Lee Bounds with a Continuous Treatment in Sample Selection

Ying-Ying Lee* Chu-An Liu†

October 29, 2025

Abstract

We study causal inference in sample selection models where a continuous or multivalued treatment affects both outcomes and their observability (e.g., employment or survey response). We generalize the widely used Lee (2009)'s bounds for binary treatment effects. Our key innovation is a "sufficient treatment value" assumption that imposes weak restrictions on selection heterogeneity and is implicit in separable threshold-crossing models, including monotone effects on selection. Our double debiased machine learning estimator enables nonparametric and high-dimensional methods, using covariates to tighten the bounds and capture heterogeneity. Applications to Job Corps and Civilian Conservation Corps (CCC) program evaluations reinforce prior findings under weaker assumptions.

Keywords: Average dose-response, debiased machine learning, multivalued treatment, nonseparable model, partial identification.
JEL Classification: C14, C21

*Corresponding author. Department of economics, University of California Irvine, Irvine, CA 92697, U.S.A. E-mail: yingying.lee@uci.edu. https://sites.google.com/site/yyleelilian. Tel: +1 9498244834. Fax: +1 9498242492.
†Institute of Economics, Academia Sinica, Taipei City 115, Taiwan.
E-mail: caliu@econ.sinica.edu.tw. https://chuanliu.weebly.com/.

1 Introduction

Sample selection is a common challenge in studying treatment effects. A classic question in empirical economics is estimating the effect of training programs on wages. There is sample selection because a training program not only affects wages via human capital accumulation, but also affects the chance that a worker is eventually employed and hence is selected into samples. We only observe wages of employed workers. If the estimation is based on a selected sample with observed outcomes, then one must isolate the effect on employment status (extensive margin) to learn about the effect on wages (intensive margin). Nonetheless in survey data, when the treatment affects response behavior, those who respond to the survey (i.e., who are selected into samples) in the treatment group are no longer comparable to respondents from the control group. Such problems of sample selection, attrition, or missing data arise in fields other than economics; for example, in medical studies, quality of life after assignment of a new drug is only observed if a patient does not die (truncation by death). In education, final test score is only observed if a student does not drop out.

Due to the non-random sample selection, the causal effect on the outcome is not point-identified without imposing further assumptions on functional forms or distributions; e.g., Heckman (1976, 1979), Imbens and Angrist (1994), Zhang et al. (2009), Honoré and Hu (2020), Chen and Roth (2023). Assuming the treatment variable is randomly assigned (conditional on observables), we build on the seminal work of Horowitz and Manski (1995) and Lee (2009), who bound the average causal effect of a binary treatment. The setup is fully non-parametric and hence allows for general heterogeneity. Many treatment or policy variables are continuous or discrete multivalued, e.g., hours in a social program, lottery prize, drug dosage, tuition subsidy, cash transfer, air pollution, etc. As Lee (2009)'s bound has been a common practice in empirical economics, we fill in the important gap to provide a corresponding tool to deal with sample selection in studying the causal effect of a continuous or multivalued treatment. The replication package of codes and data for our empirical applications is available on the authors' websites for practical implementation.

We provide the worst-case sharp bounds for the average treatment effect, or the average dose-response function, for *always-takers* who are selected into samples, or whose outcomes are observed, regardless of the treatment values they receive.[1] Note that the selected sample given a certain treatment value d consists of always-takers and *compliers* who are not selected given other treatment values $d' \neq d$. The key element of the bounds is the proportion of always-takers in the

[1] Always-takers are also known as always-observed, always-responders, always-employed, nonattriters, or survivors. The concept of always-takers is from the literature on imperfect compliance of treatment (Angrist et al., 1996), where "taking" is the taking of the treatment affected by an instrument, rather than selection into the sample affected by the treatment, considered in this paper.

selected d-treated sample, which is used to trim the observed outcomes for the worst-case lower (upper) bound when all always-takes' outcomes are smaller (or larger) than all compliers' outcomes. Recall that for a binary treatment, Lee (2009) assumes monotone treatment effects on selection, i.e., if a subject is selected in the control group, then it must be selected in the treatment group. Then only always-takers can be in the selected untreated sample and are identified. For a continuous or multivalued treatment, we propose a novel *sufficient treatment value* assumption on selection: if a subject is selected into samples when it receives the *sufficient treatment value*, then it remains selected when receiving *any* other treatment values. So we generalize the monotonicity assumption in Lee (2009) by assuming the sufficient treatment value to be zero for a binary treatment. Then the selected subjects who are treated with the sufficient treatment value are always-takers. So the probability of always-takers is identified by the minimum conditional selection probability over the treatment values.

For example, in a standard setting of survey attrition, if we find the response rate (conditional selection probability) given cash transfer is lowest at \$1,000, then our sufficient treatment value assumption is that everyone who responded to a survey question when receiving a cash transfer of \$1,000 would also have responded if there received any other values (always-takers). And there are some additional people (compliers) who only responded when they received some other transfer values and did not respond given \$1,000. We can interpret the sufficient treatment value \$1000 as the least-favored treatment (cash transfer) to induce selection into samples (responding to the survey) in the sense that if a subject is selected under the sufficient treatment value, then it is selected under any other treatment values.

In fact, the sufficient treatment value is implicit under a separable structural error in selection, or in a widely used class of latent variable threshold-crossing models (Vytlacil, 2002). This important observation suggests that the sufficient treatment value assumption is not restrictive, and the subpopulation of always-takers is a natural target under minimal assumptions. The associated always-takers are the largest subpopulation for whom we can partially identify the average effect of switching treatment over a range of values chosen by researchers for treatment intervention.

Moreover we allow for unconfoundness assumption in observational studies, and subjects with different pretreatment covariates can have different sufficient treatment values. So the information of covariates could potentially tighten the bounds and confidence intervals, or capture heterogenous effects that is not revealed without using the covariates, as supported by our empirical illustrations. Our bounds and asymptotic inference are robust to the extensive margin effect on selection, in the sense that when there is no selection bias or no extensive margin, our bounds contain the point-identified causal object. So we avoid pre-testing the treatment effect on selection.

We note that one might be interested in the partial (or marginal) effect defined as the derivative

of the average dose response function. However, we cannot bound such derivative by the same approach of Lee (2009). Instead, we bound the average effects of switching treatment values within a subset of the support of the treatment, or an average derivatives over two treatment values, which could be interesting in practice.

We illustrate our methodology by two applications. We find significant effects when incorporating covariates, while the bounds estimates without covariates show positive but insignificant effects. So incorporating covariates is useful to increases precision and captures heterogeneity. First, we revisit the Job Corps program, one of the largest federally funded job training programs in the U.S. The evaluation of these programs has been the focus of a substantive methodological literature, due to the high cost of about $14,000 on average per participant; see Schochet et al. (2008), Lee (2009), Flores et al. (2012), among others. The participants are exposed to different numbers of actual hours of academic and vocational training. Their labor market outcomes may differ if they accumulate different amounts of human capital acquired through different lengths of exposure. We try to understand whether hours of training raise wages by helping them find a job (extensive margin) or by increasing human capital that would affect the intensive margin effect on the outcome. We find that increasing the training from 1.5 weeks to 9 months increases log weekly earnings by at least 0.224 at 5% significance level for those always employed (*always-takers*).

The second application evaluates the Civilian Conservation Corps (CCC) in Aizer et al. (2024), who conduct the first lifetime evaluation of the largest federal youth employment program in U.S. history created to address high youth unemployment during the Great Depression. The Job Corps is a modern-era job training program that was modeled after the CCC and shares many features. We bound the effect of service duration on the age at death and strengthen the findings in Aizer et al. (2024). As differential attrition could bias the OLS estimates, they find that the effect of duration on longevity is consistently positive and statistically significant under various imputation approaches. Our bounds suggest that increasing duration from about 3 months to 14 months increases the average death age by at least 1.17 years at 5% significance level.

Another theoretical contribution of this paper is a weaker sufficiency assumption of a *sufficient set* of M treatment values, which is a useful approximation when the error of unobserved heterogeneity is non-separable in the selection equation. For example of $M = 2$, if a program participant is employed under *both* one week and fifteen months of training, then this participant is always employed. Such a weaker identification assumption results in a tradeoff with less informative (wider) bounds. Furthermore we utilize the well-known Fréchet-Hoeffding bounds for the discrete treatment *without* imposing any shape restrictions on selection.

We incorporate covariates, following Semenova (2024) on the generalized Lee bounds for the binary treatment. Our bound estimator is doubly debiased using an orthogonal moment func-

tion and cross-fitting, which enables nonparametric and machine learning methods to handle high-dimensional data, following the recent double debiased machine learning (DML) literature (Chernozhukov et al., 2018). Since the average dose-response function, or the mean potential outcome, and its bounds are functions of the continuous treatment, such non-regular estimand cannot be estimated at the regular root-n rate. We use a kernel function for localizing the continuous treatment as in Colangelo and Lee (2025).

The paper is organized as follows. Section 2 describes the sample selection model under the potential outcome framework, or equivalently a nonparametric non-separable structural model (e.g., Imbens and Newey (2009)). We discuss related literature. Section 3 presents the basic Lee bounds without covariates for a continuous/multivalued treatment under a sufficient value/set assumption on the treatment effect on selection. We give estimation and inference theory in Section 4. Section 6 incorporates the covariates and presents the DML inference. Section 5 and Section 7 present the empirical illustration on evaluating the Job Corps and the CCC programs. Appendix contains the main proofs of Theorems and Lemmas. In the online supplementary appendix, we present the proofs of Corollaries, and supplementary material for the empirical applications.

2 Sample selection model and related literature

The researcher chooses a compact subset of the support of the treatment variable D, denoted as $\mathcal{D}$, for treatment intervention. So we aim to learn about the intensive margin effect on the outcome of switching the treatment values over $\mathcal{D}$. For a continuous D, let $\mathcal{D} = [\underline{\mathcal{D}}, \overline{\mathcal{D}}]$; for a discrete D, let $\mathcal{D} = \left\{\underline{\mathcal{D}} =: d_1, d_2, ..., d_J := \overline{\mathcal{D}}\right\}$ with dimension J smaller or equal to the dimension of D.

For any treatment value $d \in \mathcal{D}$ that a subject receives, let Y_d be the continuous potential outcome, or the response function of d, and $S_d \in \{0, 1\}$ be the potential selection indicator for whether the subject's outcome is observed. If a subject is treated at d, i.e., $D = d$, let the selection status $S = S_d$ and the outcome $Y = Y_d$. The observed data vector $W = (D, S, S \cdot Y)$, so the outcome is recorded as zero if missing in the sample.

Following the literature and focusing on the sample selection bias, we begin with the independence Assumption 1 on treatment assignment. After the key results are established, we consider the standard conditional independence assumption given covariates in Section 6.

Assumption 1 (Independence) *D is independent of $\left\{(Y_d, S_d) : d \in \mathcal{D}\right\}$.*

Under Assumption 1, we identify the selection probability at treatment d, $\mathbb{P}(S_d = 1) = \mathbb{E}[S_d] = \mathbb{E}[S|D = d]$, and hence the average treatment effect (ATE) on selection $\mathbb{E}[S_d - S_{d'}] = \mathbb{E}[S|D = d] - \mathbb{E}[S|D = d']$, also known as the extensive margin effect of switching treatment from d' to d.

Assumption 1 also identifies the average outcome of the selected population at d, $\mathbb{E}[Y_d|S_d = 1] = \mathbb{E}[Y|S = 1, D = d]$. But $\mathbb{E}[Y_d|S_d = 1] - \mathbb{E}[Y_{d'}|S_{d'} = 1]$ is not causal if $\{S_d = 1\}$ and $\{S_{d'} = 1\}$ are different subpopulations. There are two common assumptions for $\mathbb{E}[Y_d|S_d = 1] - \mathbb{E}[Y_{d'}|S_{d'} = 1]$ to capture the intensive margin:[2] (i) Assume no ATE on selection (no extensive margin), or $\{S_d = 1\} = \{S_{d'} = 1\}$ have the same distribution for all $d, d' \in \mathcal{D}$. (ii) Assume missing at random (Rubin, 1976), so $\mathbb{E}[Y_d|S_d = 1] - \mathbb{E}[Y_{d'}|S_{d'} = 1] = \mathbb{E}[Y_d - Y_{d'}]$ is the population ATE on the outcome. But these two assumptions can be restrictive and unrealistic.

To understand the source of selection bias, note that the selected population $\{S_d = 1\}$ is composed of always-takers and d-compliers. Define *always-takers* AT:= $\{S_{d'} = 1 \text{ for all } d' \in \mathcal{D}\}$ to be those selected into samples regardless of the treatment value they receive over $\mathcal{D}$. Define *d-compliers* $\text{CP}_d := \{S_d = 1, S_{d'} = 0 \text{ for some } d' \in \mathcal{D}\}$ to be those induced to selection due to the treatment value d but are not selected at d'. Recall our goal of recovering the intensive margin effect of switching treatment between any values in $\mathcal{D}$. Always-takers are the common subpopulation that are selected into samples for *all* treatment values in $\mathcal{D}$, while d-compliers are missing in some selected samples with d', $\{D = d', S = 1\}$. As we never observe d-compliers in some samples with d', it is not possible to learn about their causal effect of switching treatment from d to d'. So without further assumptions such as functional form for extrapolation, we could only hope to learn about the causal effect for the always-takers. Therefore our target parameter is the mean potential outcome Y_d for always-takers,

$$\beta_d := \mathbb{E}[Y_d|\{S_{d'} = 1 \text{ for all } d' \in \mathcal{D}\}]$$

for $d \in \mathcal{D}$. Note that the definition of always-takers depends on the range of treatment values of interest $\mathcal{D}$. So the notation β_d should depend on $\mathcal{D}$ that is suppressed for simplicity.

When the treatment is continuous, β_d is known as always-takers' *average dose-response function*. When the treatment variable is binary, i.e., $\mathcal{D} = \{0, 1\}$, always-takers' ATE is $\beta_1 - \beta_0 = \mathbb{E}[Y_1 - Y_0 \mid S_1 = 1, S_0 = 1]$, studied in Lee (2009). Nonetheless we cannot determine whether a specific subject is an always-taker or a complier. So we take a bound/partial identification approach following Lee (2009) and Zhang and Rubin (2003). Our bounds estimation and inference are robust to the extensive margin effect on selection. Gerard et al. (2020) propose similar worst-case sharp bounds with manipulation-robust inference in regression discontinuity designs. We also discuss how the sharp bounds might be tightened by our sufficient value/set assumption or the covariates (e.g., Fan and Park (2010)).

[2]We can decompose $\mathbb{E}[Y_d|S_d = 1] - \mathbb{E}[Y_{d'}|S_{d'} = 1] = InM + ExM$, where $InM := \mathbb{E}[Y_d|S_d = 1] - \mathbb{E}[Y_{d'}|S_d = 1]$ from the intensive margin and $ExM = \mathbb{E}[Y_{d'}|S_d = 1] - \mathbb{E}[Y_{d'}|S_{d'} = 1]$ from the extensive margin that cause the selection bias. Because $\mathbb{E}[Y_{d'}|S_d = 1]$ is not identified, we cannot disentangle InM and ExM.

There are recent developments in sample selection models using the bound/partial identification approach. Honoré and Hu (2020) and Honoré and Hu (2022) consider parametric and semiparametric structural models. In addition to the concern of misspecification, the parametric selection equation often relies on the monotonicity assumption. Estrada (2024) studies spillover effects under sample selection, which can be viewed as Lee bound for the multivalued treatment effect. Heiler (2024) and Olma (2021) study Lee bounds for the conditional average binary treatment effect given a continuous covariate, which is a non-regular estimand as ours. Kroft et al. (2024) extend Lee bounds for multilayered sample selection to account for training affecting workers sorting to firms. Kline and Santos (2013) assess the sensitivity of empirical conclusions among a continuum of assumptions ordered from strongest (missing at random) to weakest (worst-case bounds). See the literature reviews on partial identification in Ho and Rosen (2017), Molinari (2020), Kline and Tamer (2023), and references therein.

Alternatively another literature on sample selection models with *exclusion restrictions* or partial randomization assumes a variable Z in the selection equation of S that does not enter Y. In Heckman's classic sample selection model ("Heckit"), the structural equations are linear in (D, Z) and are separable in the normally distributed errors. Ahn and Powell (1993), Das et al. (2003), and Escanciano et al. (2016) consider more nonparametric settings. The standard sample selection model is generally not point-identified without exclusion restrictions. Nevertheless, it has been noted to be difficult to find a credible instrument Z in practice; see, for example, Honoré and Hu (2020). DiNardo et al. (2021) proactively create an instrument by ex-ante randomizing the participants of the Moving to Opportunity experiment to differing intensity of follow-up. Behaghel et al. (2015) use information on the number of calls made to each individual before responding to the survey to identify the ATE of a binary treatment for a subpopulation of respondents, in the absence of instruments. See also Garlick and Hyman (2022) for evaluation of various sample selection correction methods and references therein. Chen and Roth (2023) discuss problems of log-like transformations with zeros and propose some solutions. We capture general heterogenous causal effects without exclusion restrictions and free from misspecification.

3 Lee Bounds

We establish the sharp worst-case bounds for β_d with a continuous/multivalued treatment, building on Horowitz and Manski (1995) and Lee (2009) for a binary treatment. The upper bound is when all always-takers' wages are larger than all d-compliers' wages. Denote the fraction of always-takers among the selected subjects with treatment d as p_d. Then all always-takers' wages are larger than the $(1 - p_d)$-quantile of the observed wage distribution at d. So we can construct the worst-case

bound by trimming the upper and lower tails of the observed outcome distribution by p_d. Next we present the well known worst-case bounds based on a given p_d, and then we propose identification strategies of p_d for the continuous and multivalued treatment, which is new to the literature.

Independent treatment Assumption 1 identifies the selection probability at d by the conditional selection probability given d, $\mathbb{P}(S_d = 1) = \mathbb{E}[S|D = d] =: s(d)$. If the proportion of always-takers $\pi_{\text{AT}} := \mathbb{P}(S_{d'} = 1 : d' \in \mathcal{D})$ is known, then the fraction of always-takers among the selected subjects with treatment d is $p_d = \mathbb{P}(\text{AT}|S_d = 1) = \pi_{\text{AT}}/s(d)$. Let $Q^d(u)$ be the u-quantile of $Y|D = d, S = 1$. Then the bounds of β_d are the trimmed means:

$$\rho_{dU}(\pi_{\text{AT}}) := \mathbb{E}[Y|Y \geq Q^d(1 - \pi_{\text{AT}}/s(d)), D = d, S = 1] \tag{1}$$

for the upper bound and $\rho_{dL}(\pi_{\text{AT}}) := \mathbb{E}[Y|Y \leq Q^d(\pi_{\text{AT}}/s(d)), D = d, S = 1]$ for the lower bound. The key element of the bounds is the proportion of always-takers π_{AT}. Once we identify π_{AT} and hence $p_d = \pi_{\text{AT}}/s(d)$, we can consistently estimate the bounds.

Note that when there is no complier, the selected sample is composed of always-takers only, so $s(d) = \mathbb{P}(S_{d'} = 1 : d' \in \mathcal{D})$ is constant and $p_d = \pi_{\text{AT}}/s(d) = 1$ for all $d \in \mathcal{D}$. That is, there is no extensive margin, and $\beta_d = \rho_{dU}(\pi_{\text{AT}}) = \rho_{dL}(\pi_{\text{AT}}) = \mathbb{E}[Y|D = d, S = 1]$ is point-identified.

Section 3.1 and Section 3.2 present novel sufficient assumptions on the treatment effect on selection to identify the proportion of always-takers π_{AT}. Section 3.3 discusses the connection with the structural selection model. For expositional ease, we focus on the upper bound.

3.1 Identification of the proportion of always-takers

The key identification Assumption 2 requires one *sufficient treatment value* d_{AT} such that if a subject is selected at d_{AT} then it will be selected at any treatment values.

Assumption 2 (Sufficient treatment value) *There exists a treatment value $d_{AT} \in \mathcal{D}$ such that $S_d \geq S_{d_{AT}}$ almost surely (a.s.) for any $d \in \mathcal{D}$.*

Assumption 2 essentially assumes that always-takers are the selected d_{AT}-receipts, $\{S_d = 1 : d \in \mathcal{D}\} = \{S_{d_{\text{AT}}} = 1\}$. Together with Assumption 1, the proportion of always-takers $\pi_{\text{AT}} = \mathbb{P}(S_{d_{\text{AT}}} = 1) = \mathbb{E}[S|D = d_{\text{AT}}] =: s(d_{\text{AT}})$ is identified.

Assuming $s(\cdot)$ to be continuous for a continuous D, the extreme value theorem implies that $d_{\text{AT}} = \arg\min_{d \in \mathcal{D}} s(d)$ exists and $\pi_{\text{AT}} = s(d_{\text{AT}}) = min_{d \in \mathcal{D}} s(d)$ can be estimated from the data. Notice that d_{AT} depends on $\mathcal{D}$ that is a range of treatment values chosen by the researcher for policy intervention. A larger $\mathcal{D}$ results in a smaller π_{AT}, which trims more observations, and wider bounds.

Importantly Assumption 2 allows any shape of the effect on selection. Consider an example of $d_{\text{AT}} = 1$. A 2-complier can have $\{S_{d_{\text{AT}}} = S_1 = 0, S_2 = 1, S_3 = 0\}$. In contrast, a stronger monotonicity assumption, which assumes $S_{d'} \geq S_d$ a.s. for any $d' > d$, rules out this event because it requires $S_3 = 1$ if $S_2 = 1$.

Lemma 1 formally presents our generalized Lee bounds with a continuous treatment or a discrete multivalued treatment.

Lemma 1 *Let Assumption 1 hold. Assuming $s(d) > 0$ for $d \in \mathcal{D}$, then $\beta_d \in [\rho_{dL}(\pi_{AT}), \rho_{dU}(\pi_{AT})]$ with $\pi_{AT} = \mathbb{P}(S_{d'} = 1 : d' \in \mathcal{D})$ given in equation (1). Further let Assumption 2 hold. Then we identify $\pi_{AT} = s(d_{AT})$, the sharp bounds for $\beta_d \in [\rho_{dL}(s(d_{AT})), \rho_{dU}(s(d_{AT}))]$, and $\beta_{d_{AT}} = \mathbb{E}[Y|D = d_{AT}, S = 1]$.*

Remark 1 (Binary treatment in Lee (2009)) Assumption 2 includes the familiar monotonicity assumption for a binary treatment in Lee (2009) that assumes $S_1 \geq S_0$ a.s., i.e., if a subject in the control group $\{D = 0\}$ is selected, then it remains selected if it was in the treated group $\{D = 1\}$, i.e, $d_{\text{AT}} = 0$. So defiers (0-compliers) are excluded (Imbens and Angrist, 1994). Then Lemma 1 implies Proposition 1a in Lee (2009): the upper bound of the always-takers' ATE, $\beta_1 - \beta_0 = \mathbb{E}[Y_1 - Y_0|S_0 = S_1 = 1]$, is $\rho_{1U}(s(0)) - \beta_{d_{\text{AT}}} = \mathbb{E}[Y|Y \geq Q^d(1 - s(0)/s(1)), D = 1, S = 1] - \mathbb{E}[Y|D = 0, S = 1]$ and the lower bound is $\rho_{1L}(s(0)) - \beta_{d_{\text{AT}}} = \mathbb{E}[Y|Y \leq Q^d(s(0)/s(1)), D = 1, S = 1] - \mathbb{E}[Y|D = 0, S = 1]$.

We remark that if we are only interested in two values of the continuous treatment, then the identification of the bounds for the binary treatment in Lee (2009) can be directly applied to the case of two continuous treatment values. But policy makers rarely consider only two values, and the always-takers at the two values could be different from the always-takers at another values; for example, $\{S_{d_1} = 1, S_{d_2} = 1\} \neq \{S_{d_3} = 1, S_{d_4} = 1\}$. New challenge in identification arises when we consider many treatment values.

3.2 Sufficient set assumption

We introduce a weaker sufficient set Assumption 3 that does not assume one sufficient treatment value d_{AT} but assumes a set $\mathcal{D}_M$ of M treatment values, which includes Assumption 2 as a special case with $M = 1$. Assumption 3 essentially assumes that always-takers $\{S_d = 1 : d \in \mathcal{D}\} = \{S_d = 1 : d \in \mathcal{D}_M\}$.

Assumption 3 (Sufficient set) *There exists a set of treatment values $\mathcal{D}_M := \{d_1, d_2, ..., d_M\} \subseteq \mathcal{D}$ such that $S_d \geq \min_{d' \in \mathcal{D}_M} S_{d'}$ a.s. for any $d \in \mathcal{D}$.*

To see how Assumption 3 is weaker than Assumption 2 or a larger M is weaker, consider an example: Under Assumption 3 with $M = 3$ and $\mathcal{D}_3 = \{1, 2, 3\}$, $\{S_1 = 0, S_2 = 1, S_3 = 0\}$ and $\{S_1 = 1, S_2 = 0, S_3 = 1\}$ are both possible values for a complier. But Assumption 2 with $d_{\text{AT}} = 1$ does not allow a complier to take $\{S_1 = 1, S_2 = 0, S_3 = 1\}$. What Assumption 3 does not allow is this event $\{S_1 = 1, S_2 = 1, S_{2.5} = 0, S_3 = 1\}$ for example. But assuming a larger $M = 4$ and $\mathcal{D}_4 = \{1, 2, 2.5, 3\}$ would allow a complier to take that.

However, the proportion of always-takers $\pi_{\text{AT}} = \mathbb{P}(S_d = 1 : d \in \mathcal{D}_M)$ for $M \geq 2$ is not point-identified as we cannot observe the M potential outcomes $\{S_d : d \in \mathcal{D}_M\}$ at the same time. We can use the lower bound of π_{AT} for trimming, because $\rho_{dU}(\pi_{\text{AT}})$ is decreasing in π_{AT} and $\rho_{dL}(\pi_{\text{AT}})$ is increasing in π_{AT}. Theorem 1 below provides the Lee bounds formally.

Specifically, consider a practical example of $M = 2$ and $\mathcal{D}_2 = \{\underline{\mathcal{D}}, \overline{\mathcal{D}}\}$. It implies that if a subject is selected at the boundary $\underline{\mathcal{D}}$ and $\overline{\mathcal{D}}$, then this subject must be selected at any treatment value in $\mathcal{D}$. This example of $M = 2$ includes the special case when the selection's response is a concave function of treatment D. The single-peaked pattern may fit well the law of marginal returns. Therefore $S_{\underline{\mathcal{D}}} = 1$ and $S_{\overline{\mathcal{D}}} = 1$ if and only if $S_d = 1$ for all $d \in \mathcal{D}$. This observation gives the insight to use the well-known Fréchet-Hoeffding bounds for $\mathbb{P}(S_{\underline{\mathcal{D}}} = 1, S_{\overline{\mathcal{D}}} = 1)$ given in Theorem 1. The bounds resemble the Fréchet-Hoeffding bounds for $\mathbb{P}(S_0 = 1, S_1 = 1)$ for the binary treatment without shape restrictions, as shown in Heckman et al. (1997). We require the sufficient set Assumption 3 due to the continuous treatment variable. It is important to note that when the treatment is multivalued discrete with support $\mathcal{D} = \mathcal{D}_M$, Assumption 3 holds by construction and is dropped. So Theorem 1 provides the sharp bounds without restricting the selection response of the multivalued treatment.

Theorem 1 *Let $\mathcal{D}_M = \{d_1, d_2, ..., d_M\} \subseteq \mathcal{D}$ with a fixed dimension M. Let Assumption 1 hold. Then*

$$\pi_L^M := \max\left(\sum_{d \in \mathcal{D}_M} s(d) - M + 1, 0\right) \leq \mathbb{P}(S_d = 1 : d \in \mathcal{D}_M) \leq \min_{d \in \mathcal{D}_M} s(d) =: \pi_U^M.$$

Further let Assumption 3 hold. Then $\pi_{AT} = \mathbb{P}(S_d = 1 : d \in \mathcal{D}_M) \in \left[\pi_L^M, \pi_U^M\right]$ and $\beta_d \in \left[\rho_{dL}(\pi_L^M), \rho_{dU}(\pi_L^M)\right]$. The bounds are sharp.

A final goal is to derive the bounds without restrictions on the selection response of a continuous treatment, i.e., dropping Assumption 3. To make progress, we may assume that the treatment effect is a piecewise constant function $\beta_d = \sum_{m=1}^{M-1} \beta_{d_m} \mathbf{1}\{d \in [d_m, d_{m+1})\}$. Then treatment can be effectively discretized and $\beta_{d_m} \in [\rho_{d_m L}(\pi_L^M), \rho_{d_m U}(\pi_L^M)]$ for $m = 1, .., M-1$. In practice, one might discretize the continuous treatment variable into M-multivalued variable for sensitivity analysis.

In general and in theory, we'd like M to be large to allow for a general non-separable nonparametric structural selection model, as discussed in Section 3.3. However, the bounds can be wide or less informative for a large M. As shown in Theorem 1, π_L^M can be small, unless $s(d)$ is close to one when there is selection bias. We illustrate this tradeoff in Section 5 by evaluating the Job Corps program and discuss how to choose M.

Remark 2 (Binary outcome) Kroft et al. (2024) provide the Lee bounds with a binary outcome and a binary treatment. We can extend their bounds for a binary outcome to a continuous treatment: $\rho_{dL}(\pi_L^M) = \max\{0, 1 - \mathbb{P}(Y = 0|S = 1, D = d)/p_d\}$ and $\rho_{dU}(\pi_L^M) = \min\{1, \mathbb{P}(Y = 1|S = 1, D = d)/p_d\}$, where $p_d = \pi_L^M/s(d)$.[3] We focus on the continuous outcome in this paper and develop the inference for the binary outcome in a separate paper.

3.3 Structural selection equation

To understand how the sufficient treatment value Assumption 2 and the sufficient set Assumption 3 impose conditions on the heterogeneity in the structural selection equation, we discuss its relationship with the threshold-crossing model in Vytlacil (2002). Recall that our potential outcome framework is equivalent to the structural equation $S = \mathbf{1}\{q(D, \eta) \geq 0\}$ with unobserved non-separable and multi-dimensional error η. The structural equation q is nonparametric and model-free. Then we can write the potential variable $S_d = \mathbf{1}\{q(d, \eta) \geq 0\}$.

Assumption 2′ (Latent index selection model) *(i) Let $S = \mathbf{1}\{q(D) \geq \eta\}$, where $q(d)$ is measurable and nontrivial function of d. (ii) For a continuous treatment with a compact $\mathcal{D}$, there exists $d_{AT} = \arg\inf_{d\in\mathcal{D}} q(d) \in \mathcal{D}$.*

Assumption 2′ implies Assumption 2. For a multivalued treatment with a finite countable $\mathcal{D}$, d_{AT} exists under Assumption 2′(i). For a continuous D, assuming $q(\cdot)$ in Assumption 2′(i) to be continuous, the extreme value theorem implies (ii). This important observation suggests our new sufficient treatment value Assumption 2 not restrictive and implied by a common threshold-crossing model with a separable error, or the latent index selection model in Vytlacil (2002).[4]

In the selected sample at d, $\{S = 1, D = d\} = \{S_d = 1, D = d\}$, the subpopulation $\{S_d = 1\} = \{\eta \leq q(d)\} = \text{AT} \cup \text{CP}_d$, where always-takers $\text{AT} = \{\eta \leq q(d_{\text{AT}})\}$ and d-compliers $\text{CP}_d =$

[3] $\mathbb{P}(Y = 1|S = 1, D = d) = \mathbb{P}(Y = 1|\text{AT}, D = d)p_d + \mathbb{P}(Y = 1|\text{CP}_d, D = d)(1 - p_d)$. The bounds on $\mathbb{P}(Y = 1|\text{AT}, D = d) = \mathbb{E}[Y_d|\text{AT}]$ are obtained by the worst-case bounds of $\mathbb{P}(Y = 1|\text{CP}_d, D = d) \in [0, 1]$.

[4] Vytlacil (2002) shows that the latent index selection model Assumption 2′(i) is equivalent to the local average treatment effect (LATE) model with Independence (as our Assumption 1) and Monotonicity assumptions in Imbens and Angrist (1994). The LATE Monotonicity assumes the orders of S_d to be the same for everyone, i.e., for all $(d, d') \in \mathcal{D} \times \mathcal{D}$, either $S_d \geq S_{d'}$ *a.s.*, or $S_d \leq S_{d'}$ *a.s.* Our Assumption 2 (or Assumption 2′(ii)) is weaker than such Monotonicity assumption and only requires the sufficient treatment value d_{AT} (or a minimizer of $q(d)$) exists.

$\{q(d_{\text{AT}}) < \eta \leq q(d)\}$. The sufficient treatment value has meaningful economic interpretation. For example, Behaghel et al. (2015) interpret η as the individual reluctance to respond to surveys and call the compliers as the marginal respondents. Then the sufficient treatment value can be the least-favored treatment value to induce responding to surveys. So if subjects are willing to respond to surveys when receiving d_{AT}, then they continue responding to surveys when receiving any other treatment values.

To further appreciate always-takers as the target population, we discuss the gettable ATE (GATE) $\mathbb{E}[Y_1 - Y_0|\eta \leq k]$ for some constant k, defined by DiNardo et al. (2021). As k increases, GATE converges to the population ATE $\mathbb{E}[Y_1 - Y_0]$. As subpopulation-specific ATEs are commonplace, DiNardo et al. (2021) show how different GATE parameters may be identified under weaker assumptions than in the traditional parametric framework. We choose $k = q(d_{\text{AT}})$ to characterize always-takers that are the largest subpopulation for whom we can partially identify the ATE of switching treatment values over $\mathcal{D}$, without imposing further assumptions on the functional forms or distributions.

Now we consider a more general non-separable nonparametric model in Assumption 3′ that implies our sufficient set Assumption 3.

Assumption 3′ (Latent index selection model with non-separable errors) *Let* $S = \mathbf{1}\{q(D, \eta) \geq 0\}$. *There exists* $d_{AT}(\eta) = \arg\inf_{d\in\mathcal{D}} q(d, \eta) \in \mathcal{D}_M$ *for each* η.

The unobserved heterogeneity is captured by η, so there could be an infinite number of types and the corresponding sufficient value $d_{\text{AT}}(\eta)$ in the most general structural model. Our sufficient set Assumption 3 restricts there to be M types of unobserved heterogeneity η, in the sense that $d_{\text{AT}}(\eta)$ belongs to $\mathcal{D}_M$.

Under Assumption 3′, Assumption 2 can be implied by further assuming $d_{\text{AT}}(\eta) = d_{\text{AT}}$ to be a constant for all individual with η, and $M = 1$. Therefore we argue that the sufficient value Assumption 2 is reasonable with a separable structural error in selection as in Assumption 2′, and the sufficient set Assumption 3 is a useful approximation under more general non-separable errors.

4 Estimation and inference

We estimate bounds over an equally spaced grid $\mathcal{D}_J = \{d_1, .., d_J\} \subset \mathcal{D}$ for a continuous treatment. We can view $\mathcal{D}_J$ as the set of treatment values where the policy maker considers treatment intervention. The estimation procedure is easy to implement, as the bounds estimates are sample analogs to the parameters defined in Theorem 1. When D is continuous, we use a kernel function $K_h(D - d) = k((D - d)/h)/h$, where the kernel function k includes a sub-population whose treatment is around d, and the size of the sub-population is controlled by the bandwidth h shrinking

to zero as the sample size grows. We provide inference on the causal effect of switching treatment between any two values in $\mathcal{D}_J$. We present the asymptotic theory for a fixed J and also for $J \to \infty$ so $\mathcal{D}_J \to \mathcal{D}$.

For a multivalued discrete treatment, it is straightforward to use the treatment indicator $\mathbf{1}\{D = d\}$ in place of the binary treatment indicator D in the estimator in Lee (2009) and let $\mathcal{D}_J$ be the (sub)support of D. We develop the inference theory for a discrete treatment in a separate paper.

We implement the estimation procedure using leave-out estimators for $s(d)$ and Q^d in Step 1 and Step 2. The leave-out estimation is similar to the cross-fitting in the recent double debiased machine learning literature (Chernozhukov et al., 2018). The leave-out preliminary estimation achieves stochastic equicontinuity without strong entropy conditions using empirical process theory. Specifically, for some fixed $L \in \{2, ..., n\}$, randomly partition the observation indices into L distinct groups $I_\ell, \ell = 1, ..., L$, such that the sample size of each group is the largest integer smaller than n/L. The number of folds L is not random and typically small, such as five or ten in practice; see, e.g., Chernozhukov et al. (2018), Velez (2025). When there is no sample splitting ($L = 1$), $\hat{s}_1(d)$ and $\hat{Q}_1^d$ use all observations in the full sample.[5]

The estimation procedure under Assumption 2 follows four steps:

Step 0. Estimate the sufficient treatment value $\hat{d}_{\mathrm{AT}_J} = \arg\min_{d \in \mathcal{D}_J} \hat{s}(d)$, where a kernel estimator $\hat{s}(d) = \sum_{i=1}^n S_i K_h(D_i - d) / \sum_{i=1}^n K_h(D_i - d)$. For $d = \hat{d}_{\mathrm{AT}_J}$, $\hat{\beta}_d = \sum_{i=1}^n Y_i S_i K_h(D_i - d) / \sum_{i=1}^n S_i K_h(D_i - d)$. Estimate the proportion of always-takers π_{AT} by $\hat{\pi} = \min_{d \in \mathcal{D}_J} \hat{s}(d) = \hat{s}(\hat{d}_{\mathrm{AT}_J})$.

For $\ell = 1, ..., L$, the estimators in Step 1 and in Step 2 use observations not in I_ℓ, denoted as $I_\ell^c := \{1, ..., n\} \setminus I_\ell$.

Step 1. Compute the leave-out kernel estimator $\hat{s}_\ell(d) = \sum_{i \in I_\ell^c} S_i K_h(D_i - d) / \sum_{i \in I_\ell^c} K_h(D_i - d)$. Estimate the trimming probability p_d by $\hat{p}_\ell = \min\{\hat{s}_\ell(\hat{d}_{\mathrm{AT}_J}) / \hat{s}_\ell(d), 1\} - \nu$ for some small positive constant ν used for robust inference that we explain in the following.

Step 2. For $d \neq \hat{d}_{\mathrm{AT}_J}$, estimate the $(1 - \hat{p}_\ell)$-quantile of $Y|D = d, S = 1$ by $\hat{Q}_\ell^d(1 - \hat{p}_\ell)$. A nonparametric estimator can be the generalized inverse function of the CDF estimate $\hat{F}_{Y|D=d,S=1_\ell}(y) = \sum_{i \in I_\ell^c} \mathbf{1}\{Y_i \leq y\} S_i K_h(D_i - d) / \sum_{i \in I_\ell^c} S_i K_h(D_i - d)$.

[5] When $L = n$, $\hat{s}_\ell$ uses all observations except for the ℓ^{th} observation, and is well-known as the leave-one-out estimator (i.e., leave the ℓ^{th} observation out), e.g., Powell et al. (1989). A large L is computationally costly.

Step 3. Compute the kernel estimator of $\mathbb{E}[Y\mathbf{1}\{Y \geq Q^d(1-p)\}|D=d, S=1]$ and obtain

$$\widehat{\rho_{dU}(\pi)} = \frac{\sum_{\ell=1}^{L}\sum_{i\in I_\ell} Y_i \mathbf{1}\{Y_i \geq \hat{Q}^d_\ell(1-\hat{p}_\ell)\} S_i K_h(D_i - d)}{\sum_{i=1}^{n} S_i K_h(D_i - d)} \frac{1}{\hat{p}},$$

where $\hat{p} = \hat{\pi}/\hat{s}(d) - \nu$ using the full sample from Step 0.

Our inference procedure is robust to extensive margin, allowing the selection probability to be of any unknown functional form with respect to treatment. The challenge for the continuous or multivalued treatment relative to the well studied binary treatment case comes from an infinite or multiple number of treatment values (J) and the corresponding potential selections and outcomes. Importantly, we allow the sufficient treatment value to be not unique in the sense that there exists a subset $\mathcal{D}_c \subseteq \mathcal{D}$ such that $d = \arg\min_{d' \in \mathcal{D}} s(d')$ for any $d \in \mathcal{D}_c$ and $\mathbb{P}(D \in \mathcal{D}_c) > 0$. So $s(d)$ is constant over $d \in \mathcal{D}_c$, which is implied by no treatment effect on selection (extensive margin) when changing d within the subset $\mathcal{D}_c$. Then we could use any $d \in \mathcal{D}_c$ as a sufficient treatment value d_{AT}. So for any $d \in \mathcal{D}_c$, $s(d_{\text{AT}})/s(d) = 1$ and $\beta_d = \rho_{dU}(\pi_{\text{AT}}) = \rho_{dL}(\pi_{\text{AT}}) = \mathbb{E}[Y|D=d, S=1]$ is point-identified. However the asymptotic distributions of the bound-estimators $\left[\widehat{\rho_{dL}(\pi_{\text{AT}})}, \widehat{\rho_{dU}(\pi_{\text{AT}})}\right]$ do not converge to the asymptotic distribution of the point-estimator $\hat{\beta}_d$ as $s(d_{\text{AT}})/s(d) \to 1$. We avoid the complication in testing $s(d) = s(d_{\text{AT}})$, e.g., if the hypothesis $s(d_{\text{AT}})/s(d) = 1$ is not rejected, then compute the point-estimator $\hat{\beta}_d$. Instead, we estimate the tight bounds using the trimming probability $\hat{p}_\ell = \min\{\hat{s}_\ell(\hat{d}_{\text{AT}_J})/\hat{s}_\ell(d), 1\} - \nu \xrightarrow{p} p_d = s(d_{\text{AT}})/s(d) - \nu \leq 1 - \nu < 1$, which contain the untrimmed point-estimator $\hat{\beta}_d = \sum_{i=1}^{n} Y_i S_i K_h(D_i - d)/\sum_{i=1}^{n} S_i K_h(D_i - d)$.[6]

Therefore we estimate bounds that are non-sharp with the trimming probability $p_d = s(d_{\text{AT}})/s(d) - \nu$ but still tight with small ν. We choose this practical and conservative strategy so that our asymptotic theorem and inference are valid regardless of the extensive margin effect on selection, and are easy to implement and interpret.

We show in Theorem 2 that the estimation errors of $\hat{d}_{\text{AT}_J}$ and grid approximation are asymptotically ignorable. That is, for any given J and for n large enough, $\hat{d}_{\text{AT}_J} = d_{\text{AT}_J} := \arg\min_{d \in \mathcal{D}_J} s(d)$. So for a any fixed set of treatment values $\mathcal{D}_J$, we can find the sufficient treatment value d_{AT_J} given a large enough sample. As $\mathcal{D}$ contains an infinite number of values for a continuous treatment, we give conditions on the grid size J going to infinity to approximate $\mathcal{D}$. We show that as $J, n \to \infty$, $d_{\text{AT}_J} \to d_{\text{AT}} := \arg\min_{d \in \mathcal{D}} s(d)$. Assumption 4 below gives conditions on J that depends on the accuracy of $\hat{s}_\ell(d)$ and the shape of $s(d)$ characterized by $\bar{M}$.

[6] In the proof of Theorem 2, we show that $\hat{d}_{\text{AT}_{J\ell}} = \arg\min_{d \in \mathcal{D}_J} \hat{s}_\ell(d) = \hat{d}_{\text{AT}_J}$ when n large enough. So $\hat{p}_\ell = \hat{s}_\ell(\hat{d}_{\text{AT}_J})/\hat{s}_\ell(d) - \nu$. However, in finite samples, it is possible that $\hat{s}_\ell(\hat{d}_{\text{AT}_J})/\hat{s}_\ell(d) > 1$ for some $d \in \mathcal{D}_J$. In such case, if $\hat{d}_{\text{AT}_J}$ and $\hat{d}_{\text{AT}_{J\ell}}$ are in $\mathcal{D}_c$, then we let $\hat{p}_\ell \leq 1 - \nu$ and estimate bounds of $\beta_{\hat{d}_{\text{AT}_{J\ell}}}$ that contain its point-estimate.

Assumption 4 *Let $s^{(m)}(d)$ be the m^{th} derivative of $s(d)$ for $m \in \{1, 2, ...\}$. Let $\mathcal{D} = \mathcal{D}_s \cup \mathcal{D}_c$, where $\mathcal{D}_c := \{d : s^{(m)}(d) = 0, \forall m \geq 1\}$ and $\mathcal{D}_s := \{d : s^{(m)}(d) \neq 0, \exists m < \infty\}$. If $\mathcal{D}_s \neq \emptyset$, let $\bar{M} = \min\{m : s^{(m)}(d) \neq 0, m = 1, 2, ..., \forall d \in \mathcal{D}_s\} < \infty$. If $\mathcal{D}_s = \emptyset$, let $\bar{M} = 0$. Let an equally spaced grid $\mathcal{D}_J = \{d_1, .., d_J\} \subseteq \mathcal{D}$ with $J = O(\mathsf{s}_n{}^{-1/\bar{M}})$ and $\mathsf{s}_n := \sup_{d \in \mathcal{D}} |\hat{s}(d) - s(d)| = o_{\mathbb{P}}(1)$.*

$\mathcal{D}_c$ is the set of treatment values not affecting the selection. For $d \in \mathcal{D}_c$, $s(d) = c$ for some generic constant c, so $s'(d) = 0$. For $\bar{M} = 0$ ($\mathcal{D} = \mathcal{D}_c$), there is no extensive margin, so $\beta_d = \mathbb{E}[Y|D = d, S = 1]$ is point-identified, and there is no restriction on $J \to \infty$. When $s(d)$ is strictly monotone over $\mathcal{D}$, $s'(d) \neq 0$, so $\bar{M} = 1$ and $\mathcal{D}_c = \emptyset$. When $s(d)$ is strictly concave, i.e., $s'(d) = 0$ for a $d \in \mathcal{D}$ and $\mathcal{D}_c = \emptyset$, we have $\bar{M} = 2$. A larger $\bar{M}$ implies that it is harder to compare $s(d_j)$ and $s(d_{j+1})$, so we need to estimate $s(d)$ more accurately, i.e., s_n needs to go to zero faster for a larger $\bar{M}$ on a given grid.

The kernel estimation is well-studied, and we use the results in Donald et al. (2012) and Hansen (2022a).

Assumption 5 *(i) The kernel function k is non-negative symmetric bounded kernel with a compact support such that $\int k(u)du = 1$, $\int uk(u)du = 0$, and $\kappa := \int u^2 k(u)du < \infty$. Let the roughness of the kernel be $R_k := \int k(u)^2 du$.*

(ii) $h \to 0$, $nh \to \infty$, $nh^5 \to c \in [0, \infty)$.

(iii) For $d \in \mathcal{D}$ and $y \in \mathcal{Y}$, $s(d) < 1$; $f_{Y|DS}(y|d, 1)$ is continuous and bounded away from 0; $F_{Y|DS}(y|d, 1)s(d)f_D(d)$ is bounded and has bounded continuous second derivative with respect to d; $var(Y|D = d, S = 1)$, $\mathbb{E}\big[|Y|^3\big|D = d, S = 1\big]$, and the second derivative of $\mathbb{E}[Y|D = d, S = 1]$ are continuous in d.

Theorem 2 *Let Assumptions 1, 4, and 5 hold.*

1. *Under Assumption 2, $\pi = s(d_{AT_J})$. Then for $d \in \mathcal{D}_J$ and $d \neq \hat{d}_{AT_J}$, as $n \to \infty$,*

$$\begin{aligned}\sqrt{nh}\Big(\widehat{\rho_{dU}(\pi)} - \rho_{dU}(\pi) - h^2 B_{dU}\Big) &\xrightarrow{d} \mathcal{N}(0, V_{dU}) \\ \sqrt{nh}\Big(\widehat{\rho_{dL}(\pi)} - \rho_{dL}(\pi) - h^2 B_{dL}\Big) &\xrightarrow{d} \mathcal{N}(0, V_{dL}),\end{aligned} \tag{2}$$

where $p = p_d = s(d_{AT_J})/s(d) - \nu$, $V_{dU} := p^{-2}\big(V_1 + V_2 + V_3 + V_{23}\big)R_k/(s(d)f_D(d))$, $V_3 := var(Y\mathbf{1}\{Y \geq Q^d(1-p)\}|D = d, S = 1), V_2 := p(1-p)Q^d(1-p)^2, V_{23} := -2p(1-p)\rho_{dU}(\pi)Q^d(1-p), V_1 := \big(V_\pi + p^2 V_{s(d)}\big)s(d)^{-1}\big(Q^d(1-p) - \rho_{dU}(\pi)\big)^2$, with $V_{s(d)} := s(d)(1-s(d))$ and $V_\pi = V_{s(d_{AT_J})}f_D(d)/f_D(d_{AT_J})$. For the lower bound, $V_{dL} := p^{-2}\big(V_1 + V_2 + V_3^L +$

$V_{23}\big)R_k/(s(d)f_D(d))$, *where* $V_3^L := var(Y\mathbf{1}\{Y \le Q^d(p)\}|D=d,S=1)$, *and* V_1, V_2, V_{23} *are defined as above with* $Q^d(p)$ *in place of* $Q^d(1-p)$ *and with* $\rho_{dL}(\pi)$ *in place of* $\rho_{dU}(\pi)$. B_{dU} *and* B_{dL} *are given explicitly in the proof in the Appendix.*

For $d = \hat{d}_{AT_J}$, $\sqrt{nh}\big(\hat{\beta}_{d_{AT_J}} - \beta_{d_{AT_J}} - h^2 B_3\big) \xrightarrow{d} \mathcal{N}\big(0, V_{d_{AT_J}}\big)$ *as* $n \to \infty$, *where* $V_{d_{AT_J}} = var(Y|D=d_{AT_J},S=1)R_k/(s(d_{AT_J})f_D(d_{AT_J}))$.

As $J \to \infty$, $d_{AT_J} \to d_{AT}$ *and the above statements hold with* d_{AT} *in place of* d_{AT_J}.

2. *Under Assumption 3, choose* $\mathcal{D}_M$ *to contain* $\hat{d}_{AT_J}$. *Let* $\hat{p}_\ell = \hat{\pi}^M_{L\ell}/\hat{s}_\ell(d)$ *in Step 1, and follow Step 2 and Step 3 to obtain the bounds* $[\widehat{\rho_{dL}(\pi)}, \widehat{\rho_{dU}(\pi)}]$, *where* $\pi = \sum_{d\in\mathcal{D}_M} s(d) - M + 1 > 0$. *Let* $V_\pi = \sum_{m=1}^M V_{s(d_m)} f_D(d)/f_D(d_m)$. *For* $d \in \mathcal{D}_M^c \cap \mathcal{D}_J$, *the above asymptotic distributions (2) hold. For* $d \in \mathcal{D}_M \cap \mathcal{D}_J$, *the above asymptotic distributions (2) hold with* $V_1 := \big(V_\pi + (p^2-2p)V_{s(d)}\big)\, s(d)^{-1}\big(Q^d(1-p) - \rho_{dU}(\pi)\big)^2$.

Notice that we allow for no extensive margin, e.g., there exists $d \neq d_{\mathrm{AT}_J}$ and $s(d) = s(d_{\mathrm{AT}_J})$. So we use $p_d = 1-\nu$ to estimate tight bounds of β_d, rather than a point-estimand.

The asymptotic variance V_{dU} can be decomposed to the three steps of the estimation procedure: V_1 is from estimating the effect on selection and the trimming probability in Step 1. V_2 is from the quantile regression in Step 2. V_3 comes from the Step 3 trimmed regression. V_{23} is from the covariance of the Step 2 and Step 3 estimation errors.

Estimation of the variances is easily carried out by replacing all of the above quantities with their sample analogs or by the sample variances of the influence functions given in equation (4) in the proof of Theorem 2 in the appendix. No additional preliminary estimators are needed. The confidence interval of at least 95% coverage can be computed by $\Big[\widehat{\rho_{dL}(\pi)} - 1.96\times\widehat{\sigma_{dL}}/\sqrt{nh}, \widehat{\rho_{dU}(\pi)} + 1.96\times\widehat{\sigma_{dU}}/\sqrt{nh}\Big]$ with $\widehat{\sigma_{dL}} := \sqrt{\widehat{V_{dL}}}$ and $\widehat{\sigma_{dU}} := \sqrt{\widehat{V_{dU}}}$. This interval will asymptotically contain the region $[\rho_{dL}(\pi), \rho_{dU}(\pi)]$ with at least 95% probability.

We can bound the ATE of increasing the treatment from d_1 to d_2, $\underline{\Delta}_{d_1d_2} := \rho_{d_2L}(\pi) - \rho_{d_1U}(\pi) \le \beta_{d_2} - \beta_{d_1} \le \rho_{d_2U}(\pi) - \rho_{d_1L}(\pi) =: \bar{\Delta}_{d_1d_2}$. We note that one might be interested in the partial (or marginal) effect defined as the derivative of β_d with respect to d, i.e., $\theta_d = \frac{\partial}{\partial d}\beta_d$. However, we could not bound such derivative by the same approach of Lee (2009). We can view $\Delta_{d_1d_2} := \beta_{d_2} - \beta_{d_1} = \int_{d_1}^{d_2} \theta_s ds$ as an average derivate over d_1 to d_2. Corollary 1 provides the asymptotic distribution of the bounds estimators for the ATE $\hat{\underline{\Delta}}_{d_1d_2} := \widehat{\rho_{d_2L}(\pi)} - \widehat{\rho_{d_1U}(\pi)}$ and $\hat{\bar{\Delta}}_{d_1d_2} := \widehat{\rho_{d_2U}(\pi)} - \widehat{\rho_{d_1L}(\pi)}$. Denote the bandwidth in $\widehat{\rho_{dL}(\pi)}$ and $\widehat{\rho_{dU}(\pi)}$ be h_{dL} and h_{dU}, respectively.

Corollary 1 (ATE) *Let the conditions in Theorem 2 hold. Then* $\sqrt{n}V_{Un}^{-1/2}\big(\hat{\bar{\Delta}}_{d_1d_2} - \bar{\Delta}_{d_1d_2} - (h^2_{d_2U}B_{d_2U} - h^2_{d_1L}B_{d_1L})\big) \xrightarrow{d} \mathcal{N}(0,1)$ *and* $\sqrt{n}V_{Ln}^{-1/2}\big(\hat{\underline{\Delta}}_{d_1d_2} - \underline{\Delta}_{d_1d_2} - (h^2_{d_2L}B_{d_2L} - h^2_{d_1U}B_{d_1U})\big) \xrightarrow{d}$

$\mathcal{N}(0,1)$, *where* $V_{Un} := \mathbb{E}[(\phi_{d_2U} - \phi_{d_1L})^2]$ *and* $V_{Ln} := \mathbb{E}[(\phi_{d_2L} - \phi_{d_1U})^2]$ *with the influence functions* ϕ_{dL}, ϕ_{dU} *given explicitly in equation (4) in the Appendix.*

Assume Assumption 2. When $h_{d2U} = h_{d_1L} = h$, $\lim_{n\to\infty,h\to 0} hV_{Un} = \mathsf{V}_{d_2U} + \mathsf{V}_{d_1L} - 2\mathsf{C}_{d_1d_2U}$. *When* $h_{d_2L} = h_{d_1U} = h$, $\lim_{n\to\infty,h\to 0} hV_{Ln} = \mathsf{V}_{d_2L} + \mathsf{V}_{d_1U} - 2\mathsf{C}_{d_1d_2L}$, *where* $p_d = \pi/s(d)$, $\mathsf{C}_{d_1d_2U} := R_k V_\pi(Q^{d_2}(1-p_{d_2}) - \rho_{d_2U}(\pi))(Q^{d_1}(p_{d_1}) - \rho_{d_1L}(\pi))/(\pi^2 f_D(d_{AT}))$, *and* $\mathsf{C}_{d_1d_2L} := R_k V_\pi(Q^{d_1}(1-p_{d_1}) - \rho_{d_1U}(\pi))(Q^{d_2}(p_{d_2}) - \rho_{d_2L}(\pi))/(\pi^2 f_D(d_{AT}))$.

The variance can be estimated by the plug-in sample analogues $\hat{\mathsf{V}}_{Un} = n^{-1}\sum_{i=1}^n(\hat{\phi}_{d_2Ui} - \hat{\phi}_{d_1Li})^2$ and $\hat{\mathsf{V}}_{Ln} = n^{-1}\sum_{i=1}^n(\hat{\phi}_{d_2Li} - \hat{\phi}_{d_1Ui})^2$. And the 95% confidence interval $\big[\hat{\underline{\Delta}}_{d_1d_2} - 1.96\times\sqrt{\hat{\mathsf{V}}_{Ln}}/\sqrt{n}, \hat{\bar{\Delta}}_{d_1d_2} + 1.96\times\sqrt{\hat{\mathsf{V}}_{Un}}/\sqrt{n}\big]$.

Next we briefly discuss how to choose an undersmoothing bandwidth h smaller than the optimal bandwidth that minimizes the asymptotic mean squared error (AMSE) such that the bias is first-order asymptotically negligible, i.e., $h^2\sqrt{nh} \to 0$, so the above confidence interval is valid.

We could estimate the leading bias B_{dU} by the method in Powell and Stoker (1996) and Colangelo and Lee (2025). Let the notation $\widehat{\rho_{dU,b}(\pi)}$ be explicit on the bandwidth b and $\widehat{B}_{dU} := \big(\widehat{\rho_{dU,b}(\pi)} - \widehat{\rho_{dU,ab}(\pi)}\big)\big/\big(b^2(1-a^2)\big)$ with a pre-specified fixed scaling parameter $a \in (0,1)$. From the proof of Theorem 3.2 in Colangelo and Lee (2025), we can choose a by minimizing the leading term of $var(\hat{B}_{dU})$, i.e., minimizing $(1-a^2)^{-2}a^{-d_T}$ for a d_T-dimensional continuous treatment. By deriving the first-order and second-order conditions, we obtain the minimizer $a^* = \sqrt{d_T/(d_T+4)} = \sqrt{1/5}$, for $d_T = 1$ in our case.

Then we propose a data-driven bandwidth $\hat{h}_{dU} := (\widehat{V_{dU}}/(4\widehat{B}_{dU}^2))^{1/5}n^{-1/5}$ to consistently estimate the AMSE optimal bandwidth h^*_{dU} by Theorem 3.2 in Colangelo and Lee (2025). For the lower bound, the same estimation applies to the leading bias $\widehat{B}_{dL}$ and the AMSE optimal bandwidth $\hat{h}_{dL}$.

5 Empirical illustration: Job Corps

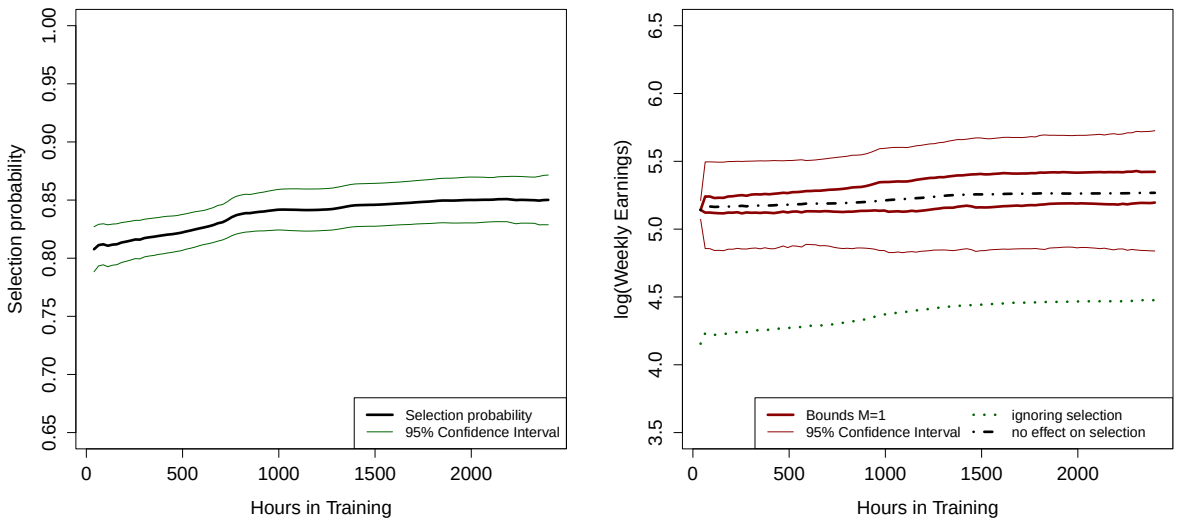
We illustrate our method by evaluating the Job Corps program. We use the Job Corps dataset in Hsu et al. (2023). The continuous treatment variable (D) is the total hours spent in academic and vocational training. The outcome variable (Y) is the weekly earnings in the fourth year. Our sample consists of 4,024 subjects who completed at least 40 hours (one week) of training. In the online appendix, Figure 8 shows the distribution of D, and Table 1 provides brief descriptive statistics. As our analysis builds on Flores et al. (2012), Hsu et al. (2020), Hsu et al. (2023), we refer the readers to the reference therein for further details of Job Corps.

We use $J = 100$ grid points over $\mathcal{D} = [40, 2400]$ that ranges from one week to fifteen months.

We use the Epanechnikov kernel with a undersmoothing bandwidth and ten-fold cross-fitting. The undersmoothing bandwidth at each d is chosen by $0.8 \times \min(\hat{h}_{dL}, \hat{h}_{dU})$, given in Section 4, where we use the rule-of-thumb bandwidth $h_1 = 1.05 \times \hat{\sigma}_D \times n_\ell^{-1/5} \times c_1$ with a constant $c_1 = 1$ and the sample standard deviation of D, $\hat{\sigma}_D$, to estimate the initial variance and bias with $a = \sqrt{0.2}$ and $b = h_1/a$. The resulting undersmoothing bandwidth ranges from 349.17 to 626.91. The estimated trimming probability is capped by $\hat{p}_\ell \leq 1 - \nu$ with $\nu = 0.01$.

To learn about the effects on selection (or the extensive margin effect), we estimate the average dose-response function of selection $\mathbb{E}[S_d] = s(d)$. The left panel of Figure 1 presents the estimates of the conditional selection probability given d, $\hat{s}(d)$. More training seems to increase the extensive margin. The minimum selection probability estimate is $\hat{\pi}_{\text{AT}} = \min_{d \in \mathcal{D}_J} \hat{s}(d) = \hat{s}(40) \approx 0.8082$. So we use the least treatment value 40 as the sufficient treatment value, i.e., if a participant is employed with one-week training, then this participant will remain employed when receiving more training between one week to fifteen months.

Figure 1: (Job Corps) Estimated selection probability and bounds without covariates



In Figure 9 in Section B.3 in the online appendix, the histograms of Y and $\log(Y)$ in the selected sample $\{Y_i > 0$, or $S_i = 1, i = 1, ..., n\}$ show that weekly earnings has a skewed distribution, and log(weekly earnings) is closer to normal. It is well-known (e.g., Chen and Roth (2023)) that averages can be heavily influenced by observations in the tail, especially when the outcome has a skewed distribution, as the weekly earnings in the Job Corps data in Figure 9. So we estimate the ATE in log, i.e., let $\beta_d = \mathbb{E}[\log(Y_d)|\text{AT}]$, a concave transformation of the outcome that is less heavily influenced by outcomes in the tail of the distribution. The right panel of Figure 1 shows

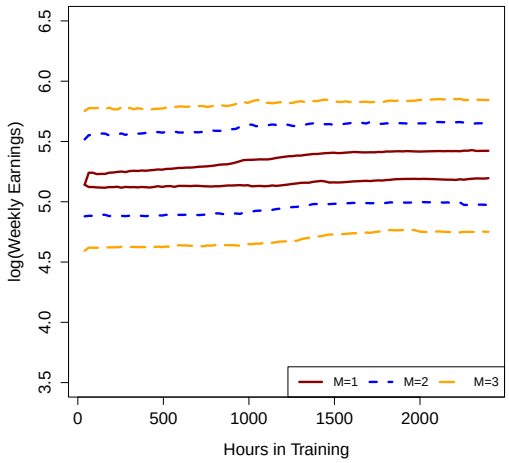
the estimated bounds under Assumption 2 with $d_{\text{AT}} = 40$ ($M = 1$). The thinner lines are the 95% confidence intervals using the sample variance of the influence function. We find the largest ATE $\beta_{2400} - \beta_{40} = \mathbb{E}[\log(Y_{2400}) - \log(Y_{40})|\text{AT}]$ bounded by $[0.054, 0.281]$ with a 90% confidence $[-0.251, 0.543]$. That is, increasing the training hours from one week to fifteen months would increase the weekly earnings by at least 5.4%, but this effect is not significant at 10% level. We also estimate the ATE in level and do not find any effects of switching hours within $[40, 2400]$; see Figure 10 in Section B.3 in the online appendix.

For comparison, the black dash-dotted line is the estimate of $\mathbb{E}[Y|D = d, S = 1]$ assuming that there is no effect on selection, i.e., the selected sample over $\mathcal{D}$ contains only the always-takers, $\{S_{d'} = 1 : d' \in \mathcal{D}\} = \{S_d = 1\}$ for all $d \in \mathcal{D}$. That is, $\beta_d = \mathbb{E}[Y|D = d, S = 1]$ is point-identified. Or under the missing at random assumption $\mathbb{E}[Y|D = d, S = 1] = \mathbb{E}[Y_d]$. We also plot the estimates of $\mathbb{E}[Y|D = d]$ ignoring selection using all observations including $\{i : S_i = 0\}$.

Sufficient set For sensitivity analysis and illustrating the sufficient set Assumption 3, Figure 2 presents the estimated bounds with $M = 1, 2, 3$ given in Theorem 1. Since $d_{\text{AT}} = \arg\min_{d' \in \mathcal{D}} s(d')$ is estimated as $\hat{d}_{\text{AT}_J} = 40$, we choose $\mathcal{D}_M$ to contain 40 such that the trimming probability $\pi_L^M / s(d) < 1$. It is natural to include the boundary points $\underline{\mathcal{D}} = 40$ and $\overline{\mathcal{D}} = 2400$ in $\mathcal{D}_M$. Following the discussion of the non-separable structural selection equation in Assumption 3′, for a subject with $d_{\text{AT}}(\eta) = \underline{\mathcal{D}}$, more training hours helps employment. On the other hand, for a subject with $d_{\text{AT}}(\eta) = \underline{\mathcal{D}}$, the largest treatment value hurts selection probability (employment). So for $M = 2$, we let $\mathcal{D}_2 = \{40, 2400\}$. Assumption 3 allows a complier to be employed with 40 training hours but unemployed with 2400 hours. And if a participant is employed with both the smallest and largest hours ($d = 40, 2400$), then this participant must be employed at any hours between one week to fifteen months, as an always-taker. The blue dashed line in Figure 2 uses the lower bound $\hat{\pi}_L^2 = \hat{s}(40) + \hat{s}(2400) - 1 \approx 0.658$.

For $M > 2$, we suggest choosing $\mathcal{D}_M$ as a equally spaced sub-grid over $\mathcal{D}_J$ and including $\hat{d}_{\text{AT}_J}$, if there is no further information on the structural selection model. So for $M = 3$, we let $\mathcal{D}_3 = \{40, 1208, 2400\}$, the orange long-dashed line uses the lower bound $\hat{\pi}_L^3 = \hat{s}(40) + \hat{s}(1208) + \hat{s}(2400) - 1 \approx 0.5$. As expected, the bounds are less informative when we assume more treatment values, i.e., a larger M. So we note that the sufficient set assumption is more of theoretical interest than practical. We focus on sufficient value Assumption 2 in empirical analysis. Next we incorporate covariates to potentially tighten the bounds and confidence intervals.

Figure 2: (Job Corps) Estimated bounds with sufficient sets



6 Estimation and inference conditional on covariates

We weaken the independent treatment assumption and the sufficient value/set assumption by conditioning on the covariates. So the covariates X can serve two purposes: First, in observational data, it is more plausible and standard to assume conditional independence, also known as unconfoundedness, selection on observables, or ignorability. Second, we allow subjects with different pretreatment covariates to have different sufficient treatment values $d_{\text{AT}x} \in \mathcal{D}$. It might be reasonable that participants with some particular covariates benefit from more training, but more training could hurt the employment of some participants with different covariates. So the bounds may be tightened by incorporating the covariates.

After conditional on the covariates, we follow Semenova (2024) to represent the generalized Lee bounds as a moment equation and derive an orthogonal moment for it. The advantage of using the orthogonality moment is that the first-stage estimation has no contribution to the asymptotic variance of the bounds. We utilize cross-fitting to remove overfitting bias without strong entropy conditions, following the recent double debiased machine learning (DML) literature (Chernozhukov et al., 2018). Then we show the asymptotic theory that accommodates low-dimensional smooth and high-dimensional sparse designs.

Notice that β_d and its bounds are functions of d over $\mathcal{D}$ and hence are of infinite dimension for a continuous treatment. Such non-regular nonparametric estimands cannot be estimated at a regular root-n rate without further assumptions. The asymptotic theory is more involved with the kernel function and bandwidth h, compared with the semiparametric inference in Chernozhukov

et al. (2018). Colangelo and Lee (2025) provide DML inference for β_d that is point-identified when there is no selection bias. In particular, if researchers assume "no effect on selection" by using the selected sample excluding those with zero outcomes, then they estimate $\mathbb{E}[\mathbb{E}[Y|S=1,D=d,X]]$. Or if researchers "ignore selection" by including all observations with zero outcomes, then they estimate $\mathbb{E}[\mathbb{E}[Y|D=d,X]]$. See also Kennedy et al. (2017), Su et al. (2019), and Chernozhukov et al. (2022b) for non-regular estimands and machine learning.

The conditional independence Assumption 6 means that conditional on observables, the treatment variable is as good as randomly assigned, or conditionally exogenous.

Assumption 6 (Conditional independence) *D is independent of $\{(Y_d, S_d) : d \in \mathcal{D}\}$ conditional on X.*

Assumption 7 relaxes Assumption 2 to allow subjects with different values of pretreatment covariates x to have different sufficient values $d_{\mathrm{AT}x} \in \mathcal{D}$.

Assumption 7 (Conditional sufficiency) *For $x \in \mathcal{X}$, there exists $d_{ATx} \in \mathcal{D}$ such that $\mathbb{P}(S_d \geq S_{d_{ATx}}, \forall d \in \mathcal{D}|X=x) = 1$.*

As discussed in Assumption 2′, a sufficient condition of Assumption 7 is to assume a separable structural error η in selection $S = \mathbf{1}\{q(D,X) \geq \eta\}$, and there exists $d_{\mathrm{AT}x} = \arg\min_{d\in\mathcal{D}} q(d,x)$.

Let the conditional average dose-response function of always-takers be $\beta_d(x) := \mathbb{E}[Y_d|\{S_{d'} = 1 : d' \in \mathcal{D}\}, X = x]$, so $\beta_d = \int_{\mathcal{X}} \beta_d(x) f_X(x|S_{d'} = 1 : d' \in \mathcal{D})dx$. Let the conditional probability of always-takers be $\pi_{\mathrm{AT}}(x) := \mathbb{P}(S_{d'} = 1 : d' \in \mathcal{D}|X = x)$. Let the corresponding conditional upper bound given in (1) be $\bar{\beta}_d(x) := \rho_{dU}(\pi_{\mathrm{AT}}(x), x) := \mathbb{E}[Y|Y \geq Q^d(1 - \pi_{\mathrm{AT}}(x)/s(d,x), x), D = d, S = 1, X = x]$, where the conditional selection probability $s(d,x) := \mathbb{P}(S = 1|D = d, X = x)$ and $Q^d(u,x)$ is the u-quantile of $Y|D = d, S = 1, X = x$. Define the aggregate upper bound for β_d as $\bar{\beta}_d := \int_{\mathcal{X}} \bar{\beta}_d(x) f_X(x|\mathrm{AT})dx$.

Lemma 2 below extends Lemma 1 in Semenova (2024) to show that the upper bound $\bar{\beta}_d$ is a ratio of two moments, by replacing the binary treatment indicator D with a kernel function $K_h(D-d)$. The moment function for the lower bound is defined analogously in the proof of Lemma 2 in the Appendix.

Lemma 2 (Moment-based representation) *Assuming Assumption 6, $s(d,x) = \mathbb{E}[S_d|X = x]$. Further assuming Assumption 7, $\pi_{AT}(x) = \min_{d\in\mathcal{D}} s(d,x) = s(d_{ATx}, x)$. Assume $\pi_{AT} = \mathbb{E}\big[\pi_{AT}(X)\big] > 0$ and the generalized propensity score $\mu_d(x) := f_{D|X}(d|x) > 0$ with probability one,*

for $d \in \mathcal{D}$. Then the sharp upper bound of β_d is

$$\bar{\beta}_d = \mathbb{E}[\bar{\beta}_d(X)\pi_{AT}(X)]/\pi_{AT} = \lim_{h\to 0} \mathbb{E}\big[m_{dU}(W,\xi)\big]/\pi_{AT}, \textit{where}$$
$$m_{dU}(W,\xi) := \frac{K_h(D-d)}{\mu_d(X)} SY \cdot \mathbf{1}\{Y \geq Q^d(1-\pi_{AT}(X)/s(d,X),X)\},$$

the nuisance parameter $\xi(d,x) = \big\{s(d,x), \mu_d(x), Q^d(1-\pi_{AT}(x)/s(d,x),x)\big\}$.

Remark 3 (Conditional sufficient set) We discuss incorporating the covariates under the sufficient set Assumption 3. We provide a non-sharp bound that can be implemented in practice. The sharp bound is out of the scope of this paper. Under Assumption 3, Theorem 1 directly implies the conditional version of the bounds on $\pi_{\text{AT}}(x)$:

$$\pi_L^M(x) := \max\left(\sum_{d\in\mathcal{D}_M} s(d,x) - M + 1, 0\right) \leq \pi_{\text{AT}}(x) = \mathbb{P}(S_d = 1 : d \in \mathcal{D}_M | X = x)$$
$$\leq \min_{d\in\mathcal{D}_M} s(d,x) =: \pi_U^M(x).$$

Then together with the proof of Lemma 2, the sharp upper bound $\mathbb{E}[\rho_{dU}(\pi_{\text{AT}}(X),X)\pi_{\text{AT}}(X)]/\pi_{\text{AT}}$ is not point-identified and is smaller than $\mathbb{E}[\rho_{dU}(\pi_L^M(X),X)\pi_U^M(X)]/\pi_L^M = \lim_{h\to 0}\mathbb{E}[m_{dU}(W,\xi)\cdot \pi_U^M(X)/\pi_L^M(X)]/\pi_L^J$ with $\pi_{\text{AT}}(X) = \pi_L^M(X)$ in $m_{dU}(W,\xi)$, which likely is a non-sharp bound. Similar, the sharp lower bound $\mathbb{E}[\rho_{dL}(\pi_{\text{AT}}(X),X)\cdot\pi_{\text{AT}}(X)]/\pi_{\text{AT}} \geq \mathbb{E}[\rho_{dL}(\pi_L^M(X),X)\cdot\pi_L^M(X)]/\pi_U^M = \lim_{h\to 0}\mathbb{E}[m_{dL}(W,\xi)\cdot\pi_L^M(X)/\pi_L^M(X)]/\pi_U^M$, with $\pi_{\text{AT}}(X) = \pi_L^M(X)$ in $m_{dL}(W,\xi)$.

6.1 Estimation and inference

We estimate the bounds over an evenly spaced grid $\mathcal{D}_J := \{d_1, ..., d_J\} \subseteq \mathcal{D}$. Let the sets $\mathcal{X}_j := \{x : s(d_j,x)/s(d,x) \leq 1, \forall d \in \mathcal{D}_J\}$ for $j = 1, ..., J$. So for $x \in \mathcal{X}_j$, $d_{\text{AT}x_J} = d_j = \arg\min_{d\in\mathcal{D}_J} s(d,x)$. We can classify the subjects into J groups $\mathcal{X} = \cup_{j=1,...,J}\mathcal{X}_j$. By consistently estimating $s(d,x)$ over the grid $\mathcal{D}_J$, we show that the mis-classification error is asymptotically first-order ignorable.

We allow the subsets $\mathcal{X}_j$ to overlap, i.e., the sufficient treatment value can be not unique, so there could be no treatment effect on selection over some range in $\mathcal{D}$. For example, if there exists $x \in \mathcal{X}_j \cap \mathcal{X}_{j+1}$ so that $s(d_j,x)/s(d_{j+1},x) = 1$, then the bounds degenerate to points at d_j and d_{j+1}, i.e., $\bar{\beta}_{d_j}(x) = \underline{\beta}_{d_j}(x) = \mathbb{E}[Y|D = d_j, S = 1, X = x]$ and $\bar{\beta}_{d_{j+1}}(x) = \underline{\beta}_{d_{j+1}}(x) = \mathbb{E}[Y|D = d_{j+1}, S = 1, X = x]$. We estimate the robust bounds using the trimming probability $\hat{p}_{d_{j+1}d_j}(x) = \hat{s}(d_j,x)/\hat{s}(d_{j+1},x) - \nu$ for some small positive ν, rather than the untrimmed point-estimator, as discussed in Section 4.

For $X_i \in \mathcal{X}_j$, define the moment function to be $m^j_{dU}(W_i, \xi) := m_{dU}(W_i, \xi)$ with $\pi_{\text{AT}}(X_i) = s(d_j, X_i)$. To construct orthogonality as $h \to 0$, we derive the correction terms for $s(d_j, x)$, $s(d, x)$, $Q^d(u, x)$, $\mu_d(x)$, respectively, collected in

$$\begin{aligned} cor^j_{dU}(W, \xi) = Q^d(1 - p_{dd_j}(X), X) \cdot \bigg(& \frac{K_h(D - d_j)}{\mu_{d_j}(X)}(S - s(d_j, X)) - \frac{K_h(D - d)}{\mu_d(X)} p_{dd_j}(X)(S - s(d, X)) \\ & + \frac{K_h(D - d)S}{\mu_d(X)} \big(- \mathbf{1}\{Y \geq Q^d(1 - p_{dd_j}(X), X)\} + p_{dd_j}(X) \big) \bigg) \\ & + \big(\mu_d(X) - K_h(D - d) \big) \cdot \mathbb{E}\big[Y \big| Y \geq Q^d(1 - p_{dd_j}(X), X), D = d, S = 1, X \big] \frac{s(d_j, X)}{\mu_d(X)} \end{aligned}$$

for $p_{dd_j}(X) \leq 1 - \nu < 1$ with $\nu > 0$. For $d = d_j$, let $\nu = 0$, $p_{dd_j}(X) = 1$, and

$$\begin{aligned} m^j_{dU}(W, \xi) = m^j_{dL}(W, \xi) &= \frac{K_h(D - d)}{\mu_d(X)} SY, \\ cor^j_{dU}(W, \xi) = cor^j_{dL}(W, \xi) &= \big(\mu_d(X) - K_h(D - d) \big) \mathbb{E}\big[Y \big| D = d, S = 1, X \big] \frac{s(d, X)}{\mu_d(X)}. \end{aligned} \tag{3}$$

Then the orthogonal moment function is defined as $g^j_{dU} := m^j_{dU} + cor^j_{dU}$. Let $g_{dU}(W, \xi) = \sum_{j=1}^J g^j_{dU}(W, \xi) \mathbf{1}\{X \in \mathcal{X}_j\} / \sum_{j=1}^J \mathbf{1}\{X \in \mathcal{X}_j\}$ that allows overlapping $\mathcal{X}_j$s.

These correction terms are derived based on equation (4.7) in Semenova (2024). Specifically let the true nuisance parameter be ξ_0 and $\xi_r := \xi_0 + r(\xi - \xi_0)$ for some ξ close to ξ_0 and $r \in (0, 1)$. Then we verify that the partial derivative of the moment function g^j_{dU} with respect to r is zero. When $p_{dd_j}(x) = 1$, $\bar{\beta}_d(x) = \underline{\beta}_d(x)$ is point-identified. So the correction term (3) is only for the nuisance function $\mu_d(x)$ as derived in Colangelo and Lee (2025).

The estimation procedure follows four steps:

Step 1. (L-fold Cross-fitting) As defined in Section 4, L-fold cross-fitting randomly partitions the observation indices into L distinct groups $I_\ell, \ell = 1, ..., L$. For $\ell = 1, ..., L$, the estimator $\hat{\xi}_\ell(W)$ for the nuisance function $\xi(W) = \big(s(d, X), Q^d(1 - p_{dd_j}(X), X), \mu_d(X), \mathbb{E}[Y | Y \geq Q^d(1 - p_{dd_j}(X), X), D = d, S = 1, X], d \in \mathcal{D} \big)$ uses observations not in I_ℓ, satisfying Assumption 8 below.

Step 2. (Double robustness) For $i \in I_\ell$ and $X_i \in \hat{\mathcal{X}}_{j\ell} = \{x : \hat{s}_\ell(d_j, x)/\hat{s}_\ell(d, x) \leq 1, \forall d \in \mathcal{D}_J\}$, estimate the orthogonal moment function by $g_{dU}(W_i, \hat{\xi}_\ell) = \sum_{j=1}^J g^j_{dU}(W_i, \hat{\xi}_\ell) \mathbf{1}\{X_i \in \hat{\mathcal{X}}_{j\ell}\} / \sum_{j=1}^J \mathbf{1}\{X_i \in \hat{\mathcal{X}}_{j\ell}\}$.

Step 3. The DML estimator in Colangelo and Lee (2025) for π_{AT}: $\hat{\pi}_{\text{AT}} = n^{-1} \sum_{\ell=1}^L \sum_{i \in I_\ell} \psi(W_i, \hat{\xi}_\ell)$, where $\psi(W_i, \xi) := K_h(D_i - d_j)(S_i - s(d_j, X_i))/\mu_{d_j}(X_i) + s(d_j, X_i)$ if $X_i \in \hat{\mathcal{X}}_{j\ell}$.

Step 4. The DML estimator $\hat{\overline{\beta}}_d = n^{-1}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} g_{dU}(W_i, \hat{\xi}_\ell)/\hat{\pi}_{\text{AT}}$.

Denote the L_2-norm $\left\|\hat{\xi}-\xi\right\|_2 = \|\Delta\hat{\xi}(W)\|_2 = \left(\int(\hat{\xi}(w)-\xi(w))^2 f_W(w)dw\right)^{1/2}$.

Assumption 8 *(i) (Strict overlap) $s(d,x) \in (c, 1-c)$ for some constant $c \in (0, 1/2)$, for all $(d,x) \in \mathcal{D}\times\mathcal{X}$. $\inf_{d\in\mathcal{D}} \operatorname{ess\,inf}_{x\in\mathcal{X}} \mu_d(x) \geq c$ for some positive constant c.*

(ii) $s(d,x)$, $f_{YSDX}(y,1,d,x)$ and $\mathbb{E}[Y|D=d,S=1,X=x]$ are two-times differentiable with respect to d with all two derivatives being bounded uniformly over $\mathcal{Y}\times\mathcal{D}\times\mathcal{X}$.

The derivative of $Q^d(u,x)$ with respect to u and $var(Y|D=d,S=1,X=x)$ is bounded uniformly over $\mathcal{D}\times\mathcal{X}$.

$\mathbb{E}[|Y|^3|S=1,D=d,X=x]$ is continuous in d uniformly over $\mathcal{X}$.

(iii) The following terms are $o_{\mathbb{P}}(1)$ for $d\in\mathcal{D}$: $\|\Delta\hat{\mu}_d(X)\|_2$, $\sup_{y\in\mathcal{Y}_0}\|\Delta\hat{\mathbb{E}}[Y|Y\geq y,S=1,D=d,X]\|_2$, $\|\Delta\hat{s}(d,X)\|_2$, $\sup_p\|\Delta\hat{Q}^d(p,X)\|_2$, where $\mathcal{Y}_0$ be a compact subset of the support of Y.

(iv) The following terms are $o_{\mathbb{P}}(1/\sqrt{nh})$ for $d\in\mathcal{D}$:
$\sup_{y\in\mathcal{Y}_0}\|\Delta\hat{\mathbb{E}}[Y|Y\geq y,S=1,D=d,X]\|_2\|\Delta\hat{\mu}_d(X)\|_2$,
$\sup_{p\in(0,1)}\|\Delta\hat{Q}^d(p,X)\|_2\Big(\sup_{p\in(0,1)}\|\Delta\hat{Q}^d(p,X)\|_2+\|\Delta\hat{\mu}_d(X)\|_2+\|\Delta\hat{s}(d,X)\|_2\Big)$,
$\|\Delta\hat{s}(d,X)\|_2\Big(\|\Delta\hat{s}(d,X)\|_2+\sup_{y\in\mathcal{Y}_0}\|\Delta\hat{\mathbb{E}}[Y|Y\geq y,S=1,D=d,X]\|_2+\|\Delta\hat{\mu}_d(X)\|_2\Big)$.

(v) $h\to 0$, $nh\to\infty$, $\sqrt{nh}h^2\to c\in[0,\infty)$.

Assumption 8(i) requires the generalized propensity score (GPS) $\mu_d(X)$ to be bounded away from zero, which is the standard overlap assumption (Semenova, 2024). We note that such common support assumption should be made with care in practice and is strong especially with many control variables (D'Amour et al., 2021). In our empirical application, we find a common support by trimming away observations whose estimated GPSs are smaller than some fixed trimming parameter. Details are in Section 7 and Section B.1 in the online appendix.

Assumption 8(iii)(iv) gives tractable high-level rate conditions on the nuisance function estimators. These are standard conditions similarly assumed in the DML literature (Chernozhukov et al., 2018; Colangelo and Lee, 2025). The rate conditions use a "partial L_2" norm in the sense that the regressor D is fixed at d and the expectation is based on the marginal distribution of X, e.g., $\|\Delta\hat{s}(d,X)\|_2 = \left(\int_{\mathcal{X}}(\hat{s}(d,x)-s(d,x))^2 f_X(x)dx\right)^{1/2}$. See Colangelo and Lee (2025) for detailed discussion on the low-level conditions of the nuisance function estimators that satisfy such high-level conditions, such as kernel, series, and neural network in low-dimensional settings with fixed dimension of X. In high-dimensional settings where the dimension of X grows with the sample

size, our inference theory is valid as long as the nuisance function estimators satisfy the high-level rate conditions, e.g., Lasso, as we illustrate in Section 6.2.

Similar to Assumption 4, Assumption 9 imposes conditions on the grid.

Assumption 9 *Let $s^{(m)}(d,x)$ be the m^{th} derivative of $s(d,x)$ w.r.t. d for $m \in \{1,2,...\}$. Let $\mathcal{D} = \mathcal{D}_{sx} \cup \mathcal{D}_{cx}$ for any $x \in \mathcal{X}$, where $\mathcal{D}_{cx} := \{d : s^{(m)}(d,x) = 0, \forall m \geq 1\}$ and $\mathcal{D}_{sx} := \{d : s^{(m)}(d,x) \neq 0, \exists m < \infty\}$. If $\mathcal{D}_{sx} \neq \emptyset$, let $\bar{M}_x = \min\{m : s^{(m)}(d,x) \neq 0, m = 1,2,..., \forall d \in \mathcal{D}_{sx}\} < \infty$. If $\mathcal{D}_{sx} = \emptyset$, let $\bar{M}_x = 0$. Let $\bar{M} = \max_{x \in \mathcal{X}} \bar{M}_x$. Let an equally spaced grid $\mathcal{D}_J = \{d_1,..,d_J\} \subseteq \mathcal{D}$ with $J = O(\mathsf{s}_n{}^{-1/\bar{M}})$ and $\mathsf{s}_n := \sup_{d \in \mathcal{D}, x \in \mathcal{X}} |\hat{s}(d,x) - s(d,x)| = o_{\mathbb{P}}(1)$.*

Theorem 3 *Let Assumptions 5, 6, 7, 8, and 9 hold. Then for $d \in \mathcal{D}$, $\hat{\overline{\beta}}_d - \overline{\beta}_d = n^{-1}\sum_{i=1}^n \phi_{dU}(W_i,\xi) - \overline{\beta}_d + o_{\mathbb{P}}(1/\sqrt{nh})$, where $\phi_{dU}(W_i,\xi) := g_{dU}(W_i,\xi)/\pi_{AT} - \psi(W_i,\xi)\overline{\beta}_d/\pi_{AT}$. And $\sqrt{nh}\big(\hat{\overline{\beta}}_d - \bar{\beta}_d - h^2\mathsf{B}_{dU}\big) \xrightarrow{d} \mathcal{N}(0,\mathsf{V}_{dU})$. Similarly for the lower bounds, $\hat{\underline{\beta}}_d - \underline{\beta}_d = n^{-1}\sum_{i=1}^n \phi_{dL}(W_i,\xi) - \underline{\beta}_d + o_{\mathbb{P}}(1/\sqrt{nh})$, where $\phi_{dL}(W_i,\xi) := g_{dL}(W_i,\xi)/\pi_{AT} - \psi(W_i,\xi)\underline{\beta}_d/\pi_{AT}$. And $\sqrt{nh}\big(\hat{\underline{\beta}}_d - \underline{\beta}_d - h^2\mathsf{B}_{dL}\big) \xrightarrow{d} \mathcal{N}(0,\mathsf{V}_{dL})$, where B_{dU}, V_{dU}, B_{dL}, and V_{dL} are given explicitly in the proof in the Appendix.*

We can estimate the asymptotic variance V_{dU} by the sample variance of the estimated influence function $\hat{\mathsf{V}}_{dU} := hn^{-1}\sum_{i=1}^n \phi_{dU}(W_i,\hat{\xi})^2$. As described in Section 4, we could estimate the leading bias B_{dU} by the method in Powell and Stoker (1996) and Colangelo and Lee (2025). Let the notation $\hat{\overline{\beta}}_{d,b}$ be explicit on the bandwidth b and $\widehat{\mathsf{B}}_{dU} := \big(\hat{\overline{\beta}}_{d,b} - \hat{\overline{\beta}}_{d,ab}\big)/\big(b^2(1-a^2)\big)$ with a pre-specified fixed scaling parameter $a \in (0,1)$. Then a data-driven bandwidth $\hat{h}_{dU} := (\widehat{V_{dU}}/(4\widehat{\mathsf{B}}_{dU}^2))^{1/5} n^{-1/5}$ consistently estimates the optimal bandwidth that minimizes the AMSE. For the lower bound, the same estimation applies to $\hat{\mathsf{V}}_{dL}$, $\widehat{\mathsf{B}}_{dL}$, and $\hat{h}_{dL}$. Then we can choose an undersmoothing bandwidth h that is smaller than $\hat{h}_{dU}$ and $\hat{h}_{dL}$ to construct the 95% confidence interval $\Big[\hat{\underline{\beta}}_d - 1.96 \times \widehat{\sigma_{dL}}/\sqrt{nh}, \hat{\overline{\beta}}_d + 1.96 \times \widehat{\sigma_{dU}}/\sqrt{nh}\Big]$ with $\widehat{\sigma_{dL}} := \sqrt{\widehat{V_{dL}}}$ and $\widehat{\sigma_{dU}} := \sqrt{\widehat{V_{dU}}}$. Under additional assumptions, we can straightforwardly show consistency of $\hat{\mathsf{V}}_{dU}$, $\widehat{\mathsf{B}}_{dU}$, and $\hat{h}_{dU}$ following Theorem 3.2 in Colangelo and Lee (2025), so we omit the repetition.

We can bound the ATE of increasing the treatment from d_1 to d_2, $\underline{\Delta}_{d_1 d_2} := \underline{\beta}_{d_2} - \overline{\beta}_{d_1} \leq \beta_{d_2} - \beta_{d_1} \leq \overline{\beta}_{d_2} - \underline{\beta}_{d_1} =: \bar{\Delta}_{d_1 d_2}$. Denote the bandwidth in $\hat{\underline{\beta}}_d$ and $\hat{\overline{\beta}}_d$ be h_{dL} and h_{dU}, respectively.

Corollary 2 (ATE) *Let the conditions in Theorem 3 hold. Let $\mathsf{V}_{Un} := \mathbb{E}[(\phi_{d_2U} - \phi_{d_1L})^2]$ and $\mathsf{V}_{Ln} := \mathbb{E}[(\phi_{d_2L} - \phi_{d_1U})^2]$. Then $\sqrt{n}\mathsf{V}_{Un}{}^{-1/2}\big(\hat{\bar{\Delta}}_{d_1 d_2} - \bar{\Delta}_{d_1 d_2} - (h_{d_2U}^2\mathsf{B}_{d_2U} - h_{d_1L}^2\mathsf{B}_{d_1L})\big) \xrightarrow{d} \mathcal{N}(0,1)$ and $\sqrt{n}\mathsf{V}_{Ln}^{-1/2}\big(\hat{\underline{\Delta}}_{d_1 d_2} - \underline{\Delta}_{d_1 d_2} - (h_{d_2L}^2\mathsf{B}_{d_2L} - h_{d_1U}^2\mathsf{B}_{d_1U})\big) \xrightarrow{d} \mathcal{N}(0,1)$.*

The variance can be estimated by $\hat{\mathsf{V}}_{Un} = n^{-1}\sum_{i=1}^n (\phi_{d_2U}(W_i,\hat{\xi}) - \phi_{d_1L}(W_i,\hat{\xi}))^2$ and $\hat{\mathsf{V}}_{Ln} = n^{-1}\sum_{i=1}^n (\phi_{d_2L}(W_i,\hat{\xi}) - \phi_{d_1U}(W_i,\hat{\xi}))^2$. The 95% confidence interval $\big[\hat{\underline{\Delta}}_{d_1 d_2} - 1.96 \times \sqrt{\hat{\mathsf{V}}_{Ln}}/\sqrt{n}, \hat{\bar{\Delta}}_{d_1 d_2} + 1.96 \times \sqrt{\hat{\mathsf{V}}_{Un}}/\sqrt{n}\big]$.

6.2 First-step nuisance parameter estimation

We illustrate our estimation procedure by applying Lasso methods in Step 1 to estimate the nuisance functions, when X is potentially high-dimensional. We provide sufficient conditions to verify the high-level Assumption 8. We follow Su et al. (2019) (SUZ, hereafter) to approximate the outcome, selection, and treatment models by varying coefficient linear regressions and logistic regressions. Particularly, the penalized kernel-smoothing least square and maximum likelihood estimations select covariates for each value of the continuous treatment. The approximation errors satisfy Assumption 10 below that imposes sparsity structures so that the number of effective covariates that can affect them is small. See Farrell (2015), Chernozhukov et al. (2022a), for example, for in-depth discussions on the specification of high-dimensional sparse models.

To estimate the $Q^d(p,x)$, we first estimate the conditional CDF $F_{Y|SDX}(y|1,d,x)$ and then compute the generalized inverse function. To estimate the conditional density $\mu_d(x) = f_{D|X}(d|x)$, we first estimate the conditional CDF $F_{D|X}$ and then take the numerical derivative. We estimate the $s(d,x)$, $F_{D|X}$, and $F_{Y|SDX}$ by the logistic distributional Lasso regression in Belloni et al. (2017) and SUZ. We modify the penalized local least squares estimators and use the conditional density estimator in SUZ. For completeness, we present the estimators and asymptotic theory in SUZ in the Appendix and refer readers to SUZ for details.

Let $b(X)$ be a $p\times 1$ vector of basis functions and Λ be the logistic CDF. Define the approximation errors $r_d(x;F_D) = F_{D|X}(d|x) - \Lambda(b(x)'\beta_d)$, $r_{dy}(x;F_Y) = F_{Y|SDX}(y|1,d,x) - \Lambda(b(x)'\alpha_{dy})$, $r_d(x;s) = s(d,x) - \Lambda(b(x)'\theta_d)$, and $r_d(x;\rho) = \rho_{dU}(\pi,x) - b(x)'\gamma_d$.

Denote the usual 1-norm $\|\alpha\|_1 = \sum_{j=1}^p |\alpha_j|$ for a vector $\alpha = (\alpha_1,...,\alpha_p)$. Denote $\|W\|_{\mathbb{P},\infty} = \sup_{w\in\mathcal{W}}|w|$ and $\|W\|_{\mathbb{P}_{N_\ell},2} = (N_\ell^{-1}\sum_{i\notin I_\ell} W_i^2)^{1/2}$ for a generic random variable W with support $\mathcal{W}$. Assumption 10 collects the conditions in Theorems 3.1 and 3.2 in SUZ.

Assumption 10 (Lasso) *Let $\mathcal{D}_0$ be a compact subset of the support of D, $\mathcal{Y}_0$ be a compact subset of the support of Y, and $\mathcal{X}$ be the support of X.*

(i) (a) $\|\max_{j\le p}|b_j(X)|\|_{\mathbb{P},\infty} \le \zeta_n$ and $\underline{C} \le E[b_j(X)^2] \le 1/\underline{C}$, for some positive constant $\underline{C}$, $j = 1,...,p$. (b) $\sup_{d\in\mathcal{D}_0,y\in\mathcal{Y}_0}\max(\|\beta_d\|_0,\|\alpha_{dy}\|_0,\|\theta_d\|_0,\|\gamma_d\|_0) \le \mathfrak{s}$ for some $\mathfrak{s}$ which possibly depends on n, where $\|\theta\|_0$ denotes the number of nonzero coordinates of θ.

(c) $\sup_{d\in\mathcal{D}_0}\|r_d(X;F_D)\|_{\mathbb{P}_n,2} = O_{\mathbb{P}}((\mathfrak{s}\log(p\vee n)/n)^{1/2})$ and $\sup_{d\in\mathcal{D}_0,y\in\mathcal{Y}_0}\Big(\big\|r_{dy}(X;F_Y)k((D-d)/h_1)^{1/2}\big\|_{\mathbb{P}_n,2} + \big\|r_d(X;s)k((D-d)/h_1)^{1/2}\big\|_{\mathbb{P}_n,2} + \big\|r_d(X;\rho)k((D-d)/h_1)^{1/2}\big\|_{\mathbb{P}_n,2}\Big) = O_{\mathbb{P}}\big(\big(\mathfrak{s}\log(p\vee n)/n\big)^{1/2}\big)$.

(d) $\sup_{d\in\mathcal{D}_0}\|r_d(X;F_D)\|_{\mathbb{P},\infty}=O((\mathfrak{s}^2\zeta_n^2\log(p\vee n)/n)^{1/2})$ *and* $\sup_{d\in\mathcal{D}_0,y\in\mathcal{Y}_0}\left(\left\|r_{dy}(X;F_Y)\right\|_{\mathbb{P},\infty}+\left\|r_d(X;s)\right\|_{\mathbb{P},\infty}+\left\|r_d(X;\rho)\right\|_{\mathbb{P},\infty}\right)=O\left((\mathfrak{s}^2\zeta_n^2\log(p\vee n)/(nh_1))^{1/2}\right)$. *(e)* $f_{D|X}(d,x)$ *is second-order differentiable w.r.t.* d *with bounded derivatives uniformly over* $(d,x)\in\mathcal{D}_0\times\mathcal{X}$. *(f)* $\zeta_n^2\mathfrak{s}^2\iota_n^2\log(p\vee n)/(nh_1)\to 0$, $nh_1^5/(\log(p\vee n))\to 0$.

(ii) Uniformly over $(d,x,y)\in\mathcal{D}_0\times\mathcal{X}\times\mathcal{Y}_0$, *(a) there exists some positive constant* $\underline{C}<1$ *such that* $\underline{C}\le f_{D|X}(d|x)\le 1/\underline{C}$, $\underline{C}\le F_{Y|SDX}(y|1,d,x)\le 1-\underline{C}$, *and* $\underline{C}\le s(d,x)\le 1-\underline{C}$; *(b)* $s(d,x)$ *and* $F_{Y|SDX}(y|1,d,x)$ *are three times differentiable w.r.t.* d *with all three derivatives being bounded.*

(iii) There exists a sequence $\iota_n\to\infty$ *such that* $0<\kappa'\le\inf_{\delta\neq 0,\|\delta\|_0\le\mathfrak{s}\iota_n}\|b(X)'\delta\|_{\mathbb{P}_n,2}/\|\delta\|_2\le\sup_{\delta\neq 0,\|\delta\|_0\le\mathfrak{s}\iota_n}\|b(X)'\delta\|_{\mathbb{P}_n,2}/\|\delta\|_2\le\kappa''<\infty$ *w.p.a.*1.

Let Assumption 10 hold. Then Theorems 3.1 and 3.2 in SUZ imply $\sup_{(d,x)\in\mathcal{D}_0\times\mathcal{X}}|\hat{s}_\ell(d,x)-s(d,x)|=O_{\mathbb{P}}(A_n)$, $\sup_{(d,x)\in\mathcal{D}_0\times\mathcal{X},u\in(0,1)}|\hat{Q}_\ell^d(u,x)-Q^d(u,x)|=O_{\mathbb{P}}(A_n)$, $\sup_{(d,x,y)\in\mathcal{D}_0\times\mathcal{X}\times\mathcal{Y}_0}|\Delta\hat{\mathbb{E}}[Y|Y\ge y,S=1,D=d,X=x]|O_{\mathbb{P}}(A_n)$, where $A_n=\iota_n(\log(p\vee n)\mathfrak{s}^2\zeta_n^2/(nh_1))^{1/2}$. And $\sup_{(d,x)\in\mathcal{D}_0\times\mathcal{X}}|\hat{\mu}_{d\ell}(x)-\mu_d(x)|=O_{\mathbb{P}}(R_n)$, where $R_n=h_1^{-1}(\log(p\vee n)\mathfrak{s}^2\zeta_n^2/n)^{1/2}$. Then we can obtain the L_2 rates $\|\hat{\xi}_\ell-\xi\|_2$ to verify Assumption 8. In particular a sufficient condition of Assumption 8(iii) is $A_n\to 0$ and $R_n\to 0$. And a sufficient condition of Assumption 8(iv) is $\sqrt{nh}A_nR_n\to 0$ and $\sqrt{nh}A_n^2\to 0$.

7 Empirical illustration with covariates

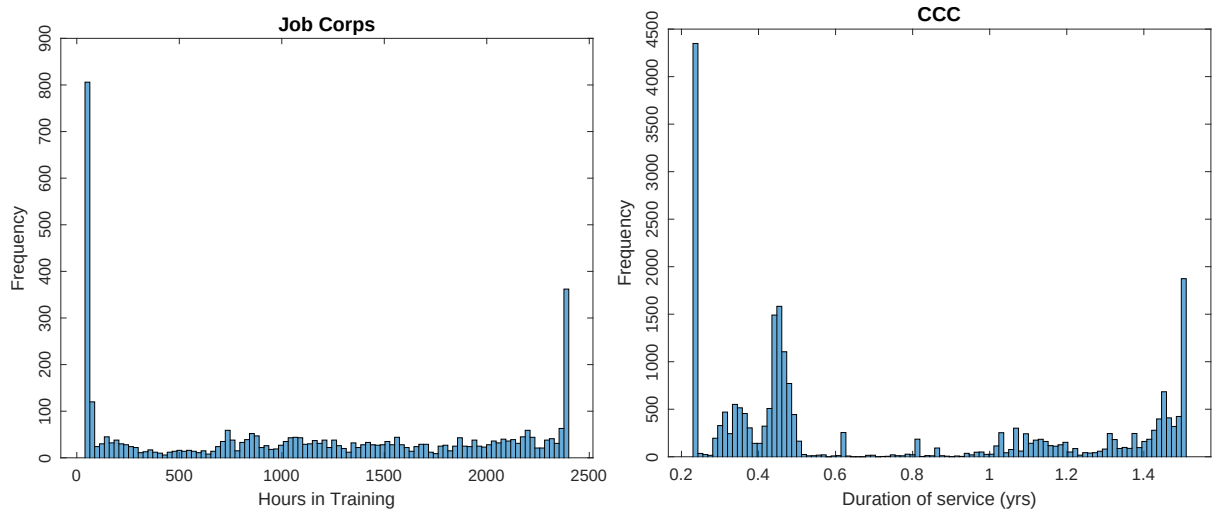
We illustrate our DML estimator by evaluating the Job Corps and CCC program. First in Step 0, we prepare a sub-sample that satisfies Assumption 8(i) for estimating the bounds. We use the full sample to estimate the GPS by $\tilde{\mu}_d(X_i)$ and the selection probability by $\tilde{s}(d,X_i)$. We obtain a sub-sample where $\tilde{\mu}_d(X_i)\ge trim_{GPS}$ and $\tilde{s}(d,X_i)\ge 5\%$ for *all* i in this sub-sample and for *all* $d\in\mathcal{D}_J$, where the trimming parameter $trim_{GPS}$ is based on Imbens (2004) and detailed in Section B.1 in the online appendix. Then we use this sub-sample to estimate the bounds following the procedure described in Section 6.1. For JC, we remove 11 observations whose $\tilde{s}(d,X_i)<5\%$ for some $d\in\mathcal{D}_J$, and no observation is removed by $trim_{GPS}$. For CCC, we remove 94 observation whose $\tilde{\mu}_d(X_i)<trim_{GPS}$, and all observations have $\tilde{s}(d,X_i)\ge 5\%$. So we only trim a small number of observations relative to the sample size.

In Step 1, we use the Lasso estimation given in Section 6.2. We present the results with a linear basis function $b(X)$ for the regularized varying coefficient linear regression, e.g., $\rho_{dU}(\pi,X)=X'\gamma_d$. Results with a quadratic basis function for specification robustness and implementation details are

in Section B in the online appendix. Same as the setting in Section 5, we use the Epanechnikov kernel with a undersmoothing bandwidth and ten-fold cross-fitting. Let $h_1 = 1.05 \times \hat{\sigma}_D \times n_\ell^{-1/5} \times c_1$, where we use a range of constants c_1 for robustness check.

We allow subjects to have different sufficient treatment values depending on X, i.e., the minimizer of the conditional selection probability given X and $D = d$. So we might tighten the bounds and the confidence intervals, or capture heterogenous causal effects that are not revealed without the covariates. Figure 3 reports histograms of the sufficient treatment values, $\{\hat{d}_{\text{AT}x i}, i = 1, ...n\}$. For Job Corps in the left panel, most sufficient treatment values are 40 or 2400, but we also see values over $\mathcal{D}$. In particular, about 20% of the participants have $d_{\text{AT}x} = 40$, i.e., 40 training hours give them the lowest likelihood of employment. For about 9% of the participants who have $d_{\text{AT}x} = 2400$, any hours smaller than 2400 would help employment. For the CCC data in the right panel, about 59.2% of the participants have $d_{\text{AT}x} < 0.5$ years and about 26.19% of the participants have $d_{\text{AT}x} > 1.2$ years.

Figure 3: Histograms of the sufficient treatment values



7.1 Job Corps

We follow the literature to assume the conditional independence Assumption 6, which is indirectly assessed in Flores et al. (2012). It means that receiving different levels of the treatment is random, conditional on a rich set of observed covariates measured at the baseline survey. We may further use a control function with an instrumental variable to address the concern that the conditional independence assumption might not hold (Imbens and Newey, 2009; Lee, 2015).

The top left panel of Figure 4 shows the estimated selection probability $\hat{\mathbb{E}}[S_d]$ with covariates

and without covariates.[7] The bottom left panel of Figure 4 presents the estimated bounds $[\hat{\underline{\beta}}_d, \hat{\bar{\beta}}_d]$ and the 95% confidence intervals. We find a positive intensive margin effect: increasing the training from 1.5 week to 9 months increases log weekly earnings by at least 0.224, at 5% significance level. This is from the largest lower-bound estimate for the ATE of switching hours over $\mathcal{D}$. That is, we find the largest ATE $\beta_{1446.465} - \beta_{63.838}$ bounded by $[0.224, 0.718]$ with the 95% confidence interval $[0.127, 0.865]$.[8] We also not that we do not find significant positive effects on level of weakly earnings.

Consistent with prior empirical research on the Job Corps, e.g., Flores et al. (2012) and Hsu et al. (2023) have found inverted-U shaped average dose-response functions on labor outcomes. The concave shape is reasonable to consider the optimal treatment intensity in other settings. Our sufficient treatment values assumption does not assume any shape restrictions and hence includes a concave or monotone selection response.

For comparison, the top right panel presents the bounds without X, $[\widehat{\rho_{dL}}, \widehat{\rho_{dU}}]$ (red solid line) given in Figure 4. We also compute two point-estimators of the average dose-response function "ignoring selection" and "no effect on selection" by the DML estimator in Colangelo and Lee (2025) incorporating X. The "ignoring selection" estimates use the full sample including those with zero outcomes, while the "no effect on selection" estimates use the selected sample with positive outcomes. Our bounds are around the "no effect on selection" estimates, as expected.

To demonstrate the usefulness of the DML method, the bottom right panel presents two alternative estimators from Lemma 2: the regression-type estimator of $\mathbb{E}[\bar{\beta}_d(X)\pi_{\text{AT}}(X)]/\mathbb{E}\big[\pi_{\text{AT}}(X)\big]$ (Bounds $\rho(\cdot)$) and the inverse-probability-weighting estimator of $\lim_{h\to 0}\mathbb{E}\big[m_{dU}(W,\xi)\big]/\pi_{AT}$ (Bounds $m(\cdot)$). These two estimators are not doubly robust, so we see that they are both biased downward, compared with our bounds (Bounds $g(\cdot)$) using a doubly robust moment function.

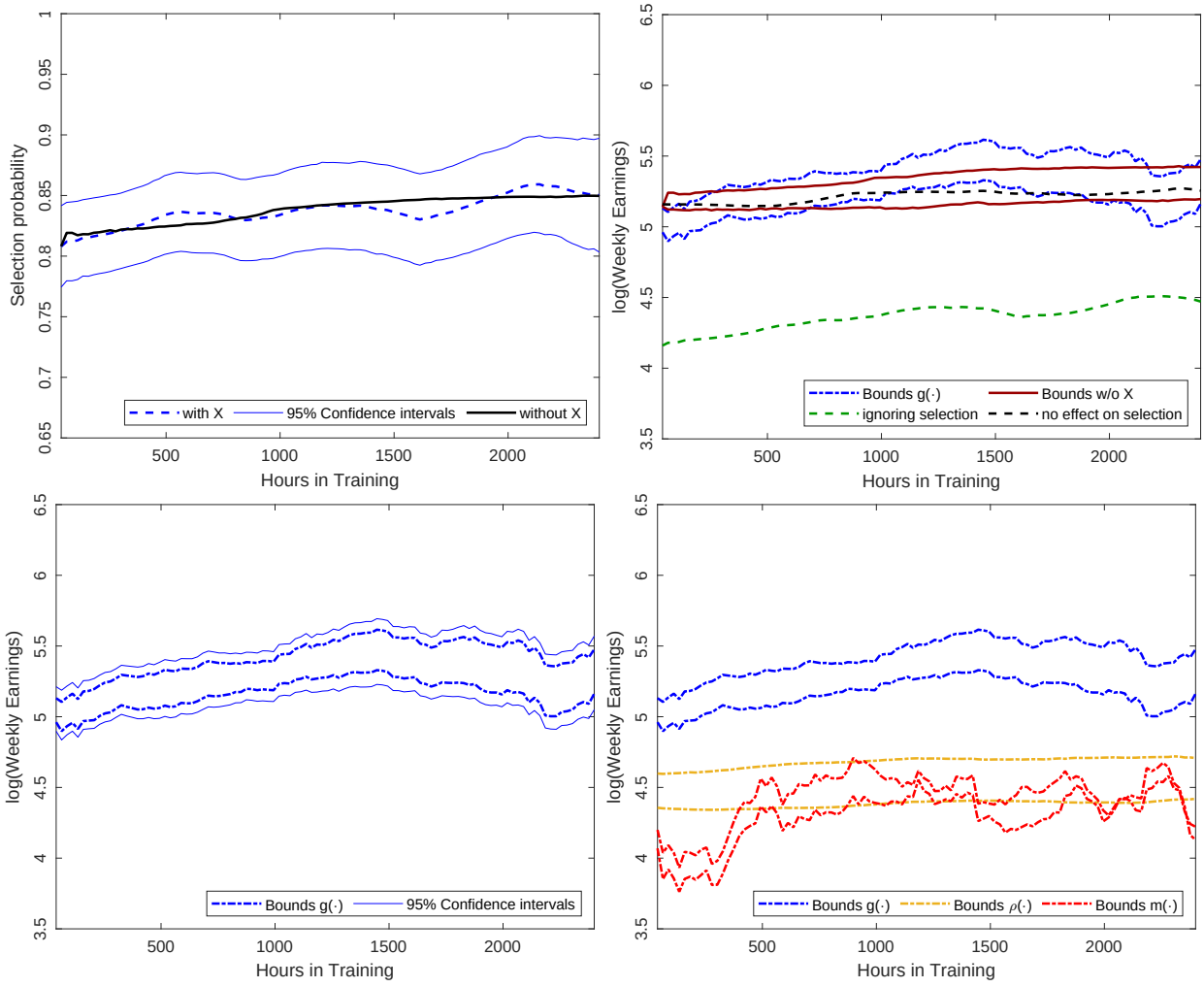
7.2 Civilian Conservation Corps (CCC)

We study the effect of service duration on the age at death, using the data for panel A of Table III in Aizer et al. (2024). They include the 17,639 men (75% of the original sample size 23,722) who have information on death age and who dies after age 45. Aizer et al. (2024) investigate the extent of sample selection and the effects of missing data on their estimates, and conclude modest bias from non-random attrition. They estimate an accelerated failure time model with added controls

[7] We estimate the selection probability $\mathbb{E}[S_d]$ with covariates X by the DML estimator in Colangelo and Lee (2025), $\hat{\mathbb{E}}[S_d] = n^{-1}\sum_{\ell=1}^{L}\sum_{i\in I_\ell} K_h(D_i - d)(S_i - \hat{s}_\ell(d, X_i))/\hat{\mu}_{d\ell}(X_i) + \hat{s}_\ell(d, X_i)$.

[8] $c_1 = 1$. For robustness check, we find positive ATE at 5% significance level over $c_1 \in \{0.75, 1, 1.25, 1.5, 1.75\}$ and using a linear or quadratic basis function $b(X)$ in Section B.3 in the online appendix. The undersmoothing bandwidths range from 159.008 to 259.490.

Figure 4: (Job Corps) Estimated selection probability and bounds with covariates



for the characteristics of the enrollees and the camps to show that the estimates remain stable. They find one more year of training increases the death age by one year.

From the histograms of the death age and log(death age) in Figure 13 in Section B.4 in the online appendix, we see that the distribution of log(death age) is more skewed. So we use the level of *death age* as our outcome variable Y and estimate the bounds for $\beta_d = \mathbb{E}[death\ age|\text{AT}]$.

We consider one hundred equally spaced grid points $\mathcal{D}_{100} \subset \mathcal{D} = [0.238, 1.5]$, which are from the 15% to 85% quantiles of duration in years. Figure 5 reports the estimates without using covariates. The largest ATE is when increasing duration from 0.238 to 1.488 years with the bounds $[0.747, 1.765]$ and 90% confidence interval $[-1.031, 3.476]$.[9] The proportion of always-takers, or

[9]The undersmoothing bandwidths range from 0.1909 to 0.4409. $\nu = 0.01$ and $c_1 = 1$. For robustness check, the lower bound estimates of the largest ATE with $c_1 \in \{0.75, 1, 1.25, 1.5, 1.75, 2\}$ are positive but insignificant at 10% level.

the minimum selection probability, is $\hat{\pi}_{\text{AT}} = \hat{s}(0.2382) = 0.8155$. Our sufficient treatment value Assumption 2 means that if the participants' death ages are observed when they received 0.2382 years of service, then they would remained observed for any duration over these one hundred grid points.

Next we include covariates as in column (3) Add Indiv Controls in Panel A in Table III in Aizer et al. (2024). Figure 6 reports the estimates with covariates. The largest ATE is when increasing duration from 0.238 to 1.156 years with the bounds $[1.169, 1.755]$ and 95% confidence interval $[0.366, 2.795]$. That is, increasing duration from 0.238 to 1.156 years would at least increase the average death age by 1.169 years and the effect is significant at 5% level, which is consistent with the findings in Aizer et al. (2024).[10] However, the effects are not significant for other choices of c_1. By the histogram in Figure 12 and summary statistics in Table 2 in the online appendix, the distribution of duration might have mass points at 0.5, 1, 1.5, 2 years. So the caveat of implementing our method on this dataset is that the smoothness conditions could be violated. This may explain why the estimates in Figure 6 are not smooth. Nevertheless, we are able to implement our method at a coarser grid with 100 grid points and assume the distributions are smooth locally at these grid points. We also consider thirty equally spaced grid points $\mathcal{D}_{30} \subset \mathcal{D} = [0.238, 1.5]$ in Figure 7 and obtain similar results as $\mathcal{D}_{100}$.[11]

8 Conclusion

We study causal effects of a treatment/policy variable that could be either continuous, multivalued discrete or binary, in a sample selection model where the treatment affects the outcome and also researchers' ability to observe the outcome. To account for the non-random selection into samples, we provide sharp bounds for the mean potential outcome of always-takers whose outcomes are observed regardless of their treatment value, generalizing Lee (2009)'s bound. We propose a novel sufficient treatment value (set) assumption to (partially) identify the share of always-takers in the observed selected d-receipts for each treatment value d.

By incorporating pretreatment covariates X, we allow for unconfoundedness and allow subjects with different values of X to have different sufficient treatment values, which might tighten the bounds, increase precision (tighten the confidence intervals), and reveal heterogeneity, as we illustrate in our empirical analysis of the Job Corps and CCC programs. The inference procedure is

[10] $\nu = 0.01$ and $c_1 = 1.5$. For robustness check, the lower bounds with $c_1 \in \{1.75, 2\}$ and linear/quadratic basis functions range from 0.288 to 0.54 but are not significant at 10% level. The undersmoothing bandwidths range from 0.162 to 0.313.

[11] The largest ATE is when increasing the duration from 0.238 to 1.109 years with the bounds $[1.255, 1.822]$ and 95% confidence interval $[0.43, 2.918]$. $c_1 = 1.5$. The undersmoothing bandwidths range from 0.158 to 0.294.

Figure 5: (CCC) Estimated selection probability and bounds without covariates

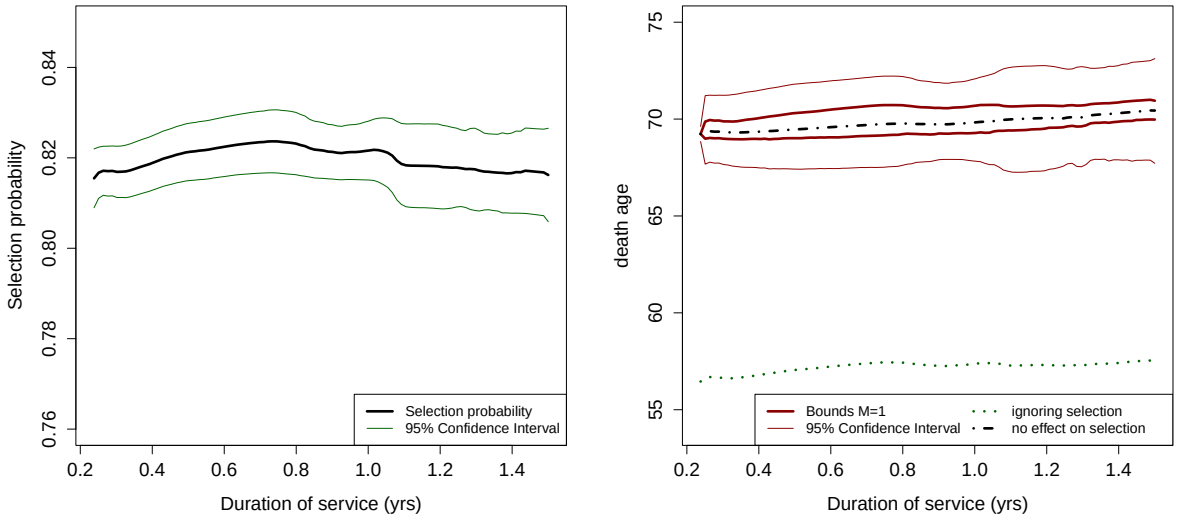
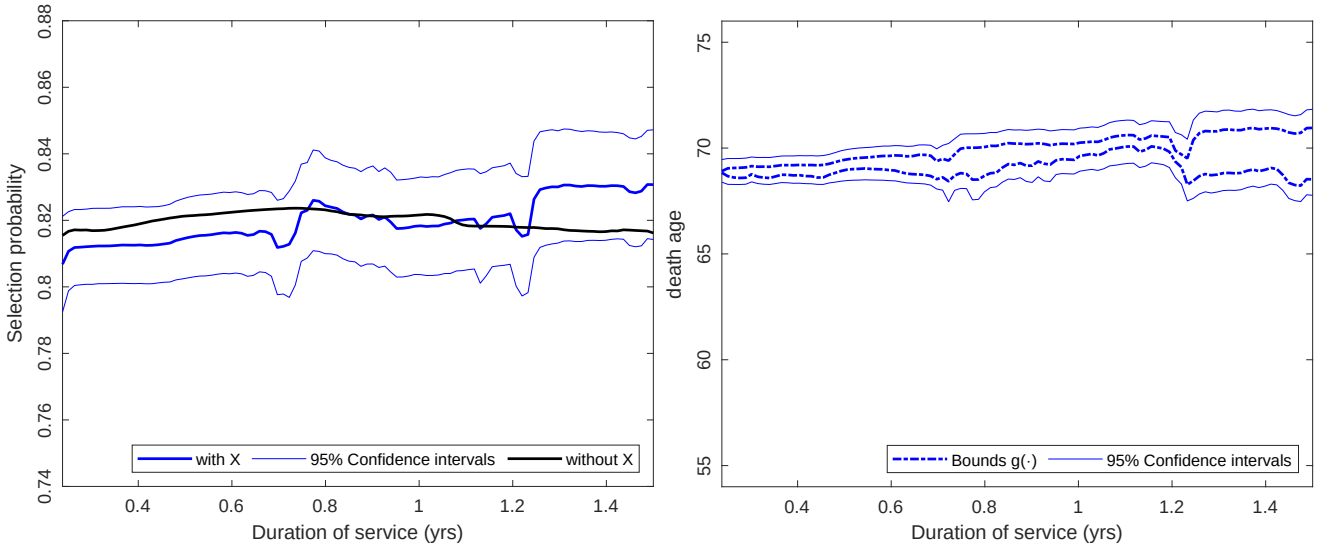


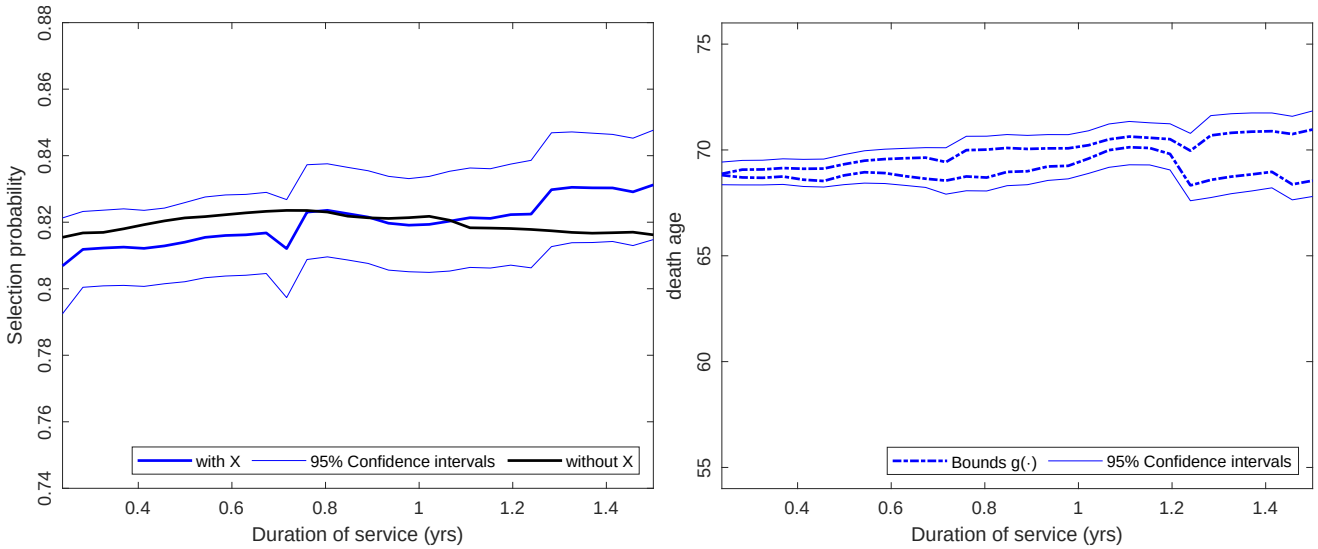
Figure 6: (CCC) Estimated selection probability and bounds with covariates over $\mathcal{D}_{100}$



robust to extensive margin, allowing for any unknown functional form of the selection probability with respect to treatment.

There are many potential applications of our bounds for continuous/multivalued treatments. For example, Ao et al. (2021) study the multivalued treatments of the Workforce Investment Act program on participants' earnings. Cesarini et al. (2017) use Swedish lotteries data to study the effect of wealth on labor supply.

Figure 7: (CCC) Estimated selection probability and bounds with covariates over $\mathcal{D}_{30}$



Appendices

Proof of Lemma 1: The proof is implied by Theorem 1 with $M = 1$ and $d_1 = d_{\text{AT}}$. □

Proof of Theorem 1: $\mathbb{P}(S_{d_1} = S_{d_2} = ... = S_{d_M} = 1) = \mathbb{P}(S_{d_1} = 1, \{S_{d_2} = ... = S_{d_M} = 1\}) = \mathbb{P}(S_{d_1} = 1) + \mathbb{P}(\{S_{d_2} = ... = S_{d_M} = 1\}) - \mathbb{P}(S_{d_1} = 1 \text{ or } \{S_{d_2} = ... = S_{d_M} = 1\}) \geq \mathbb{P}(S_{d_1} = 1) + \mathbb{P}(\{S_{d_2} = ... = S_{d_M} = 1\}) - 1$ that is the Fréchet-Hoeffding bound. The same argument gives $\mathbb{P}(\{S_{d_m} = ... = S_{d_M} = 1\}) \geq \mathbb{P}(S_{d_m} = 1) + \mathbb{P}(\{S_{d_{m+1}} = ... = S_{d_M} = 1\}) - 1$ for $m = 2, ..., M-1$. Under Assumption 1, $\mathbb{P}(S_{d_m} = 1) = s(d_m)$. So we obtain π_L^M by induction.

Denote the always-takers $\text{AT} = \{S_d = 1 : d \in \mathcal{D}\}$ and the d-complier $\text{CP}_d = \{S_d = 1, S_{d'} = 0 \text{ for some } d' \in \mathcal{D}\}$. By the law of iterated expectations,

$$\begin{aligned}\mathbb{E}[Y|S=1, D=d] = \mathbb{E}[Y_d|S_d=1] &= \mathbb{E}[Y_d|S_d=1, \text{AT}] \cdot \mathbb{P}(S_d=1, \text{AT}|S_d=1) \\ &\quad + \mathbb{E}[Y_d|S_d=1, \text{CP}_d] \cdot (1 - \mathbb{P}(S_d=1, \text{AT}|S_d=1)) \\ &= \beta_d \cdot \pi_{\text{AT}}/s(d) + \mathbb{E}[Y_d|S_d=1, \text{CP}_d] \cdot (1 - \pi_{\text{AT}}/s(d)).\end{aligned}$$

Then we apply Proposition 1a in Lee (2009). Specifically, we replace Δ_0^{UB} in Lee (2009) with $\mathbb{E}[Y|D=d, S=1, Y \geq Q^d(1-\pi_{\text{AT}}/s(d))]$, replace p_0 with $1 - \pi_{\text{AT}}/s(d)$, replace $D=1$ with $D=d$, replace Y_1^* with Y_d, replace $\{S_0=1, S_1=1\}$ with AT , replace $\{S_0=1, S_1=1\}$ with CP_d, and remove $\mathbb{E}[Y|D=0, S=1]$. Then the same arguments in the proof of Proposition 1a in Lee (2009) yields that $\beta_d \leq \mathbb{E}[Y|Y \geq Q^d(1-\pi_{\text{AT}}/s(d)), D=d, S=1] =: \rho_{dU}(\pi_{\text{AT}})$ that is sharp.

Since $\rho_{dU}(\pi_{\text{AT}})$ is decreasing in π_{AT}, $\rho_{dU}(\pi_{\text{AT}}) \leq \rho_{dU}(\pi_L^M)$ that is sharp. A similar argument for the sharp lower bound follows. □

Proof of Theorem 2: We prove the result using a given d_{AT} and $\mathcal{D}_M$. In the later section **Misclassification** of the proof, we show that using $\hat{d}_{\text{AT}_J}$ is equivalent asymptotically. Let W_i be a generic random variable and $\mathcal{W}_\ell^c$ denote the observations W_i for $i \notin I_\ell$. Denote the sample average operator over all observations as $\mathbb{E}_n[W] := n^{-1}\sum_{i=1}^n W_i$, the sample average operator over the observations in group I_ℓ as $\mathbb{E}_{n_\ell}[W] := n_\ell^{-1}\sum_{i\in I_\ell} W_i$, the sample average operator over the observations in group I_ℓ^c as $\mathbb{E}_{n_\ell}^c[W] := (n-n_\ell)^{-1}\sum_{i\in I_\ell^c} W_i$, and the conditional mean operator given the observations in group I_ℓ^c as $\mathbb{E}_\ell[W] := \mathbb{E}[W|\mathcal{W}_\ell^c]$. Denote as $\mathbf{1}_i := \mathbf{1}\{Y_i \geq Q^d(1-p)\}$ and $\hat{\mathbf{1}}_i^\ell := \mathbf{1}\{Y_i \geq \hat{Q}_\ell^d(1-\hat{p}_\ell)\}$ for $i \in I_\ell$.

We first derive the asymptotically linear representation $\widehat{\rho_{dU}(\pi)} - \rho_{dU}(\pi) = \mathbb{E}_n[\phi_{dU}] + o_{\mathbb{P}}(1/\sqrt{nh})$, where for $p = p_d < 1$,[12]

$$
\begin{aligned}
\phi_{dU} &:= (\phi_{12} + \phi_3 - \rho_{dU}(\pi)\phi_1)p^{-1}, \qquad (4)\\
\phi_3 &:= \frac{K_h(D-d)S}{s(d)f_D(d)}\left(Y\mathbf{1} - \mathbb{E}[Y\mathbf{1}|D=d,S=1]\right),\\
\phi_{12} = \phi_{12}^U &:= \phi_1 Q^d(1-p) - \phi_2(1-p)Q^d(1-p)f_{Y|DS}(Q^d(1-p)|d,1),\\
\phi_1 &:= \left(\phi_\pi - p\phi_s(d)\right)/s(d), \text{with}\\
\phi_s(d) &:= (S-s(d))\frac{K_h(D-d)}{f_D(d)}, \hat{\pi}_\ell - \pi = \mathbb{E}_{n_\ell}^c[\phi_\pi] + o_{\mathbb{P}}(1/\sqrt{nh}),\\
\hat{Q}_\ell^d(\tau) - Q^d(\tau) &= \mathbb{E}_{n_\ell}^c[\phi_2(\tau)] + o_{\mathbb{P}}(1/\sqrt{nh}), \text{with } \phi_2(\tau) := \frac{K_h(D-d)S\left(\tau - \mathbf{1}\left\{Y \leq Q^d(\tau)\right\}\right)}{s(d)f_D(d)f_{Y|DS}(Q^d(\tau)|d,1)}.
\end{aligned}
$$

For $\pi = \sum_{d\in\mathcal{D}_M} s(d) - M + 1$, $\phi_\pi = \sum_{d\in\mathcal{D}_M}\phi_s(d)$.

For the lower bound $\widehat{\rho_{dL}(\pi)} - \rho_{dL}(\pi) = \mathbb{E}_n[\phi_{dL}] + o_{\mathbb{P}}(1/\sqrt{nh})$, let $\mathbf{1} = \mathbf{1}\{Y \leq Q(p)\}$, $\hat{\mathbf{1}}_i^\ell = \mathbf{1}\{Y_i \leq \hat{Q}_\ell^d(\hat{p}_\ell)\}$, and $\phi_{12} = \phi_{12}^L := \phi_1 Q^d(p) + \phi_2(p)Q^d(p)f_{Y|DS}(Q^d(p)|d,1)$. For $p = 1$, $\phi_{dU} = \phi_{dL} := \frac{K_h(D-d)S}{s(d)f_D(d)}\left(Y - \mathbb{E}[Y|D=d,S=1]\right)$.

We focus on the upper bound $\rho_{dU}(\pi)$ in the proof, and similar arguments for the lower bound $\rho_{dL}(\pi)$ are given later.

Let $\rho_{dU}(\pi) =: num/p$, where $num := \mathbb{E}[Y\mathbf{1}\{Y \geq Q^d(1-p)\}|D=d,S=1]$. By a linearization of the estimator,

$$
\widehat{\rho_{dU}(\pi)} - \rho_{dU}(\pi) = \frac{\widehat{num} - num}{p} - \frac{\rho_{dU}(\pi)}{p}(\hat{p}-p) + O_{\mathbb{P}}\left(\|\widehat{num} - num\|\|\hat{p}-p\| + \|\hat{p}-p\|^2\right). \quad (5)
$$

Decompose the numerator estimator $\widehat{num} = L^{-1}\sum_{\ell=1}^L \widehat{num}^\ell$ in (5) to the Step1&Step2 estimation error and the Step 3 estimation,

$$
\widehat{num}^\ell = \frac{\mathbb{E}_{n_\ell}\left[Y\hat{\mathbf{1}}^\ell K_h(D-d)S\right]}{\mathbb{E}_n\left[K_h(D-d)S\right]} = \widehat{num}_{12}^\ell + \widehat{num}_3^\ell, \text{ where}
$$

[12] To simplify notation and without loss of clarity, we suppress the subscript of d in p_d.

$\widehat{num}_{12}^{\ell} := \dfrac{\mathbb{E}_{n_\ell}\left[Y(\hat{\mathbf{1}}^{\ell}-\mathbf{1})K_h(D-d)S\right]}{\widehat{den}}$, $\widehat{num}_{3}^{\ell} := \dfrac{\mathbb{E}_{n_\ell}\left[Y\mathbf{1}K_h(D-d)S\right]}{\widehat{den}}$, $\widehat{den} := \mathbb{E}_n\left[K_h(D-d)S\right]$.

To derive the influence function at each step, we use the following four Claims, whose proofs are in the online appendix.

Claim-Step1: For any $d \in \mathcal{D}$, (i) $\hat{s}(d) - s(d) = \mathbb{E}_n\left[\phi_s(d)\right] + O_{\mathbb{P}}(1/(nh) + h^4) = O_{\mathbb{P}}(1/\sqrt{nh} + h^2)$.
(ii) $\lim_{h\to 0} h\mathbb{E}[\phi_s^2(d)] = V_{s(d)}R_k/f_D(d)$.
(iii) For any $d' \neq d$, $\lim_{h\to 0} h\mathbb{E}[\phi_{s(d')}\phi_{s(d)}] = 0$.
(iv) Let $\hat{p} = \hat{\pi}/\hat{s}(d) - \nu$. $\hat{p} - p = \mathbb{E}_n[\phi_1] + O_{\mathbb{P}}(1/(nh) + h^4)$.

Claim-Step2: Theorems 3.7 and 4.1 in Donald et al. (2012) provide that for any $d \in \mathcal{D}$, $\hat{Q}^d(\tau) - Q^d(\tau) = \mathbb{E}_n[\phi_2(\tau)] + o_{\mathbb{P}}(1/\sqrt{nh})$, uniformly in $\tau \in [0,1]$. And $V_Q(\tau) := \lim_{h\to 0} h\mathbb{E}[\phi_2(\tau)^2] = \frac{\tau(1-\tau)R_k}{s(d)f_D(d)f_{Y|DS}(Q^d(\tau)|d,1)^2}$.

Claim-Step3: $\widehat{den} - den = O_{\mathbb{P}}(1/\sqrt{nh} + h^2)$, where $den := s(d)f_D(d)$. $\widehat{num}_3^{\ell} - \mathbb{E}[Y\mathbf{1}|D = d, S = 1] = \mathbb{E}_{n_\ell}[\phi_3] + o_{\mathbb{P}}(1/\sqrt{nh})$. $\lim_{h\to 0} h\mathbb{E}[\phi_3^2] = V_3R_k/(f_D(d)s(d))$.

Claim-SE: $\widehat{num_{12}}^{\ell} = num_{12}^{\ell} + o_{\mathbb{P}}(1/\sqrt{nh})$, where

$$num_{12}^{\ell} := (\hat{p}_\ell - p)Q^d(1-p) - \left(\hat{Q}_\ell^d(1-p) - Q^d(1-p)\right) Q^d(1-p) f_{Y|DS}(Q^d(1-p)|d,1)$$

for $\rho_{UB}(\pi)$, and $num_{12}^{\ell} := (\hat{p}_\ell - p)Q^d(p) + \left(\hat{Q}_\ell^d(p) - Q^d(p)\right) Q^d(p) f_{Y|DS}(Q^d(p)|d,1)$ for $\rho_{LB}(\pi)$.

$$\begin{aligned}
&\widehat{num}^{\ell} - num = num_{12}^{\ell} + \widehat{num}_3^{\ell} - \mathbb{E}[Y\mathbf{1}|D = d, S = 1] + o_{\mathbb{P}}((nh)^{-1/2}) \\
&= (\hat{p}_\ell - p)Q^d(1-p) - \left(\hat{Q}_\ell^d(1-p) - Q^d(1-p)\right) Q^d(1-p) f_{Y|DS}(Q^d(1-p)|d,1) \\
&\quad + \mathbb{E}_{n_\ell}[\phi_3] + o_{\mathbb{P}}((nh)^{-1/2}) \\
&= \mathbb{E}_{n_\ell}^c[\phi_1]Q^d(1-p) - \mathbb{E}_{n_\ell}^c[\phi_2]Q^d(1-p) f_{Y|DS}(Q^d(1-p)|d,1) + \mathbb{E}_{n_\ell}[\phi_3] + o_{\mathbb{P}}((nh)^{-1/2}),
\end{aligned}$$

where the first and second equalities is by Claim-SE and Claim-Step3, and the third equality is by Claim-Step1 and Claim-Step2.

Note that $L^{-1}\sum_{\ell=1}^L \mathbb{E}_{n_\ell}^c[W] = L^{-1}\sum_{\ell=1}^L (n-n_\ell)^{-1}\sum_{i\in I_\ell^c} W_i = (L(n-n/L))^{-1}(L-1)\sum_{i=1}^n W_i = \mathbb{E}_n[W]$. And $L^{-1}\sum_{\ell=1}^L \mathbb{E}_{n_\ell}[W] = \mathbb{E}_n[W]$. Then $\widehat{num} - num = L^{-1}\sum_{\ell=1}^L \widehat{num}^{\ell} - num = \mathbb{E}_n\big[\phi_1 Q^d(1-p) - \phi_2 Q^d(1-p) f_{Y|DS}(Q^d(1-p)|d,1) + \phi_3\big] + o_{\mathbb{P}}(1/\sqrt{nh}) = \mathbb{E}_n[\phi_{12} + \phi_3] + o_{\mathbb{P}}(1/\sqrt{nh})$. By (5), we obtain ϕ.

Variance. We derive the asymptotic variance V_{dU} by $\mathbb{E}\left[\left(\phi_{12}^U + \phi_3 - \rho_{dU}(\pi)\phi_1\right)^2\right] = \mathbb{E}\big[{\phi_{12}^U}^2 + \phi_3^2 + \rho_{dU}(\pi)^2\phi_1^2 + 2\phi_{12}^U\phi_3 - 2\rho_{dU}(\pi)\phi_1\phi_{12}^U - 2\rho_{dU}(\pi)\phi_1\phi_3\big]$. Since ϕ_1 does not involve Y, the law of iterated expectations yields $\lim_{h\to 0} h\mathbb{E}[\phi_1\phi_2] = 0$ and $\lim_{h\to 0} h\mathbb{E}[\phi_1\phi_3] = 0$.

$\lim_{h\to 0} h\mathbb{E}[\phi_1^2] = \lim_{h\to 0}(h\mathbb{E}[\phi_\pi^2] + p^2 \cdot h\mathbb{E}[\phi_s^2(d)])/s(d)^2 = \left(V_\pi + p^2 V_{s(d)}\right) R_k/(f_D(d)s(d)^2)$, by Claim-Step1 (ii)(iii), where $V_\pi = V_{s(d_{\mathrm{AT}_J})} f_D(d)/f_D(d_{\mathrm{AT}_J})$ if $\pi = s(d_{\mathrm{AT}_J})$. Under Assumption 3, when $d \in \mathcal{D}_M$ so that π contains $s(d)$, there is an additional term in $\lim_{h\to 0} h\mathbb{E}[\phi_1^2]$, $-2p \lim_{h\to 0} h\mathbb{E}[\phi_\pi \phi_s(d)]/s(d)^2 = -2pV_{s(d)}R_k/(f_D(d)s(d)^2)$.

$\lim_{h\to 0} h\mathbb{E}\left[{\phi_{12}^U}^2\right] = \lim_{h\to 0} h\mathbb{E}\left[\phi_1^2\right] Q^d(1-p)^2 + h\mathbb{E}\left[\phi_2(1-p)^2\right] Q^d(1-p)^2 f_{Y|DS}(Q^d(1-p)|d,1)^2$.

By Claim-Step2, $\lim_{h\to 0} h\mathbb{E}[\phi_2(1-p)^2] = V_Q(1-p) = \frac{p(1-p)R_k}{s(d)f_D(d)f_{Y|DS}(Q^d(1-p)|d,1)^2}$. We obtain V_2.

$\lim_{h\to 0} h\mathbb{E}[\phi_1 \phi_{12}^U] = \lim_{h\to 0} h\mathbb{E}[\phi_1^2] Q^d(1-p)$.

For $\mathbb{E}[\phi_{12}^U \phi_3]$,

$$
\begin{aligned}
&h\mathbb{E}[\phi_2(\tau)\phi_3] \times s(d)^2 f_D(d)^2 f_{Y|DS}(Q^d(\tau)|d,1) \\
&= h\mathbb{E}\left[SK_h(D-d)^2 \mathbb{E}\left[(Y\mathbf{1} - \mathbb{E}[Y\mathbf{1}|D=d,S=1])\left(\tau - \mathbf{1}\{Y \le Q^d(\tau)\}\right)|D,S\right]\right] \\
&= h\mathbb{E}\Big[SK_h(D-d)^2\Big(\mathbb{E}[Y\mathbf{1}|D,S]\tau - \mathbb{E}[Y\mathbf{1}|D=d,S=1]\tau \\
&\quad + \mathbb{E}[Y\mathbf{1}|D=d,S=1]\mathbb{E}[\mathbf{1}\{Y \le Q^d(\tau)\}|D,S]\Big)\Big] \qquad (6) \\
&\to R_k s(d) f_D(d) \tau \mathbb{E}[Y\mathbf{1}|D=d,S=1]
\end{aligned}
$$

as $h \to 0$. So $\lim_{h\to 0} h\mathbb{E}[\phi_{12}^U \phi_3] = \lim_{h\to 0} -h\mathbb{E}[\phi_2(1-p)\phi_3] Q^d(1-p) f_{Y|DS}(Q^d(1-p)|d,1) = V_{23}R_k/(2f_D(d)s(d))$.

For $\rho_{LB}(p)$ with $\mathbf{1} = \mathbf{1}\{Y \le Q^d(p)\}$, (6) becomes $h\mathbb{E}\Big[SK_h(D-d)^2\Big(\mathbb{E}[Y\mathbf{1}|D,S]p - \mathbb{E}[Y\mathbf{1}|D=d,S=1]p + \mathbb{E}[Y\mathbf{1}|D=d,S=1]\mathbb{E}[\mathbf{1}\{Y \le Q^d(p)\}|D,S]\Big) - \mathbb{E}[Y\mathbf{1}|D,S]\Big] \to R_k s(d) f_D(d)(p-1)\mathbb{E}[Y\mathbf{1}|D=d,S=1]$. So $\lim_{h\to 0} h\mathbb{E}[\phi_{12}^L \phi_3] = \lim_{h\to 0} h\mathbb{E}[\phi_2(p)\phi_3] Q^d(p) f_{Y|DS}(Q^d(p)|d,1) = -(1-p)\mathbb{E}[Y\mathbf{1}|D=d,S=1] Q^d(p) R_k/(s(d)f_D(d)) = V_{23}R_k/(2f_D(d)s(d))$.

Putting together the above results yields $V_{dU} = \lim_{h\to 0} p^{-2}\Big\{h\mathbb{E}[\phi_1^2] \cdot \left(Q^d(1-p)^2 + \rho_{dU}(\pi)^2 - 2\rho_{dU}(\pi)Q^d(1-p)\right) + h\mathbb{E}[\phi_2(1-p)^2] \cdot Q^d(1-p)^2 f_{Y|DS}(Q^d(1-p)|d,1)^2 + h\mathbb{E}[\phi_3^2] + h2\mathbb{E}[\phi_{12}\phi_3]\Big\}$.

Bias. We show $\mathbb{E}[\phi_{dU}] = h^2 B_{dU} + o_{\mathbb{P}}(h^2)$, where $B_{dU} := \left(B_\pi - pB_s(d)\right)Q^d(1-p)/s(d) - B_2(1-p)Q^d(1-p)f_{Y|DS}(Q^d(1-p)|d,1) + B_3 - \rho_{dU}(\pi)B_1)p^{-1}$ derived below. So the bias is first-order asymptotically ignorable by assuming $\sqrt{nh}h^2 = o(1)$.

Assuming the second derivatives of $s(d)$ and $f_D(d)$ are bounded continuous,

$$
\begin{aligned}
&\mathbb{E}[\phi_s(d)] = \mathbb{E}[(s(D) - s(d))K_h(D-d)/f_D(d)] = h^2 B_s(d) + o_{\mathbb{P}}(h^2), \text{where} \\
&B_s(d) := \frac{\kappa}{2}\frac{d^2}{dv^2}\left\{(s(v) - s(d))f_D(v)\right\}\Big|_{v=d}/f_D(d).
\end{aligned}
$$

Under Assumption 2, $B_\pi = B_s(d_{\mathrm{AT}})$. Under Assumption 3, $\mathbb{E}[\phi_\pi] = \sum_{d\in\mathcal{D}_J} \mathbb{E}[\phi_s(d)]$, so $B_\pi =$

$\sum_{d\in\mathcal{D}_J} B_s(d)$. We obtain $B_1 = \big(B_\pi - pB_s\big)/s(d)$. By the same argument,

$$\begin{aligned}\mathbb{E}[\phi_2(1-p)] &= \mathbb{E}\left[(\pi - F_{Y|DS}(Q^d(1-p)|D,1))\frac{K_h(D-d)s(D)}{s(d)f_D(d)f_{Y|DS}(Q^d(1-p)|d,1)}\right]\\ &= h^2 B_2(1-p) + o_{\mathbb{P}}(h^2), \text{where}\end{aligned}$$

$$B_2(1-p) := \frac{\kappa}{2}\frac{d^2}{dv^2}\big\{(\pi - F_{Y|DS}(Q^d(1-p)|v,1))f_D(v)s(v)\big\}\big|_{v=d}/(s(d)f_D(d)f_{Y|DS}(Q^d(1-p)|d,1));$$

$$\mathbb{E}[\phi_3] = \mathbb{E}\left[\frac{K_d(D-d)s(D)}{s(d)f_D(d)}\big(\mathbb{E}[Y\mathbf{1}|D,S=1] - \mathbb{E}[Y\mathbf{1}|D=d,S=1]\big)\right] = h^2 B_3 + o_{\mathbb{P}}(h^2), \text{where}$$

$$B_3 := \frac{\kappa}{2}\frac{d^2}{dv^2}\big\{f_D(v)s(v)\big(\mathbb{E}[Y\mathbf{1}|D=v,S=1] - \mathbb{E}[Y\mathbf{1}|D=d,S=1]\big)\big\}\big|_{v=d}/(s(d)f_D(d)).$$

Similarly for the lower bound, $B_{dL} := \big(B_\pi - pB_s(d)\big)Q^d(p)/s(d) - B_2(p)Q^d(p)f_{Y|DS}(Q^d(p)|d,1) + B_3 - \rho_{dU}(\pi)B_1)p^{-1}$, where $\mathbf{1} = \mathbf{1}\{Y \le Q^d(p)\}$.

Asymptotic normality. The asymptotic normality follows from the Lyapunov central limit theorem with the third absolute moment. Specifically, the Lyapunov condition holds if $\lim_{n\to\infty} n^{-1/2}h^{3/2}$ $\mathbb{E}[|\phi|^3] = 0$. Under the condition that $\mathbb{E}[|Y|^3\mathbf{1}|D=d,S=1]$ is continuous in $d\in\mathcal{D}$, we can show that $\mathbb{E}[|\phi_3|^3] = O(h^{-2})$ by a similar algebra as for the variance in Claim-Step 1(ii).

Misclassification. Let the true $\tau_j = s(d_{j+1}) - s(d_j)$ with the estimate $\hat\tau_j = \hat s(d_{j+1}) - \hat s(d_j)$. First consider $\hat d_{\text{AT}_J} = \arg\min_{d\in\mathcal{D}_J}\hat s(d) = d_{j'} \in \mathcal{D}_c$ and $\tau_{j'} = 0$. As both $d_{j'}$ and $d_{d'+1}$ are minimizers of $s(d)$, $\beta_{d_{j'}}$ and $\beta_{d_{j'+1}}$ are point-identified. There is no misclassification problem, as we can use $\hat d_{\text{AT}_J} = d_{j'}$ and $\hat p_\ell = \min\{\hat s_\ell(d_{j'})/\hat s_\ell(d_{j'+1}), 1\} - \nu$ in finite samples to estimate valid bounds at $d = d_{j'+1}$ that contains the point estimates $\hat\beta_{d_{j'+1}}$.

We consider misclassification when $\hat\tau_j \le 0 < \tau_j$ or $\hat\tau_j \ge 0 > \tau_j$ for some $j \in \{1,...,J-1\}$. Suppose $\tau_j \ne 0$, $d_{\text{AT}} = \arg\min_{d\in\mathcal{D}} s(d) \in \mathcal{D}_s$, and $d_{\text{AT}_J} = \arg\min_{d\in\mathcal{D}_J} s(d) \in \mathcal{D}_s$. Because $|\hat\tau_j - \tau_j| \le |\hat s(d_{j+1}) - s(d_{j+1})| + |\hat s(d_j) - s(d_j)| \le 2\mathsf{s}_n$, misclassification implies that $0 < |\tau_j| \le 2\mathsf{s}_n$. For a discrete multivalued treatment, $|\tau_j| > 2\mathsf{s}_n$ for n large enough, which implies correct classification. For a continuous treatment, let $\mathsf{D} = \overline{\mathcal{D}} - \underline{\mathcal{D}}$, so $d_{j+1} - d_j = \mathsf{D}/(J-1)$ for all equally spaced grid point $d_j \in \mathcal{D}_J$. By the Taylor expansion and mean-value theorem, for $d_j \in \mathcal{D}_s$, for some $\bar d_j$ between d_j and d_{j+1}, and for some generic constant C,

$$|\tau_j| = \left|\sum_{m=1}^{\bar M - 1} s^{(m)}(d_j)\mathsf{D}^m/(m!(J-1)^m) + s^{(\bar M)}(\bar d_j)\mathsf{D}^{\bar M}/(\bar M!(J-1)^{\bar M})\right| \ge C/J^{\bar M}.$$

Assuming for n large enough, $2\mathsf{s}_n < CJ^{-\bar M}$, and hence $|\tau_j| > 2\mathsf{s}_n$, which implies correct classification, for all $j \in \{1,..,J-1\}$ and $J \ge 2$. So for n large enough, there is no misclassification and $\hat d_{\text{AT}_J} = d_{\text{AT}_J} := \arg\min_{d\in\mathcal{D}_J} s(d)$. Similarly in each leave-out group ℓ, $\hat d_{\text{AT}J\ell} = \arg\min_{d\in\mathcal{D}_J}\hat s_\ell(d) = \hat d_{\text{AT}_J} = d_{\text{AT}_J}$ for n large enough.

Since $s(d)$ is continuous, as $J \to \infty$, $|d_{\text{AT}_J} - d_{\text{AT}}| \leq \mathsf{D}/(J-1)$. So $d_{\text{AT}_J} \to d_{\text{AT}}$ as $J \to \infty$. □

Proof of Lemma 2: Let $p_d(X) := \pi/s(d,X)$. $\mathbb{E}[m_{dU}(W,\xi)|X=x]$

$$
\begin{aligned}
&= \mathbb{E}[\mathbb{E}[Y\mathbf{1}\{Y \geq Q^d(1-p_d(x),x)\}|S=1,D,X=x]K_h(D-d)|S=1,X=x]\frac{\mathbb{P}(S=1|X=x)}{\mu_d(x)} \\
&= \int \mathbb{E}[Y\mathbf{1}\{Y \geq Q^d(1-p_d(x),x)\}|S=1,D=d+uh,X=x]k(u)f_{D|X=x,S=1}(d+uh)du \\
&\quad \times \frac{\mathbb{P}(S=1|X=x)}{\mu_d(x)} \\
&= \mathbb{E}[Y\mathbf{1}\{Y \geq Q^d(1-p_d(x),x)\}|S=1,D=d,X=x] \cdot f_{D|X=x,S=1}(d)\frac{\mathbb{P}(S=1|X=x)}{\mu_d(x)} + o(h) \\
&= \mathbb{E}[Y|Y \geq Q^d(1-p_d(x),x)\},S=1,D=d,X=x]p_d(x) \cdot s(d,x) + o(h) = \rho_{dU}(\pi,x) \cdot \pi + o(h).
\end{aligned}
$$

Then the aggregate sharp upper bound is $\int_{\mathcal{X}} \rho_{dU}(\pi_{\text{AT}}(x),x)f_X(x|S_{d'}=1:d' \in \mathcal{D})dx = \int_{\mathcal{X}} \rho_{dU}(\pi_{\text{AT}}(x),x)\pi_{\text{AT}}(x)f_X(x)\frac{f_X(x|\text{AT})}{\pi_{\text{AT}}(x)f_X(x)}dx = \lim_{h\to 0}\mathbb{E}\big[m_{dU}(W,\xi)\big]/\pi_{\text{AT}}$.

Similarly for the lower bound, define

$$
m_{dL}(W,\xi) := \frac{K_h(D-d)}{\mu_d(X)} \cdot S \cdot Y \cdot \mathbf{1}\{Y \leq Q^d(p_d(X),X)\}.
$$

For $X_i \in \mathcal{X}_j$, define the moment function to be $m^j_{dL}(W,\xi) := m_{dL}(W,\xi)$ with $\pi_{\text{AT}}(X) = s(d_j,X)$. To construct orthogonality as $h \to 0$, $cor^j_{dL}(W,\xi) = Q^d(p_{dd_j}(X),X)\Big(\frac{K_h(D-d_j)}{\mu_{d_j}(X)}(S - s(d_j,X)) - \frac{K_h(D-d)}{\mu_d(X)}p_{dd_j}(X)(S-s(d,X)) + \frac{K_h(D-d)S}{\mu_d(X)}\big(1\{Y > Q^d(p_{dd_j}(X),X)\} - 1 + p_{dd_j}(X)\big)\Big) + (\mu_d(X) - K_h(D-d))\mathbb{E}[Y|Y \leq Q^d(p_{dd_j}(X),X),D=d,S=1,X]\frac{s(d_j,X)}{\mu_d(X)}$. Then the orthogonal moment function is $g^j_{dL} := m^j_{dL} + cor^j_{dL}$. □

Proofs of Theorem 3: For ease of exposition, we collect the notations below.

Notations. $\kappa := \int_{-\infty}^{\infty} u^2k(u)du$. $\mathsf{K}_d := K_h(D-d)$, $\mathsf{s}_d := s(d,X)$, $\lambda_d := 1/\mu_d(X)$.
For the upper bound, $\mathsf{Q} := Q^d(1-p_{dd_j}(X),X)$, $\mathsf{1} := \mathbf{1}\{Y \geq Q^d(1-p_{dd_j}(X),X)\}$, $\rho := \mathbb{E}[Y|Y \geq \mathsf{Q},D=d,S=1,X]$.

$\mathsf{B}_{dU} := \big(\mathsf{B}_{gU} - \mathsf{B}_\psi \bar\beta_d\big)/\pi_{\text{AT}}$, where $\mathsf{B}_\psi := \mathbb{E}\Big[2\frac{\partial}{\partial d}s(d,X)\frac{\partial}{\partial d}f_{D|X}(d|X)/f_{D|X}(d|X) + \frac{\partial^2}{\partial d^2}s(d,X)\Big]\kappa/2$,

$$\begin{aligned}\mathsf{B}_{gU} := \frac{\kappa}{2}\mathbb{E}\Bigg[&\frac{\partial^2}{\partial d^2}\big(f_{D|X}(d|X)s(d,x)\mathbb{E}[Y|Y\geq \mathsf{Q}, S=1, D=d, X]\big(1-F_{Y|SDX}(\mathsf{Q}|1,d,X)\big)\big)\\ &+\mathsf{Q}\lambda_{d_j}\frac{\partial^2}{\partial d_j^2}\big(f_{D|X}(d_j,X)s(d_j,X)\big) - \mathsf{Q}\lambda_{d_j}\mathsf{s}_{d_j}\frac{\partial^2}{\partial d_j^2}f_{D|X}(d_j|X)\\ &+(\mathsf{Q}-\rho)\lambda_d\mathsf{s}_{d_j}\frac{\partial^2}{\partial d^2}f_{D|X}(d,X) - \mathsf{Q}\lambda_d\frac{\partial^2}{\partial d^2}\big(f_{D|X}(d|X)s(d,X)(1-F_{Y|SDX}(\mathsf{Q}|S=1,d,X))\big)\Bigg].\end{aligned}$$

$\mathsf{V}_{dU} := \big(\mathsf{V}_{gU} + \mathsf{V}_\psi\bar\beta_d^2 - 2\bar\beta_d\mathsf{V}_{gU\psi}\big)/\pi_{\text{AT}}^2$, where $\mathsf{V}_\psi := R_k\mathbb{E}[\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})\lambda_{d_j}]$, $\mathsf{V}_{gU} := \mathbb{E}\big[var(Y|Y\geq \mathsf{Q}, S=1, D=d, X)\mathsf{s}_{d_j}\lambda_d + \big(\mathsf{Q}^2\lambda_{d_j} + (\rho-\mathsf{Q})^2\lambda_d\big)\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})\big]R_k$, and $\mathsf{V}_{gU\psi} = R_k\mathbb{E}[\mathsf{Q}\lambda_{d_j}\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})]$.
$\mathsf{C}_{d_1d_2U} := R_k\mathbb{E}\big[Q^{d_1}(p_{d_1d_j}(X),X)Q^{d_2}(1-p_{d_2d_j}(X),X)\lambda_{d_j}\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})\big]$.

For the lower bound, $\mathsf{Q} := Q^d(p_{dd_j}(X),X)$, $\mathsf{1} := \mathbf{1}\{Y\leq Q^d(p_{dd_j}(X),X)\}$, and $\rho := \mathbb{E}[Y|Y\leq \mathsf{Q}, D=d, S=1, X]$.
$\mathsf{B}_{dL} := \big(\mathsf{B}_{gL} - \mathsf{B}_\psi\underline{\beta_d}\big)/\pi_{\text{AT}}$, where

$$\begin{aligned}\mathsf{B}_{gL} := \frac{\kappa}{2}\mathbb{E}\Bigg[&\frac{\partial^2}{\partial d^2}\big(f_{D|X}(d|X)s(d,x)\mathbb{E}[Y|Y\leq \mathsf{Q}, S=1, D=d, X]F_{Y|SDX}(\mathsf{Q}|1,d,X)\big)\\ &+\mathsf{Q}\lambda_{d_j}\frac{\partial^2}{\partial d_j^2}\big(f_{D|X}(d_j,X)s(d_j,X)\big) - \mathsf{Q}\lambda_{d_j}\mathsf{s}_{d_j}\frac{\partial^2}{\partial d_j^2}f_{D|X}(d_j|X)\\ &+(\mathsf{Q}-\rho)\lambda_d\mathsf{s}_{d_j}\frac{\partial^2}{\partial d^2}f_{D|X}(d,X) - \mathsf{Q}\lambda_d\frac{\partial^2}{\partial d^2}\big(f_{D|X}(d|X)s(d,X)F_{Y|SDX}(\mathsf{Q}|S=1,d,X)\big)\Bigg].\end{aligned}$$

$\mathsf{V}_{dL} := \big(\mathsf{V}_{gL} + \mathsf{V}_\psi\underline{\beta_d}^2 - 2\underline{\beta_d}\mathsf{V}_{gL\psi}\big)/\pi_{\text{AT}}^2$, where $\mathsf{V}_{gL} := \mathbb{E}\big[var(Y|Y\leq \mathsf{Q}, S=1, D=d, X)\mathsf{s}_{d_j}\lambda_d + \big(\mathsf{Q}^2\lambda_{d_j} + (\rho-\mathsf{Q})^2\lambda_d\big)\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})\big]R_k$, and $\mathsf{V}_{gL\psi} = R_k\mathbb{E}[\mathsf{Q}\lambda_{d_j}\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})]$.
$\mathsf{C}_{d_1d_2L} := R_k\mathbb{E}\big[Q^{d_1}(1-p_{d_1d_j}(X),X)Q^{d_2}(p_{d_2d_j}(X),X)\lambda_{d_j}\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})\big]$.

We can write $\hat{\bar\beta}_d = \hat A/\hat B$, where $A = \mathbb{E}[g_{dU}(W,\xi)]$ and $B = \pi_{\text{AT}}$. By linearization, $\hat A/\hat B = A/B + (\hat A - A)/B - (\hat B - B)A/B^2 + O_{\mathbb{P}}\big(|(\hat A - A)(\hat B - B) + (\hat B - B)^2|\big)$.

By Theorem 3.1 in Colangelo and Lee (2025), $\hat\pi_{\text{AT}} = n^{-1}\sum_{i=1}^n \psi(W_i,\hat\xi_\ell) = n^{-1}\sum_{i=1}^n\psi(W_i,\xi) + o_{\mathbb{P}}(1/\sqrt{nh})$, the bias $\mathbb{E}[\hat\pi_{\text{AT}}] - \pi_{\text{AT}} = h^2\mathsf{B}_\psi + o_{\mathbb{P}}(h^2)$, and $\sqrt{nh}\big(\hat\pi_{\text{AT}} - \pi_{\text{AT}} - h^2\mathsf{B}_\psi\big) \xrightarrow{d} \mathcal{N}(0,\mathsf{V}_\psi)$

The proof focuses on deriving the asymptotically linear representation of the numerator $\hat A$: $n^{-1}\sum_{i=1}^n g_{dU}(W_i,\hat\xi_\ell) = n^{-1}\sum_{i=1}^n g_{dU}(W_i,\xi) + o_{\mathbb{P}}(1/\sqrt{nh})$. By linearization, we obtain the asymptotically linear representation $\hat{\bar\beta}_d - \bar\beta_d = n^{-1}\sum_{i=1}^n\phi(W_i,\xi) - \bar\beta_d + o_{\mathbb{P}}(1/\sqrt{nh})$ and the bias $\mathbb{E}[\hat{\bar\beta}_d] - \bar\beta_d = h^2\mathsf{B}_d + o_{\mathbb{P}}(h^2)$. Then we will show $\sqrt{nh}\big(\hat{\bar\beta}_d - \bar\beta_d - h^2\mathsf{B}_d\big) \xrightarrow{d} \mathcal{N}(0,\mathsf{V}_d)$. We first show the special case under Assumption 7 with $J=1$, i.e., everyone has the same known unique sufficient treatment value d_{AT}. The results for $J\geq 2$ are implied if there is no classification error. Finally we show that the classification error is asymptotically ignorable.

The proof follows the arguments in the proof of Theorem 3.1 in Colangelo and Lee (2025) for the DML estimator of β_d under unconfoundedness without selection. The new challenges for $\bar{\beta}_d$ arise due to more nuisance functions ξ that enter the orthogonal moment function nonlinearly and complicate the notations. We re-define the nuisance functions so that the orthogonal moment function is linear in each of the re-defined nuisance functions.

The nuisance function estimator $\hat{\xi}_\ell$ uses observations $\mathcal{W}_\ell^c$. We will show $L^{-1}\sum_{\ell=1}^L \mathbb{E}_{n_\ell} g(W,\hat{\xi}_\ell) = \mathbb{E}_n g(W,\xi) + o_{\mathbb{P}}(1/\sqrt{nh})$ by showing $\mathbb{E}_{n_\ell} g(W,\hat{\xi}_\ell) - \mathbb{E}_{n_\ell} g(W,\xi) = o_{\mathbb{P}}(1/\sqrt{nh})$. Below we suppress the subscript ℓ in $\hat{\xi}_\ell$ to simplify notation as $\hat{\xi}$. We focus on the upper bound as the same arguments apply to the lower bound.

The orthogonal moment function $g_{dU}^j := m_{dU}^j + cor_{dU}^j$ is denoted as

$$g(W,\xi) = \mathsf{K}_d\lambda_d SY\mathsf{1} + \mathsf{Q}\bigg(\mathsf{K}_{d_j}\lambda_{d_j}(S-\mathsf{s}_{d_j}) + \mathsf{K}_d\lambda_d\big(\mathsf{s}_{d_j} - S\mathsf{1}\big)\bigg) + (1-\mathsf{K}_d\lambda_d)\rho\mathsf{s}_{d_j},$$

where the nuisance parameter $\xi = (\mathsf{s}_{d_j}, \mathsf{Q}, \lambda_d, \lambda_{d_j}, \rho, \mathsf{1})$.

Let $\xi_{-\iota} = \xi \setminus \xi_\iota$ for $\iota = 1,..,6$. As $g(W,\xi)$ is linear in each element in ξ, we can write $g(W,\hat{\xi}_\iota,\xi_{-\iota}) - g(W,\xi_\iota,\xi_{-\iota}) = \Delta\hat{\xi}_\iota \cdot g_{-\iota}(\xi_{-\iota})$. And $g_{-\iota}(\xi_{-\iota})$ is also linear in each of the remaining elements in $\xi_{-\iota}$.

For example, as $\xi_1 = \mathsf{s}_{d_j}$ and $\xi_{-1} = (\mathsf{Q}, \lambda_d, \lambda_{d_j}, \rho, \mathsf{1})$, we can write $g(W,\hat{\xi}_1,\xi_{-1}) - g(W,\xi_1,\xi_{-1}) = \Delta\hat{\mathsf{s}}_{d_j} \cdot g_{-1}(\xi_{-1})$, where $g_{-1}(\xi_{-1}) = \mathsf{Q}\big(-\mathsf{K}_{d_j}\lambda_{d_j} + \mathsf{K}_d\lambda_d\big) + (1-\mathsf{K}_d\lambda_d)\rho$.

As $\xi_2 = \mathsf{Q}$ and $\xi_{-12} = \xi \setminus (\xi_1,\xi_2) = \xi_{-1} \setminus \xi_2 = (\lambda_d, \lambda_{d_j}, \rho, \mathsf{1})$, and $g_{-12}(\xi_{-12}) = -\mathsf{K}_{d_j}\lambda_{d_j} + \mathsf{K}_d\lambda_d$, we can write $g_{-1}(\hat{\xi}_2,\xi_{-12}) - g_{-1}(\xi_2,\xi_{-12}) = \Delta\hat{\xi}_2 \cdot g_{-12}(\xi_{-12}) = \Delta\hat{\mathsf{Q}}\big(-\mathsf{K}_{d_j}\lambda_{d_j} + \mathsf{K}_d\lambda_d\big)$.

Further as $\xi_3 = \lambda_d$ and $g_{-123}(\xi_{-123}) = \mathsf{K}_d$, we can write $g_{-12}(\hat{\xi}_3,\xi_{-123}) - g_{-12}(\xi_3,\xi_{-123}) = \Delta\hat{\xi}_3 \cdot g_{-123}(\xi_{-123}) = \Delta\lambda_d\mathsf{K}_d$.

We decompose the remainder term $\mathbb{E}_{n_\ell}\big[g(W,\hat{\xi}) - g(W,\xi)\big]$ for $\ell = 1,..,L$ to the following and control each term to be $o_{\mathbb{P}}(1/\sqrt{nh})$:

$$\mathbb{E}_{n_\ell}\big[g(W,\hat{\xi}) - g(W,\xi)\big] = \sum_{\iota=1}^6 \mathbb{E}_{n_\ell}\big[g(W,\hat{\xi}_\iota,\xi_{-\iota}) - g(W,\xi)\big] - \mathbb{E}_\ell\big[g(W,\hat{\xi}_\iota,\xi_{-\iota}) - g(W,\xi)\big] \quad \text{(SE)}$$

$$+\mathbb{E}_\ell\big[g(W,\hat{\xi}_\iota,\xi_{-\iota}) - g(W,\xi)\big] \quad \text{(DR)}$$

$$+O_{\mathbb{P}}\left(\sum_{\iota\neq j\in\{1,...,6\}} \|\hat{\xi}_\iota - \xi_\iota\|_2 \|\hat{\xi}_j - \xi_j\|_2\right) + ... \quad \text{(Rem)}$$

(DR) We show the sufficient conditions for (DR)$= o_p(1/\sqrt{nh})$ are $\sqrt{nh}h^2\mathbb{E}[|\Delta\hat{\xi}_\iota|] = o_{\mathbb{P}}(1)$ for $\iota = 1,2,3,4,5$ and $\sqrt{nh}\|\hat{\mathsf{Q}} - \mathsf{Q}\|_2^2 = o_{\mathbb{P}}(1)$.

For $\iota = 1,...,5$, ξ_ι is a function of X. We can show $\mathbb{E}[g_{-\iota}(\xi_\iota)|X] = O_{\mathbb{P}}(h^2)$ by Assumption 8. This is the key benefit of the doubly robust moment function. Then by the law of iterated expectations, $\big|\mathbb{E}_\ell\big[g(W,\hat{\xi}_\iota,\xi_{-\iota}) - g(W,\xi)\big]\big| = \big|\mathbb{E}_\ell\big[\Delta\hat{\xi}_\iota \cdot g_{-\iota}(\xi_{-\iota})\big]\big| \leq \mathbb{E}_\ell\big[\big|\Delta\hat{\xi}_\iota\big| \cdot \big|\mathbb{E}[g_{-\iota}(\xi_{-\iota})|X]\big|\big] = O_{\mathbb{P}}(\mathbb{E}_\ell\big[\big|\Delta\hat{\xi}_\iota\big|\big]h^2)$. So it suffices to assume $\sqrt{nh}h^2\mathbb{E}[|\Delta\hat{\xi}_\iota|] = o_{\mathbb{P}}(1)$.

For example, by a standard algebra for kernel, we can show that $\mathbb{E}[\mathsf{K}_d|X] = \mu_d(X) + O_{\mathbb{P}}(h^2)$. For $\xi_1 = \mathsf{s}_{d_j}$, $\big|\mathbb{E}_\ell\big[g(W,\hat{\xi}_1,\xi_{-1}) - g(W,\xi)\big]\big| \leq \mathbb{E}_\ell\big[\big|\hat{\mathsf{s}}_{d_j} - \mathsf{s}_{d_j}\big| \cdot \big|\mathbb{E}[\mathsf{Q}(-\mathsf{K}_{d_j}\lambda_{d_j} + \mathsf{K}_d\lambda_d) + (1-\mathsf{K}_d\lambda_d)\rho|X]\big|\big] =$

$O_{\mathbb{P}}\big(\mathbb{E}_\ell\big[\big|\Delta\hat{\mathsf{s}}_{d_j}\big|\big]h^2\big)$.

The same argument applies to $\xi_\iota = \mathsf{Q}, \lambda_d, \lambda_{d_j}, \rho$. We can verify $\mathbb{E}[g_{-\iota}(\xi_{-\iota})|X] = O_{\mathbb{P}}(h^2)$ by Assumption 8(ii), where $g_{-2}(\xi_{-2}) = \mathsf{K}_{d_j}\lambda_{d_j}(S - \mathsf{s}_{d_j}) + \mathsf{K}_d\lambda_d\big(\mathsf{s}_{d_j} - S1\big)$, $g_{-3}(\xi_{-3}) = \mathsf{K}_d SY1 + \mathsf{Q}\mathsf{K}_d\big(\mathsf{s}_{d_j} - S1\big) - \mathsf{K}_d\rho\mathsf{s}_{d_j}$, $g_{-4}(\xi_{-4}) = \mathsf{Q}\mathsf{K}_{d_j}(S - \mathsf{s}_{d_j})$, $g_{-5}(\xi_{-5}) = (1 - \mathsf{K}_d\lambda_d)\mathsf{s}_{d_j}$.

Now consider (DR) with $\xi_6 = 1$. The following second equality uses $\int_{\hat{\mathsf{Q}}}^{\infty}(y - \mathsf{Q})f_{Y|SDX}dy = \int_{\mathsf{Q}}^{\infty}(y - \mathsf{Q})f_{Y|SDX}dy - \frac{1}{2}\mathsf{Q}f_{Y|SDX}(\mathsf{Q})(\hat{\mathsf{Q}} - \mathsf{Q})^2 + o_{\mathbb{P}}((\Delta\hat{\mathsf{Q}})^2)$, by Taylor expansion, Leibniz rule, and uniformly bounded $f'_{Y|SDX}$. So $\mathbb{E}_\ell\big[\Delta\hat{\xi}_6 \cdot g_{-6}(\xi_{-6})\big] = \mathbb{E}_\ell\big[(\hat{1} - 1) \cdot \mathsf{K}_d\lambda_d S(Y - \mathsf{Q})\big]$

$$
\begin{aligned}
&= \mathbb{E}_\ell\Big[\mathsf{K}_d\lambda_d S \cdot \Big(\int_{\hat{\mathsf{Q}}}^{\infty}(y - \mathsf{Q})f_{Y|SDX}dy - \int_{\mathsf{Q}}^{\infty}(y - \mathsf{Q})f_{Y|SDX}dy\Big)\Big] \\
&= \mathbb{E}_\ell\Big[\big(s(d, X) + O_{\mathbb{P}}(h^2)\big) \cdot \Big(-\frac{1}{2}\mathsf{Q}f_{Y|SDX}(\mathsf{Q})(\hat{\mathsf{Q}} - \mathsf{Q})^2 + o_{\mathbb{P}}((\Delta\hat{\mathsf{Q}})^2)\Big)\Big] \\
&= O_{\mathbb{P}}(\|\Delta\hat{\mathsf{Q}}\|_2^2).
\end{aligned}
$$

So it suffices to assume $\sqrt{nh}\|\hat{\mathsf{Q}} - \mathsf{Q}\|_2^2 = o_{\mathbb{P}}(1)$.

(SE) We show that the sufficient condition for (SE)$= o_p(1/\sqrt{nh})$ is $\|\Delta\hat{\xi}_\iota\|_2 = o_{\mathbb{P}}(1)$ for $\iota = 1, ..., 5$, due to cross-fitting.

Define $\Delta_{i\ell} := g(W_i, \hat{\xi}_\iota, \xi_{-\iota}) - g(W_i, \xi) - \mathbb{E}_\ell\big[g(W_i, \hat{\xi}_\iota, \xi_{-\iota}) - g(W_i, \xi)\big] = \Delta\xi_\iota g_{-\iota}(\xi_{-\iota}) - \mathbb{E}_\ell\big[\Delta\xi_\iota g_{-\iota}(\xi_{-\iota})\big]$. By construction and independence of $\mathcal{W}_\ell^c$ and W_i for $i \in I_\ell$, $\mathbb{E}_\ell[\Delta_{i\ell}] = 0$ and $\mathbb{E}_\ell[\Delta_{i\ell}\Delta_{j\ell}] = 0$ for $i, j \in I_\ell$. If $\mathbb{E}_\ell[(\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta_{i\ell})^2] = h\mathbb{E}_\ell[\Delta_{i\ell}^2] = o_p(1)$, then the conditional Markov's inequality implies that $\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta_{i\ell} = o_{\mathbb{P}}(1)$. So it suffices to show that $h\mathbb{E}_\ell[\Delta_{i\ell}^2] = o_p(1)$.

Because $\mathbb{E}_\ell\big[\Delta\xi_\iota g_{-\iota}(\xi_{-\iota})\big] = o_{\mathbb{P}}(1/\sqrt{nh})$ as shown above for (DR), $h\mathbb{E}_\ell[\Delta_{i\ell}^2] = O_{\mathbb{P}}\Big(h\mathbb{E}_\ell\big[(\Delta\xi_\iota)^2 \times g_{-\iota}(\xi_{-\iota})^2\big]\Big) = O_{\mathbb{P}}\Big(h\mathbb{E}_\ell\big[(\Delta\xi_\iota)^2\mathbb{E}\big[g_{-\iota}(\xi_{-\iota})^2|X\big]\big]\Big) = O_{\mathbb{P}}\Big(\mathbb{E}_\ell\big[(\Delta\xi_\iota)^2\big]\Big)$, by showing $h\mathbb{E}\big[g_{-\iota}(\xi_{-\iota}^2)|X\big] = O_{\mathbb{P}}(1)$.

Specifically $h\mathbb{E}[\mathsf{K}_d^2|X] = \int h^{-1}k((D - d)/h)^2 f_{D|X}(D|X)dD = \int k(u)^2 f_{D|X}(d + uh|X)du = R_k f_{D|X}(d|X) + o_{\mathbb{P}}(1) = O_{\mathbb{P}}(1)$. For example of $\xi_1 = \mathsf{s}_{d_j}$, $h\mathbb{E}_\ell[\Delta_{i\ell}^2] = O_{\mathbb{P}}\Big(h\mathbb{E}_\ell[(\Delta\hat{\mathsf{s}}_{d_j})^2 \cdot \mathbb{E}[(\mathsf{Q}(-\mathsf{K}_{d_j}\lambda_{d_j} + \mathsf{K}_d\lambda_d) + (1 - \mathsf{K}_d\lambda_d)\rho)^2|X]]\Big) = O_{\mathbb{P}}(\|\Delta\hat{\mathsf{s}}_{d_j}\|_2^2)$. The same argument applies to $\xi_\iota = \mathsf{Q}, \rho, \lambda_d, \lambda_{d_j}$.

For $\xi_6 = 1$, it suffices to show that $h\mathbb{E}_\ell\big[(\hat{1} - 1)^2 \cdot \mathsf{K}_d^2\lambda_d^2 S(Y - \mathsf{Q})^2\big] = hE_\ell\big[\mathsf{K}_d^2\lambda_d^2 S \cdot \mathbb{E}[\mathbf{1}\{\mathsf{Q} < Y \le \hat{\mathsf{Q}}\} \cdot (Y - \mathsf{Q})^2|S, D, X]\big] = o_{\mathbb{P}}(1)$, considering $\hat{\mathsf{Q}} > \mathsf{Q}$ without loss of generality. By integration by parts, $\mathbb{E}[\mathbf{1}\{\mathsf{Q} < Y \le \hat{\mathsf{Q}}\} \cdot (Y - \mathsf{Q})^2|S, D, X] = \int_{\mathsf{Q}}^{\hat{\mathsf{Q}}}(y - \mathsf{Q})^2 f_{Y|SDX}(y)dy = (\hat{\mathsf{Q}} - \mathsf{Q})^2 F_{Y|SDX}(\hat{\mathsf{Q}}) - 2\int_{\mathsf{Q}}^{\hat{\mathsf{Q}}}(y - \mathsf{Q})F_{Y|SDX}(y)dy = O_{\mathbb{P}}(|\Delta\hat{\mathsf{Q}}|^2) = o_{\mathbb{P}}(1)$, because the last term $\int_{\mathsf{Q}}^{\hat{\mathsf{Q}}}(y - \mathsf{Q})F_{Y|SDX}(y)dy = \frac{1}{2}(\hat{\mathsf{Q}} - \mathsf{Q})^2 F_{Y|SDX}(\bar{\mathsf{Q}}) = O_{\mathbb{P}}(|\Delta\hat{\mathsf{Q}}|^2)$ for some $\bar{\mathsf{Q}}$ between Q and $\hat{\mathsf{Q}}$, by Taylor expansion and Leibniz rule. By the same algebra of kernel as above, we can show $hE_\ell[\mathsf{K}_d^2\lambda_d^2 S] = O_{\mathbb{P}}(1)$. Then we obtain the result.

(Rem) We show that the sufficient condition for (Rem)$= o_{\mathbb{P}}(1/\sqrt{nh})$ is $\|\Delta\hat{\xi}_\iota\|_2\|\Delta\hat{\xi}_j\|_2 = o_{\mathbb{P}}(1/\sqrt{nh})$ for $\iota \neq j$.

Consider a simplified case when there are three elements in $\xi = (\xi_1, \xi_2, \xi_3)$. The same arguments apply to the general result with more complicated notations. We suppress W in the function $g(W, \xi) = g(\xi_1, \xi_2, \xi_3)$ when there is no confusion. We show the remainder terms in (Rem)

$$
\begin{aligned}
&g(\hat{\xi}) - g(\xi) - \big(g(\hat{\xi}_1, \xi_{-1}) - g(\xi)\big) - \big(g(\hat{\xi}_2, \xi_{-2}) - g(\xi)\big) - \big(g(\hat{\xi}_3, \xi_{-3}) - g(\xi)\big) \\
&- g(\xi_1, \hat{\xi}_2, \hat{\xi}_3) + g(\xi_1, \hat{\xi}_2, \hat{\xi}_3) \\
&= \Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_3) + \Delta\hat{\xi}_1\Delta\hat{\xi}_3 g_{-13}(\xi_2) + \Delta\hat{\xi}_2\Delta\hat{\xi}_3 g_{-23}(\xi_1) + \Delta\hat{\xi}_1\Delta\hat{\xi}_2\Delta\hat{\xi}_3 g_{-123}, \text{where} \qquad (7) \\
&g(\hat{\xi}) - g(\xi_1, \hat{\xi}_2, \hat{\xi}_3) - \big(g(\hat{\xi}_1, \xi_{-1}) - g(\xi)\big) = \Delta\hat{\xi}_1 g_{-1}(\hat{\xi}_2, \hat{\xi}_3) - \Delta\hat{\xi}_1 g_{-1}(\xi_2, \xi_3) + \Delta\hat{\xi}_1 g_{-1}(\xi_2, \hat{\xi}_3) - \Delta\hat{\xi}_1 g_{-1}(\xi_2, \hat{\xi}_3) \\
&= \Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\hat{\xi}_3) + \Delta\hat{\xi}_1\Delta\hat{\xi}_3 g_{-13}(\xi_2) + \Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_3) - \Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_3) \\
&= \Delta\hat{\xi}_1\Delta\hat{\xi}_3 g_{-13}(\xi_2) + \Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_3) + \Delta\hat{\xi}_1\Delta\hat{\xi}_2\big(g_{-12}(\hat{\xi}_3) - g_{-12}(\xi_3)\big) \\
&= \Delta\hat{\xi}_1\Delta\hat{\xi}_3 g_{-13}(\xi_2) + \Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_3) + \Delta\hat{\xi}_1\Delta\hat{\xi}_2\Delta\hat{\xi}_3 g_{-123}, \text{ and} \\
&- \big(g(\hat{\xi}_2, \xi_{-2}) - g(\xi)\big) - g(\hat{\xi}_3, \xi_{-3}) + g(\xi_1, \hat{\xi}_2, \hat{\xi}_3) = -\Delta\hat{\xi}_2 g_{-2}(\xi_1, \xi_3) + \Delta\hat{\xi}_2 g_{-2}(\xi_1, \hat{\xi}_3) = \Delta\hat{\xi}_2\Delta\hat{\xi}_3 g_{-23}(\xi_1).
\end{aligned}
$$

We start with the term $\Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_{-12})$ in (7). We show $\mathbb{E}_\ell\left[\left|\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta\hat{\xi}_1\Delta\hat{\xi}_2 g_{-12}(\xi_{-12})\right|\right] = o_{\mathbb{P}}(1)$ by focusing on the second term in $g_{-12}(\xi_{-12}) = -\mathsf{K}_{d_j}\lambda_{d_j} + \mathsf{K}_d\lambda_d$ below, as the same argument applies to the first term. So (Rem) is $o_{\mathbb{P}}(1/\sqrt{nh})$ follows by the conditional Markov's inequality and triangle inequalities.

$$
\begin{aligned}
&\mathbb{E}_\ell\left[\left|\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta\hat{\xi}_1\Delta\hat{\xi}_2\mathsf{K}_d\lambda_d\right|\right] \le \sqrt{n_\ell h}\int_{\mathcal{X}}\int_{\mathcal{D}}\left|\Delta\hat{\mathsf{s}}_{d_j}\right|\left|\Delta\hat{\mathsf{Q}}\right| K_d(D-d)\lambda_d f_{DX}(D,X)dDdX \\
&\le \sqrt{n_\ell h}\left(\int_{\mathcal{X}}\int_{\mathcal{D}}\left(\Delta\hat{\mathsf{s}}_{d_j}\right)^2 K_d(D-d)\lambda_d f_{DX}(D,X)dDdX\right)^{1/2}\left(\int_{\mathcal{X}}\int_{\mathcal{D}}\left(\Delta\hat{\mathsf{Q}}\right)^2 K_d(D-d)\lambda_d f_{DX}(D,X)dDdX\right)^{1/2} \\
&= O_{\mathbb{P}}\left(\sqrt{n_\ell h}\left(\int_{\mathcal{X}}\left(\Delta\hat{\mathsf{s}}_{d_j}\right)^2 f_X(X)dX\right)^{1/2}\left(\int_{\mathcal{X}}\left(\Delta\hat{\mathsf{Q}}\right)^2 f_X(X)dX\right)^{1/2}\right) = o_{\mathbb{P}}(1)
\end{aligned}
$$

by Cauchy-Schwarz inequality and Assumption 8. The same argument applies to other terms in (7). Specifically, $g_{-13}(\xi_{-13}) = \mathsf{K}_d(\mathsf{Q}-\rho)$, $g_{-14}(\xi_{-14}) = -\mathsf{K}_{d_j}\mathsf{Q}$, $g_{-15}(\xi_{-15}) = 1-\mathsf{K}_d\lambda_d$, $g_{-16}(\xi_{-16}) = 0$. By the law of iterated expectations and assuming $\mathbb{E}\big[|g_{-1\iota}(W,\xi_{-1\iota})|\big|X\big]$ is uniformly bounded, for $\iota = 3,4,5$, $\mathbb{E}_\ell\left[\left|\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta\hat{\xi}_1\Delta\hat{\xi}_\iota g_{-1\iota}(W,\xi_{-1\iota})\right|\right] \le \sqrt{n_\ell h}\mathbb{E}\Big[\left|\Delta\hat{\xi}_1\right|\left|\Delta\hat{\xi}_\iota\right|\mathbb{E}\big[|g_{-1\iota}(W,\xi_{-1\iota})|\big|X\big]\Big] \le \sqrt{n_\ell h}\left(\int_{\mathcal{X}}\left(\Delta\hat{\xi}_1\right)^2\mathbb{E}\big[|g_{-1\iota}(W,\xi_{-1\iota})|\big|X\big]f(X)dX\right)^{1/2}\left(\int_{\mathcal{X}}\left(\Delta\hat{\xi}_\iota\right)^2\mathbb{E}\big[|g_{-1\iota}(W,\xi_{-1\iota})|\big|X\big]f_X(X)dX\right)^{1/2} = O_{\mathbb{P}}\left(\sqrt{n_\ell h}\left(\int_{\mathcal{X}}\left(\Delta\hat{\xi}_1\right)^2 f_X(X)dX\right)^{1/2}\left(\int_{\mathcal{X}}\left(\Delta\hat{\xi}_\iota\right)^2 f_X(X)dX\right)^{1/2}\right) = o_{\mathbb{P}}(1)$.

The same argument applies to ξ_ι, $\iota = 2,3,4,5$. Specifically we can show that $g_{-23} = \mathsf{K}_d(\mathsf{s}_{d_j} - S1)$, $g_{-24} = \mathsf{K}_{d_j}(S - \mathsf{s}_{d_j})$, $g_{-25} = 0$, $g_{-34} = 0$, $g_{-35} = -\mathsf{K}_{d_j}\mathsf{s}_{d_j}$, $g_{-45} = 0$.

Now consider $\xi_6 = 1 = \mathbf{1}\{Y \ge \mathsf{Q}\}$. Then $g_{-6}(\xi_{-6}) = \mathsf{K}_d\lambda_d S(Y-\mathsf{Q})$, $g_{-26}(\xi_{-26}) = -\mathsf{K}_d\lambda_d S$, $g_{-36}(\xi_{-36}) = \mathsf{K}_d S(Y-\mathsf{Q})$, and $g_{-\iota 6} = 0$ for $\iota = 1,4,5$.

We first compute $\mathbb{E}_\ell\left[\left(\Delta\hat{1}\right)^2\Big|S,D,X\right] = \mathbb{E}_\ell\left[\left(\mathbf{1}\{Y \ge \hat{\mathsf{Q}}\} - \mathbf{1}\{Y \ge \mathsf{Q}\}\right)^2\Big|S,D,X\right] = \big|F_{Y|SDX}(\hat{\mathsf{Q}}) -$

$F_{Y|SDX}(\mathsf{Q})\big| = f_{Y|SDX}(\mathsf{Q})\big|\hat{\mathsf{Q}} - \mathsf{Q}\big| + O_{\mathbb{P}}((\Delta\hat{\mathsf{Q}})^2)$. For $\Delta\hat{\xi}_2\Delta\hat{\xi}_6 g_{-26}(\xi_{-26})$ in (7),

$$\begin{aligned}
&\mathbb{E}_\ell\left[\left|\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta\hat{\mathsf{Q}}\Delta\hat{\mathsf{1}}\mathsf{K}_d\lambda_d S\right|\right] \le \sqrt{n_\ell h}\mathbb{E}_\ell\left[\left|\Delta\hat{\mathsf{Q}}\right|\left|\Delta\hat{\mathsf{1}}\right| K_d(D-d)\lambda_d S\right] \\
&\le \sqrt{n_\ell h}\left(\int_{\mathcal{X}}\int_{\mathcal{D}}\left(\Delta\hat{\mathsf{Q}}\right)^2 K_d(D-d)\lambda_d s(D,X) f_{DX}(D,X)dDdX\right)^{1/2} \\
&\times\left(\int_{\mathcal{X}}\int_{\mathcal{D}}\mathbb{E}_\ell\left[\left(\Delta\hat{\mathsf{1}}\right)^2|S=1,D,X\right] K_d(D-d)\lambda_d s(d,X) f_{DX}(D,X)dDdX\right)^{1/2} \\
&= O_{\mathbb{P}}\left(\sqrt{n_\ell h}\left(\int_{\mathcal{X}}\left(\Delta\hat{\mathsf{Q}}\right)^2 f_X(X)dX\right)^{1/2}\left(\int_{\mathcal{X}}\left|\Delta\hat{\mathsf{Q}}\right| f_X(X)dX\right)^{1/2}\right),
\end{aligned}$$

which is $o_{\mathbb{P}}(1)$ by assuming $\sqrt{nh}\|\Delta\hat{\mathsf{Q}}\|_2^2 = o_{\mathbb{P}}(1)$.

Now we turn to $\Delta\hat{\xi}_3\Delta\hat{\xi}_6 g_{-36}(\xi_{-36})$ in (7). We first compute $\mathbb{E}_\ell\left[\left(\Delta\hat{\mathsf{1}}\right)^2|Y-\mathsf{Q}||S=1,D,X\right] = \mathbb{E}_\ell\left[\left(\mathbf{1}\{\mathsf{Q}>Y\ge\hat{\mathsf{Q}}\}+\mathbf{1}\{\mathsf{Q}\le Y<\hat{\mathsf{Q}}\}\right)|Y-\mathsf{Q}||S=1,D,X\right] = \mathbf{1}\{\mathsf{Q}>\hat{\mathsf{Q}}\}\int_{\hat{\mathsf{Q}}}^{\mathsf{Q}}(\mathsf{Q}-Y)f_{Y|SDX}(y)dy + \mathbf{1}\{\mathsf{Q}<\hat{\mathsf{Q}}\}\int_{\mathsf{Q}}^{\hat{\mathsf{Q}}}(Y-\mathsf{Q})f_{Y|SDX}(y)dy = \frac{1}{2}f_{Y|SDX}(\mathsf{Q})(\Delta\hat{\mathsf{Q}})^2 + o_{\mathbb{P}}((\Delta\hat{\mathsf{Q}})^2)$. Then

$$\begin{aligned}
&\mathbb{E}_\ell\left[\left|\sqrt{h/n_\ell}\sum_{i\in I_\ell}\Delta\hat{\lambda}_d\Delta\hat{\mathsf{1}}\mathsf{K}_d S(Y-\mathsf{Q})\right|\right] \le \sqrt{n_\ell h}\mathbb{E}_\ell\left[\left|\Delta\hat{\lambda}_d\right|\left|\Delta\hat{\mathsf{1}}\right|\mathsf{K}_d S|Y-\mathsf{Q}|\right] \\
&\le \sqrt{n_\ell h}\left(\int_{\mathcal{X}}\int_{\mathcal{D}}\left(\Delta\hat{\lambda}_d\right)^2\mathsf{K}_d s(D,X)\mathbb{E}[|Y-\mathsf{Q}||S=1,D,X]f_{DX}(D,X)dDdX\right)^{1/2} \\
&\quad\times\left(\int_{\mathcal{X}}\int_{\mathcal{D}}\mathbb{E}_\ell\left[\left(\Delta\hat{\mathsf{1}}\right)^2|Y-\mathsf{Q}|\Big|S=1,D,X\right]\mathsf{K}_d s(D,X)f_{DX}(D,X)dDdX\right)^{1/2} \\
&= O_{\mathbb{P}}\Big(\sqrt{n_\ell h}\Big(\int_{\mathcal{X}}\left(\Delta\hat{\lambda}_d\right)^2 s(d,X)\mathbb{E}[|Y-\mathsf{Q}||S=1,D=d,X]f_{DX}(d,X)dX\Big)^{1/2} \\
&\quad\times\Big(\frac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{D}}f_{Y|SDX}(\mathsf{Q})(\Delta\hat{\mathsf{Q}})^2 s(d,X)f_{DX}(d,X)dX\Big)^{1/2}\Big) + o_{\mathbb{P}}(h^2),
\end{aligned}$$

which is $o_{\mathbb{P}}(1)$ by assuming $\sqrt{nh}\|\Delta\hat{\lambda}_d\|_2\|\Delta\hat{\mathsf{Q}}\|_2 = o_{\mathbb{P}}(1)$ and $\mathbb{E}[|Y-\mathsf{Q}||S=1,D=d,X]$ and $f_{Y|SDX}$ are uniformly bounded.

In sum, to control the reminder term $\mathbb{E}_{n_\ell}\big[g(W,\hat{\xi})-g(W,\xi)\big] = o_{\mathbb{P}}(1/\sqrt{nh})$, the sufficient rate conditions are $\|\Delta\hat{\xi}_\iota\|_2\|\Delta\hat{\xi}_j\|_2 = o_{\mathbb{P}}(1/\sqrt{nh})$ for $\iota\neq j$, $\|\Delta\hat{\xi}_\iota\|_2 = o_{\mathbb{P}}(1)$ for $\iota = 1,2,3,4,5$, and $\sqrt{nh}\|\hat{\mathsf{Q}}-\mathsf{Q}\|_2^2 = o_{\mathbb{P}}(1)$.

Assuming $\sqrt{nh}h^2\to C\in[0,\infty)$ and $\|\Delta\hat{\xi}_\iota\|_2 = o_p(1)$ yields $\sqrt{nh}h^2\mathbb{E}_\ell\big[\big|\Delta\hat{\mathsf{s}}_{d_j}\big|\big] = o_p(1)$.

We derive the sufficient rate conditions in Assumption 8(iii)(iv) based on $\|\Delta\hat{\xi}\|_2$. For $\hat{\mathsf{Q}} = \hat{Q}^d(1-\hat{p}_{dd_j}(X),X)$, $\Delta\hat{\mathsf{Q}} = \hat{Q}^d(1-\hat{p}_{dd_j}(X),X) - Q^d(1-\hat{p}_{dd_j}(X),X) + Q^d(1-\hat{p}_{dd_j}(X),X) + Q^d(1-p_{dd_j}(X),X) = O_{\mathbb{P}}\Big(\sup_{p\in(0,1)}\Big|\hat{Q}^d(p,X)-Q^d(p,X)\Big|\Big) + O_{\mathbb{P}}\big(|\hat{s}(d,X)-s(d,X)|+|\hat{s}(d_j,X)-s(d_j,X)|\big)$.

For $\hat{\rho} = \hat{\mathbb{E}}[Y|Y\ge\hat{\mathsf{Q}},S=1,D=d,X]$, $\|\Delta\hat{\rho}\|_2 \le \|\hat{\mathbb{E}}[Y|Y\ge\hat{\mathsf{Q}},S=1,D=d,X] - \mathbb{E}[Y|Y\ge$

$\hat{\mathsf{Q}}, S=1, D=d, X]\|_2 + \|\mathbb{E}[Y|Y \geq \hat{\mathsf{Q}}, S=1, D=d, X] - \mathbb{E}[Y|Y \geq \mathsf{Q}, S=1, D=d, X]\|_2 \leq \sup_{y\in\mathcal{Y}_0} \|\Delta\hat{\mathbb{E}}[Y|Y\geq y, S=1, D=d, X]\|_2 + O_{\mathbb{P}}(\sup_{p\in(0,1)} \|\Delta\hat{\mathsf{Q}}\|_2)$.

Bias We use the law of iterated expectations, the dominated convergence theorem, and standard algebra of kernel in the following.

$$
\begin{aligned}
&\mathbb{E}\Big[\mathsf{K}_d\lambda_d S\mathbb{E}[Y1|S,D,X] + \mathsf{Q}\Big(\mathsf{K}_{d_j}\lambda_{d_j}(s(D,X)-\mathsf{s}_{d_j}) + \mathsf{K}_d\lambda_d\big(\mathsf{s}_{d_j} - S(1-F_{Y|SDX}(\mathsf{Q}|S,D,X))\big)\Big)\\
&\quad + (1-\mathbb{E}[\mathsf{K}_d|X]\lambda_d)\rho\mathsf{s}_{d_j}\Big]\\
&= \mathbb{E}\Big[\mathsf{K}_d\lambda_d s(D,X)\mathbb{E}[Y|Y\geq\mathsf{Q},S=1,D,X]\big(1-F_{Y|SDX}(\mathsf{Q}|1,D,X)\big) + \mathsf{Q}\Big(\lambda_{d_j}\Big\{f_{D|X}(d_j,X)s(d_j,X)\\
&\quad + \frac{h^2}{2}\kappa\frac{\partial^2}{\partial d_j^2}\big(f_{D|X}(d_j,X)s(d_j,X)\big) - f_{D|X}(d_j|X)\mathsf{s}_{d_j} - \frac{h^2}{2}\kappa\mathsf{s}_{d_j}\frac{\partial^2}{\partial d_j^2}f_{D|X}(d_j|X)\Big\}\\
&\quad + \lambda_d\Big\{f_{D|X}(d|X)\mathsf{s}_{d_j} + \frac{h^2}{2}\kappa\mathsf{s}_{d_j}\frac{\partial^2}{\partial d^2}f_{D|X}(d,X) - \mathsf{K}_d s(D,X)(1-F_{Y|SDX}(\mathsf{Q}|S=1,D,X))\Big\}\Big)\\
&\quad + \big(1-f_{D|X}(d|X)\lambda_d - \frac{h^2}{2}\kappa\lambda_d\frac{\partial^2}{\partial d^2}f_{D|X}(d,X)\big)\rho\mathsf{s}_{d_j}\Big] + o_{\mathbb{P}}(h^2)\\
&= \mathbb{E}\Big[\rho s(d_j,X) + \frac{h^2}{2}\kappa\frac{\partial^2}{\partial d^2}\Big(f_{D|X}(d|X)s(d,x)\mathbb{E}[Y|Y\geq\mathsf{Q},S=1,D=d,X]\big(1-F_{Y|SDX}(\mathsf{Q}|1,d,X)\big)\Big)\\
&\quad + \mathsf{Q}\Big(\mathsf{s}_{d_j} + \frac{h^2}{2}\kappa\lambda_{d_j}\frac{\partial^2}{\partial d_j^2}\big(f_{D|X}(d_j,X)s(d_j,X)\big) - \mathsf{s}_{d_j} - \frac{h^2}{2}\kappa\lambda_{d_j}\mathsf{s}_{d_j}\frac{\partial^2}{\partial d_j^2}f_{D|X}(d_j|X) + \mathsf{s}_{d_j}\\
&\quad + \frac{h^2}{2}\kappa\lambda_d\mathsf{s}_{d_j}\frac{\partial^2}{\partial d^2}f_{D|X}(d,X) - \mathsf{s}_{d_j} - \frac{h^2}{2}\kappa\lambda_d\frac{\partial^2}{\partial d^2}\Big(f_{D|X}(d|X)s(d,X)\big(1-F_{Y|SDX}(\mathsf{Q}|S=1,d,X)\big)\Big)\Big)\\
&\quad - \frac{h^2}{2}\kappa\lambda_d\rho\mathsf{s}_{d_j}\frac{\partial^2}{\partial d^2}f_{D|X}(d,X)\Big] + o_{\mathbb{P}}(h^2)\\
&= \mathbb{E}\big[\rho\mathsf{s}_{d_j}\big] + h^2\mathsf{B}_{gU} + o_{\mathbb{P}}(h^2).
\end{aligned}
$$

The same arguments apply to B_{gL}.

Variance. Note that $\mathbb{E}[S|D=d_j,X] = s(d_j,X) = \mathsf{s}_{d_j}$ and hense $var(S|D=d_j,X) = \mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})$. Theorem 3.1 in Colangelo and Lee (2025) gives $hvar(\psi(W_i,\xi)) \to R_k\mathbb{E}[\mathbb{E}[(S-\mathsf{s}_{d_j})^2|D=d_j,X]\lambda_{d_j}] = R_k\mathbb{E}[var(S|D=d_j,X)\lambda_{d_j}] = \mathsf{V}_\psi$.

First compute $h\mathbb{E}[\mathsf{K}_{d_j}^2(S-\mathsf{s}_{d_j})^2|X] = h\int_{\mathcal{D}} h^{-2}k((v-d_j)/h)^2\mathbb{E}[(S-\mathsf{s}_{d_j})^2|D=v,X]f_{D|X}(v)dv = \int k(u)^2\mathbb{E}[(S-\mathsf{s}_{d_j})^2|D=d_j+hu,X]f_{D|X}(d_j+uh)du = R_k var(S|D=d_j,X)f_{D|X}(d_j)+o(1) = R_k\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})\lambda_{d_j}^{-1}+o(1)$.

In $h\mathbb{E}[g(W,\xi)\psi(W_i,\xi)]$, the product term with $\mathsf{K}_{d_j}\mathsf{K}_d$ results in a convolution kernel $\int k((d-d_j)/h+u)k(u)du$ and hence is $o(1)$. So $h\mathbb{E}[g(W,\xi)\psi(W_i,\xi)] = h\mathbb{E}[\mathsf{Q}\mathsf{K}_{d_j}^2\lambda_{d_j}^2(S-\mathsf{s}_{d_j})^2]+o(1)\to\mathsf{V}_{g\psi}$.

Next we compute $hvar(g(W_i,\xi)) = h\mathbb{E}[(g(W,\xi)-\mathbb{E}[\rho\mathsf{s}_{d_j}])^2]+o(1)$ with

$$
g(W,\xi)-\mathbb{E}[\rho\mathsf{s}_{d_j}] = \mathsf{K}_d\lambda_d(SY1-\rho\mathsf{s}_{d_j}) + \mathsf{Q}\mathsf{K}_{d_j}\lambda_{d_j}(S-\mathsf{s}_{d_j}) + \mathsf{Q}\mathsf{K}_d\lambda_d\big(\mathsf{s}_{d_j}-S1\big) + \big(\rho\mathsf{s}_{d_j}-\mathbb{E}[\rho\mathsf{s}_{d_j}]\big), \quad (8)
$$

For the first term in (8), $h\mathbb{E}[\mathsf{K}_d^2\lambda_d^2(SY\mathsf{1}-\rho\mathsf{s}_{d_j})^2] = h\mathbb{E}[\mathbb{E}[\mathsf{K}_d^2(SY^2\mathsf{1}-2SY\mathsf{1}\rho\mathsf{s}_{d_j}+\rho^2\mathsf{s}_{d_j}^2)|X]\lambda_d^2]$

$$
\begin{aligned}
&= h\mathbb{E}[\mathbb{E}[\mathsf{K}_d^2(\mathbb{E}[Y^2|Y\geq \mathsf{Q},D,X](1-F_{Y|S,D,X}(\mathsf{Q}|1,D,X))s(D,X)\\
&- 2\mathbb{E}[Y|Y\geq \mathsf{Q},D,X](1-F_{Y|S,D,X}(\mathsf{Q}|1,D,X))s(D,X)\rho\mathsf{s}_{d_j}+\rho^2\mathsf{s}_{d_j}^2)|X]\lambda_d^2]\\
&= R_k\mathbb{E}[\mathbb{E}[Y^2|Y\geq \mathsf{Q},D=d,X]\mathsf{s}_{d_j}-2\rho^2\mathsf{s}_{d_j}^2+\rho^2\mathsf{s}_{d_j}^2)f_{D|X}(d,X)\lambda_d^2]+o(1)\\
&= R_k\mathbb{E}[\mathbb{E}[Y^2|Y\geq \mathsf{Q},D=d,X]\mathsf{s}_{d_j}-\rho^2\mathsf{s}_{d_j}^2+\rho^2\mathsf{s}_{d_j}-\rho^2\mathsf{s}_{d_j})\lambda_d]+o(1)\\
&= R_k\mathbb{E}[(var(Y^2|Y\geq \mathsf{Q},D=d,X]\mathsf{s}_{d_j}+\rho^2\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j}))\lambda_d]+o(1).
\end{aligned}
$$

Similar arguments apply to the second term in (8), so $h\mathbb{E}[\mathsf{Q}^2\mathsf{K}_{d_j}^2\lambda_{d_j}^2(S-\mathsf{s}_{d_j})^2] = R_k\mathbb{E}[\mathsf{Q}^2\lambda_{d_j}\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})] + o(1)$. And for the third term, $h\mathbb{E}[\mathsf{Q}^2\mathsf{K}_d^2\lambda_d^2(\mathsf{s}_{d_j}-S\mathsf{1})^2] = R_k\mathbb{E}[\mathsf{Q}^2\lambda_d\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})] + o(1)$. For the last term, $h\mathbb{E}[(\rho\mathsf{s}_{d_j}-\mathbb{E}[\rho\mathsf{s}_{d_j}])^2] = O(h) = o(1)$.

For the product of the first and third terms, $h\mathbb{E}[\mathsf{Q}\mathsf{K}_d^2\lambda_d^2(SY\mathsf{1}-\rho\mathsf{s}_{d_j})(\mathsf{s}_{d_j}-S\mathsf{1})] = -R_k\mathbb{E}[\mathsf{Q}\lambda_d\rho\mathsf{s}_{d_j}(1-\mathsf{s}_{d_j})] + o(1)$. The other cross products are $o(1)$ by the law of iterated expectations. Therefore, we obtain $hvar(g(W_i,\xi)) = \mathsf{V}_g + o(1)$.

Asymptotic normality. Asymptotic normality follows from the Lyapunov central limit theorem with the third absolute moment. Let $s_n^2 := \sum_{i=1}^n var\big(\sqrt{nh}n^{-1}\phi(W_i,\xi)\big) = hvar(\phi) = \mathsf{V}_d + o(1)$ as shown above. If $\mathbb{E}\big[\big|\sqrt{nh}n^{-1}\phi(W,\xi)\big|^3\big] = O((n^3h)^{-1/2})$, then the Lyapunov condition holds: $\sum_{i=1}^n \mathbb{E}\big[\big|\sqrt{nh}n^{-1}\phi(W_i,\xi)\big|^3\big]/s_n^3 = O((nh)^{-1/2}) = o(1)$. That is, it suffices to show that $\mathbb{E}[|\phi|^3] = O(h^{-2})$, which holds by assuming that $\mathbb{E}[|Y|^3\mathbf{1}|D=d,S=1,X]$ is continuous in d uniformly over $\mathcal{X}$ and other assumed conditions.

Misclassification. Following the proof of Theorem 2, let the true $\tau_j(x) = s(d_{j+1},x) - s(d_j,x)$ with the estimate $\hat{\tau}_j(x) = \hat{s}(d_{j+1},x) - \hat{s}(d_j,x)$. Let $\mathcal{M}_j = \{x : \hat{\tau}_j(x) < 0 < \tau_j(x)\} \cup \{x : \hat{\tau}_j(x) > 0 > \tau_j(x)\}$. As we have discussed in the proof of Theorem 2, there is no mis-classified problem when $\tau_j(x) = 0$. We define the correctly-classified set $\mathcal{C}_j = \{x : \hat{\tau}_j(x)\times\tau_j(x) > 0\} \cup \{x : \tau_j(x) = 0\}$.

Let $\mathbf{1}_{\mathcal{M}_J} = \mathbf{1}\{X \in \cup_{j\in\{1,..,J-1\}}\mathcal{M}_j\}$ for the mis-classified set and $\mathbf{1}_{\mathcal{C}_J} = \mathbf{1}\{X \in \cap_{j\in\{1,..,J-1\}}\mathcal{C}_j\}$ for the correctly classified set. So $\mathbf{1}_{\mathcal{M}_J} + \mathbf{1}_{\mathcal{C}_J} = 1$. We have shown that $\mathbb{E}_n\big[g(W,\hat{\xi})\mathbf{1}_{\mathcal{C}_J}\big] - \mathbb{E}_n\big[g(W,\xi)\big] = o_{\mathbb{P}}(1/\sqrt{nh})$. The goal is to show $\mathbb{E}_n\big[g(W,\hat{\xi})\mathbf{1}_{\mathcal{M}_J}\big] = o_{\mathbb{P}}(1/\sqrt{nh})$.

Because $|\hat{\tau}_j(x) - \tau_j(x)| \leq |\hat{s}(d_{j+1},x) - s(d_{j+1},x)| + |\hat{s}(d_j,x) - s(d_j,x)| \leq 2\mathsf{s}_n$, $\mathcal{M}_j \subseteq \mathcal{M}_j^1 := \{x : 0 < |\tau_j(x)| \leq 2\mathsf{s}_n\}$. Thus it suffices to show $\mathbb{E}_n\big[g(W,\hat{\xi})\mathbf{1}_{\mathcal{M}_j^1}\big] = o_{\mathbb{P}}(1/\sqrt{nh})$ for all $j \in \{1,..,J-1\}$.

Let $\mathsf{D} = \overline{\mathcal{D}} - \underline{\mathcal{D}}$, so $d_{j+1} - d_j = \mathsf{D}/(J-1)$. By the Taylor expansion and mean-value theorem, for $d_j \in \mathcal{D}_{sx}$, for some $\bar{d}_j$ between d_j and d_{j+1}, and for some generic constant C,

$$
|\tau_j(x)| = \left|\sum_{m=1}^{\bar{M}_x-1} s^{(m)}(d_j,x)\mathsf{D}^m/(m!(J-1)^m) + s^{(\bar{M}_x)}(\bar{d}_j,x)\mathsf{D}^{\bar{M}_x}/(\bar{M}_x!(J-1)^{\bar{M}_x})\right| \geq C/J^{\bar{M}}.
$$

So $\inf_{j\in\{1,..,J-1\},x\in\mathcal{X}}|\tau_j(x)| \geq CJ^{-\bar{M}}$. Assuming $2\mathsf{s}_n < CJ^{-\bar{M}}$ for n large enough, $\mathbf{1}_{\mathcal{M}_j^1} = 0$.

The above calculation of variance $h\mathbb{E}[g(W,\hat{\xi}_\ell)^2|X,\mathcal{W}_\ell^c] = O_{\mathbb{P}}(1)$ uniformly in X. By the conditional Markov inequality, for any $\epsilon > 0$ and for n large enough, $\mathbb{P}\big(\big|\sqrt{nh}\mathbb{E}_{n\ell}[g(W,\hat{\xi}_\ell)\mathbf{1}_{\mathcal{M}_j^1}]\big| >$

$\epsilon\big|\mathcal{W}_\ell^c\big) < \mathbb{E}\big[hg(W,\hat{\xi}_\ell)^2\mathbf{1}_{\mathcal{M}_j^1}\big|\mathcal{W}_\ell^c\big]/\epsilon^2 \leq \mathbb{E}\big[h\mathbb{E}[g(W,\hat{\xi}_\ell)^2|X,\mathcal{W}_\ell^c]\mathbf{1}_{\mathcal{M}_j^1}\big|\mathcal{W}_\ell^c\big]/\epsilon^2 = 0.$ □

Acknowledgements We thank Adriana Lleras-Muney for providing the Civilian Conservation Corps (CCC) data. We thank Vira Semenova, David McKenzie, and participants in the seminars in Academia Sinica, Emory University, UCI, UCSD, UCLA, 2024 California Econometrics Conference, and 2025 Econometric Society World Congress for helpful comments.

References

Ahn, H. and J. L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1), 3–29.

Aizer, A., N. Early, S. Eli, G. Imbens, K. Lee, A. Lleras-Muney, and A. Strand (2024). The lifetime impacts of the new deal's youth employment program. *Quarterly Journal of Economics 139*(4), 2579–2635.

Angrist, J., G. Imbens, and D. Rubin (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association 91*, 444–455.

Ao, W., S. Calonico, and Y.-Y. Lee (2021). Multivalued treatments and decomposition analysis: An application to the WIA program. *Journal of Business & Economic Statistics 39*, 358–371.

Behaghel, L., B. Crépon, M. Gurgand, and T. Le Barbanchon (2015). Please call again: Correcting nonresponse bias in treatment effect models. *The Review of Economics and Statistics 97*(5), 1070–1080.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Cesarini, D., E. Lindqvist, M. J. Notowidigdo, and R. Östling (2017). The effect of wealth on individual and household labor supply: Evidence from Swedish lotteries. *American Economic Review 107*(12), 3917–46.

Chen, J. and J. Roth (2023). Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics 139*(2), 891–936.

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. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., W. Newey, and R. Singh (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica 90*(3), 967–1027.

Chernozhukov, V., W. K. Newey, and R. Singh (2022b). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal 25*(3), 576–601.

Colangelo, K. and Y.-Y. Lee (2025). Double debiased machine learning nonparametric inference with continuous treatments. *Journal of Business & Economic Statistics*, 1–13.

Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies 70*(1), 33–58.

DiNardo, J., J. Matsudaira, J. McCrary, and L. Sanbonmatsu (2021). A practical proactive proposal for dealing with attrition: Alternative approaches and an empirical example. *Journal of Labor Economics 39*(S2).

Donald, S. G., Y.-C. Hsu, and G. F. Barrett (2012). Incorporating covariates in the measurement of welfare and inequality: methods and applications. *The Econometrics Journal 15*(1), C1–C30.

D'Amour, A., P. Ding, A. Feller, L. Lei, and J. Sekhon (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics 221*(2), 644–654.

Escanciano, J. C., D. T. Jacho-Chávez, and A. Lewbel (2016). Identification and estimation of semiparametric two-step models. *Quantitative Economics 7*(2), 561–589.

Estrada, P. (2024). Spillover effects with nonrandom sample selection. Working paper, Emory University.

Fan, Y. and S. S. Park (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory 26*(3), 931–51.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics 189*(1), 1–23.

Flores, C. A., A. Flores-Lagunes, A. Gonzalez, and T. C. Neumann (2012). Estimating the effects of length of exposure to instruction in a training program: The case of Job Corps. *The Review of Economics and Statistics 94*(1), 153–171.

Garlick, R. and J. Hyman (2022). Quasi-experimental evaluation of alternative sample selection corrections. *Journal of Business & Economic Statistics 40*(3), 950–964.

Gerard, F., M. Rokkanen, and C. Rothe (2020). Bounds on treatment effects in regression discontinuity designs with a manipulated running variable. *Quantitative Economics 11*(3), 839–870.

Hansen, B. E. (2022a). *Econometrics*. Princeton University Press.

Hansen, B. E. (2022b). *Probability and Statistics for Economists*. Princeton University Press.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pp. 475–492. National Bureau of Economic Research, Inc.

Heckman, J., J. Smith, and N. Clements (1997). Making the most out of program evaluations and social experiments: accounting for heterogeneity in program impacts. *Review of Economic Studies 64*, 487–535.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Heiler, P. (2024). Heterogeneous treatment effect bounds under sample selection with an application to the effects of social media on political polarization. *Journal of Econometrics 244*(1), 105856.

Ho, K. and A. M. Rosen (2017). *Partial Identification in Applied Research: Benefits and Challenges*, pp. 307–359. Econometric Society Monographs. Cambridge University Press.

Honoré, B. E. and L. Hu (2020). Selection without exclusion. *Econometrica 88*(3), 1007–1029.

Honoré, B. E. and L. Hu (2022). Sample selection models without exclusion restrictions: Parameter heterogeneity and partial identification. *Journal of Econometrics*, 105360.

Horowitz, J. L. and C. F. Manski (1995). Identification and robustness with contaminated and corrupted data. *Econometrica 63*, 281–302.

Hsu, Y.-C., M. Huber, Y.-Y. Lee, and L. Lettry (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics 35*(7), 814–840.

Hsu, Y.-C., M. Huber, Y.-Y. Lee, and C.-A. Liu (2023). Testing monotonicity of mean potential outcomes in a continuous treatment with high-dimensional data. *The Review of Economics and Statistics*, forthcoming.

Hsu, Y.-C., M. Huber, C.-A. Liu, and Y.-Y. Lee (2023). Replication data for: Testing Monotonicity of Mean Potential Outcomes in a Continuous Treatment with High-Dimensional Data.

Imbens, G. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics 86*(1), 4–29.

Imbens, G. W. and J. D. Angrist (1994). Identification and estimation of local average treatment effects. *Econometrica 62*(2), 467–475.

Imbens, G. W. and W. K. Newey (2009). Identification and estimation of triangular simultaneous equations models without additivity. *Econometrica 77*(5), 1481–1512.

Kennedy, E. H., Z. Ma, M. D. McHugh, and D. S. Small (2017). Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B 79*(4), 1229–1245.

Kline, B. and E. Tamer (2023). Recent developments in partial identification. *Annual Review of Economics 15*, 125–150.

Kline, P. and A. Santos (2013). Sensitivity to missing data assumptions: Theory and an evaluation of the U.S. wage structure. *Quantitative Economics 4*(2), 231–267.

Kroft, K., I. Mourifié, and A. Vayalinkal (2024). Lee bounds with multilayered sample selection. NBER working paper 32952.

Lee, D. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies 76*(3), 1071–1102.

Lee, Y.-Y. (2015). Partial mean processes with generated regressors: Continuous treatment effects and nonseparable models. Working paper.

Molinari, F. (2020). Microeconometrics with partial identification. In S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin (Eds.), *Handbook of Econometrics, Volume 7A*, pp. 355–486. Elsevier.

Olma, T. (2021). Nonparametric estimation of truncated conditional expectation functions. arxiv:2109.06150.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*(6), 1403–30.

Powell, J. L. and T. M. Stoker (1996). Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics 75*(2), 291–316.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Schochet, P. Z., J. Burghardt, and S. McConnell (2008). Does Job Corps work? impact findings from the national Job Corps study. *American Economic Review 98*(5).

Semenova, V. (2024). Generalized Lee bounds. *Journal of Econometrics*, forthcoming.

Su, L., T. Ura, and Y. Zhang (2019). Non-separable models with high-dimensional data. *Journal of Econometrics 212*(2), 646–677.

Velez, A. (2025). On the asymptotic properties of debiased machine learning estimators. Working paper.

Vytlacil, E. (2002). Independence, monotonicity, and latent index models: An equivalence result. *Econometrica 70*(1), 331–341.

Zhang, J. L. and D. B. Rubin (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics 28*(4), 353–368.

Zhang, J. L., D. B. Rubin, and F. Mealli (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association 104*(485), 166–176.

Online Supplementary Appendix for Lee Bounds with a Continuous Treatment in Sample Selection

Ying-Ying Lee[†] Chu-An Liu[‡]

Section A presents the proofs of Claim-Step1, 2, 3 in Proof of Theorem 2 and Corollaries 1 and 2. Section B presents the details of the first-step Lasso estimation in Section 6.2 and supplementary material for the empirical applications.

A Additional proofs

A.1 Proofs of Claim-Step1, 2, 3 in Proof of Theorem 2:

Proof of Claim-Step1: (i) The kernel regression estimator $\hat{s}_\ell(d) = \mathbb{E}_{n_\ell}[SK_h(D-d)]/\hat{f}_{D\ell}(d)$, where the denominator $\hat{f}_{D\ell}(d) = \mathbb{E}_{n_\ell}[K_h(D-d)]$, is well studied. For example, Theorems 19.1 and 19.2 Hansen (2022a) provide that $\hat{s}_\ell(d) - s(d) = O_{\mathbb{P}}(1/\sqrt{nh} + h^2)$. By a linearization as in (5),

$$\begin{aligned}\hat{s}_\ell(d) &= \frac{\mathbb{E}_{n_\ell}[SK_h(D-d)] - s(d)f_D(d)}{f_D(d)} - \frac{s(d)}{f_D(d)}\left(\hat{f}_{D\ell}(d) - f_D(d)\right) \\ &\quad + O_{\mathbb{P}}\left(\|\hat{f}_{D\ell} - f_D\|^2 + \|\hat{f}_{D\ell} - f_D\|\|\mathbb{E}_{n_\ell}[SK_h(D-d)] - s(d)f_D(d)\|\right) \\ &= \frac{1}{f_D(d)}\left(\mathbb{E}_{n_\ell}[SK_h(D-d)] - s(d)\mathbb{E}_{n_\ell}[K_h(D-d)]\right) + o_{\mathbb{P}}(1/\sqrt{nh}) \\ &= \mathbb{E}_{n_\ell}[\phi_2(d)] + o_{\mathbb{P}}(1/\sqrt{nh}).\end{aligned}$$

(ii) A standard algebra (e.g., Theorem 19.2 in Hansen (2022a)) yields $V_{s(d)}$. Specifically, $h\mathbb{E}[\phi^2_{s(d)}] = h\mathbb{E}[(S-s(d))^2K_h(D-d)^2/f_D(d)^2] = h\int \mathbb{E}[(S-s(d))^2|D=v]K_h(v-d)^2f_D(v)dv/f_D(d)^2 = \int \mathbb{E}[(S-s(d))^2|D=d+uh]k(u)^2f_D(d+uh)du/f_D(d)^2 = \mathbb{E}[(S-s(d))^2|D=d]R_k/f_D(d) + O(h) \to V_{s(d)}R_k/s(d)$, under the condition that $s(d)f_D(d)$ and its first derivative are bounded uniformly.

(iii) Let the convolution kernel $\bar{k}(x) = \int k(u)k(u-x)du$. $h\mathbb{E}[\phi_{s(d')}\phi_{s(d)}] = h\mathbb{E}\big[(S-s(d'))(S-s(d))K_h(D-d')K_h(D-d)\big]/(f_D(d')f_D(d)) = \int \mathbb{E}[(S-s(d'))(S-s(d))|D=d+uh]k(u)k\big(\frac{d-d'}{h}+u\big)f_D(d+uh)du/(f_D(d')f_D(d)) = \mathbb{E}[(S-s(d'))(S-s(d))|D=d]\bar{k}\big(\frac{d'-d}{h}\big)/f_D(d') + o(1) = s(d)(1-s(d))\bar{k}\left(\frac{d'-d}{h}\right)/f_D(d') + o(1) = o(1)$, as $h \to 0$.

(iv) Similar arguments and a linearization as in (5) give the result for $\hat{p}_\ell$.

Proof of Claim-Step3: The kernel regression estimator of $\mathbb{E}[Y\mathbf{1}|D=d, S=1, \mathcal{W}^c_\ell]$ is standard as in Claim-Step1.

[†]Department of economics, University of California Irvine, Irvine, CA 92697, U.S.A.
E-mail: yingying.lee@uci.edu. https://sites.google.com/site/yyleelilian. Tel: +1 9498244834. Fax: +1 9498242492.
[‡]Institute of Economics, Academia Sinica, Taipei City 115, Taiwan.
E-mail: caliu@econ.sinica.edu.tw. https://chuanliu.weebly.com/.

Proof of Claim-SE: Let $W_i := Y_i(\hat{\mathbf{1}}_i^\ell - \mathbf{1}_i)K_h(D_i - d)S_i$.

$$\begin{aligned}\mathbb{E}_\ell[W] &= \mathbb{E}\left[\int_{\hat{Q}_\ell^d(1-\hat{p}_\ell)}^{Q^d(1-p)} y f_{Y|DS}(y|D,S)K_h(D-d)S \middle| \mathcal{W}_\ell^c\right] \\ &= -\left(\hat{Q}_\ell^d(1-\hat{p}_\ell) - Q^d(1-p)\right) Q^d(1-p) f_{Y|DS}(Q^d(1-p)|d,1)s(d)f_D(d) + O_{\mathbb{P}}(h^2)\end{aligned}$$

by a Taylor series expansion and Leibniz rule. By Claim-Step1 and Claim-Step2, we show

$$\begin{aligned}\hat{Q}_\ell^d(1-\hat{p}_\ell) - Q^d(1-p) &= \hat{Q}_\ell^d(1-\hat{p}_\ell) - Q_\ell^d(1-\hat{p}_\ell) + Q_\ell^d(1-\hat{p}_\ell) - Q^d(1-p) \\ &= \mathbb{E}_{n\ell}^c[\phi_2(1-p)] - (\hat{p}_\ell - p)\partial Q^d(\tau)/\partial\tau|_{\tau=1-p} + o_{\mathbb{P}}(1/\sqrt{nh}) + O_{\mathbb{P}}\big(\|\hat{p}_\ell - p\|^2\big),\end{aligned}$$

where the second inequality $\hat{Q}_\ell^d(1-\hat{p}_\ell) - Q_\ell^d(1-\hat{p}_\ell) - \big(\hat{Q}_\ell^d(1-p) - Q^d(1-p)\big) = o_{\mathbb{P}}(1/\sqrt{nh})$ that we show below.

Claim-Step2, Theorem 4.1 in Donald et al. (2012), and the functional delta method imply that $\sqrt{nh}\big(\hat{Q}_\ell^d(p_1) - Q_\ell^d(p_1) - \big(\hat{Q}_\ell^d(p_0) - Q^d(p_0)\big)\big) = \sqrt{nh}\mathbb{E}_{n\ell}^c[\phi_2(p_1) - \phi_2(p_0)] + o_{\mathbb{P}}(1)$ weakly converges to a Gaussian process indexed by $(p_0, p_1) \in [0,1]^2$, which has mean zero and variance $\lim_{n\to\infty}\mathbb{E}\big[h\big(\phi_2(p_1) - \phi_2(p_0)\big)^2\big] = \lim_{n\to\infty}\mathbb{E}\Big[hK_h(D-d)^2 S\Big(\frac{p_1 - \mathbf{1}\{Y \le Q^d(p_1)\}}{f_{Y|DS}(Q^d(p_1)|d,1)} - \frac{p_0 - \mathbf{1}\{Y \le Q^d(p_0)\}}{f_{Y|DS}(Q^d(p_0)|d,1)}\Big)^2\Big]$ $\times (s(d)^2 f_D(d)^2)^{-1} = O\big(\|p_1 - p_0\|\big)$. So the condition in Theorem 18.5 in Hansen (2022b) holds, i.e., for all $\delta > 0$ and $(p_0, p_1) \in [0,1]^2$, $\Big(\mathbb{E}\big[\sup_{\|p_1 - p_0\| \le \delta} \|\sqrt{h}(\phi_2(p_1) - \phi_2(p_0))\|^2\big]\Big)^{1/2} \le C\delta^\psi$ for some $C < \infty$ and $0 < \psi < \infty$. It follows that $\sqrt{nh}\mathbb{E}_{n\ell}^c[\phi_2(p)]$ is stochastic equicontinuous, i.e., $\forall \eta, \epsilon > 0$, there exists some $\delta > 0$ such that $\limsup_{n\to\infty}\mathbb{P}\Big(\sup_{\|p_1 - p_0\| \le \delta}\big\|\sqrt{nh}\mathbb{E}_{n\ell}^c\big[\phi_2(p_1) - \phi_2(p_0)\big]\big\| > \eta\Big) \le \epsilon$. We obtain $\sqrt{nh}\mathbb{E}_{n\ell}^c[\phi_2(p_1) - \phi_2(p_0)] = o_{\mathbb{P}}(1)$ uniformly over $\|p_1 - p_0\| \le \delta$, and hence $\hat{Q}_\ell^d(1-\hat{p}_\ell) - Q_\ell^d(1-\hat{p}_\ell) - \big(\hat{Q}_\ell^d(1-p) - Q^d(1-p)\big) = \mathbb{E}_{n\ell}^c[\phi_2(1-\hat{p}_\ell) - \phi_2(1-p)] + o_{\mathbb{P}}(1/\sqrt{nh}) = o_{\mathbb{P}}(1/\sqrt{nh})$ as $\|\hat{p}_\ell - p\| = o_{\mathbb{P}}(1)$ by Claim-Step1.

Further assuming $\sqrt{nh}h^2 = o(1)$ and $\sqrt{nh}\|\hat{p}_\ell - p\|^2 = o_{\mathbb{P}}(1)$, we obtain $\mathbb{E}_\ell[W] = num_{12}^\ell \times den + o_{\mathbb{P}}(1/\sqrt{nh})$.

$$\begin{aligned}h\mathbb{E}_\ell\left[W^2\right] &= h\mathbb{E}\left[Y^2(\hat{\mathbf{1}}^\ell - \mathbf{1})^2 K_h^2(D-d)S \middle| \mathcal{W}_\ell^c\right] \\ &= \mathbb{E}[Y^2(\hat{\mathbf{1}}^\ell - \mathbf{1})^2 | D = d, S = 1, \mathcal{W}_\ell^c]s(d)f_D(d)R_k + O_{\mathbb{P}}(h) \\ &= o_{\mathbb{P}}(1)\end{aligned}$$

by the consistency of Step 1 and Step 2. By the conditional Markov inequality, $\sqrt{n_\ell h}(\mathbb{E}_{n_\ell}[W] - \mathbb{E}_\ell[W]) = \sqrt{n_\ell h}(\mathbb{E}_{n_\ell}[W] - num_{12}^\ell \times den) + o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$.

By a linearization as (5),

$$\widehat{num}_{12}^\ell = num_{12}^\ell + \frac{\mathbb{E}_{n_\ell}[W] - num_{12}^\ell \times den}{den} - \frac{num_{12}^\ell}{den}(\widehat{den} - den) + o_{\mathbb{P}}(1/\sqrt{nh}).$$

By the consistency of Step 1 and Step 2, $num_{12}^{\ell} = o_{\mathbb{P}}(1)$, and by Claim-Step3, the above third term $\frac{num_{12}^{\ell}}{den}(\widehat{den} - den) = o_{\mathbb{P}}(1/\sqrt{nh} + h^2)$. Note that $\partial Q^d(\tau)/\partial\tau|_{\tau=1-p} \times f_{Y|DS}(Q^d(1-p)|d,1) = 1$.

For $\rho_{LB}(\tau)$, $\hat{\mathbf{1}} = \mathbf{1}\{Y \leq \hat{Q}^d(\hat{p}_\ell)\}$.

$$\begin{aligned}\mathbb{E}_\ell[W] &= \mathbb{E}\left[\int_{Q^d(p)}^{\hat{Q}^d(\hat{p}_\ell)} y f_{Y|DS}(y|D,S)K_h(D-d)S \middle| \mathcal{W}_\ell^c\right] \\ &= \left(\hat{Q}_\ell^d(\hat{p}_\ell) - Q^d(p)\right) Q^d(p) f_{Y|DS}(Q^d(p)|d,1)s(d)f_D(d) + O_{\mathbb{P}}(h^2), \text{ where}\end{aligned}$$

$$\begin{aligned}\hat{Q}_\ell^d(\hat{p}_\ell) - Q^d(p) &= \hat{Q}_\ell^d(\hat{p}_\ell) - Q_\ell^d(\hat{p}_\ell) + Q_\ell^d(\hat{p}_\ell) - Q^d(p) \\ &= \hat{Q}_\ell^d(p) - Q^d(p) + (\hat{p}_\ell - p)\partial Q^d(\tau)/\partial\tau|_{\tau=p} + O_{\mathbb{P}}(\|\hat{Q}_\ell^d - Q^d\|^2 + \|\hat{p}_\ell - p\|^2).\end{aligned}$$

A.2 Proofs of Corrollaries

Proof of Corollary 1: By (4) in the proof of Theorem 2, $\hat{\bar{\Delta}}_{d_1d_2} - \bar{\Delta}_{d_1d_2} = n^{-1}\sum_{i=1}^n \left(\phi_{d_2Ui} - \phi_{d_1Li}\right) + o_{\mathbb{P}}((nh)^{-1/2})$, where $h = \min\{h_{d_2U}, h_{d_1L}\}$. We next verify the third-absolute-moment condition for the Lyapunov CLT. By the variance calculation in the proof of Theorem 2, let $s_n^2 = \sum_{i=1}^n \mathbb{E}\left[\left(\phi_{d_2Ui} - \phi_{d_1Li}\right)/\sqrt{n}\right)^2\right] = \mathsf{V}_{Un} = O(h^{-1})$ and $\mathbb{E}[|\phi|^3] = O(h^{-2})$. So $\sum_{i=1}^n \mathbb{E}\left[\left|\left(\phi_{d_2Ui} - \phi_{d_1Li}\right)/\sqrt{n}\right|^3\right]/s_n^3 = O(nh^{-2}n^{-3/2}/h^{-3/2}) = O((nh)^{-1/2}) = o(1)$. Then by the Lyapunov CLT, $s_n^{-1}\sum_{i=1}^n \left(\phi_{d_2Ui} - \phi_{d_1Li} - (h_{d_2U}^2 B_{d_2U} - h_{d_1L}^2 B_{d_1L})\right)/\sqrt{n} = \mathsf{V}_{Un}^{-1/2}\sqrt{n}n^{-1}\sum_{i=1}^n \left(\phi_{d_2Ui} - \phi_{d_1Li} - (h_{d_2U}^2 B_{d_2U} - h_{d_1L}^2 B_{d_1L})\right) \xrightarrow{d} \mathcal{N}(0,1)$.

The remainder term $\mathsf{V}_{Un}^{-1/2}\sqrt{n}o_{\mathbb{P}}((nh)^{-1/2}) = o_{\mathbb{P}}(1)$, so we obtain $\mathsf{V}_{Un}^{-1/2}\sqrt{n}\left(\hat{\bar{\Delta}}_{d_1d_2} - \bar{\Delta}_{d_1d_2} - (h_{d_2U}^2 B_{d_2U} - h_{d_1L}^2 B_{d_1L})\right) \xrightarrow{d} \mathcal{N}(0,1)$. The same arguments apply to $\hat{\underline{\Delta}}_{d_1d_2}$.

We show that as $n \to \infty, h \to 0$, $hV_{Un} = h\mathbb{E}[\phi_{d_2U}^2 + \phi_{d_1L}^2 - 2\phi_{d_2U}\phi_{d_1L}] = V_{d_2U} + V_{d_1L} - 2C_{d_1d_2U} + o(1)$. In $h\mathbb{E}[\phi_{d_2U}\phi_{d_1L}]$, the cross-product terms with $K_h(D-d_1)K_h(D-d_2)$ result in a convolution kernel $\int k((d_2-d_1)/h + u)k(u)du$ and hence is $o(1)$. From the proof of Theorem 2, $h\mathbb{E}[\phi_\pi^2] = R_k V_\pi / f_D(d_{\text{AT}}) + o(1)$. Then $h\mathbb{E}[\phi_{d_2U}\phi_{d_1L}] = h\mathbb{E}[\phi_\pi^2(Q^{d_2}(1-p_{d_2}) - \rho_{d_2U}(\pi))(Q^{d_1}(p_{d_1}) - \rho_{d_1L}(\pi))/(s(d_1)s(d_2)p_{d_1}p_{d_2})] + o(1) = C_{d_1d_2U} + o(1)$. The same arguments apply to $C_{d_1d_2L}$ for the lower bound $\underline{\Delta}_{d_1d_2}$. □

Proof of Corollary 2: By Theorem 3, $\hat{\bar{\Delta}}_{d_1d_2} - \bar{\Delta}_{d_1d_2} = n^{-1}\sum_{i=1}^n \left(\phi_{d_2U}(W_i,\xi) - \phi_{d_1L}(W_i,\xi)\right) + o_{\mathbb{P}}((nh)^{-1/2})$, where $h = \min\{h_{d_2U}, h_{d_1L}\}$. We next verify the third-absolute-moment condition for the Lyapunov CLT. By the variance calculation in the proof of Theorem 3, let $s_n^2 = \sum_{i=1}^n \mathbb{E}\left[\left(\phi_{d_2U}(W_i,\xi) - \phi_{d_1L}(W_i,\xi)\right)/\sqrt{n}\right)^2\right] = \mathsf{V}_{Un} = O(h^{-1})$ and $\mathbb{E}[|\phi|^3] = O(h^{-2})$. Therefore $\sum_{i=1}^n \mathbb{E}\left[\left|\left(\phi_{d_2U}(W_i,\xi) - \phi_{d_1L}(W_i,\xi)\right)/\sqrt{n}\right|^3\right]/s_n^3 = O(nh^{-2}n^{-3/2}/h^{-3/2}) = O((nh)^{-1/2}) = o(1)$. By the Lyapunov CLT, $s_n^{-1}\sum_{i=1}^n \left(\phi_{d_2U}(W_i,\xi) - \phi_{d_1L}(W_i,\xi) - (h_{d_2U}^2 \mathsf{B}_{d_2U} - h_{d_1L}^2 \mathsf{B}_{d_1L})\right)/\sqrt{n} = \mathsf{V}_{Un}^{-1/2}\sqrt{n}n^{-1}\sum_{i=1}^n \left(\phi_{d_2U}(W_i,\xi) - \phi_{d_1L}(W_i,\xi) - (h_{d_2U}^2 \mathsf{B}_{d_2U} - h_{d_1L}^2 \mathsf{B}_{d_1L})\right) \xrightarrow{d} \mathcal{N}(0,1)$.

The remainder term $\mathsf{V}_{Un}^{-1/2}\sqrt{n}o_{\mathbb{P}}((nh)^{-1/2}) = o_{\mathbb{P}}(1)$, so we obtain $\mathsf{V}_{Un}^{-1/2}\sqrt{n}\left(\hat{\bar{\Delta}}_{d_1d_2} - \bar{\Delta}_{d_1d_2} -\right.$

$(h_{d_2U}^2 \mathsf{B}_{d_2U} - h_{d_1L}^2 \mathsf{B}_{d_1L})) \xrightarrow{d} \mathcal{N}(0,1)$. The same arguments apply to $\underline{\hat{\Delta}}_{d_1 d_2}$. □

B Supplements and details for empirical applications

B.1 Step 0 strict overlap sub-sample

We prepare a sub-sample that satisfies Assumption 8(i) for estimating the bounds. In Step 0, we use the full sample and bandwidth h_1 to estimate the GPS by $\tilde{\mu}_d(X_i)$ and the selection probability by $\tilde{s}(d, X_i)$.

We choose the fixed trimming parameter $trim_{GPS}$ by extending the idea of Imbens (2004) for a binary treatment to a continuous treatment. We limit the "importance weight" $\bar{k}/(\mu_d(X_i) n_\ell h) \leq 5\%$, where n_ℓ is the floor of n/L (the largest integer less than or equal to n/L) and the kernel function is bounded by $\bar{k} := \max_{u \in \mathcal{R}} k(u)$ (for the Epanechnikov kernel, $\bar{k} = 0.75/\sqrt{5}$). So the trimming rule is to drop the observation i if

$$\tilde{\mu}_d(X_i) < \bar{k}/(5\% n_\ell h_1) =: trim_{GPS}$$

for *some* $d \in \mathcal{D}_J$. So we obtain an overlap sample, denoted as $Sample_{GPS}$, where $\tilde{\mu}_d(X_i) \geq trim_{GPS}$ for *all* $i \in Sample_{GPS}$ and for *all* $d \in \mathcal{D}_J$. For JC, $trim_{GPS} = 0.000055$. For CCC, $trim_{GPS} = 0.01205$.

For the selection probability and sufficient always-takers, we drop observation i in the full sample if $\min_{d' \in \mathcal{D}_J} \tilde{s}(d', X_i) < 5\%$. We obtain a sample $Sample_S$ where $\tilde{s}(d, X_i) \geq 5\%$ for *all* $i \in Sample_S$ and for *all* $d \in \mathcal{D}_J$.

Finally, we obtain a sub-sample that is an intersection of $Sample_{GPS}$ and $Sample_S$ for estimating the bounds.

But in the cross-fitting sub-samples in Step 1, it is possible to obtain a GPS estimate below $trim_{GPS}$ and a small proportion of always-takers. So we set the GPS estimate below $trim_{GPS}$ to $trim_{GPS}$, following Hsu et al. (2023) to obtain a more stable estimator. That is, replace $\hat{\mu}_{d\ell}(X_i)$ with $\max\{\hat{\mu}_{d\ell}(X_i), trim_{GPS}\}$.

To further address possible small proportion of always-takers $\hat{p}_{dd_i\ell}(X_i)$ in estimating the conditional bounds, we restrict the conditional bounds estimates by the maximum and minimum of the outcome variable Y in the sample. That is, we estimate $\rho_{dU}(\pi_{\text{AT}}(X_i), X_i)$ by $\min\big\{\widehat{\mathbb{E}}_\ell[Y|Y \geq \hat{Q}_\ell^d(1 - \hat{p}_{dd_j\ell}(X_i), X_i)_\ell, D = d, S = 1, X = X_i], \max\{Y_i, i = 1, ..., n\}\big\}$, for $i \in I_\ell$. Similarly, estimate $\rho_{dL}(\pi_{\text{AT}}(X_i), X_i)$ by $\max\big\{\widehat{\mathbb{E}}_\ell[Y|Y \leq \hat{Q}_\ell^d(\hat{p}_{dd_j\ell}(X_i), X_i)_\ell, D = d, S = 1, X = X_i], \min\{Y_i, i = 1, ..., n\}\big\}$.

B.2 First-step Lasso estimation in Section 6.2

Let the logistic likelihood $M(y, x; g) = -\big(y \log(\Lambda(b(x)'g)) + (1-y)\log(1 - \Lambda(b(x)'g))\big)$, where Λ is the logistic CDF. Penalty loading matrix $\hat{\Psi}_{d\ell}$, $\hat{\Psi}_{dy\ell}$, $\hat{\Xi}_{d\ell}$ are computed by Algorithm 1 below from the iterative Algorithms 3.1 and 3.2 in SUZ. Let the final penalty loading matrix $\hat{\Psi}_{d\ell}$ be $\hat{\Psi}_{d\ell}^m$ from Algorithm 1 for some fixed positive integer M.

For $\ell \in \{1, ..., L\}$,

- $\hat{F}_{D|X_\ell}(D|X) = \Lambda(b(X)'\hat{\beta}_{d\ell})$, where

$$\hat{\beta}_{d\ell} = \arg\min_{\beta} \frac{1}{N_\ell} \sum_{i \notin I_\ell} M(\mathbf{1}\{D_i \leq d\}, X_i; \beta) + \frac{\tilde{\lambda}}{N_\ell} \|\hat{\Psi}_{d\ell}\beta\|_1, \tag{S.1}$$

  $N_\ell = n - n_\ell$, the penalty $\tilde{\lambda} = 1.1\Phi^{-1}(1 - r/\{p \vee nh_1\})n^{1/2}$, for some $r \to 0$ and $h_1 \to 0$, with the standard normal CDF Φ. We follow SUZ and set $r = 1/\log(n)$.

  Compute $\hat{F}_{D|X_\ell}(d|x) = \Lambda(b(X)'\hat{\beta}_{d\ell})$ from (S.1). Then the conditional density estimator

$$\hat{\mu}_{d\ell}(x) = \frac{\hat{F}_{D|X_\ell}(d + h_1|x) - \hat{F}_{D|X_\ell}(d - h_1|x)}{2h_1}.$$

- $\hat{Q}^d_\ell(p, x) = \inf\{y : \hat{F}_{Y|SDX_\ell}(y|1, d, x) \geq p\}$, where $\hat{F}_{Y|SDX_\ell}(y|1, d, x) = \Lambda(x'\hat{\alpha}_{dy\ell})$ with

$$\hat{\alpha}_{dy\ell} := \arg\min_{\alpha} \frac{1}{N_\ell} \sum_{i \notin I_\ell} M(\mathbf{1}\{Y_i \leq y\}, X_i, \alpha) S_i k\left(\frac{D_i - d}{h_1}\right) + \frac{\lambda}{N_\ell} \|\hat{\Psi}_{dy\ell}\alpha\|_1, \tag{S.2}$$

  $N_\ell = \sum_{i \notin I_\ell} S_i$, and the penalty $\lambda = \ell_n(\log(p \vee nh_1)nh_1)^{1/2}$ and $\ell_n = \sqrt{\log(\log(nh_1))}$.

- $\hat{s}_\ell(d, x) = \Lambda(b(x)'\hat{\theta}_{d\ell})$, where

$$\hat{\theta}_{d\ell} := \arg\min_{\theta} \frac{1}{N_\ell} \sum_{i \notin I_\ell} M(S_i, X_i; \theta) k\left(\frac{D_i - d}{h_1}\right) + \frac{\lambda}{N_\ell} \|\hat{\Upsilon}_{d\ell}\theta\|_1, \tag{S.3}$$

  $N_\ell = n - n_\ell$.

- $\hat{\rho}_{dU_\ell}(\pi, x) = \hat{\mathbb{E}}[Y|Y \geq \hat{Q}^d_\ell(1 - \hat{\pi}_\ell/\hat{s}_\ell(d, x), x), S = 1, D = d, X = x] = b(x)'\hat{\gamma}_{d\ell}$, where

$$\begin{aligned}\hat{\gamma}_{d\ell} := \arg\min_{\gamma} \frac{1}{2N_\ell} \sum_{i \notin I_\ell} (Y_i - b(X_i)'\gamma)^2 k\left(\frac{D_i - d}{h_1}\right) S_i \mathbf{1}\{Y_i \geq \hat{Q}^d_\ell(1 - \hat{\pi}_\ell/\hat{s}_\ell(d, X_i), X_i)\} \\ + \frac{\lambda}{N_\ell} \|\hat{\Xi}_{d\ell}\gamma\|_1,\end{aligned} \tag{S.4}$$

  with $N_\ell = \sum_{i \notin I_\ell} S_i \mathbf{1}\{Y_i \geq \hat{Q}^d_\ell(1 - \hat{\pi}_\ell/\hat{s}_\ell(d, X_i), X_i)\}$.

Algorithm 1 (SUZ Algorithm 3.1 and 3.2) *For $\ell \in \{1, ..., L\}$,*

- *For $\hat{\mu}_{d\ell}(x)$,*

  1. *Let $\hat{\Psi}^0_{d\ell} = diag(l^0_{d\ell,1}, ..., l^0_{d\ell,p})$, where $l^0_{d\ell,j} = \|\mathbf{1}\{D \leq d\}b_j(X)\|_{\mathbb{P}_{N_\ell},2}$. Compute $\hat{\beta}^0_{d\ell}$ by (S.1) with $\hat{\Psi}^0_{d\ell}$ in place of $\hat{\Psi}_{d\ell}$. Let $\hat{F}^0_{D|X_\ell}(D|X) = \Lambda(b(x)'\hat{\beta}^0_{d\ell})$.*

2. *Compute* $\hat{\Psi}^m_{d\ell} = diag(l^m_{d\ell,1}, ..., l^m_{d\ell,p})$, *where* $l^m_{d\ell,j} = \left\| \left(\mathbf{1}\{D \leq d\} - \hat{F}^{m-1}_{D|X_\ell}(d|X) \right) b_j(X) \right\|_{\mathbb{P}_{N_\ell},2}$, *for* $m = 1, ..., M$. *Compute* $\hat{\beta}^m_{d\ell}$ *by (S.1) with* $\hat{\Psi}^m_{d\ell}$ *in place of* $\hat{\Psi}_{d\ell}$. *Let* $\hat{F}^m_{D|X_\ell}(d|x) = \Lambda(b(x)'\hat{\beta}^m_{d\ell})$.

- *For* $\hat{Q}^d_\ell(p, x)$,

  1. *Let* $\hat{\Psi}^0_{dy\ell} = diag(l^0_{dy\ell,1}, ..., l^0_{dy\ell,p})$, *where* $l^0_{dy\ell,j} = \left\| \mathbf{1}\{Y \leq y\} S b_j(X) k((D-d)/h_1) h_1^{-1/2} \right\|_{\mathbb{P}_{N_\ell},2}$. *Compute* $\hat{\alpha}^0_{dy\ell}$ *by (S.2) with* $\hat{\Psi}^0_{dy\ell}$ *in place of* $\hat{\Psi}_{dy\ell}$. *Let* $\hat{F}^0_{Y|SDX_\ell}(y|1, d, x) = \Lambda(b(x)'\hat{\alpha}^0_{dy\ell})$.
  2. *Compute* $\hat{\Psi}^m_{dy\ell} = diag(l^m_{dy\ell,1}, ..., l^m_{dy\ell,p})$, *where* $l^m_{dy\ell,j} = \left\| \left(\mathbf{1}\{Y \leq y\} - \hat{F}^{m-1}_{Y|SDX_\ell}(y|1, d, X) \right) S \times b_j(X) K((D-d)/h_1) h_1^{-1/2} \right\|_{\mathbb{P}_{N_\ell},2}$, *for* $m = 1, ..., M$. *Compute* $\hat{\alpha}^m_{dy\ell}$ *by (S.2) with* $\hat{\Psi}^m_{dy\ell}$ *in place of* $\hat{\Psi}_{dy\ell}$. *Let* $\hat{F}^m_{Y|SDX_\ell}(y|1, d, x) = \Lambda(b(x)'\hat{\alpha}^m_{dy\ell})$.

- *For* $\hat{s}_\ell(d, x)$,

  1. *Let* $\hat{\Upsilon}^0_{d\ell} = diag(l^0_{d\ell,1}, ..., l^0_{d\ell,p})$, *where* $l^0_{d\ell,j} = \| S b_j(X) k((D-d)/h_1) h_1^{-1/2} \|_{\mathbb{P}_{N_\ell},2}$. *Compute* $\hat{\theta}^0_{d\ell}$ *by (S.3) with* $\hat{\Upsilon}^0_{d\ell}$ *in place of* $\hat{\Upsilon}_{d\ell}$. *Let* $\hat{s}^0_\ell(d, x) = \Lambda(b(x)'\hat{\theta}^0_{d\ell})$.
  2. *Compute* $\hat{\Upsilon}^m_{d\ell} = diag(l^m_{d\ell,1}, ..., l^m_{d\ell,p})$, *where* $l^m_{d\ell,j} = \left\| \left(S - \hat{s}^{m-1}_\ell(d, X) \right) b_j(X) K((D-d)/h_1) h_1^{-1/2} \right\|_{\mathbb{P}_{N_\ell},2}$, *for* $m = 1, ..., M$. *Compute* $\hat{\theta}^m_{d\ell}$ *by (S.3) with* $\hat{\Upsilon}^m_{d\ell}$ *in place of* $\hat{\Upsilon}_{d\ell}$. *Let* $\hat{s}^m_\ell(d, x) = \Lambda(b(x)'\hat{\theta}^m_{d\ell})$.

- *For* $\hat{\rho}_{dU_\ell}(\pi, x)$,

  1. *Let* $\hat{\Xi}^0_{d\ell} = diag(l^0_{d\ell,1}, ..., l^0_{d\ell,p})$, *where* $l^0_{d\ell,j} = \| Y b_j(X) S \mathbf{1}\{Y \geq \hat{Q}^d_\ell(1 - \hat{\pi}_\ell / \hat{s}_\ell(d, X), X)\} k((D-d)/h_1) h_1^{-1/2} \|_{\mathbb{P}_{N_\ell},2}$. *Compute* $\hat{\gamma}^0_{d\ell}$ *by (S.4) with* $\hat{\Xi}^0_{d\ell}$ *in place of* $\hat{\Xi}_{d\ell}$. *Let* $\hat{\rho}^0_{dUU_\ell}(\pi, x) = b(x)'\hat{\gamma}^0_{d\ell}$.
  2. *Compute* $\hat{\Xi}^m_{d\ell} = diag(l^m_{d\ell,1}, ..., l^m_{d\ell,p})$, *where* $l^m_{d\ell,j} = \left\| \left(Y - \hat{\rho}^{m-1}_{dU_\ell}(\pi, X) \right) S \mathbf{1}\{Y \geq \hat{Q}^d_\ell(1 - \hat{\pi}_\ell / \hat{s}_\ell(d, X), X)\} b_j(X) K((D-d)/h_1) h_1^{-1/2} \right\|_{\mathbb{P}_{N_\ell},2}$, *for* $m = 1, ..., M$. *Compute* $\hat{\gamma}^m_{d\ell}$ *by (S.4) with* $\hat{\Xi}^m_{d\ell}$ *in place of* $\hat{\Xi}_{d\ell}$. *Let* $\hat{\rho}^m_{dU\ell}(\pi, x) = b(x)'\hat{\gamma}^m_{d\ell}$.

Following SUZ, we choose $M = 5$.

B.3 Supplements for empirical applications: Job Corps

Let the covariates $X = (X_c, X_b, X_{ca})$, with continuous X_c, binary X_b, and categorical X_{ca}. Figure 11 presents the estimated bounds with a quadratic basis function $b(X) = (X_c, X_c^2, X_b, X_{ca}, X_{ca}^2)$ for the Job Corps. The bounds estimates for the largest effect of switching training hours from

87.677 to 1112.727 are bounded by $[0.276, 0.787]$ with the 95% confidence interval $[0.172, 0.928]$.[13] These results are similar to the estimates using the linear basis function in Figure 4.

Table 1: (Job Corps) Descriptive statistics

Variable	Mean	Median	Std Dev	Min	Max	Nonmissing
weekly earnings in fourth year (Y)	215.52	194.31	202.62	0.00	1879.17	4024
hours in training (D)	1195.32	966.43	965.93	0.86	5142.86	4024
selection status (S)	0.83	1.00	0.37	0.00	1.00	4024
female	0.43	0.00	0.50	0.00	1.00	4024
age	18.33	18.00	2.14	16.00	24.00	4024
White	0.25	0.00	0.43	0.00	1.00	4024
Black	0.50	1.00	0.50	0.00	1.00	4024
Hispanic	0.17	0.00	0.38	0.00	1.00	4024
years of education	9.91	10.00	1.93	0.00	20.00	3968
native English	0.85	1.00	0.36	0.00	1.00	4024
has children	0.18	0.00	0.38	0.00	1.00	4024
ever worked	0.14	0.00	0.35	0.00	1.00	4024
mean gross weekly earnings	19.59	0.00	98.67	0.00	2000.00	4024
household size	3.48	3.00	2.03	0.00	15.00	3968
household gross income brackets	2.21	1.00	2.44	0.00	7.00	2529
personal gross income brackets	0.49	0.00	0.63	0.00	7.00	1789
mother's years of education	9.40	12.00	5.03	0.00	20.00	3288
father's years of education	7.17	10.00	6.00	0.00	20.00	2519
welfare receipt during childhood	1.93	1.00	1.26	0.00	4.00	3753
poor or fair general health	0.12	0.00	0.33	0.00	1.00	4024
physical or emotional problems	0.04	0.00	0.20	0.00	1.00	4024
ever arrested	0.24	0.00	0.43	0.00	1.00	4024
extent of recruiter support	1.56	1.00	1.07	0.00	5.00	3934
idea about the desired training	0.84	1.00	0.37	0.00	1.00	4024
expected hourly wage after Job Corps	4.50	0.00	6.73	0.00	96.00	1808
expected to be training for a job	1.03	1.00	0.27	0.00	3.00	3945
expected stay in Job Corps	6.60	0.00	9.78	0.00	36.00	4024

Note: We use missing dummies for missing observations in covariates.

B.4 Supplements for empirical applications: CCC

We include the following covariates as in column (3) of Aizer et al. (2024).

- X4-X7: X4 and X6 are family and individual characteristics, and X5 and X7 are imputation indicators. There are 8 continuous variables and 24 dummy variables.
- Year dummies: The birth year is varied from 1870 to 1929. After dropping the birth year dummies with the observations less than 10, we have 23 year dummies.

[13] The undersmoothing bandwidths range from 160.298 to 274.602. $\nu = 0.01$ and $c_1 = 1$.

Figure 8: (Job Corps) Histogram of hours of training

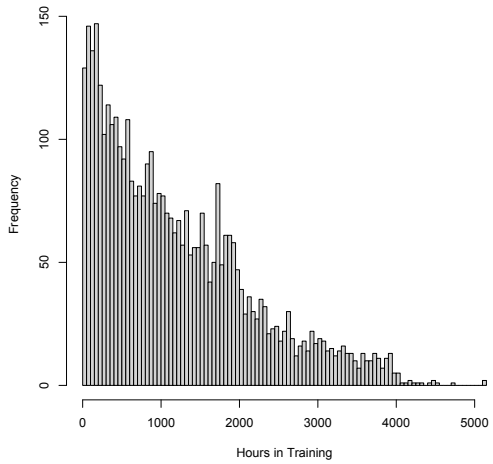


Figure 9: (Job Corps) Histograms of Y and $\log(Y)$ in $\{Y_i > 0, i = 1, ..., n\}$

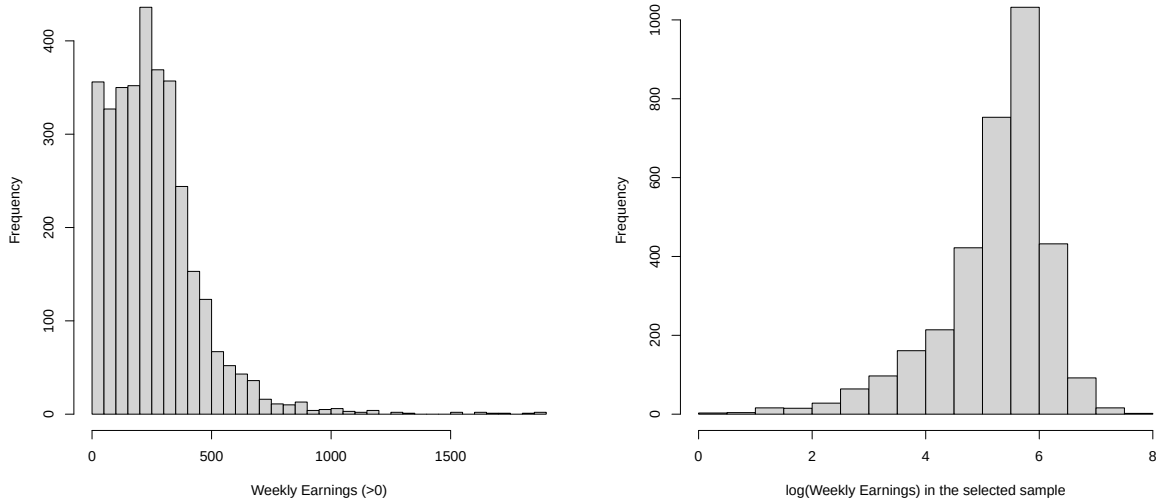


Figure 10: (Job Corps) Estimated bounds and 95% confidence intervals for Weekly Earnings without X

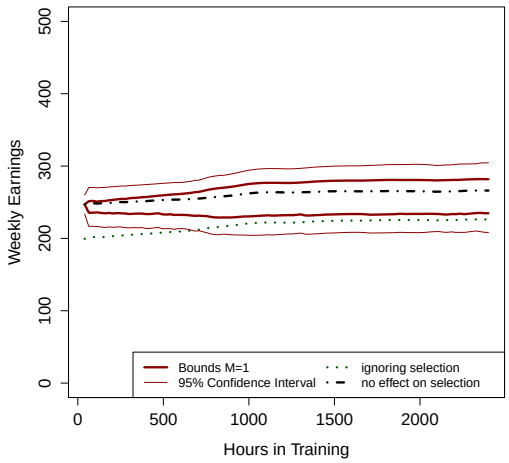


Figure 11: (Job Corps) Estimated bounds and 95% confidence intervals with quadratic $b(X)$

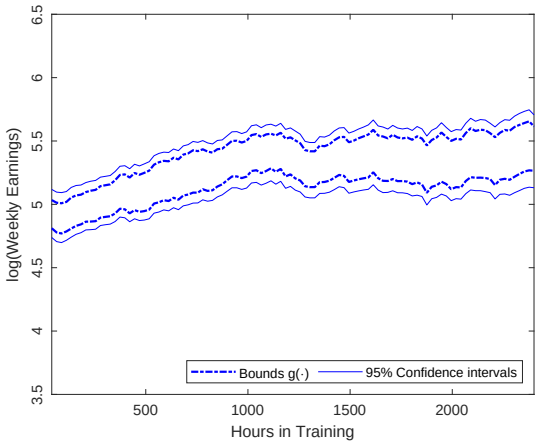


Table 2: (CCC) Descriptive statistics

Variable	Mean	Median	Std Dev	Minimum	Maximum	Non-imputing
death age (Y)	57.03	68.79	31.00	0.00	102.80	23722
duration of service (D)	0.82	0.51	0.71	0.00	8.37	23722
selection status (S)	0.82	1.00	0.39	0.00	1.00	23722
birth yerar	1919.78	1920.00	3.71	1870.00	1929.00	23722
ever rejected	0.02	0.00	0.14	0.00	1.00	23722
disabled	0.01	0.00	0.09	0.00	1.00	23722
non-junior	0.01	0.00	0.08	0.00	1.00	23722
reported age younger than DMF	0.09	0.00	0.28	0.00	1.00	23722
reported age older than DMF	0.17	0.00	0.37	0.00	1.00	23722
not eligible	0.02	0.00	0.12	0.00	1.00	23722
age is 17 or 18	0.56	1.00	0.49	0.00	1.00	23722
first allottee amount	21.63	22.00	3.71	0.00	30.00	22970
allottee is father	0.33	0.00	0.47	0.00	1.00	23722
allottee is mother	0.47	0.00	0.50	0.00	1.00	23722
gap in service	0.16	0.00	0.37	0.00	1.00	23722
log distance from home to camp	4.25	4.40	1.66	-9.40	8.04	23722
hispanic	0.48	0.00	0.50	0.00	1.00	23722
highest grade completed	8.60	8.61	1.65	0.00	17.00	14506
household size excluding applicant	4.74	4.74	1.50	0.00	22.00	7870
live on farm	0.25	0.25	0.25	0.00	1.00	8101
height	67.79	67.79	1.81	49.00	91.00	8141
weight	1.38	1.38	0.10	0.68	2.90	8234
father living	0.80	0.80	0.23	0.00	1.00	7943
mother living	0.85	0.85	0.21	0.00	1.00	8006
tenure in county	12.66	12.66	3.10	0.00	35.00	5432

Note: We use imputation dummies for imputed observations in covariates.

Figure 12: (CCC) Histogram of duration of service

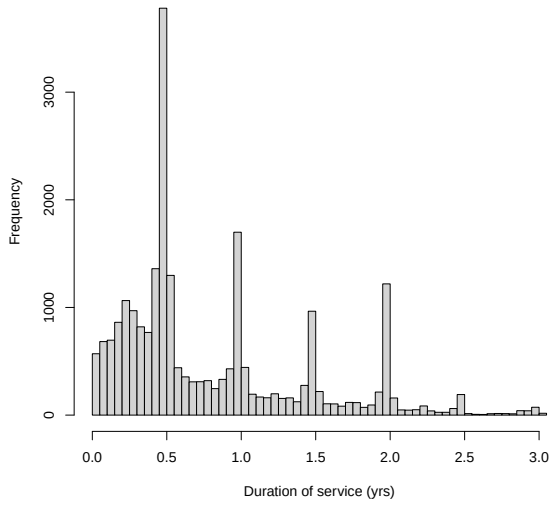


Figure 13: (CCC) Histograms of Y and $\log(Y)$ in $\{Y_i > 0, i = 1, ..., n\}$

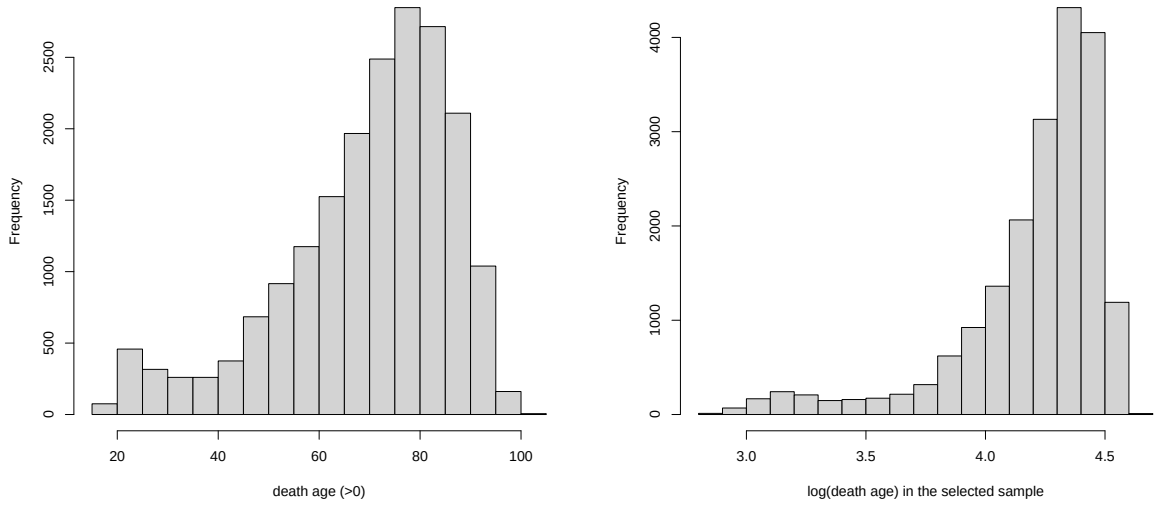


Figure 14 presents the estimated bounds with a quadratic basis function $b(X)$ for the CCC. For the estimates with quadratic $b(X)$ on $\mathcal{D}_{100}$, the largest ATE is when increasing the duration from 0.276 to 1.156 years with the bounds $[0.836, 2.205]$ and 95% confidence interval $[0.119, 3.230]$.[14]

[14]The undersmoothing bandwidths range from 0.166 to 0.356. $c_1 = 1.5, \nu = 0.01$.

Figure 14: (CCC) Estimated bounds and 95% confidence intervals with quadratic $b(X)$