z - 1/(1+e^{-\theta^\top x}))$. The quantile regression estimating equation eventually lives in the product of the VC class $\mathcal{F}$ and the smooth parametric class $\mathcal{G}$:

$$\mathcal{F} = \{(x,y,z) \mapsto (\tau - \mathbb{I}\{\gamma^\top h(x) < 0\} \,:\, \gamma \in \mathbb{R}^d,\}$$
$$\mathcal{G} = \{(x,y,z) \mapsto h(x)ze^{\theta^\top x} \,:\, ||\theta - \theta_0|| \leq 1\}.$$

Therefore, Assumption (C) follows from Theorem 9.15 in [32] and the dominated convergence theorem. The same arguments verify this condition for the final estimating equation.

Thus, we have shown that if $Q(x) = \beta_0^\top h(x)$ for some $\beta_0 \in \mathbb{R}^d$, then all the requirements for the proof of Theorem 10.16 in [32] are satisfied, and hence the percentile bootstrap based on $\mathring{\psi}_*^+$ will be asymptotically valid.

Finally, it remains to remove the assumption that the linear quantile model is correctly specified. If $Q(x) \neq \beta^\top h(x)$ for any $\beta \in \mathbb{R}^d$, then we may once again lower bound $\hat{\psi}_*^+$ by the estimator that balances $h(x)$ *and* the true quantile $Q(x)$. This brings us back to the well-specified case, and the preceding arguments imply the validity of the bootstrap upper confidence bound. □