Causal Inference on Networks under Misspecified Exposure Mappings: A Partial Identification Framework

Maresa Schröder [1 2] **Miruna Oprescu** [3] **Stefan Feuerriegel** [1 2] **Nathan Kallus** [3 4]

Abstract

Estimating treatment effects in networks is challenging, as each potential outcome depends on the treatments of all other nodes in the network. To overcome this difficulty, existing methods typically impose an *exposure mapping* that compresses the treatment assignments in the network into a low-dimensional summary. However, if this mapping is misspecified, standard estimators for direct and spillover effects can be severely biased. We propose a novel *partial identification framework for causal inference on networks* to assess the robustness of treatment effects under misspecifications of the exposure mapping. Specifically, we derive sharp upper and lower bounds on direct and spillover effects under such misspecifications. As such, our framework presents a novel application of causal sensitivity analysis to exposure mappings. We instantiate our framework for three canonical exposure settings widely used in practice: (i) weighted means of the neighborhood treatments, (ii) threshold-based exposure mappings, and (iii) truncated neighborhood interference in the presence of higher-order spillovers. Furthermore, we develop *orthogonal estimators* for these bounds and prove that the resulting bound estimates are valid, sharp, and efficient. Our experiments show the bounds remain informative and provide reliable conclusions under misspecification of exposure mappings.

1. Introduction

Estimating treatment effects in network settings is crucial for evaluating policy effectiveness and designing personalized interventions (Viviano, 2025). However, classic methods from causal inference assume *no* interference between units, meaning that the outcome of each unit is independent of the treatments from other units, but this assumption is often violated in social networks (Forastiere et al., 2021; Matthay & Glymour, 2022; Ogburn et al., 2024).

Example: *Consider a public health intervention that targets individuals aged 60 and above to encourage COVID-19 vaccination (Freedman et al., 2026). Individuals targeted by the interventions may be more likely to get vaccinated themselves (direct effect), while also influencing decisions of their friends through social interactions (spillover effect).*

Causal inference in network settings is fundamentally challenging due to *interference* (Anselin, 1988; Forastiere et al., 2021). In such settings, outcomes depend not only on a unit's own treatment but also on the treatments of connected units, which leads to spillover effects. More importantly, this directly violates the stable unit treatment variable assumption (SUTVA), so that standard treatment effects are no longer identified. Instead, causal inference on networks must consider *all treatments of the complete network.*

A naïve approach to handle interference is to condition on *all* treatment assignments in the network. However, this grows exponentially in the number of units N and is thus highly impractical. A common workaround is to simplify the problem and impose an *exposure mapping* that compresses the networks' treatment assignment into a low-dimensional summary (e.g., number or share of treated neighbors in the network) (Aronow & Samii, 2017). This approach is widely used in the literature (e.g., Forastiere et al., 2021; 2022; Ogburn et al., 2024) (see Sec. 2 for a detailed overview). *However*, the exposure mapping must be specified *a priori* based on domain knowledge of how spillovers propagate through the network. In many applications, this mechanism is only partially understood, so the exposure mapping is likely to be misspecified, resulting in biased effect estimates.

As a remedy, we develop a novel **partial identification framework for inference on networks** to assess the robustness of treatment effects under misspecifications of the exposure mapping through sensitivity analysis (see Fig. 1). Specifically, we derive sharp upper and lower bounds on the (conditional) potential outcomes and treatment effects when the exposure mapping is misspecified. We instantiate our framework for three canonical exposures used in practice: (i) weighted means of neighborhood treatments,

[1]LMU Munich [2]Munich Center for Machine Learning [3]Cornell University [4]Netflix. Correspondence to: Maresa Schröder <maresa.schroeder@lmu.de>.

Preprint. March 12, 2026.

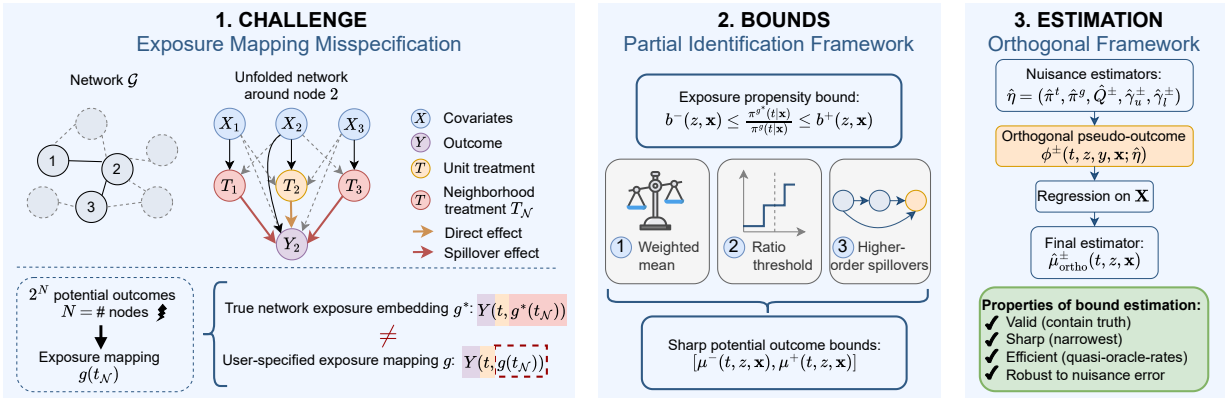


Figure 1. **Overview and contribution.** (1) **Challenge**: Each unit's outcome depends on the entire treatment assignments in the network. Existing methods compress the treatment assignments into a low-dimensional summary via an exposure mapping g. However, g might differ from the true mapping, thus leading to biased effect estimates. (2) **Bounds**: We model misspecification through an exposure-propensity bound and derive treatment effect bounds for 3 common exposure mappings. (3) **Estimation**: We estimate the bounds via an orthogonal two-stage framework. Our estimated bounds are valid, sharp, efficient, and robust to nuisance estimation errors.

(ii) threshold-based exposure mappings, and (iii) truncated neighborhood interference in the presence of higher-order spillovers. We make the following contributions:[1]

1. Our bounds fulfill several desirable theoretical properties. In particular, our bounds are *sharp* and *valid*, in that they provide the narrowest possible intervals that contain the true outcome given a specific level of exposure mapping misspecification.
2. We provide a model-agnostic *orthogonal bound estimator* that achieves quasi-oracle convergence rates and remains *robust* to nuisance model misspecification.
3. We provide guarantees for the *estimated* bounds in that they remain *sharp*, *valid*, *efficient*.

Our experiments show the bounds are informative and foster reliable decision-making under exposure mapping misspecification.

2. Related work[2]

Interference: Literature allowing for interference between different units broadly considers two distinct scenarios: (i) *partial or cluster-based interference*, and (ii) *network interference*. We focus on the latter, namely, network interference. In contrast, partial interference assumes that interference happens within groups or clusters but not across different groups (e.g., Bargagli-Stoffi et al., 2025; Tchetgen Tchetgen & VanderWeele, 2012; Qu et al., 2024; Fang & Forastiere, 2025; VanderWeele et al., 2014). However, this group-level interference is unlikely to hold in real-world settings, making the assumption restrictive and often invalid.

Network interference is less explored, where the majority of methods only apply to randomized controlled trials (RCTs), instead of observational data (e.g., Alzubaidi & Higgins, 2024; Aronow & Samii, 2017; Leung, 2020; Sävje et al., 2021). Methods targeted to network interference generally assume correct knowledge of a network treatment-summarizing function g, called *exposure mapping* (e.g., Chen et al., 2024a; Forastiere et al., 2021; 2022; Liu et al., 2023; Ogburn et al., 2024; Sengupta et al., 2025). There are three exposure mappings that are commonly applied in the literature: (i) the (weighted) neighborhood mean exposure, (ii) a thresholding function, and (iii) one-step neighborhood exposure. We provide a formal definition in Section 3. *However, these methods fail to provide correct estimates of treatment effect if the exposure mapping is misspecified.*

Misspecification in network inference: Only a few works consider causal effect estimation under misspecification. Of those, one stream focuses on causal effect estimation when the network structure is unknown, i.e., when there is uncertainty about the existence of certain edges in the network (e.g., Egami, 2021; Bhattacharya et al., 2020; Sävje, 2024; Weinstein & Nevo, 2023; 2025; Zhang et al., 2023; 2025). In contrast, our work assumes the *network is fully known*, but the *exposure mapping is misspecified*.

We are aware of two works that allow for a potential misspecification of the exposure mapping, but in a simplified setting (see Supplement B for details). Leung (2022) considers approximate neighborhood interference, by allowing treatments assigned to units further from the unit of interest to have potentially nonzero, but decreasing, effects on the unit's outcome. Belloni et al. (2022) consider causal effect estimation under an unknown neighborhood radius. Unlike our work, both methods are restricted to the specific type of misspecification and are only applicable to *average* causal effects. *In contrast, our proposed framework incorporates misspecification not only of the neighborhood radius, but also covers other types of misspecification in exposure*

[1] Code available at GitHub: https://github.com/m-schroder/ExposureMisspecification

[2] We provide an extended overview of related work in Supp. B.

mappings as well as a broad set of causal estimands.

Research gap: In sum, there is no general framework for bounding potential outcomes and treatment effects under various types of exposure mapping misspecifications for experimental and observational data. This is our contribution.

3. Setup

Notation: We use capital letters X to denote random variables, with realizations x (lowercase letters). The probability distribution of X is represented by $\mathbb{P}_X$, though we omit the subscript when the context makes it clear. For discrete variables the probability mass function is written as $P(x) = P(X = x)$ and for continuous variables, the probability density function as $p(x)$. In our work, we build upon the potential outcomes framework (Rubin, 2005). We provide an overview of the notation in Supplement A.

3.1. Network setting

We follow the standard setting for causal inference on networks (Chen et al., 2024b; Forastiere et al., 2021). We consider an (undirected) network of known structure given by the sets of nodes $\mathcal{N}_\mathcal{G}$ with $|\mathcal{G}| = N$ and edges $\mathcal{E}$ with $(i, j) = (j, i)$ for $i, j \in \mathcal{G}$. For each node i, we define a partition of the network as $(i, \mathcal{N}_i, \mathcal{N}_{-i})$, where $\mathcal{N}_i$ defines the *neighborhood* of node i, i.e., all nodes j connected to i by an edge $(i, j) \in \mathcal{E}$ and $\mathcal{N}_{-i}$ the complement of $\mathcal{N}_i$ in $\mathcal{G}$. We refer to $|\mathcal{N}_i| = n_i$ as the *degree* of node i. We omit the subscript whenever it is obvious from the context.

Every unit i consists of the following variables: a treatment $T_i \in \{0, 1\}$, confounders $X_i \in \mathcal{X}^d$, and an outcome $Y_i \in \mathcal{Y}$. We allow the treatment assignment to depend on (i) the unit's own covariates $\mathbf{X}_i = X_i$ [*homogeneous peer influence*], or (ii) both unit i's and it's neighbors covariates $\mathbf{X}_i = (X_i, X_{\mathcal{N}_i})$ [*heterogeneous peer influence*], where we additionally assume that every node has the same degree n. The treatment assignment of unit i is independent of the other units' treatment assignments given the covariates $\mathbf{X}$. We denote the unit *propensity score* $P(t \mid \mathbf{X} = \mathbf{x})$ as $\pi^t(\mathbf{x})$.

3.2. Exposure mappings

As standard in treatment effect estimation on networks (e.g., Chen et al., 2024a; Forastiere et al., 2021), we assume the existence of an *exposure mapping* $g : [0, 1]^{n_i} \to \mathcal{Z}$ with $z_i := g(t_{\mathcal{N}_i})$, which a summary function of the treatments assigned to the neighbors of node i. Here, z_i is assumed to be a sufficient representation to capture how neighbors' treatments affect the outcome Y_i. Therefore, the potential outcome is fully represented by $Y_i(t_i, z_i)$ and depends on the binary treatment t_i as well as the discrete or continuous neighborhood treatment z_i. We denote the network propensity by $\pi^g(z \mid \mathbf{x}) := p(g(t_{\mathcal{N}_i}) = z \mid \mathbf{X_i} = \mathbf{x_i})$.

Prior literature commonly builds upon three different types of *exposure mappings* (see Section 2):

(1) Weighted mean of neighborhood treatments: A large body of existing works assume the summary function g to represent the mean of the neighbors' treatments (e.g., Belloni et al., 2022; Chen et al., 2024a; 2025; Forastiere et al., 2021; 2022; Jiang & Sun, 2022; Leung, 2020; Ma & Tresp, 2021). However, in many cases, such as social network or spatial interference settings, it is reasonable to assume that the neighbors' effects vary, e.g., with closeness in friendship or spatial distance (e.g., Giffin et al., 2023). Hence, this motivates to formalize the underlying exposure mapping g^* as a *weighted* mean of treatments.

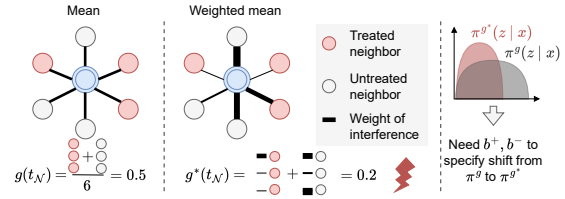


Figure 2. Exposure misspecification: weighted mean of treatments

(2) Thresholding: Similarly, the mapping g can be assumed to be a function of the sum or the proportion of treated neighbors (e.g., Aronow & Samii, 2017; McNealis et al., 2024; Ogburn et al., 2024; Qu et al., 2024). For example, the mapping could result in a binary variable Z indicating if more than half of the neighbors were treated.

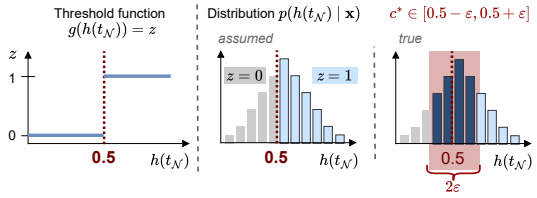


Figure 3. Exposure misspecification: thresholding function

(3) Higher-order spillovers: The assumption that only the treatment of direct neighbors results in spillover effects can be too strong (e.g., Belloni et al., 2022; Leung, 2022; Ogburn et al., 2024; Weinstein & Nevo, 2023). *Higher-order neighbors* without a direct connecting edge in the network potentially confound the causal relationship as well.

Figure 4. Exposure misspecification: higher-order spillovers

3.3. Causal inference on networks

We are interested in estimating the *average potential outcome* (APO) under individual and neighborhood treatments

$T = t$ and $Z = z$, where $z = g(t_{\mathcal{N}})$, given by

$$\psi(t, z) := \mathbb{E}[Y(t, z)], \tag{1}$$

and the *conditional average potential outcome* (CAPO)

$$\mu(t, z, \mathbf{x}) := \mathbb{E}[Y(t, z) \mid \mathbf{X} = \mathbf{x}]. \tag{2}$$

The *overall effect* can be decomposed into a *direct effect* (capturing the impact of a unit's own treatment on its outcome) and a *spillover effect* (capturing the indirect impact of neighbors' treatments on that unit's outcome).

Definition 3.1 (Direct effects (ADE / IDE))**.** *The average (ADE) and individual (IDE) direct effects between individual treatment assignments $T = t$ and $T = t^{'}$ while keeping the neighborhood treatment $Z = z$ constant are defined as*

$$\tau_d^{(t,z),(t^{'},z)} := \psi(t, z) - \psi(t^{'}, z) \tag{3}$$

$$\tau_{d_i}^{(t,z),(t^{'},z)}(\mathbf{x}_i) := \mu(t, z, \mathbf{x}_i) - \mu(t^{'}, z, \mathbf{x}_i). \tag{4}$$

Definition 3.2 (Spillover effects (ASE / ISE))**.** *The average (ASE) and individual (ISE) spillover effects between neighborhood treatment $Z = z$ and $Z = z^{'}$ while keeping the individual treatment $T = t$ constant are defined as*

$$\tau_s^{(t,z),(t,z^{'})} := \psi(t, z) - \psi(t, z^{'}) \tag{5}$$

$$\tau_{s_i}^{(t,z),(t,z^{'})}(\mathbf{x}_i) := \mu(t, z, \mathbf{x}_i) - \mu(t, z^{'}, \mathbf{x}_i). \tag{6}$$

Definition 3.3 (Overall effects (AOE / IOE))**.** *The average (AOE) and individual (IOE) total effects between individual treatment assignments $T = t$ and $T = t^{'}$ and neighborhood treatment assignments $Z = z$ and $Z = z^{'}$ are defined as*

$$\tau_o^{(t,z),(t^{'},z^{'})} := \psi(t, z) - \psi(t^{'}, z^{'}) \tag{7}$$

$$\tau_{o_i}^{(t,z),(t^{'},z^{'})}(\mathbf{x}_i) := \mu(t, z, \mathbf{x}_i) - \mu(t^{'}, z^{'}, \mathbf{x}_i). \tag{8}$$

As standard in causal inference on networks (e.g., Chen et al., 2024a; Forastiere et al., 2021), we make the standard assumptions on consistency, unconfoundedness, and positivity, but adapted to the network interference setting.

Assumption 3.4 (Network consistency)**.** The potential outcome equals the observed outcome given the same unit and neighborhood treatment exposure, i.e., $y_i = y_i(t_i, t_{\mathcal{N}_i})$ if i receives treatment t_i and neighborhood treatment $t_{\mathcal{N}_i}$.

Assumption 3.5 (Network interference)**.** A unit's treatment only affects its own as well as its neighbors' outcomes. The interference of the treatment with the neighbors outcomes is given by a (potentially unknown) summary function g^*, i.e., $\forall t_{\mathcal{N}_i}, t^{'}_{\mathcal{N}_i}$ which satisfy $g^*(t_{\mathcal{N}_i}) = g^*(t^{'}_{\mathcal{N}_i})$, it holds $y_i(t_i, t_{\mathcal{N}_i}) = y_i(t_i, t^{'}_{\mathcal{N}_i})$.

Assumption 3.6 (Network unconfoundedness)**.** Given the individual and the features of the neighborhood, the potential outcome is independent of the individual and the neighborhood treatment, i.e., $\forall t, t_{\mathcal{N}} : y_i(t, t_{\mathcal{N}}) \perp\!\!\!\perp t_i, t_{\mathcal{N}_i} \mid \mathbf{x}_i$. If the summary function g^* is correctly specified, it also holds $\forall t, g^*(t_{\mathcal{N}}) : y_i(t, g^*(t_{\mathcal{N}})) \perp\!\!\!\perp t_i, g^*(t_{\mathcal{N}_i}) \mid \mathbf{x}_i$.

Assumption 3.7 (Network positivity)**.** Given the individual and neighbors' features, every treatment pair (t, z) is observed with a positive probability, i.e., $0 < p(t, z \mid \mathbf{x}) < 1$ for all $\mathbf{x}, t, z$.

Under Assumptions 3.4–3.7 and correctly specified exposure mapping g^*, the (conditional) potential outcomes can be identified from observational data. However, if the exposure mapping $g \neq g^*$ is misspecified, Assumptions 3.5 and 3.6 are *not* satisfied. Therefore, the potential outcomes is *not* point-identified from the existing data.

3.4. Objective: partial identification under misspecification of the exposure mapping

We propose to move to partial identification and compute upper and lower bounds $\mu^{\pm}(t, z^*, \mathbf{x})$ on the potential outcomes and treatment effects under such misspecification.[3]

We formalize the partial identification as a distribution shift in the *exposure mapping propensity* $\pi^g(z \mid \mathbf{x}) := p(g(t_{\mathcal{N}}) = z \mid \mathbf{x})$ between the employed and true but unknown mapping g and g^*. For given shifts between g and g^*, we aim to construct upper and lower bounds $b^-(z, \mathbf{x}) \leq b^+(z, \mathbf{x})$ with $b^-(z, \mathbf{x}) \in (0, 1]$ and $b^+(z, \mathbf{x}) \in [1, \infty)$ on the generalized propensity ratio, such that, for all $z, \mathbf{x}$, we have

$$b^-(z, \mathbf{x}) \leq \frac{p(g^*(t_{\mathcal{N}}) = z \mid \mathbf{x})}{p(g(t_{\mathcal{N}}) = z \mid \mathbf{x})} \leq b^+(z, \mathbf{x}). \tag{9}$$

Of note, we do not impose any parametric assumption on the data-generating process. The specific interpretation and construction of $b^{\pm}$ depends on the definition of g and g^*.

We now formalize how our framework quantifies misspecifications in order to obtain bounds for the different exposure mappings. We provide justifications in Supplement D.

(1) *Weighted mean of neighborhood treatments:* Let the exposure mapping $g(t_{\mathcal{N}}) := \sum_{j \in \mathcal{N}} \frac{t_j}{n} = \frac{N_T}{n}$ be specified as the proportion of treated neighbors, where N_T denotes the number of treated neighbors and n denotes the neighborhood size. We assume the true exposure mapping $g^*(t_{\mathcal{N}})$ is given by a weighted proportion of treated neighbors, where each weight is allowed to differ from $\frac{1}{n}$ for at most a value $\frac{1}{n} \geq \varepsilon \geq 0$. Then, the upper and lower bound are

$$b^-(z, \mathbf{x}) = \inf_{s \in \mathcal{Z}} \frac{P(\frac{ns}{1-\varepsilon n} \leq N_T \leq \frac{nz}{1+\varepsilon n} \mid \mathbf{x})}{P(ns \leq N_T \leq nz \mid \mathbf{x})}, \tag{10}$$

$$b^+(z, \mathbf{x}) = \sup_{s \in \mathcal{Z}} \frac{P(\frac{ns}{1+\varepsilon n} \leq N_T \leq \frac{nz}{1-\varepsilon n} \mid \mathbf{x})}{P(ns \leq N_T \leq nz \mid \mathbf{x})}. \tag{11}$$

(2) *Thresholding:* Let $h(t_{\mathcal{N}}) := \sum_{j \in \mathcal{N}} \frac{t_j}{n}$ denote the proportion of treated neighbors. Assume the exposure mapping

[3] In the main paper, we focus on bounds for potential outcomes. We provide extensions to the treatment effects in Supplement C.

is specified through a threshold as $g(t_{\mathcal{N}}) = f(h(t_{\mathcal{N}})) := \mathbf{1}_{[h(t_{\mathcal{N}}) \geq c]}$. Then, $P(g(t_{\mathcal{N}}) = 1 \mid \mathbf{x}) = P(N_T \geq nc \mid \mathbf{x})$, where N_T denotes the number of treated neighbors. We now allow the true threshold c^* defining $g^*(t_{\mathcal{N}}) := \mathbf{1}_{[h(t_{\mathcal{N}}) \geq c^*]}$ to differ by an amount $\varepsilon \in [0, \min\{c, 1-c\}]$ from c, i.e., $c^* \in [c \pm \varepsilon]$. By straightforward computation, we receive

$$\frac{P(N_T \geq n(c+\varepsilon) \mid \mathbf{x})}{P(P(N_T \geq nc \mid \mathbf{x})} \leq \frac{P(g^*(t_{\mathcal{N}}) = 1 \mid \mathbf{x})}{P(g(t_{\mathcal{N}}) = 1 \mid \mathbf{x})} \tag{12}$$

$$\leq \frac{P(N_T \geq n(c-\varepsilon) \mid \mathbf{x})}{P(P(N_T \geq nc \mid \mathbf{x})}, \tag{13}$$

where the bounds for $z = 0$ follow with the complement probabilities.

(3) *Higher-order spillovers:* Here, g is misspecified in that it is not merely a function of $t_{\mathcal{N}_i}$, but also a function of other treatments $t_U \subset t_{\mathcal{N}_{-i}}$ for the respective node i. As a result, t_U biases the exposure summary z in an unobserved manner. We thus apply sensitivity bounds from the unobserved confounding literature (Dorn & Guo, 2022; Frauen et al., 2023), in that we require user-specified functions $b^{\pm}$, such that

$$b^-(z, \mathbf{x}) \leq \frac{p(g(t_{\mathcal{N} \cup U}) = z \mid \mathbf{x})}{p(g(t_{\mathcal{N}}) = z \mid \mathbf{x})} \leq b^+(z, \mathbf{x}), \tag{14}$$

where g can be any exposure mapping, such as the mean or the thresholding function in (1) and (2).

4. Our partial identification framework

We now present our framework. We derive sharp and valid bounds $\mu^{\pm}(t, z, \mathbf{x})$ on the potential outcomes following ideas from causal sensitivity analysis (Section 4.1). In Supplement C, we translate these into corresponding bounds for the direct, spillover, and overall effects. Next, we develop an orthogonal estimator $\hat{\mu}^{\pm}_{\text{ortho}}(t, z, \mathbf{x})$ based on orthogonal statistical learning theory (Section 4.2). Finally, we derive the theoretical properties of our estimator, including convergence rates, sharpness, and validity guarantees (Section 4.3). All proofs are in Supplement D.

4.1. Derivation of the bounds $\mu^{\pm}(t, z, \mathbf{x})$

We now introduce our *sharp* upper and lower bounds of the CAPO with respect to the misspecification $b^{\pm}$.

Definition 4.1. *Let $\tilde{\mathbb{P}}$ denote a distribution on $(\mathbf{X}, T, Z, Y(T, Z))$, such that (i) $\tilde{\mathbb{P}}$ matches the observed distribution $\mathbb{P}$ on $(\mathbf{X}, T, Z, Y)$, and (ii) the corresponding conditional distribution $\tilde{\pi}^g(z \mid \mathbf{x})$ satisfies $b^-(z, \mathbf{x}) \leq \frac{\tilde{\pi}^g(z|\mathbf{x})}{\pi^g(z|\mathbf{x})} \leq b^+(z, \mathbf{x})$ almost surely. Let $\mathcal{M}$ denote the set of such distributions $\tilde{P}$. Then, the sharp bounds of the CAPO with respect to the misspecification bounds $b^{\pm}(z, \mathbf{x})$ are given by*

$$\mu^+(t, z, \mathbf{x}) = \sup_{\tilde{P} \in \mathcal{M}} \mathbb{E}_{\tilde{P}}[Y(t, z) \mid \mathbf{X} = \mathbf{x}], \tag{15}$$

$$\mu^-(t, z, \mathbf{x}) = \inf_{\tilde{P} \in \mathcal{M}} \mathbb{E}_{\tilde{P}}[Y(t, z) \mid \mathbf{X} = \mathbf{x}]. \tag{16}$$

Intuition: To obtain the CAPO bounds, we need to bound $\mathbb{E}[Y \mid t, z, \mathbf{x}] = \int_{\mathcal{Y}} y p(y \mid t, z, \mathbf{x}) \, \mathrm{d}y$. We can construct *valid* bounds based on Eq. 9 by simply setting $\mu^{\pm}(y \mid t, z, \mathbf{x}) = \frac{1}{b^{\mp}(z, \mathbf{x})} \mathbb{E}[Y \mid t, z, \mathbf{x}]$. However, the resulting bounds are *not sharp*, but these are conservative and potentially uninformative. To obtain the equalities in Definition 4.1, we follow ideas from sensitivity analysis (Dorn et al., 2025; Frauen et al., 2023) and find a *cut-off value* C, such that, in a very simplified notation, we have

$$\mu^{\pm} = \frac{1}{b^{\mp}} \int_{-\infty}^{C^{\pm}} y p(y \mid \cdot) \, \mathrm{d}y + \frac{1}{b^{\pm}} \int_{C^{\pm}}^{\infty} y p(y \mid \cdot) \, \mathrm{d}y. \tag{17}$$

Let $F_Y(y) := F_Y(y \mid t, z, \mathbf{x})$ denote the conditional cumulative distribution function (CDF) of Y. We define the *conditional quantile function* of the outcome at level $\alpha^{\pm} = \frac{(1 - b^{\mp}(z, \mathbf{x})) b^{\pm}(z, \mathbf{x})}{b^{\pm}(z, \mathbf{x}) - b^{\mp}(z, \mathbf{x})}$ as

$$Q^{\pm}(t, z, \mathbf{x}) := \begin{cases} \inf\left\{ y \mid F_Y(y), \geq \alpha^{\pm} \right\}, & \text{if } b^- < 1 < b^+, \\ \inf\left\{ y \mid F_Y(y), \geq \frac{1}{2} \right\}, & \text{if } b^- = b^+, \end{cases} \tag{18}$$

where we abbreviated the notation for $b^{\pm}(z, \mathbf{x})$ through $b^{\pm}$.

The bounds are *sharp* under interference by employing $Q^{\pm}(t, z, \mathbf{x})$ as the cut-off value C. We formalize this in the following theorem, where we present the closed-form solution that facilitates estimation:

Theorem 4.2. *Let $Q^{\pm}(t, z, \mathbf{x})$ be defined as in Eq. (18) and let $(u)_+ = \max\{u, 0\}$. The sharp CAPO upper and lower bounds are given by*

$$\begin{aligned} \mu^{\pm}(t, z, \mathbf{x}) = {} & Q^{\pm}(t, z, \mathbf{x}) \\ & + \frac{1}{b^{\mp}(z, \mathbf{x})} \mathbb{E}\big[(Y - Q^{\pm}(t, z, \mathbf{x}))_+ \mid t, z, \mathbf{x}\big] \\ & - \frac{1}{b^{\pm}(z, \mathbf{x})} \mathbb{E}\big[(Q^{\pm}(t, z, \mathbf{x}) - Y)_+ \mid t, z, \mathbf{x}\big]. \end{aligned} \tag{19}$$

Remark 4.3 (Limits of the sensitivity model). If $b^-(z, \mathbf{x}) = b^+(z, \mathbf{x}) = 1$ (no exposure-mapping shift), then the identified set collapses and $\mu^{\pm}(t, z, \mathbf{x}) = \mathbb{E}[Y \mid t, z, \mathbf{x}]$. As $b^+(z, \mathbf{x}) \to \infty$ with $b^-(z, \mathbf{x})$ fixed, the upper bound concentrates on the top $b^-(z, x)$-tail of $Y \mid (t, z, \mathbf{x})$ (the lower bound on the bottom $b^-(z, x)$-tail). In the extreme limit $b^-(z, \mathbf{x}) \to 0$ and $b^+(z, \mathbf{x}) \to \infty$, the bounds become vacuous and converge to the conditional support $\mu^+(t, z, \mathbf{x}) \to \operatorname{ess\,sup}\big(Y \mid t, z, \mathbf{x}\big)$; $\mu^-(t, z, \mathbf{x}) \to \operatorname{ess\,inf}\big(Y \mid t, z, \mathbf{x}\big)$.[4]

[4]The essential supremum (essential infimum) is abbreviated by ess sup (ess inf) and defines the supremum (infimum) of the essential upper (lower) bound of the conditional distribution $Y \mid t, z, \mathbf{x}$, i.e., the upper (lower) bound with non-zero measure.

4.2. Orthogonal estimatior

• **Disadvantages of plug-in estimation.** The characterization in Theorem 4.2 immediately suggests a *plug-in* estimation strategy: estimate the cut-off $Q^{\pm}(t,z,\mathbf{x})$ and the two conditional moment functions

$$\gamma_u^{\pm}(t,z,\mathbf{x}) := \mathbb{E}\big[(Y-Q^{\pm}(\cdot))_+ \mid T=t, Z=z, \mathbf{X}=\mathbf{x}\big],$$
$$\gamma_l^{\pm}(t,z,\mathbf{x}) := \mathbb{E}\big[(Q^{\pm}(\cdot)-Y)_+ \mid T=t, Z=z, \mathbf{X}=\mathbf{x}\big].$$

Then, we obtain $\hat{\mu}^{\pm}(t,z,\mathbf{x})$ by substituting $(\hat{Q}^{\pm},\hat{\gamma}_u^{\pm},\hat{\gamma}_l^{\pm})$ into Thm. 4.2. However, such plug-in estimators suffer from substantial finite-sample bias due to nuisance estimation error, especially when the nuisance functions are more complex than the bound function itself (Kennedy, 2019). We therefore apply orthogonalization strategies (Dorn et al., 2025; Oprescu et al., 2023) and derive orthogonal pseudo-outcomes for the bounds to then estimate $\mu^{\pm}(t,z,\mathbf{x})$ by regressing the pseudo-outcomes on $\mathbf{X}$.

• **Orthogonal pseudo-outcome.** Recall that $T \in \{0,1\}$ while Z is discrete or continuous. This is relevant because our bounds involve evaluation at a fixed neighborhood exposure level z. When Z is *binary or discrete*, the functional $\mathbb{P} \mapsto \mu^{\pm}(t,z,\mathbf{x})$ is regular (pathwise differentiable) for each fixed (t,z), so we can construct an efficient influence-function-based pseudo-outcome. When Z is *continuous*, evaluation at $Z=z$ is not path-wise differentiable; we therefore replace point evaluation by a locally smoothed target using a kernel $K_h(Z-z)$, which introduces a nonparametric bias-variance tradeoff governed by the bandwidth h.

For ease of presentation, we now present the orthogonal upper bound μ^+ and defer the lower bound μ^- to Supplement C. We refer to $(\pi^t,\pi^g,Q^{\pm},\gamma_u^{\pm},\gamma_l^{\pm})$ as *nuisances*.

Theorem 4.4. *Let $S=(\mathbf{X},Y,T,Z)$. Fix (t,z). Define*

$$\omega_{z,h}(Z) := \begin{cases} \mathbf{1}_{[Z=z]}, & \text{if } Z \text{ binary/discrete}, \\ K_h(Z-z), & \text{if } Z \text{ continuous}, \end{cases}$$

and let $\pi^g(Z \mid \mathbf{X})$ denote the conditional pmf (discrete Z) or density (continuous Z). Let $\widehat{\eta}=(\widehat{\pi}^t,\widehat{\pi}^g,\widehat{Q}^+,\widehat{\gamma}_u^+,\widehat{\gamma}_l^+)$ be a set of estimated nuisances. Then, an orthogonal pseudo-outcome for the CAPO upper bound $\mu^+(t,z,\mathbf{x})$ is:

$$\phi_{t,z}^+(S;\widehat{\eta}) = \widehat{Q}^+(t,z,\mathbf{X}) + \frac{\widehat{\gamma}_u^+(t,z,\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\widehat{\gamma}_l^+(t,z,\mathbf{X})}{b^+(z,\mathbf{X})} + \frac{\mathbf{1}_{[T=t]}\,\omega_{z,h}(Z)}{\widehat{\pi}^t(\mathbf{X})\,\widehat{\pi}^g(Z\mid\mathbf{X})}\left[\frac{(Y-\widehat{Q}^+(t,Z,\mathbf{X}))_+ - \widehat{\gamma}_u^+(t,Z,\mathbf{X})}{b^-(Z,\mathbf{X})} - \frac{(\widehat{Q}^+(t,Z,\mathbf{X})-Y)_+ - \widehat{\gamma}_l^+(t,Z,\mathbf{X})}{b^+(Z,\mathbf{X})}\right]. \tag{20}$$

Moreover, when $\widehat{\eta}=\eta$, the pseudo-outcome is unbiased for its target bound functional (see Remark 4.5).

Algorithm 1 Orthogonal estimator for the bounds

Input: data $\{S_i=(\mathbf{X}_i,Y_i,T_i,Z_i)\}_{i=1}^n$, target (t,z), bandwidth h (if Z continuous), folds $\{\mathcal{I}_k\}_{k=1}^K$, nuisance estimators, regression learner $\widehat{\mathbb{E}}_n$
for $k=1,\dots,K$ **do**
  Fit nuisances $\widehat{\eta}^{(-k)}$ on $\{S_i : i \notin \mathcal{I}_k\}$
  for $i \in \mathcal{I}_k$ **do**
    $\widehat{\phi}_{t,z,i}^+ \leftarrow \phi_{t,z}^+(S_i;\widehat{\eta}^{(-k)})$
  end for
end for
$\widehat{\psi}^+(t,z) \leftarrow \frac{1}{n}\sum_{i=1}^n \widehat{\phi}_{t,z,i}^+$ (sample average)
$\widehat{\mu}^+(t,z,\mathbf{x}) \leftarrow \widehat{\mathbb{E}}_n[\phi_{t,z}^+(S;\widehat{\eta}) \mid \mathbf{X}=\mathbf{x}]$ (regression fit)
Output: $\widehat{\mu}^+(t,z,\cdot)$, $\widehat{\psi}^+(t,z)$

Remark 4.5 (Unbiasedness of the pseudo-outcome). When $\widehat{\eta}=\eta$ and Z is *discrete*, we have $\mathbb{E}\big[\phi_{t,z}^+(S;\eta) \mid \mathbf{X}=\mathbf{x}\big] = \mu^+(t,z,\mathbf{x})$ and $\mathbb{E}\big[\phi_{t,z}^+(S;\eta)\big] = \psi^+(t,z)$. When Z is *continuous*, the kernel-localized pseudo-outcome targets a bandwidth-indexed functional (μ_h^+,ψ_h^+); under standard smoothness in z, $\mu_h^+(t,z,\mathbf{x}) \to \mu^+(t,z,\mathbf{x})$ and $\psi_h^+(t,z) \to \psi^+(t,z)$ as $h \downarrow 0$.

• **Bound estimation algorithm.** Motivated by Theorem 4.4, we estimate the bounds via a *two-stage procedure* (Algorithm 1): we first learn the nuisance functions $\widehat{\eta}$, then evaluate the orthogonal pseudo-outcome $\phi_{t,z}^+(S;\widehat{\eta})$ and finally obtain (i) the CAPO bound $\widehat{\mu}^+(t,z,\mathbf{x}) = \widehat{\mathbb{E}}_n[\phi_{t,z}^+(S;\widehat{\eta}) \mid \mathbf{X}=\mathbf{x}]$ by regressing the pseudo-outcome on $\mathbf{X}$ and (ii) the APO bound $\widehat{\psi}^+(t,z) = \widehat{\mathbb{E}}_n\big[\phi_{t,z}^+(S;\widehat{\eta})\big]$ via sample averaging. To mitigate overfitting bias and enable standard orthogonalization guarantees, we compute $\widehat{\phi}_{t,z,i}^+$ using K-fold cross-fitting (Chernozhukov et al., 2018): each $\widehat{\phi}_{t,z,i}^+$ uses nuisance estimates trained on data not containing i.

4.3. Theoretical properties of our bound estimator

Theorem 4.2 shows that the identified CAPO bounds $\mu^{\pm}(t,z,\mathbf{x})$ are sharp. We establish *three additional guarantees* for our orthogonal estimator (Theorem 4.4). (1) Orthogonality yields *second-order* sensitivity to nuisance estimation error, implying quasi-oracle rates for the CAPO bounds and (for discrete Z) root-n inference for the APO bounds. (2) If $Q^{\pm}$ is consistently estimated and either the propensity models (π^t,π^g) *or* the moment functions $(\gamma_u^{\pm},\gamma_l^{\pm})$ are consistently estimated, then our estimated endpoints converge (in $L_2(P_{\mathbf{X}})$) to the *sharp* bounds. (3) If $Q^{\pm}$ is misspecified, but either (π^t,π^g) or $(\gamma_u^{\pm},\gamma_l^{\pm})$ is consistently estimated, the (C)APO intervals remain asymptotically *valid*, though potentially conservative. We present results for discrete Z in the main text; the case with continuous Z is deferred to Supplement C.4. All proofs are in Supplement D.

• (1) **Quasi-oracle learning via orthogonality.** The next theorem is the key guarantee: under standard regularity conditions, orthogonality implies that nuisance errors contribute

only through error *products*.[5]

Assumption 4.6 (Regularity and overlap)**.** There exist $\varepsilon > 0$ and $M < \infty$ such that, a.s.: (i) $\varepsilon \le \pi^t(\mathbf{X}), \widehat{\pi}^t(\mathbf{X}) \le 1 - \varepsilon$; (ii) if Z is discrete, $\varepsilon \le \pi^g(z \mid \mathbf{X}), \widehat{\pi}^g(z \mid \mathbf{X})$ for all relevant $(z, \mathbf{X})$; if Z is continuous, there exists a neighborhood $\mathcal{N}_z$ of z such that for all $u \in \mathcal{N}_z$, $\varepsilon \le \pi^g(u \mid \mathbf{X}), \widehat{\pi}^g(u \mid \mathbf{X}) \le M$; (iii) $|Y|, |\widehat{\gamma}_u^{\pm}|, |\widehat{\gamma}_l^{\pm}|, |\widehat{Q}^{\pm}| \le M$.

Theorem 4.7 (Second-order nuisance error (discrete Z))**.** *Assume Z is discrete and Assumption 4.6 holds. Let $\widehat{\eta} = (\widehat{\pi}^t, \widehat{\pi}^g, \widehat{Q}^+, \widehat{\gamma}_u^+, \widehat{\gamma}_l^+)$ be the cross-fitted nuisances used in $\phi_{t,z}^+(S; \widehat{\eta})$ from Theorem 4.4. Define $r_{n,\pi} := \|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2$, $r_{n,Q} := \|\widehat{Q}^+ - Q^+\|_2$, and $r_{n,\gamma} := \|\widehat{\gamma}_u^+ - \gamma_u(\widehat{Q}^+; \cdot)\|_2 + \|\widehat{\gamma}_l^+ - \gamma_l(\widehat{Q}^+; \cdot)\|_2$, where $\gamma_u(\widehat{Q}^+; \mathbf{X}) := \mathbb{E}[(Y - \widehat{Q}^+(\mathbf{X}))_+ \mid T = t, Z = z, \mathbf{X}]$ and $\gamma_l(\widehat{Q}^+; \mathbf{X}) := \mathbb{E}[(\widehat{Q}^+(\mathbf{X}) - Y)_+ \mid T = t, Z = z, \mathbf{X}]$. Then*

$$\left\|\mathbb{E}\left[\phi_{t,z}^+(S; \widehat{\eta}) - \phi_{t,z}^+(S; \eta) \mid \mathbf{X}\right]\right\|_2 = O_p\left(r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2\right). \tag{21}$$

Thus, the contribution of nuisances to the estimation error is only second order. Next, we show that the final-stage regression can achieve the same rate as if the true pseudo-outcomes were observed (a quasi-oracle property), even if the nuisances converge more slowly.

Assumption 4.8 (Second-stage regression rate)**.** Fix (t, z) and let $\widehat{\phi}_{t,z,i}^+$ denote the cross-fitted pseudo-outcome. Let $m_{t,z}^+(\mathbf{x}) := \mathbb{E}[\widehat{\phi}_{t,z}^+ \mid \mathbf{X} = \mathbf{x}]$. Assume the regression learner used to form $\widehat{\mu}^+(t, z, \cdot)$ satisfies

$$\|\widehat{\mu}^+(t, z, \cdot) - m_{t,z}^+(\cdot)\|_2 = O_p(\delta_n), \tag{22}$$

for some (possibly model-dependent) rate δ_n.

Remark 4.9 (Second-stage regression assumption). Assumption 4.8 treats the final-stage regression step as a black box: it assumes that, when regressing the cross-fitted pseudo-outcomes on $\mathbf{X}$, the learner attains an L_2 error rate δ_n uniformly over the admissible nuisance estimates $\widehat{\eta} \in \Xi$. A broad class of learners satisfy this, including nonparametric least-squares/ERM estimators over a bounded function class $\mathcal{F}$ with bracketing entropy $\log N_{[]}(\mathcal{F}, \epsilon) \lesssim \epsilon^{-r}$ $(0 < r < 2)$, which yields the usual regression rate $\delta_n \asymp n^{-1/(2+r)}$ (up to approximation error); in particular, for d-dimensional Hölder(β) classes, $\delta_n = n^{-\beta/(2\beta+d)}$. More generally, black-box regressors satisfying standard stability/oracle-inequality properties (e.g., linear smoothers) also fit this template (Kennedy, 2023). We therefore state our results in terms of δ_n, which separates orthogonalization from the choice of final-stage regression method.

[5] *Notation:* Let $\|f\|_2 := \{\mathbb{E}[f(\mathbf{X})^2]\}^{1/2}$ denote the $L_2(P_{\mathbf{X}})$ norm. We write $W_n = o_p(a_n)$ if $W_n/a_n \to 0$ in probability and $W_n = O_p(a_n)$ if W_n/a_n is bounded in probability.We write $\rightsquigarrow$ for convergence in distribution.

Corollary 4.10 (Quasi-oracle rates and inference (discrete Z))**.** *Suppose Assumptions 4.6 and 4.8 hold, and let $r_{n,\pi}, r_{n,\gamma}, r_{n,Q}$ be as in Theorem 4.7.*

CAPO rates: The CAPO upper-bound estimator satisfies

$$\|\widehat{\mu}^+(t, z, \cdot) - \mu^+(t, z, \cdot)\|_2 = O_p\left(\delta_n + r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2\right).$$

In particular, if $r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2 = o_p(\delta_n)$, then $\|\widehat{\mu}^+(t, z, \cdot) - \mu^+(t, z, \cdot)\|_2 = O_p(\delta_n)$.

APO rates: The APO upper-bound estimator $\widehat{\psi}^+(t, z) = \mathbb{E}_n[\widehat{\phi}_{t,z}^+]$ satisfies

$$\left|\widehat{\psi}^+(t, z) - \psi^+(t, z)\right| = O_p\left(n^{-1/2} + r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2\right). \tag{23}$$

If moreover $r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2 = o_p(n^{-1/2})$, then

$$\sqrt{n}\left(\widehat{\psi}^+(t, z) - \psi^+(t, z)\right) \rightsquigarrow \mathcal{N}\left(0,\, V^+(t, z)\right), \tag{24}$$

i.e., the APO bound estimator is asymptotically normal with variance $V^+(t, z) := \mathrm{Var}(\phi_{t,z}^+(S; \eta))$ (efficiency bound).

Corollary 4.10 establishes a *quasi-oracle* property: if the nuisance estimators converge at rate $o_p(\delta_n^{1/2})$ for CAPO bounds (or $o_p(n^{-1/4})$ for APO), the estimator achieves the oracle rate $O_p(\delta_n)$, as if the nuisances were known. This follows, since nuisance errors enter only through the second-order remainder $r_{n,\pi} r_{n,\gamma} + r_{n,Q}^2$. For APOs, this additionally enables valid and tight inference.

• (2) **Sharpness of the *estimated* bounds.** Next, we state conditions under which our estimates converge to the *sharp* identified bounds from Theorem 4.2, so that the *estimated* bounds are also sharp.

Proposition 4.11 (Consistency for sharp bounds (discrete Z))**.** *Assume the conditions of Corollary 4.10 hold. Suppose $\delta_n = o_p(1)$ and $r_{n,Q} = o_p(1)$, and, in addition, either $r_{n,\pi} = o_p(1)$ or $r_{n,\gamma} = o_p(1)$. Then, $\|\widehat{\mu}^{\pm}(t, z, \cdot) - \mu^{\pm}(t, z, \cdot)\|_2 = o_p(1)$ and $|\widehat{\psi}^{\pm}(t, z) - \psi^{\pm}(t, z)| = o_p(1)$, Consequently, the estimated CAPO and APO intervals converge to the* sharp *identified intervals.*

Proposition 4.11 shows that, if $\widehat{Q}^{\pm}$ is consistent and either the propensity or conditional-moment nuisance is consistent, then Algorithm 1 consistently estimates the sharp CAPO/APO bounds in Theorem 4.2.

• (3) **Validity under $\widehat{Q}^{\pm}$ misspecification.** Sharpness guarantees require that $Q^{\pm}$ is estimated consistently. We show that, even when $Q^{\pm}$ is misspecified, our estimator yields conservative but valid bounds, provided one of the two (first-stage) nuisance "blocks" is consistently learned.

Corollary 4.12 (Asymptotic validity under misspecified cutoffs (discrete Z))**.** *Assume the conditions of Corollary 4.10*

hold. Let $\overline{Q}^{\pm}(t,z,\mathbf{x})$ be any measurable cut-off and define the induced (possibly non-sharp) bounds

$$\overline{\mu}^{\pm}(t,z,\mathbf{x};\overline{Q}^{\pm}) = \overline{Q}^{\pm}(t,z,\mathbf{x}) + \frac{1}{b^{\mp}(z,\mathbf{x})}\mathbb{E}\big[(Y-\overline{Q}^{\pm}(t,z,\mathbf{x}))_+ \mid t,z,\mathbf{x}\big] - \frac{1}{b^{\pm}(z,\mathbf{x})}\mathbb{E}\big[(\overline{Q}^{\pm}(t,z,\mathbf{x})-Y)_+ \mid t,z,\mathbf{x}\big]. \tag{25}$$

(and analogously $\overline{\psi}^{\pm}(t,z) := \mathbb{E}[\overline{\mu}^{\pm}(t,z,\mathbf{X})]$). Then, $[\overline{\mu}^{-}(t,z,\mathbf{x}),\overline{\mu}^{+}(t,z,\mathbf{x})]$ is a valid (not necessarily sharp) CAPO interval, and likewise for $[\overline{\psi}^{-}(t,z),\overline{\psi}^{+}(t,z)]$.

Moreover, if $\widehat{Q}^{\pm} \to \overline{Q}^{\pm}$ in L_2 and either (i) $(\widehat{\pi}^t,\widehat{\pi}^g)$ is consistent, or (ii) $(\widehat{\gamma}_u^{\pm},\widehat{\gamma}_l^{\pm})$ is consistent for the tail-moment targets induced by $\overline{Q}^{\pm}$, then the resulting estimated (C)APO intervals converge to $[\overline{\mu}^{-},\overline{\mu}^{+}]$ and $[\overline{\psi}^{-},\overline{\psi}^{+}]$ and are asymptotically valid, though potentially conservative. If $\overline{Q}^{\pm} = Q^{\pm}$, the bounds coincide with the sharp bounds.

Remark 4.13 (Continuous Z). When Z is continuous, evaluation at a point z requires kernel localization, leading to the usual bias-variance tradeoff in the bandwidth. We defer the corresponding rates and inference results to Supplement C.4.

5. Experiments

Data:[6] We follow common practice in causal partial identification and evaluate our framework on synthetic datasets. We generate simple networks with $N = 1000$ nodes and a 1-dimensional covariate (small dataset) and more complex networks with $N = 6000$ nodes and 6-dimensional covariates (large dataset).

Evaluation: Our goal is to *demonstrate the theoretical properties* of our framework: (1) We evaluate our bounds in terms of *validity*, i.e., we show that our bounds contain the true outcome whenever the constraints given by $b^{\pm}$ are satisfied. (2) We compare the *convergence* of our orthogonal bound estimator against the plug-in estimator. (3) We assess the *informativeness* of our bounds in terms of the widths of the resulting intervals. We report all results over 10 runs.[7]

Results: *Research question (1): Are our bounds valid?* $\Rightarrow$ We assess validity by visualizing our bounds for exposure mapping (1) on both the small (Fig. 5) and the large dataset (Fig. 6). We compare our bounds over various specifications of $b^{\pm}$ (= "factor" $\times$ true misspecification). We observe that, for a too small sensitivity bound assumption (factor 0.5), the bounds do not completely contain the true effect. For a sufficient bound assumption (factor ≥ 1), our bounds are valid.

[6] Data and implementation details are in Supplement F.

[7] Our goal is to show the validity and advantages of our model-agnostic partial identification framework. We thus refrain from comparing different instantiations of the nuisances or second-stage models.

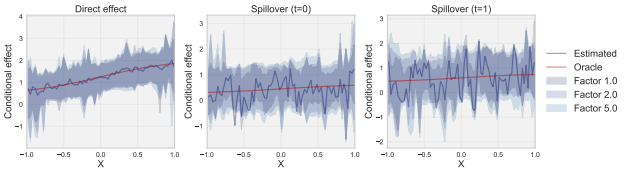


Figure 5. **Conditional effect bounds:** Visualization of our bounds around the true effect for the weighted mean exposure mapping. The width of the bounds is increasing in the sensitivity factor. Starting from factor 1.0, our bounds contain the true effect.

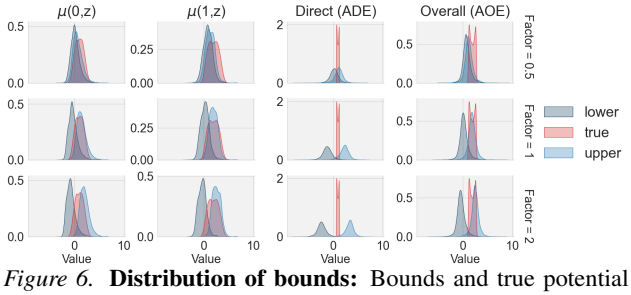
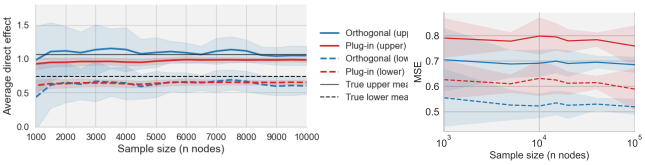


Figure 6. **Distribution of bounds:** Bounds and true potential outcomes and effects for the weighted mean exposure mapping on the large dataset. For sufficiently large sensitivity factor (≥ 1), the distributions of upper and lower bounds enclose the true PO/effect, thus confirming that the bounds are valid.

Research question (2): How does the convergence of our orthogonal estimator compare to a simple plug-in estimator? $\Rightarrow$ We compare the convergence and the behavior of the coverage of our orthogonal estimator for increasing network size N for setting (2) in Fig. 7. As expected: (a) our bounds are *valid even for small sample sizes and are quickly approaching the sharp oracle bounds*, whereas the plug-in bounds fail to provide correct coverage; (b) due to orthogonality, our framework benefits from *faster convergence*.



(a) *(b)*

Figure 7. **Coverage & convergence:** *(a)* Bounds on the ADE under a threshold exposure mapping increasing number of nodes (6-dim covariates). The orthogonal bounds are valid, approaching the sharp oracle bounds, whereas the plug-in bounds are not valid. *(b)* Our orthogonal bounds show faster convergence than the plug-in bounds.

Research question (3): How informative are our bounds? $\Rightarrow$ We assess the width of our intervals under exposure mapping (3). For decision-making, informative bounds (i) are narrow compared to the outcome range and (ii) are either strictly positive or negative. Our bounds fulfill both desiderata: (i) The average width of the ADE intervals over all z with correctly specified sensitivity factors corresponds to merely 8.71% ($\pm 0.37\%$) of the overall outcome range. (ii) All of our intervals are strictly bounded away from zero, correctly recognizing the positive treatment effect.

Conclusion: We proposed a flexible and model-agnostic framework for partial identification of potential outcomes and treatment effects on networks in the presence of exposure mapping misspecification. We derived a robust estimation framework with quasi-oracle rate properties and showed that the estimated bounds remain valid and sharp. Finally, we instantiated our framework with three commonly employed exposure mappings and highlighted the interpretability of our bounds in extensive experiments.

Acknowledgments

Miruna Oprescu was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Award DE-SC0023112.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Adhikari, S. and Zheleva, E. Inferring individual direct causal effects under heterogeneous peer influence. *Machine Learning*, 114(4):113, 2025.

Ali, S., Faruque, O., and Wang, J. Estimating direct and indirect causal effects of spatiotemporal interventions in presence of spatial interference. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2024.

Alzubaidi, S. H. and Higgins, M. J. Detecting treatment interference under k-nearest-neighbors interference. *Journal of Causal Inference*, 12(1):20230029, 2024.

Anselin, L. *Spatial Econometrics: Methods and Models*. Springer Science & Business Media, 1988.

Aronow, P. M. and Samii, C. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017.

Bargagli-Stoffi, F. J., Tortú, C., and Forastiere, L. Heterogeneous treatment and spillover effects under clustered network interference. *The Annals of Applied Statistics*, 19(1):28–55, 2025.

Belloni, A., Fang, F., and Volfovsky, A. Neighborhood adaptive estimators for causal inference under network interference. *arXiv preprint*, arXiv:2212.03683, 2022.

Bhattacharya, R., Malinsky, D., and Shpitser, I. Causal inference under interference and network uncertainty. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

Chen, W., Cai, R., Yang, Z., Qiao, J., Yan, Y., Li, Z., and Hao, Z. Doubly robust causal effect estimation under networked interference via targeted learning. In *International Conference on Machine Learning (ICML)*, 2024a.

Chen, W., Cai, R., Qiao, J., Yan, Y., and Hernández-Lobato, J. M. Causal effect estimation under networked interference without networked unconfoundedness assumption. *arXiv preprint*, arXiv:2502.19741, 2025.

Chen, Z., Guo, R., Ton, J.-F., and Liu, Y. Conformal counterfactual inference under hidden confounding. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2024b.

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

Dorn, J. and Guo, K. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, 118(544):2645–2657, 2022.

Dorn, J., Guo, K., and Kallus, N. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *Journal of the American Statistical Association*, 120(549):331–342, 2025.

Egami, N. Spillover effects in the presence of unobserved networks. *Political Analysis*, 29(3):287–316, 2021.

Fang, F. and Forastiere, L. Design-based weighted regression estimators for average and conditional spillover effects. *arXiv preprint*, arXiv:2512.12452, 2025.

Forastiere, L., Airoldi, E. M., and Mealli, F. Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association*, 116(534):901–918, 2021.

Forastiere, L., Mealli, F., Wu, A., and Airoldi, E. M. Estimating causal effects under network interference with bayesian generalized propensity scores. *Journal of Machine Learning Research*, 23:1–61, 2022.

Frauen, D., Melnychuk, V., and Feuerriegel, S. Sharp bounds for generalized causal sensitivity analysis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Frauen, D., Imrie, F., Curth, A., Melnychuk, V., Feuerriegel, S., and van der Schaar, M. A neural framework for generalized causal sensitivity analysis. In *International Conference on Learning Representations (ICLR)*, 2024.

Freedman, S., Sacks, D. W., Simon, K., and Wing, C. Direct and indirect effects of vaccines: Evidence from COVID-19. *American Economic Journal: Applied Economics*, 18 (1):1–43, 2026.

Giffin, A., Reich, B. J., Yang, S., and Rappold, A. G. Generalized propensity score approach to causal inference with spatial interference. *Biometrics*, 79(3):2220–2231, 2023.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254, 2015.

Hess, K., Frauen, D., Melnychuk, V., and Feuerriegel, S. Efficient and sharp off-policy learning under unobserved confounding. In *International Conference on Learning Representations (ICLR)*, 2026.

Hoshino, T. and Yanagi, T. Causal inference with noncompliance and unknown interference. *Journal of the American Statistical Association*, 119(548):2869–2880, 2024.

Jesson, A., Douglas, A., Manshausen, P., Solal, M., Meinshausen, N., Stier, P., Gal, Y., and Shalit, U. Assessing sensitivity to an unobserved scalable sensitivity and uncertainty analyses for causal-scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Jiang, S. and Sun, Y. Estimating causal effects on networked observational data via representation learning. In *International Conference on Information & Knowledge Management (CIKM)*, 2022.

Kallus, N. and Zhou, A. Confounding-robust policy improvement. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Kallus, N., Mao, X., and Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Kennedy, E. H. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.

Kennedy, E. H. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Khot, A., Oprescu, M., Schröder, M., Kagawa, A., and Luo, X. Spatial deconfounder: Interference-aware deconfounding for spatial causal inference. *arXiv preprint*, arXiv:2510.08762, 2025.

Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., and Silva, R. The sensitivity of counterfactual fairness to unmeasured confounding. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Leung, M. P. Treatment and spillover effects under network interference. *Review of Economics and Statistics*, 102(2): 368–380, 2020.

Leung, M. P. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022.

Lin, X., Bao, H., Cui, Y., Takeuchi, K., and Kashima, H. Scalable individual treatment effect estimator for large graphs. *Machine Learning*, 114(1):23, 2025.

Liu, J., Ye, F., and Yang, Y. Nonparametric doubly robust estimation of causal effect on networks in observational studies. *Stat*, 12(1):e549, 2023.

Ma, Y. and Tresp, V. Causal inference under networked interference and intervention policy enhancement. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Matthay, E. C. and Glymour, M. M. Causal inference challenges and new directions for epidemiologic research on the health effects of social policies. *Current Epidemiology Reports*, 9(1):22–37, 2022.

McNealis, V., Moodie, E. E. M., and Dean, N. Revisiting the effects of maternal education on adolescents' academic performance: Doubly robust estimation in a network-based observational study. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 73(3):715–734, 2024.

Melnychuk, V., Frauen, D., and Feuerriegel, S. Partial counterfactual identification of continuous outcomes with a curvature sensitivity model. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Ogburn, E. L., Sofrygin, O., Díaz, I., and van der Laan, M. J. Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611, 2024.

Ohnishi, Y., Karmakar, B., and Sabbaghi, A. Degree of interference: A general framework for causal inference under interference. *Journal of Machine Learning Research*, 26 (120):1–37, 2025.

Oprescu, M., Dorn, J., Ghoummaid, M., Jesson, A., Kallus, N., and Shalit, U. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. In

International Conference on Machine Learning (ICML), 2023.

Oprescu, M., Park, D. K., Luo, X., Yoo, S., and Kallus, N. GST-UNet: A neural framework for spatiotemporal causal inference with time-varying confounding. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Papadogeorgou, G. and Samanta, S. Spatial causal inference in the presence of unmeasured confounding and interference. *arXiv preprint*, arXiv:2303.08218, 2023.

Qu, Z., Xiong, R., Liu, J., and Imbens, G. Semiparametric estimation of treatment effects in observational studies with heterogeneous partial interference. *arXiv preprint*, arXiv:2107.12420, 2024.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. *Statistical Models in Epidemiology*, 116:1–92, 1999.

Rosenbaum, P. R. and Rubin, D. B. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, pp. 212–218, 1983.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Sävje, F. Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):1–15, 2024.

Sävje, F., Aronow, P., and Hudgens, M. Average treatment effects in the presence of unknown interference. *Annals of Statistics*, 49(2):673–701, 2021.

Schröder, M., Frauen, D., Schweisthal, J., Heß, K., Melnychuk, V., and Feuerriegel, S. Conformal prediction for causal effects of continuous treatments. *arXiv preprint*, arXiv:2407.03094, 2024.

Sengupta, S., Imai, K., and Papadogeorgou, G. Low-rank covariate balancing estimators under interference. *arXiv preprint*, arXiv:2512.13944, 2025.

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

Tchetgen Tchetgen, E. J. and VanderWeele, T. J. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.

VanderWeele, T. J., Tchetgen Tchetgen, E. J., and Halloran, M. E. Interference and sensitivity analysis. *Statistical Science : A Review Journal of the Institute of Mathematical Statistics*, 29(4):687–706, 2014.

Vansteelandt, S., Goetghebeur, E., Kenward, M. G., and Molenberghs, G. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, 16:953–979, 2006.

Viviano, D. Policy targeting under network interference. *Review of Economic Studies*, 92(2):1257–1292, 2025.

Wang, Y., Frauen, D., Schweisthal, J., Schröder, M., and Feuerriegel, S. Assessing the robustness of heterogeneous treatment effects in survival analysis under informative censoring. *arXiv preprint*, arXiv:2510.13397, 2025.

Weinstein, B. and Nevo, D. Causal inference with misspecified network interference structure. *arXiv preprint*, arXiv:2302.11322, 2023.

Weinstein, B. and Nevo, D. Bayesian estimation of causal effects using proxies of a latent interference network. *arXiv preprint*, arXiv:2505.08395, 2025.

Yin, M., Shi, C., Wang, Y., and Blei, D. M. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, 119(545):122–135, 2024.

Zhang, C., Mohan, K., and Pearl, J. Causal inference under interference and model uncertainty. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2023.

Zhang, Y., Onnela, J.-P., Sun, S., and Wang, R. Identification and estimation of heterogeneous interference effects under unknown network. *arXiv preprint*, arXiv:2510.10508, 2025.

A. Notation

$\mathcal{G}$	Network consisting of N nodes
$\mathcal{N}, \mathcal{E}$	Sets of nodes and edges in $\mathcal{G}$
$\mathcal{N}_i, \mathcal{N}_{-i}$	Network partition respective to individual i, where $\mathcal{N}_i$ defines the neighborhood of node i, i.e., the set of nodes j connected to i by an edge, and $\mathcal{N}_{-i}$ the complement of $\mathcal{N}_i$ in $\mathcal{N}$
n_i	Degree of node i, i.e., number of neighbors of i
$T_i, T_{\mathcal{N}_i}$	Binary unit and neighborhood treatments
$\mathbf{X}$	Confounders in domain $\mathcal{X}$
Y	Outcome in domain $\mathcal{Y}$
g	Exposure mapping, $g : [0,1]^n \mapsto \mathcal{Z}$
Z	Scalar summarizing the neighborhood exposure, i.e., $Z = g(T_{\mathcal{N}})$
$Y(t,z)$	Potential outcome under unit treatment t and neighborhood exposure z
$\psi(t,z)$	Average PO estimator
$\mu(t,z,\mathbf{x})$	CAPO estimator
$Q^+(t,z,\mathbf{x}), Q^-(t,z,\mathbf{x})$	Upper and lower quantiles on the conditional CDF wrt. $b^+(z,\mathbf{x}), b^-(z,\mathbf{x})$
$b^+(z,\mathbf{x}), b^-(z,\mathbf{x})$	Upper and lower bound on the exposure mapping shift due to misspecification
$\gamma_u^{\pm}(t,z,\mathbf{x})$	$\mathbb{E}\big[(Y - Q^{\pm})_+ \mid t,z,\mathbf{x}\big]$, where $(u)_+ = \max\{u,0\}$
$\gamma_l^{\pm}(t,z,\mathbf{x})$	$\mathbb{E}\big[(Q^{\pm} - Y)_+ \mid t,z,\mathbf{x}\big]$, where $(u)_+ = \max\{u,0\}$
$\pi^t(\mathbf{x}), \pi^g(z \mid \mathbf{x})$	Unit node and neighborhood propensity functions
η	Nuissance functions
$\phi_{t,z}^+(S,\eta), \phi_{t,z}^-(S,\eta)$	Upper and lower orthogonal pseudo-outcome
$\mu_{\mathrm{DR}}^{\pm}(t,z,\mathbf{x})$	Orthogonal upper and lower CAPO bound estimator
$\psi_{\mathrm{DR}}^+(t,z), \psi_{DR}^-(t,z)$	Orthogonal upper and lower average PO bound estimator

B. Extended related work

Below, we discuss related work on (i) network interference (Appendix B.1) and (ii) partial identification methods (Appendix B.2). In Appendix B.1, we first provide an overview of the related but non-discussed fields on graph neural networks (GNNs) for interference modeling and causal methods for spatial interference. Then, we give a detailed overview of other works addressing misspecified exposure mappings and highlight how our work differentiates itself from these works. In Appendix B.2, we give a brief overview of sensitivity methods for partial identification.

B.1. Network interference

GNNs: Standard graph ML models fail to estimate causal effects on networks as they follow a different optimization goal (Jiang & Sun, 2022). Furthermore, these methods are computationally inefficient, thereby rendering the application to large network data challenging or impossible (Lin et al., 2025). Therefore, multiple methods for learning a neighborhood representation of the covariates through a GNN have been proposed (e.g., Adhikari & Zheleva, 2025; Jiang & Sun, 2022; Lin et al., 2025; Ma & Tresp, 2021). However, these methods commonly assume a known exposure mapping of the neighborhood treatments.

Spatial interference: In environmental science, treatment effect estimation often faces the challenge of spatial interference, i.e., treatments from different locations affecting the outcomes at other locations. Here, data are often assumed to stem from a spatial grid, meaning that distances between nodes (=spatial cells) and the number of neighboring nodes are fixed for the entire network. Spatial causal inference approaches commonly assume a correctly specified exposure mapping as in approaches targeting network interference (e.g., Anselin, 1988; Giffin et al., 2023; Hanks et al., 2015; Papadogeorgou & Samanta, 2023). More recent approaches (e.g., Ali et al., 2024; Khot et al., 2025; Oprescu et al., 2025) assume *localized interference* based on a specified neighborhood radius and employ deep learning methods to capture the latent interference structure. Overall, all methods rely on various types of *correctly specified* exposure mappings.

Misspecified exposure mappings: Only very few works consider causal effect estimation under a misspecified exposure mapping or network uncertainty. Most works target estimation on unknown networks, i.e., when there is uncertainty about the existence of certain edges in the network. Egami (2021) provides bounds on average causal effects under network misspecification in RCTs. Sävje et al. (2021) further shows that, under unknown but limited interference in RCTs, average effects can be identified by certain standard estimators. In a follow-up work, Sävje (2024) assesses the bias of treatment effect estimators when there is a mismatch between the exposure mappings at experiment and inference time. Ohnishi et al. (2025) learn the latent structure of interference under a Bayesian prior to estimate causal effects under an arbitrary, unknown interference structure. However, the method is only applicable to randomized control trials (RCTs) and targets specific sub-effects different from the standard CATE and ATE.

In the more general setting of observational data, Weinstein & Nevo (2023) derive bounds on the bias arising from estimating causal effects under a misspecified network. In follow-up work, Weinstein & Nevo (2025) and Zhang et al. (2025) propose frameworks for estimating causal effects when only proxy networks are available. Similarly, Zhang et al. (2023) models uncertain interaction using linear graphical causal models, quantifies bias when iid (SUTVA) is incorrectly assumed, and presents a procedure to remove such bias and derive bounds for *average* causal effects.

Other works more similar to our work focus on uncertainty in the neighborhood radius. Leung (2022) considers approximate neighborhood interference, allowing treatments assigned to units further from the unit of interest to have potentially nonzero, but smaller, effects on the unit's outcome. In contrast to our work, the proposed method is restricted to the specific type of misspecification and only targets the average overall effect. Belloni et al. (2022) consider estimation under an unknown neighborhood radius, similar to our third use-case. However, the proposed method needs strong modeling assumptions and only applies to the *average* direct effect.

Orthogonal to our work, Hoshino & Yanagi (2024) propose an *instrumental exposure mapping* to summarize the spillover effects into a low-dimensional variable in instrumental variable regression settings. They show that the resulting estimands for *average* effects are interpretable even if the neighborhood radius is misspecified.

Overall, there does **not** exist a general framework for bounding potential outcomes and treatment effects under various types of exposure mapping misspecification for both experimental and observational data. This is our contribution.

B.2. Partial identification

Sensitivity analysis as partial identification: A commonly employed tool for partial identification is causal sensitivity analysis (CSA). Instead of point-identifying an estimand under the strong assumptions of *no unobserved confounding*, CSA allows unobserved confounding up to a specified confounding strength and derives bounds for causal quantities. A broad range of sensitivity models has been proposed, differing in what aspect of the data-generating process is perturbed and how deviations are parameterized (e.g., Rosenbaum & Rubin, 1983; Robins et al., 1999; Vansteelandt et al., 2006).

Marginal sensitivity model (MSM) and extensions: Much of the recent literature centers on the MSM (Tan, 2006), where bounds are obtained by optimizing over admissible propensity reweightings. Recent works show that naïve procedures can be conservative and derive *sharp* bound characterizations and estimators (Dorn & Guo, 2022; Dorn et al., 2025), which also enables efficient learning of *CATE bounds* via meta-learning (Oprescu et al., 2023). Beyond binary treatments and standard treatment effect queries, other works propose continuous-treatment marginal sensitivity models (Jesson et al., 2022), generalized sensitivity models with sharp bounds for broader causal queries (Frauen et al., 2023), and neural frameworks that automate generalized sensitivity analysis across model classes and treatment types (Frauen et al., 2024). Sensitivity-style partial identification has also been used in adjacent ML problems such as confounding-robust policy learning (Hess et al., 2026; Kallus & Zhou, 2018; Kallus et al., 2019), partial identification of counterfactual queries (Melnychuk et al., 2023), survival analysis (Wang et al., 2025), sensitivity auditing of causal fairness (Kilbertus et al., 2019; Schröder et al., 2024), and modern uncertainty quantification, e.g., conformal-style intervals for ITEs at a given sensitivity level (Yin et al., 2024).

C. Extended theory

C.1. Summary of bounds

Potential outcomes	$\mu^+(t,z,\mathbf{x}) = Q^+(t,z,\mathbf{x}) + \frac{1}{b^-(z,\mathbf{x})}\gamma_u^+(t,z,\mathbf{X}) - \frac{1}{b^+(z,\mathbf{x})}\gamma_u^-(t,z,\mathbf{X})$ $\mu^-(t,z,\mathbf{x}) = Q^-(t,z,\mathbf{x}) + \frac{1}{b^+(z,\mathbf{x})}\gamma_u^-(t,z,\mathbf{X}) - \frac{1}{b^-(z,\mathbf{x})}\gamma_l^-(t,z,\mathbf{X})$
Pseudo-outcomes (discrete)	$\phi_{t,z}^+(S;\widehat{\eta}) = \widehat{Q}^+(t,z,\mathbf{X}) + \frac{\widehat{\gamma}_u^+(t,z,\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\widehat{\gamma}_l^+(t,z,\mathbf{X})}{b^+(z,\mathbf{X})}$ $+\frac{\mathbf{1}_{[T=t]}\mathbf{1}_{[Z=z]}}{\widehat{\pi}^t(\mathbf{X})\widehat{\pi}^g(Z\|\mathbf{X})}\left[\frac{(Y-\widehat{Q}^+(t,Z,\mathbf{X}))_+-\widehat{\gamma}_u^+(t,Z,\mathbf{X})}{b^-(Z,\mathbf{X})} - \frac{(\widehat{Q}^+(t,Z,\mathbf{X})-Y)_+-\widehat{\gamma}_l^+(t,Z,\mathbf{X})}{b^+(Z,\mathbf{X})}\right]$ $\phi_{t,z}^-(S;\widehat{\eta}) = \widehat{Q}^-(t,z,\mathbf{X}) + \frac{\widehat{\gamma}_u^-(t,z,\mathbf{X})}{b^+(z,\mathbf{X})} - \frac{\widehat{\gamma}_l^-(t,z,\mathbf{X})}{b^-(z,\mathbf{X})}$ $+\frac{\mathbf{1}_{[T=t]}\mathbf{1}_{[Z=z]}}{\widehat{\pi}^t(\mathbf{X})\widehat{\pi}^g(Z\|\mathbf{X})}\left[\frac{(Y-\widehat{Q}^-(t,Z,\mathbf{X}))_+-\widehat{\gamma}_u^-(t,Z,\mathbf{X})}{b^+(Z,\mathbf{X})} - \frac{(\widehat{Q}^-(t,Z,\mathbf{X})-Y)_+-\widehat{\gamma}_l^-(t,Z,\mathbf{X})}{b^-(Z,\mathbf{X})}\right]$
Pseudo-outcomes (continuous)	$\phi_{t,z}^+(S;\widehat{\eta}) = \widehat{Q}^+(t,z,\mathbf{X}) + \frac{\widehat{\gamma}_u^+(t,z,\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\widehat{\gamma}_l^+(t,z,\mathbf{X})}{b^+(z,\mathbf{X})}$ $+\frac{\mathbf{1}_{[T=t]}K_h(Z-z)}{\widehat{\pi}^t(\mathbf{X})\widehat{\pi}^g(Z\|\mathbf{X})}\left[\frac{(Y-\widehat{Q}^+(t,Z,\mathbf{X}))_+-\widehat{\gamma}_u^+(t,Z,\mathbf{X})}{b^-(Z,\mathbf{X})} - \frac{(\widehat{Q}^+(t,Z,\mathbf{X})-Y)_+-\widehat{\gamma}_l^+(t,Z,\mathbf{X})}{b^+(Z,\mathbf{X})}\right]$ $\phi_{t,z}^-(S;\widehat{\eta}) = \widehat{Q}^-(t,z,\mathbf{X}) + \frac{\widehat{\gamma}_u^-(t,z,\mathbf{X})}{b^+(z,\mathbf{X})} - \frac{\widehat{\gamma}_l^-(t,z,\mathbf{X})}{b^-(z,\mathbf{X})}$ $+\frac{\mathbf{1}_{[T=t]}K_h(Z-z)}{\widehat{\pi}^t(\mathbf{X})\widehat{\pi}^g(Z\|\mathbf{X})}\left[\frac{(Y-\widehat{Q}^-(t,Z,\mathbf{X}))_+-\widehat{\gamma}_u^-(t,Z,\mathbf{X})}{b^+(Z,\mathbf{X})} - \frac{(\widehat{Q}^-(t,Z,\mathbf{X})-Y)_+-\widehat{\gamma}_l^-(t,Z,\mathbf{X})}{b^-(Z,\mathbf{X})}\right]$

Table 1. Summary of our bounds from the main paper.

C.2. Orthogonal lower bound

In Section 4, we provided an orthogonal estimation framework for the upper bound of the potential outcomes and treatment effects. For completeness, we now also provide the formulation for the lower bounds:

Let $S = (\mathbf{X}, Y, T, Z)$. Fix (t, z). Define the localization weight

$$\omega_{z,h}(Z) := \begin{cases} \mathbf{1}_{[Z=z]}, & \text{if } Z \text{ binary/discrete,} \\ K_h(Z-z), & \text{if } Z \text{ continuous,} \end{cases}$$

and let $\pi^g(Z \mid \mathbf{X})$ denote the conditional probability mass function (discrete Z) or density (continuous Z). Let $\widehat{\eta} = (\widehat{\pi}^t, \widehat{\pi}^g, \widehat{Q}^-, \widehat{\gamma}_u^-, \widehat{\gamma}_l^-)$ be a set of estimated nuisances. Then, an orthogonal pseudo-outcome for the CAPO lower bound $\mu^-(t,z,\mathbf{x})$ is:

$$\begin{aligned}\phi_{t,z}^-(S;\widehat{\eta}) =& \widehat{Q}^-(t,z,\mathbf{X}) + \frac{\widehat{\gamma}_u^-(t,z,\mathbf{X})}{b^+(z,\mathbf{X})} - \frac{\widehat{\gamma}_l^-(t,z,\mathbf{X})}{b^-(z,\mathbf{X})} \\ &+ \frac{\mathbf{1}_{[T=t]}\,\omega_{z,h}(Z)}{\widehat{\pi}^t(\mathbf{X})\,\widehat{\pi}^g(Z \mid \mathbf{X})}\left[\frac{(Y-\widehat{Q}^-(t,Z,\mathbf{X}))_+ - \widehat{\gamma}_u^-(t,Z,\mathbf{X})}{b^+(Z,\mathbf{X})} - \frac{(\widehat{Q}^-(t,Z,\mathbf{X})-Y)_+ - \widehat{\gamma}_l^-(t,Z,\mathbf{X})}{b^-(Z,\mathbf{X})}\right],\end{aligned} \tag{26}$$

where $\gamma_u^-(t,z,\mathbf{x}) := \mathbb{E}[(Y - Q^-(\cdot))_+ \mid t,z,\mathbf{x}]$ and $\gamma_l^-(t,z,\mathbf{x}) := \mathbb{E}[(Q^-(\cdot) - Y)_+ \mid t,z,\mathbf{x}]$.

C.3. Bounds on the treatment effects

Recall the definition of the average and individual direct effects

$$\tau_d^{(t,z),(t',z)} := \psi(t,z) - \psi(t',z) \text{ (ADE)} \quad \text{and} \quad \tau_{d_i}^{(t,z),(t',z)}(\mathbf{x}_i) := \mu(t,z,\mathbf{x}_i) - \mu(t',z,\mathbf{x}_i) \text{ (IDE)},$$

spillover/indirect effects

$$\tau_s^{(t,z),(t,z')} := \psi(t,z) - \psi(t,z') \text{ (ASE)} \quad \text{and} \quad \tau_{s_i}^{(t,z),(t,z')}(\mathbf{x}_i) := \mu(t,z,\mathbf{x}_i) - \mu(t,z',\mathbf{x}_i) \text{ (ISE)},$$

and overall effects

$$\tau_o^{(t,z),(t^{'},z^{'})} := \psi(t,z) - \psi(t^{'},z^{'}) \text{ (AOE)} \quad \text{and} \quad \tau_{o_i}^{(t,z),(t^{'},z^{'})}(\mathbf{x}_i) := \mu(t,z,\mathbf{x}_i) - \mu(t^{'},z^{'},\mathbf{x}_i) \text{ (IOE)}.$$

Based on the CAPO bounds $\mu^{\pm}(t,z,\mathbf{x})$ from Theorem 4.2, we thus obtain the treatment effects through the general formula $\tau^{+\,(a,b)} = f^{+}(a,\cdot) - f^{-}(b,\cdot)$ and $\tau^{-\,(a,b)} = f^{-}(a,\cdot) - f^{+}(b,\cdot)$, where with a slight abuse of notation τ refers to any of the (conditional) effects above, f refers to either μ or ψ, and (a,b) denotes the change in t and/or z. Specifically, the conditional treatment effects IDE / ISE / IOE are identified as follows.

Direct effect:

$$\tau_{d_i}^{+\,(t,z),(t^{'},z)}(\mathbf{x}) = Q^{+}(t,z,\mathbf{x}) - Q^{-}(t^{'},z,\mathbf{x}) + \frac{1}{b^{-}(z,\mathbf{x})}(\gamma_u^{+}(t,z,\mathbf{x}) - \gamma_l^{-}(t^{'},z,\mathbf{x})) \tag{27}$$

$$- \frac{1}{b^{+}(z,\mathbf{x})}(\gamma_l^{+}(t,z,\mathbf{x}) - \gamma_u^{-}(t^{'},z,\mathbf{x})) \tag{28}$$

$$\tau_{d_i}^{-\,(t,z),(t^{'},z)}(\mathbf{x}) = Q^{-}(t,z,\mathbf{x}) - Q^{+}(t^{'},z,\mathbf{x}) + \frac{1}{b^{+}(z,\mathbf{x})}(\gamma_u^{-}(t,z,\mathbf{x}) - \gamma_l^{+}(t^{'},z,\mathbf{x})) \tag{29}$$

$$- \frac{1}{b^{-}(z,\mathbf{x})}(\gamma_l^{-}(t,z,\mathbf{x}) - \gamma_u^{+}(t^{'},z,\mathbf{x})) \tag{30}$$

Indirect/spillover effect:

$$\tau_{d_s}^{+\,(t,z),(t,z^{'})}(\mathbf{x}) = Q^{+}(t,z,\mathbf{x}) - Q^{-}(t,z^{'},\mathbf{x}) + \frac{1}{b^{-}(z,\mathbf{x})}\gamma_u^{+}(t,z,\mathbf{x}) \tag{31}$$

$$+ \frac{1}{b^{+}(z^{'},\mathbf{x})}\gamma_u^{-}(t,z^{'},\mathbf{x}) - \frac{1}{b^{-}(z^{'},\mathbf{x})}\gamma_l^{-}(t,z^{'},\mathbf{x}) - \frac{1}{b^{+}(z,\mathbf{x})}\gamma_l^{+}(t,z,\mathbf{x}) \tag{32}$$

$$\tau_{d_s}^{-\,(t,z),(t,z^{'})}(\mathbf{x}) = Q^{-}(t,z,\mathbf{x}) - Q^{+}(t,z^{'},\mathbf{x}) + \frac{1}{b^{+}(z,\mathbf{x})}\gamma_u^{-}(t,z,\mathbf{x}) \tag{33}$$

$$+ \frac{1}{b^{-}(z^{'},\mathbf{x})}\gamma_u^{+}(t,z^{'},\mathbf{x}) - \frac{1}{b^{+}(z^{'},\mathbf{x})}\gamma_l^{+}(t,z^{'},\mathbf{x}) - \frac{1}{b^{-}(z,\mathbf{x})}\gamma_l^{-}(t,z,\mathbf{x}) \tag{34}$$

Overall effect:

$$\tau_{d_o}^{+\,(t,z),(t^{'},z^{'})}(\mathbf{x}) = Q^{+}(t,z,\mathbf{x}) - Q^{-}(t^{'},z^{'},\mathbf{x}) + \frac{1}{b^{-}(z,\mathbf{x})}\gamma_u^{+}(t,z,\mathbf{x}) \tag{35}$$

$$+ \frac{1}{b^{+}(z^{'},\mathbf{x})}\gamma_u^{-}(t^{'},z^{'},\mathbf{x}) - \frac{1}{b^{-}(z^{'},\mathbf{x})}\gamma_l^{-}(t^{'},z^{'},\mathbf{x}) - \frac{1}{b^{+}(z,\mathbf{x})}\gamma_l^{+}(t,z,\mathbf{x}) \tag{36}$$

$$\tau_{d_o}^{-\,(t,z),(t^{'},z^{'})}(\mathbf{x}) = Q^{-}(t,z,\mathbf{x}) - Q^{+}(t^{'},z^{'},\mathbf{x}) + \frac{1}{b^{+}(z,\mathbf{x})}\gamma_u^{-}(t,z,\mathbf{x}) \tag{37}$$

$$+ \frac{1}{b^{-}(z^{'},\mathbf{x})}\gamma_u^{+}(t^{'},z^{'},\mathbf{x}) - \frac{1}{b^{+}(z^{'},\mathbf{x})}\gamma_l^{+}(t^{'},z^{'},\mathbf{x}) - \frac{1}{b^{-}(z,\mathbf{x})}\gamma_l^{-}(t,z,\mathbf{x}) \tag{38}$$

The bounds on the average effects ADE, AIE, and AOE are then identified by the expectation of the individual effects over the covariates $\mathbf{X}$.

C.4. Continuous neighborhood exposure

This section gives the continuous-Z analogues of Theorem 4.7 and Corollary 4.10, as well as the corresponding *sharpness* and *validity* guarantees for the estimated bounds (complementary to the the discrete-Z results in the main text).

When Z is continuous, point evaluation at $Z = z$ is non-regular. Following the standard approach in orthogonal learning for continuous exposures, we therefore target a *kernel-localized* (bandwidth-indexed) version of the bound functional. Under smoothness in z, these localized targets converge to the original (pointwise) bounds as $h \downarrow 0$, at the usual bias–variance tradeoff governed by (n,h).

Assumption C.1 (Kernel localization)**.** Let Z be continuous and let $K_h(u) = \frac{1}{h}K(u/h)$, where K is bounded, integrates to 1, and $\int K(u)^2\,du < \infty$. Let $h = h_n \downarrow 0$ with $nh_n \to \infty$.

Kernel-localized targets. Fix (t, z) and let $h > 0$. For continuous Z, define the localized selection weight

$$\kappa_{t,z,h}(S) := \frac{\mathbf{1}_{[T=t]}\, K_h(Z - z)}{\pi^t(\mathbf{X})\, \pi^g(Z \mid \mathbf{X})}, \tag{39}$$

as in the continuous-Z modification of the proof of Theorem 4.4. Define the kernel-localized pseudo-outcome $\phi^+_{t,z,h}(S; \widehat{\eta})$ as Eq. (20) with $\omega_{z,h}(Z) = K_h(Z - z)$.

When $\widehat{\eta} = \eta$, define the associated bandwidth-indexed functionals by

$$\mu^+_h(t, z, \mathbf{x}) := \mathbb{E}\Big[\phi^+_{t,z,h}(S; \eta) \mid \mathbf{X} = \mathbf{x}\Big], \qquad \psi^+_h(t, z) := \mathbb{E}\Big[\phi^+_{t,z,h}(S; \eta)\Big]. \tag{40}$$

Under standard smoothness in z, $\mu^+_h(t, z, \mathbf{x}) \to \mu^+(t, z, \mathbf{x})$ and $\psi^+_h(t, z) \to \psi^+(t, z)$ as $h \downarrow 0$ (see Remark 4.5).

Relative to the discrete-Z case, kernel localization inflates the second-order remainder by a factor $h^{-1/2}$ (reflecting $\int K_h^2 = O(1/h)$). This propagates to the final-stage CAPO rate and yields the usual $\sqrt{nh}$ scaling for the (smoothed) APO.

Theorem C.2 (Second-order nuisance error (continuous Z))**.** *Assume Z is continuous and Assumptions 4.6 and C.1 hold. Let $\widehat{\eta} = (\widehat{\pi}^t, \widehat{\pi}^g, \widehat{Q}^+, \widehat{\gamma}^+_u, \widehat{\gamma}^+_l)$ be the cross-fitted nuisances used in $\phi^+_{t,z,h}(S; \widehat{\eta})$ (Eq. (20) with $\omega_{z,h}(Z) = K_h(Z - z)$).*

Define nuisance error rates (in L_2 norms over the appropriate arguments) by

$$r_{n,\pi} := \|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2, \qquad r_{n,Q} := \|\widehat{Q}^+ - Q^+\|_2, \tag{41}$$

$$r_{n,\gamma} := \|\widehat{\gamma}^+_u - \gamma_u(\widehat{Q}^+; \cdot)\|_2 + \|\widehat{\gamma}^+_l - \gamma_l(\widehat{Q}^+; \cdot)\|_2, \tag{42}$$

where the norms are taken over the random variables that the corresponding nuisance is evaluated on (e.g., $(Z, \mathbf{X})$ for $\pi^g(Z \mid \mathbf{X})$, $Q^+(t, Z, \mathbf{X})$, and $\gamma^\pm(t, Z, \mathbf{X})$).

Then, the conditional bias induced by nuisance estimation satisfies

$$\Big\|\mathbb{E}\Big[\phi^+_{t,z,h}(S; \widehat{\eta}) - \phi^+_{t,z,h}(S; \eta) \mid \mathbf{X}\Big]\Big\|_2 = O_p\left(\frac{r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}}{\sqrt{h}}\right). \tag{43}$$

Corollary C.3 (Quasi-oracle rates and inference (continuous Z))**.** *Assume the conditions of Theorem C.2 and that the second-stage regression learner $\widehat{\mathbb{E}}_n[\cdot \mid \mathbf{X} = \mathbf{x}]$ satisfies Assumption 4.8 with rate δ_n when regressing $\phi^+_{t,z,h}(S; \eta)$ on $\mathbf{X}$.*

Then:

CAPO rates: The CAPO upper-bound estimator satisfies

$$\|\widehat{\mu}^+_h(t, z, \cdot) - \mu^+_h(t, z, \cdot)\|_2 = O_p\left(\delta_n + \frac{r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}}{\sqrt{h}}\right). \tag{44}$$

APO rates: The APO upper-bound estimator $\widehat{\psi}^+(t, z) = \mathbb{E}_n[\widehat{\phi}^+_{t,z,h}]$ satisfies

$$|\widehat{\psi}^+_h(t, z) - \psi^+_h(t, z)| = O_p\left(\frac{1}{\sqrt{nh}} + \frac{r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}}{\sqrt{h}}\right). \tag{45}$$

$\sqrt{nh}$-CLT (central limit theorem) for the (smoothed) APO. If $r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q} = o_p(n^{-1/2})$, then

$$\sqrt{nh}\left(\widehat{\psi}^+_h(t, z) - \psi^+_h(t, z)\right) \rightsquigarrow \mathcal{N}\big(0,\ V^+_h(t, z)\big), \tag{46}$$

where one valid asymptotic variance target is $V^+_h(t, z) := \mathrm{Var}\big(\sqrt{h}\, \phi^+_{t,z,h}(S; \eta)\big)$.

Finally, if the smoothing bias satisfies $|\psi^+_h(t, z) - \psi^+(t, z)| = o((nh)^{-1/2})$ (e.g., via undersmoothing under z-smoothness), then the same CLT holds with $\psi^+(t, z)$ in place of $\psi^+_h(t, z)$.

Sharpness and validity of the estimated bounds. The previous results control the second-order remainder and deliver quasi-oracle rates for the localized targets (μ_h^+, ψ_h^+). We now record the two complementary guarantees from the main text in their continuous-Z versions: (i) consistency for the *sharp* identified bounds, and (ii) *validity* of the resulting intervals under potentially misspecified cutoffs. As before, the statements hold for both endpoints $(+/-)$; we write them for the upper endpoint for brevity, with the lower endpoint following analogously by sign-swapping in the pseudo-outcome.

Proposition C.4 (Consistency for sharp bounds (continuous Z))**.** *Assume the conditions of Corollary C.3 and consider the corresponding lower-bound estimator $\widehat{\mu}_h^-(t, z, \cdot)$ constructed from the lower-bound pseudo-outcome (defined analogously to Eq. (20)). Suppose $\delta_n = o_p(1)$ and*

$$\frac{r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2}{\sqrt{h}} = o_p(1). \tag{47}$$

Then,

$$\|\widehat{\mu}_h^{\pm}(t, z, \cdot) - \mu_h^{\pm}(t, z, \cdot)\|_2 = o_p(1), \qquad |\widehat{\psi}_h^{\pm}(t, z) - \psi_h^{\pm}(t, z)| = o_p(1). \tag{48}$$

Consequently, the estimated CAPO and APO intervals converge to the sharp *kernel-localized identified intervals for the bandwidth-indexed targets.*

Moreover, if the smoothing bias vanishes at the appropriate rate (e.g., $|\psi_h^{\pm}(t, z) - \psi^{\pm}(t, z)| = o((nh)^{-1/2})$), then the estimated intervals are asymptotically sharp for the original pointwise bounds as $h \downarrow 0$.

Corollary C.5 (Asymptotic validity under misspecified cutoffs (continuous Z))**.** *Fix measurable cutoffs $\overline{Q}^{\pm}(t, z, \mathbf{x})$ (not necessarily equal to the sharp cut-offs) and let $\overline{\mu}_h^{\pm}(t, z, \mathbf{x}; \overline{Q}^{\pm})$ and $\overline{\psi}_h^{\pm}(t, z; \overline{Q}^{\pm})$ denote the resulting (possibly non-sharp) kernel-localized bound functionals induced by these cutoffs (i.e., the targets obtained by replacing $Q^{\pm}$ in the pseudo-outcomes and taking the conditional/unconditional expectations as in Eq. (40)). Then, the induced intervals*

$$\left[\overline{\mu}_h^-(t, z, \mathbf{x}; \overline{Q}^-),\ \overline{\mu}_h^+(t, z, \mathbf{x}; \overline{Q}^+)\right] \quad \text{and} \quad \left[\overline{\psi}_h^-(t, z; \overline{Q}^-),\ \overline{\psi}_h^+(t, z; \overline{Q}^+)\right] \tag{49}$$

are (not necessarily sharp) valid CAPO and APO intervals for the kernel-localized targets.

Moreover, if $\widehat{Q}^{\pm} \to \overline{Q}^{\pm}$ in L_2 and either

(i) $(\widehat{\pi}^t, \widehat{\pi}^g)$ is consistent, or

(ii) the corresponding tail-moment regressions $(\widehat{\gamma}_u^{\pm}, \widehat{\gamma}_l^{\pm})$ are consistent for the targets induced by $\overline{Q}^{\pm}$,

then the estimated endpoints converge to the induced (conservative) targets and the resulting (C)APO intervals remain asymptotically valid, though potentially conservative. If $\overline{Q}^{\pm}$ equals the sharp cut-offs, then the induced bounds coincide with the sharp bounds, and the intervals are asymptotically sharp as well.

Conclusion. For continuous neighborhood exposure, our estimation and theory proceed exactly as in the discrete-Z case, except that (i) the indicator $\mathbf{1}_{[Z=z]}$ in the selection weight is replaced by kernel localization $K_h(Z - z)$ and (ii) the conditional pmf $\pi^g(z \mid \mathbf{X})$ is replaced by the conditional density $\pi^g(Z \mid \mathbf{X})$. This replacement yields an effective sample size nh around z, which inflates the second-order remainder by a factor $h^{-1/2}$ and leads to $\sqrt{nh}$ scaling for APO inference. Under smoothness in z, the bandwidth-indexed targets $(\mu_h^{\pm}, \psi_h^{\pm})$ converge to the point-wise bounds $(\mu^{\pm}, \psi^{\pm})$ as $h \downarrow 0$, yielding the usual bias-variance tradeoff in (n, h). The proofs in Supplement D show that all continuous-Z results follow from the discrete-Z proofs by replacing $\mathbf{1}_{[Z=z]}$ by $K_h(Z - z)$ and tracking $\int K_h^2 = O(1/h)$.

D. Proofs

D.1. Justification of the setting-specific b^+, b^-

Below we give a justification for the specification of b^+, b^- for exposure mappings (1) and (2).

(1) **Weighted mean exposure:** Define $g(t_{\mathcal{N}}) := \sum_{j\in\mathcal{N}} \frac{t_j}{n} = \frac{N_T}{n}$, where N_T denotes the number of treated neighbors and n denotes the neighborhood size. We assume $g^*(t_{\mathcal{N}}) = \sum_{j\in\mathcal{N}} w_j t_j$, where $|\frac{1}{n} - w_j| \geq \varepsilon$ for all j.

First observe that

$$b^-(z,\mathbf{x}) \leq \frac{P(\sum_{j\in\mathcal{N}} w_j t_j = z \mid \mathbf{x})}{P(\sum_{j\in\mathcal{N}} \frac{t_j}{n} = z \mid \mathbf{x})} \leq b^+(z,\mathbf{x}) \iff b^-(z,\mathbf{x}) \leq \frac{G(z \mid \mathbf{x}) - G(s \mid \mathbf{x})}{F(z \mid \mathbf{x}) - F(s \mid \mathbf{x})} \leq b^+(z,\mathbf{x}) \tag{50}$$

for all $s \in \mathcal{Z}$, where $G(\cdot)$ and $F(\cdot)$ denote the conditional cumulative distribution function of $g^*(T_{\mathcal{N}})$ and $g(T_{\mathcal{N}})$. Since $|\frac{1}{n} - w_j| \geq \varepsilon$, it holds that, for all $k \in \mathcal{Z}$, we have

$$P\left((\frac{1}{n}+\varepsilon)\sum_{j\in\mathcal{N}} t_j \leq k \mid \mathbf{x}\right) \leq P\left(\sum_{j\in\mathcal{N}} w_j t_j \leq k \mid \mathbf{x}\right) \leq P\left((\frac{1}{n}-\varepsilon)\sum_{j\in\mathcal{N}} t_j \leq k \mid \mathbf{x}\right) \tag{51}$$

$$\iff P\left(\sum_{j\in\mathcal{N}} \frac{t_j}{n} \leq \frac{k}{1+n\varepsilon} \mid \mathbf{x}\right) \leq P\left(\sum_{j\in\mathcal{N}} w_j t_j \leq k \mid \mathbf{x}\right) \leq P\left(\sum_{j\in\mathcal{N}} \frac{t_j}{n} \leq \frac{k}{1-n\varepsilon} \mid \mathbf{x}\right). \tag{52}$$

Therefore, we can bound the enumerator $G(z \mid \mathbf{x}) - G(s \mid \mathbf{x})$ by

$$P\left(\sum_{j\in\mathcal{N}} \frac{t_j}{n} \leq \frac{z}{1+n\varepsilon} \mid \mathbf{x}\right) - P\left(\sum_{j\in\mathcal{N}} \frac{t_j}{n} \leq \frac{s}{1-n\varepsilon} \mid \mathbf{x}\right) \leq G(z \mid \mathbf{x}) - G(s \mid \mathbf{x}) \tag{53}$$

$$\leq P\left(\sum_{j\in\mathcal{N}} \frac{t_j}{n} \leq \frac{z}{1-n\varepsilon} \mid \mathbf{x}\right) - P\left(\sum_{j\in\mathcal{N}} \frac{t_j}{n} \leq \frac{s}{1+n\varepsilon} \mid \mathbf{x}\right). \tag{54}$$

Then, it follows that

$$b^-(z,\mathbf{x}) = \inf_{s\in\mathcal{Z}} \frac{P(\frac{ns}{1-\varepsilon n} \leq N_T \leq \frac{nz}{1+\varepsilon n} \mid \mathbf{x})}{P(ns \leq N_T \leq nz \mid \mathbf{x})}, \qquad b^+(z,\mathbf{x}) = \sup_{s\in\mathcal{Z}} \frac{P(\frac{ns}{1+\varepsilon n} \leq N_T \leq \frac{nz}{1-\varepsilon n} \mid \mathbf{x})}{P(ns \leq N_T \leq nz \mid \mathbf{x})}. \tag{55}$$

(2) **Thresholding function:**

Let $h(t_{\mathcal{N}}) := \sum_{j\in\mathcal{N}} \frac{t_j}{n}$ and assume the exposure mapping is specified through a threshold as $g(t_{\mathcal{N}}) = f(h(t_{\mathcal{N}})) := \mathbf{1}_{[h(t_{\mathcal{N}})\geq c]}$, i.e., $P(g(t_{\mathcal{N}}) = 1 \mid \mathbf{x}) = P(N_T \geq nc \mid \mathbf{x})$, where N_T denotes the number of treated neighbors. We allow the true threshold c^* to differ by an amount $\varepsilon \in [0, \min\{c, 1-c\}]$ from c, i.e., $c^* \in [c \pm \varepsilon]$. Thus, $P(g^*(t_{\mathcal{N}}) = 1 \mid \mathbf{x}) = P(N_T \geq nc^* \mid \mathbf{x})$, and, therefore, by straightforward computation, we yield

$$\frac{P(N_T \geq n(c+\varepsilon) \mid \mathbf{x})}{P(N_T \geq nc \mid \mathbf{x})} \leq \frac{P(g^*(t_{\mathcal{N}}) = 1 \mid \mathbf{x})}{P(g(t_{\mathcal{N}}) = 1 \mid \mathbf{x})} \leq \frac{P(N_T \geq n(c-\varepsilon) \mid \mathbf{x})}{P(N_T \geq nc \mid \mathbf{x})}, \tag{56}$$

and

$$\frac{1 - P(N_T \geq n(c-\varepsilon) \mid \mathbf{x})}{1 - P(N_T \geq nc \mid \mathbf{x})} \leq \frac{P(g^*(t_{\mathcal{N}}) = 0 \mid \mathbf{x})}{P(g(t_{\mathcal{N}}) = 0 \mid \mathbf{x})} \leq \frac{1 - P(N_T \geq n(c+\varepsilon) \mid \mathbf{x})}{1 - P(N_T \geq nc \mid \mathbf{x})}. \tag{57}$$

D.2. Auxiliary theory

Our bounds employ a sensitivity method proposed in Frauen et al. (2023). However, the original contribution proposes bounds in the presence of unobserved confounding, whereas we are targeting a different setting. Below, we present Theorem 1 in (Frauen et al., 2023) adapted to our setting.

Theorem D.1. *Let $b^-(z,\mathbf{x}) \leq b^+(z,\mathbf{x})$ with $b^-(z,\mathbf{x}) \in (0,1]$ and $b^+(z,\mathbf{x}) \in [1,\infty)$, such that for all $z, \mathbf{x}$*

$$b^-(z,\mathbf{x}) \leq \frac{p(g^*(t_{\mathcal{N}}) = z \mid \mathbf{x})}{p(g(t_{\mathcal{N}}) = z \mid \mathbf{x})} \leq b^+(z,\mathbf{x}) \tag{58}$$

and define $\alpha^{\pm}(z,\mathbf{x}) := \frac{(1-b^{\mp}(z,\mathbf{x}))b^{\pm}(z,\mathbf{x})}{b^{\pm}(z,\mathbf{x})-b^{\mp}(z,\mathbf{x})}$. Furthermore, let $F_Y(y) := F_Y(y \mid t,z,\mathbf{x})$ denote the conditional cumulative distribution function (CDF) of Y. For $Y \in \mathbb{R}$ continuous, we define

$$p^+(y \mid t,z,\mathbf{x}) = \begin{cases} \frac{1}{b^+(z,\mathbf{x})} p(y \mid t,z,\mathbf{x}), \text{ if } F(y) \leq \alpha^+(z,\mathbf{x}), \\ \frac{1}{b^-(z,\mathbf{x})} p(y \mid t,z,\mathbf{x}), \text{ if } F(y) > \alpha^+(z,\mathbf{x}), \end{cases} \tag{59}$$

and for $Y \in \mathbb{R}$ discrete, we define the probability mass function

$$P^+(y \mid t,z,\mathbf{x}) = \begin{cases} \frac{1}{b^+(z,\mathbf{x})} P(y \mid t,z,\mathbf{x}), & \text{if } F(y) < \alpha^+(z,\mathbf{x}), \\ \frac{1}{b^-(z,\mathbf{x})} P(y \mid t,z,\mathbf{x}), & \text{if } F(y-1) > \alpha^+(z,\mathbf{x}), \\ \frac{1}{b^+(z,\mathbf{x})}(\alpha^+(z,\mathbf{x}) - F(y-1)) + \frac{1}{b^-(z,\mathbf{x})}(F(y) - \alpha^+(z,\mathbf{x})), & \text{otherwise.} \end{cases} \tag{60}$$

The lower bound $p^-(y \mid t,z,\mathbf{x})$ is defined through exchanging the signs in α and b. Let $F^{\pm}(y)$ denote the conditional CDF with regard to $p^{\pm}(y \mid t,z,\mathbf{x})$. Then, for all $y \in \mathcal{Y}$

$$F^+(y) \leq \inf_{\tilde{P} \in \mathcal{M}} F_{\tilde{P}}(y), F^-(y) \geq \inf_{\tilde{P} \in \mathcal{M}} F_{\tilde{P}}(y), \tag{61}$$

i.e., the bounds are valid, and

$$F^+(y) = \inf_{\tilde{P} \in \mathcal{M}} F_{\tilde{P}}(y), F^-(y) = \inf_{\tilde{P} \in \mathcal{M}} F_{\tilde{P}}(y), \tag{62}$$

i.e., the bounds are sharp, if Z is continuous or if Z is discrete and $\frac{1}{b^+(z,\mathbf{x})} \geq \pi^g(z \mid \mathbf{x})$.

D.3. Proof of Theorem 4.2

Theorem 4.2. *Let $Q^{\pm}(t,z,\mathbf{x})$ be defined as in Eq. (18) and let $(u)_+ = \max\{u,0\}$. The sharp CAPO upper and lower bounds are given by*

$$\begin{aligned} \mu^{\pm}(t,z,\mathbf{x}) = & Q^{\pm}(t,z,\mathbf{x}) \\ & + \frac{1}{b^{\mp}(z,\mathbf{x})} \mathbb{E}\big[(Y - Q^{\pm}(t,z,\mathbf{x}))_+ \mid t,z,\mathbf{x}\big] \\ & - \frac{1}{b^{\pm}(z,\mathbf{x})} \mathbb{E}\big[(Q^{\pm}(t,z,\mathbf{x}) - Y)_+ \mid t,z,\mathbf{x}\big]. \end{aligned} \tag{19}$$

Proof. Throughout the proof we focus on the upper bound for continuous outcomes. The other cases follow analogously. Recall the definition of

$$Q^{\pm}(t,z,\mathbf{x}) := \inf\left\{ y \mid F_Y(y \mid t,z,\mathbf{x}) \geq \frac{(1-b^{\mp}(z,\mathbf{x}))b^{\pm}(z,\mathbf{x})}{b^{\pm}(z,\mathbf{x}) - b^{\mp}(z,\mathbf{x})} \right\}, \tag{63}$$

when $b^-(z,\mathbf{x}) < 1 < b^+(z,\mathbf{x})$, and $Q^{\pm}(t,z,\mathbf{x}) = Q(t,z,\mathbf{x}) := \inf\{y \mid F_Y(y \mid t,z,\mathbf{x}) \geq \frac{1}{2}\}$ otherwise.

By applying Theorem D.1, the sharp upper and lower bounds on the conditional potential outcome $\mu(t,z,\mathbf{x})$ are given by

$$\mu^{\pm}(t,z,\mathbf{x}) = \frac{1}{b^{\pm}(z,\mathbf{x})} \int_{-\infty}^{Q^{\pm}(z,\mathbf{x})} y \, \mathrm{d}\mu + \frac{1}{b^{\mp}(z,\mathbf{x})} \int_{Q^{\pm}(z,\mathbf{x})}^{\infty} y \, \mathrm{d}\mu \tag{64}$$

$$= \frac{1}{b^{\pm}(z,\mathbf{x})} \cdot \alpha^{\pm} \mathrm{LCTE}_{\alpha}^{\pm}(t,z,\mathbf{x}) + \frac{1}{b^{\mp}(z,\mathbf{x})} \cdot (1-\alpha^{\pm}) \mathrm{CVaR}_{\alpha}^{\pm}(t,z,\mathbf{x}) \tag{65}$$

where we define $\alpha^{\pm} := \frac{(1-b^{\mp}(z,\mathbf{x}))b^{\pm}(z,\mathbf{x})}{b^{\pm}(z,\mathbf{x})-b^{\mp}(z,\mathbf{x})}$. Here, the CVaR$^{\pm}$ denotes the *conditional value at risk* at level $\alpha^{\pm}$ with corresponding quantiles $Q^{+}(t,z,\mathbf{x})/Q^{-}(t,z,\mathbf{x})$ defined as

$$\text{CVaR}^{+}_{\alpha}(t,z,\mathbf{x}) := \min_{q\in\mathbb{R}}\Big\{ q + \frac{1}{1-\alpha^{+}}\,\mathbb{E}\big[\,(Y-q)_{+} \mid t,z,\mathbf{x}\big]\Big\} = Q^{+}(t,z,\mathbf{x}) + \frac{b^{-}-b^{+}}{(1-b^{+})b^{-}}\mathbb{E}\big[(Y-Q^{+}(t,z,\mathbf{x}))_{+} \mid t,z,\mathbf{x}\big], \quad (66)$$

$$\text{CVaR}^{-}_{\alpha}(t,z,\mathbf{x}) := \min_{q\in\mathbb{R}}\Big\{ q + \frac{1}{1-\alpha^{-}}\,\mathbb{E}\big[\,(Y-q)_{+} \mid t,z,\mathbf{x}\big]\Big\} = Q^{-}(t,z,\mathbf{x}) + \frac{b^{+}-b^{-}}{(1-b^{-})b^{+}}\mathbb{E}\big[(Y-Q^{-}(t,z,\mathbf{x}))_{+} \mid t,z,\mathbf{x}\big] \quad (67)$$

where $(u)_{+} = \max\{u,0\}$, and LCTE$^{\pm}$ the *lower conditional tail expectation* at level $\alpha^{\pm}$ with corresponding quantiles $Q^{+}(t,z,\mathbf{x}))/Q^{-}(t,z,\mathbf{x}))$ defined as

$$\begin{aligned}\text{LCTE}^{+}_{\alpha}(t,z,\mathbf{x}) :=& \sup_{q\in\mathbb{R}}\Big\{ q \;-\; \frac{1}{\alpha^{+}}\,\mathbb{E}\big[(\,q-Y\,)_{+} \mid t,z,\mathbf{x}\big]\Big\} \\ =& Q^{+}(t,z,\mathbf{x})) - \frac{b^{+}(z,\mathbf{x})-b^{-}(z,\mathbf{x})}{(1-b^{-}(z,\mathbf{x}))b^{+}(z,\mathbf{x})}\mathbb{E}\big[(Q^{+}(t,z,\mathbf{x}))-Y)_{+} \mid t,z,\mathbf{x}\big],\end{aligned} \quad (68)$$

$$\begin{aligned}\text{LCTE}^{-}_{\alpha}(t,z,\mathbf{x}) :=& \sup_{q\in\mathbb{R}}\Big\{ q \;-\; \frac{1}{\alpha^{-}}\,\mathbb{E}\big[(\,q-Y\,)_{+} \mid t,z,\mathbf{x}\big]\Big\} \\ =& Q^{-}(t,z,\mathbf{x})) - \frac{b^{-}(z,\mathbf{x})-b^{+}(z,\mathbf{x})}{(1-b^{+}(z,\mathbf{x}))b^{-}(z,\mathbf{x})}\mathbb{E}\big[(Q^{-}(t,z,\mathbf{x}))-Y)_{+} \mid t,z,\mathbf{x}\big].\end{aligned} \quad (69)$$

With these reformulations of CVaR and LCTE then follows the desired result

$$\mu^{\pm}(t,z,\mathbf{x}) = Q^{\pm}(t,z,\mathbf{x}) + \frac{1}{b^{\mp}(z,\mathbf{x})}\mathbb{E}\big[(Y-Q^{\pm}(t,z,\mathbf{x}))_{+} \mid t,z,\mathbf{x}\big] - \frac{1}{b^{\pm}(z,\mathbf{x})}\mathbb{E}\big[(Q^{\pm}(t,z,\mathbf{x})-Y)_{+} \mid t,z,\mathbf{x}\big]. \quad (70)$$

□

D.4. Proof of Theorem 4.4

Theorem 4.4. *Let $S=(\mathbf{X},Y,T,Z)$. Fix (t,z). Define*

$$\omega_{z,h}(Z) := \begin{cases} \mathbf{1}_{[Z=z]}, & \text{if } Z \text{ binary/discrete},\\ K_h(Z-z), & \text{if } Z \text{ continuous},\end{cases}$$

and let $\pi^{g}(Z \mid \mathbf{X})$ denote the conditional pmf (discrete Z) or density (continuous Z). Let $\widehat{\eta} = (\widehat{\pi}^{t}, \widehat{\pi}^{g}, \widehat{Q}^{+}, \widehat{\gamma}^{+}_{u}, \widehat{\gamma}^{+}_{l})$ be a set of estimated nuisances. Then, an orthogonal pseudo-outcome for the CAPO upper bound $\mu^{+}(t,z,\mathbf{x})$ is:

$$\begin{aligned}&\phi^{+}_{t,z}(S;\widehat{\eta}) = \\ &\widehat{Q}^{+}(t,z,\mathbf{X}) + \frac{\widehat{\gamma}^{+}_{u}(t,z,\mathbf{X})}{b^{-}(z,\mathbf{X})} - \frac{\widehat{\gamma}^{+}_{l}(t,z,\mathbf{X})}{b^{+}(z,\mathbf{X})} \\ &+ \frac{\mathbf{1}_{[T=t]}\,\omega_{z,h}(Z)}{\widehat{\pi}^{t}(\mathbf{X})\,\widehat{\pi}^{g}(Z \mid \mathbf{X})}\left[\frac{(Y-\widehat{Q}^{+}(t,Z,\mathbf{X}))_{+} - \widehat{\gamma}^{+}_{u}(t,Z,\mathbf{X})}{b^{-}(Z,\mathbf{X})}\right. \\ &\qquad\left. - \frac{(\widehat{Q}^{+}(t,Z,\mathbf{X})-Y)_{+} - \widehat{\gamma}^{+}_{l}(t,Z,\mathbf{X})}{b^{+}(Z,\mathbf{X})}\right].\end{aligned} \quad (20)$$

Moreover, when $\widehat{\eta}=\eta$, the pseudo-outcome is unbiased for its target bound functional (see Remark 4.5).

Proof. We begin with the case where Z is discrete (including binary), so $\omega_{z,h}(Z) = \mathbf{1}_{[Z=z]}$. We discuss the continuous-Z modification at the end.

Fix (t,z) and abbreviate

$$p(t,z \mid \mathbf{X}) := \pi^{t}(\mathbf{X})\,\pi^{g}(z \mid \mathbf{X}), \qquad \alpha := \alpha^{+}(z,\mathbf{X}), \qquad Q := Q^{+}(t,z,\mathbf{X}). \quad (71)$$

Define the (unnormalized) conditional tail moments

$$\gamma_u := \gamma_u^+(t,z,\mathbf{X}) := \mathbb{E}[(Y-Q)_+ \mid T=t, Z=z, \mathbf{X}], \qquad \gamma_l := \gamma_l^+(t,z,\mathbf{X}) := \mathbb{E}[(Q-Y)_+ \mid T=t, Z=z, \mathbf{X}]. \tag{72}$$

The sharp upper bound can be written as

$$\mu^+(t,z,\mathbf{X}) = Q + \frac{\gamma_u}{b^-(z,\mathbf{X})} - \frac{\gamma_l}{b^+(z,\mathbf{X})}, \tag{73}$$

which matches the first line of Eq. (20) when $\widehat{\eta} = \eta$.

Step 1: Reparameterization of μ^+ as a convex combination of CVaR/LCTE functionals. Define the upper-tail and lower-tail pseudo-outcomes at level α (see, e.g., Dorn et al. (2025); Oprescu et al. (2023))

$$H_u(y,q) := q + \frac{1}{1-\alpha}(y-q)_+, \qquad H_l(y,q) := q - \frac{1}{\alpha}(q-y)_+. \tag{74}$$

Their conditional expectations at the true quantile Q are the conditional upper CVaR and lower conditional tail expectation (LCTE), respectively:

$$\theta_u(\mathbf{X}) := \mathbb{E}[H_u(Y,Q) \mid T=t, Z=z, \mathbf{X}] = Q + \frac{1}{1-\alpha}\gamma_u, \qquad \theta_l(\mathbf{X}) := \mathbb{E}[H_l(Y,Q) \mid T=t, Z=z, \mathbf{X}] = Q - \frac{1}{\alpha}\gamma_l. \tag{75}$$

Now set the weights

$$w_u(\mathbf{X}) := \frac{1-\alpha}{b^-(z,\mathbf{X})}, \qquad w_l(\mathbf{X}) := \frac{\alpha}{b^+(z,\mathbf{X})}. \tag{76}$$

By the definition of $\alpha^+(z,\mathbf{X})$, one has

$$w_u(\mathbf{X}) + w_l(\mathbf{X}) = \frac{1-\alpha}{b^-(z,\mathbf{X})} + \frac{\alpha}{b^+(z,\mathbf{X})} = 1. \tag{77}$$

Therefore,

$$w_u(\mathbf{X})\theta_u(\mathbf{X}) + w_l(\mathbf{X})\theta_l(\mathbf{X}) = (w_u + w_l)Q + \frac{w_u}{1-\alpha}\gamma_u - \frac{w_l}{\alpha}\gamma_l = Q + \frac{1}{b^-}\gamma_u - \frac{1}{b^+}\gamma_l = \mu^+(t,z,\mathbf{X}), \tag{78}$$

so μ^+ is a (convex) linear combination of the two tail functionals.

Step 2: Recentered efficient influence function for μ^+. Our orthogonal pseudo-outcome is the recentered efficient influence function (REIF) of $\mu^+(t,z,\mathbf{X})$. Since $w_u(\mathbf{X}), w_l(\mathbf{X})$ are known functions of $(b^\pm, \alpha)$ (hence fixed with respect to the data-generating distribution), linearity of REIFs implies

$$\phi^+_{t,z}(S;\eta) := \mathrm{REIF}(\mu^+(t,z,\mathbf{X})) = w_u(\mathbf{X})\,\phi_u(S;\eta) + w_l(\mathbf{X})\,\phi_l(S;\eta), \tag{79}$$

where $\phi_u(S;\eta) := \mathrm{REIF}(\theta_u(\mathbf{X}))$ and $\phi_l(S;\eta) := \mathrm{REIF}(\theta_l(\mathbf{X}))$.

Define the selection weight

$$\kappa_{t,z}(S) := \frac{\mathbf{1}_{[T=t]}\,\mathbf{1}_{[Z=z]}}{\pi^t(\mathbf{X})\,\pi^g(z \mid \mathbf{X})}. \tag{80}$$

By the known REIFs for conditional CVaR/LCTE functionals (e.g., Dorn et al. (2025); Oprescu et al. (2023)),

$$\phi_u(S;\eta) = \theta_u(\mathbf{X}) + \kappa_{t,z}(S)\big(H_u(Y,Q) - \theta_u(\mathbf{X})\big), \quad \phi_l(S;\eta) = \theta_l(\mathbf{X}) + \kappa_{t,z}(S)\big(H_l(Y,Q) - \theta_l(\mathbf{X})\big). \tag{81}$$

Moreover, these REIFs are *orthogonal with respect to Q*: the cutoff Q is characterized as the optimizer of the corresponding tail objective (equivalently, the Rockafellar–Uryasev CVaR variational form), so the envelope/first-order condition yields $\partial_q \mathbb{E}[H_u(Y,q) \mid t,z,\mathbf{X}]|_{q=Q} = 0$ and $\partial_q \mathbb{E}[H_l(Y,q) \mid t,z,\mathbf{X}]|_{q=Q} = 0$ (see Dorn et al. (2025); Oprescu et al. (2023)).

Finally, substituting (81) into (79), using $\theta_u(\mathbf{X}) = Q + \gamma_u/(1-\alpha)$ and $\theta_l(\mathbf{X}) = Q - \gamma_l/\alpha$, and simplifying with $w_u/(1-\alpha) = 1/b^-$ and $w_l/\alpha = 1/b^+$ yields exactly Eq. (20).

Step 3: Unbiasedness and orthogonality. Orthogonality (Neyman-orthogonality) follows because $\phi^+_{t,z}$ is a linear combination of orthogonal REIFs for θ_u and θ_l (linearity preserves orthogonality), and because θ_u, θ_l themselves are orthogonal both to the selection nuisance (π^t, π^g) and to the regression nuisances via the standard conditional-mean EIF from Eq. ((81)). Orthogonality with respect to Q is guaranteed by the envelope/first-order condition (FOC) argument above.

Unbiasedness follows by iterated expectations: conditional on $\mathbf{X}$,

$$\mathbb{E}\left[\frac{\mathbf{1}_{[T=t]}\mathbf{1}_{[Z=z]}}{p(t,z \mid \mathbf{X})}\left\{\frac{(Y-Q)_+ - \gamma_u}{b^-} - \frac{(Q-Y)_+ - \gamma_l}{b^+}\right\} \Big| \mathbf{X}\right] = \mathbb{E}\left[\frac{(Y-Q)_+ - \gamma_u}{b^-} - \frac{(Q-Y)_+ - \gamma_l}{b^+} \Big| T=t, Z=z, \mathbf{X}\right] = 0, \tag{82}$$

so $\mathbb{E}[\phi^+_{t,z}(S;\eta) \mid \mathbf{X}] = \mu^+(t,z,\mathbf{X})$. This completes the proof for discrete Z.

Continuous Z. When Z is continuous, evaluation at $Z = z$ is not pathwise differentiable. We instead use kernel localization: replace $\mathbf{1}_{[Z=z]}$ in $\kappa_{t,z}$ by $\omega_{z,h}(Z) = K_h(Z-z)$ and replace the pmf $\pi^g(z \mid \mathbf{X})$ by the conditional density $\pi^g(Z \mid \mathbf{X})$ to define the localized weight

$$\kappa_{t,z,h}(S) := \frac{\mathbf{1}_{[T=t]}\, K_h(Z-z)}{\pi^t(\mathbf{X})\, \pi^g(Z \mid \mathbf{X})}. \tag{83}$$

Then Eq. (81) and the linearity relation from Eq. (79) hold verbatim with $\kappa_{t,z}$ replaced by $\kappa_{t,z,h}$, yielding the localized pseudo-outcome in Eq. (20). The same iterated-expectations argument gives $\mathbb{E}[\phi^+_{t,z,h}(S;\eta) \mid \mathbf{X}] = \mu^+_h(t,z,\mathbf{X})$, and under standard smoothness in z, $\mu^+_h(t,z,\mathbf{X}) \to \mu^+(t,z,\mathbf{X})$ as $h \downarrow 0$. □

D.5. Proof of Theorem 4.7

Theorem 4.7 (Second-order nuisance error (discrete Z))**.** *Assume Z is discrete and Assumption 4.6 holds. Let $\widehat{\eta} = (\widehat{\pi}^t, \widehat{\pi}^g, \widehat{Q}^+, \widehat{\gamma}^+_u, \widehat{\gamma}^+_l)$ be the cross-fitted nuisances used in $\phi^+_{t,z}(S;\widehat{\eta})$ from Theorem 4.4. Define $r_{n,\pi} := \|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2$, $r_{n,Q} := \|\widehat{Q}^+ - Q^+\|_2$, and $r_{n,\gamma} := \|\widehat{\gamma}^+_u - \gamma_u(\widehat{Q}^+;\cdot)\|_2 + \|\widehat{\gamma}^+_l - \gamma_l(\widehat{Q}^+;\cdot)\|_2$, where $\gamma_u(\widehat{Q}^+;\mathbf{X}) := \mathbb{E}[(Y-\widehat{Q}^+(\mathbf{X}))_+ \mid T=t, Z=z, \mathbf{X}]$ and $\gamma_l(\widehat{Q}^+;\mathbf{X}) := \mathbb{E}[(\widehat{Q}^+(\mathbf{X})-Y)_+ \mid T=t, Z=z, \mathbf{X}]$. Then*

$$\left\|\mathbb{E}\left[\phi^+_{t,z}(S;\widehat{\eta}) - \phi^+_{t,z}(S;\eta) \mid \mathbf{X}\right]\right\|_2 = O_p\left(r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}\right). \tag{21}$$

Proof. We prove the statement for discrete Z. Throughout, fix (t,z) and suppress (t,z) in the notation whenever clear. Because we use K-fold cross-fitting, for any observation in a held-out fold the nuisance estimates $\widehat{\eta} = (\widehat{\pi}^t, \widehat{\pi}^g, \widehat{Q}^+, \widehat{\gamma}^+_u, \widehat{\gamma}^+_l)$ are functions of the training folds only; hence, when taking expectations over the held-out fold, we may treat $\widehat{\eta}$ as fixed (formally, condition on the training sample).

Let $A := \mathbf{1}_{[T=t]}\mathbf{1}_{[Z=z]}$ and write the true and estimated joint propensities as

$$\pi(\mathbf{X}) := \pi^t(\mathbf{X})\pi^g(z \mid \mathbf{X}), \qquad \widehat{\pi}(\mathbf{X}) := \widehat{\pi}^t(\mathbf{X})\widehat{\pi}^g(z \mid \mathbf{X}). \tag{84}$$

Also denote the (population) conditional means at an arbitrary cutoff $\widehat{Q}$:

$$\gamma_u(\widehat{Q};\mathbf{X}) := \mathbb{E}\Big[(Y-\widehat{Q}(\mathbf{X}))_+ \mid T=t, Z=z, \mathbf{X}\Big], \quad \gamma_l(\widehat{Q};\mathbf{X}) := \mathbb{E}\Big[(\widehat{Q}(\mathbf{X})-Y)_+ \mid T=t, Z=z, \mathbf{X}\Big] \tag{85}$$

with $\gamma_u(Q^+;\mathbf{X}) = \gamma_u(\mathbf{X})$ and $\gamma_l(Q^+;\mathbf{X}) = \gamma_l(\mathbf{X})$.

Step 1: Conditional expectation of the estimated pseudo-outcome. For discrete Z, the pseudo-outcome simplifies (since A forces $Z = z$ inside the square bracket) to

$$\phi(S;\widehat{\eta}) = \widehat{Q}(\mathbf{X}) + \frac{\widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})} + \frac{A}{\widehat{\pi}(\mathbf{X})}\left[\frac{(Y-\widehat{Q}(\mathbf{X}))_+ - \widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{(\widehat{Q}(\mathbf{X})-Y)_+ - \widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})}\right]. \tag{86}$$

Taking conditional expectations given $\mathbf{X}$ and using $\mathbb{E}[A \mid \mathbf{X}] = \pi(\mathbf{X})$ yields

$$\mathbb{E}[\phi(S;\widehat{\eta}) \mid \mathbf{X}] = \widehat{Q}(\mathbf{X}) + \frac{\widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})} + \frac{\pi(\mathbf{X})}{\widehat{\pi}(\mathbf{X})}\left[\frac{\gamma_u(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})}\right] \tag{87}$$

$$= \underbrace{\widehat{Q}(\mathbf{X}) + \frac{\gamma_u(\widehat{Q}^+;\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X})}{b^+(z,\mathbf{X})}}_{=:\,\mu_{\widehat{Q}^+}(\mathbf{X})} + \left(\frac{\pi(\mathbf{X})}{\widehat{\pi}(\mathbf{X})} - 1\right)\left[\frac{\gamma_u(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})}\right] \tag{88}$$

Moreover, by Theorem 4.4 (applied with true nuisances), we yield

$$\mathbb{E}[\phi(S;\eta) \mid \mathbf{X}] = \mu^+(\mathbf{X}) := Q(\mathbf{X}) + \frac{\gamma_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\mathbf{X})}{b^+(z,\mathbf{X})}. \tag{89}$$

Thus, we arrive at

$$\mathbb{E}[\phi(S;\widehat{\eta}) - \phi(S;\eta) \mid \mathbf{X}] = \mu_{\widehat{Q}^+}(\mathbf{X}) - \mu^+(\mathbf{X}) + \left(\frac{\pi(\mathbf{X})}{\widehat{\pi}(\mathbf{X})} - 1\right)\left[\frac{\gamma_u(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})}\right], \tag{90}$$

and

$$\|\mathbb{E}[\phi(S;\widehat{\eta}) - \phi(S;\eta) \mid \mathbf{X}]\|_2 \le \underbrace{\left\|\mu_{\widehat{Q}}(\mathbf{X}) - \mu^+(\mathbf{X})\right\|_2}_{\text{cutoff-induced error}} + \underbrace{\left\|\frac{\pi(\mathbf{X})}{\widehat{\pi}(\mathbf{X})} - 1\right\|_2 \left\|\frac{\gamma_u(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})}\right\|_2}_{\text{propensity} \times \text{regression product term}} \tag{91}$$

where the last inequality is due to the triangle inequality and Cauchy–Schwarz inequality.

Step 2: Bounding the product term by $O_p(r_{n,\pi}r_{n,\gamma})$. By Assumption 4.6, $\widehat{\pi}(\mathbf{X}) \ge \varepsilon$ a.s., hence

$$\left\|\frac{\pi(\mathbf{X})}{\widehat{\pi}(\mathbf{X})} - 1\right\|_2 = \left\|\frac{\pi(\mathbf{X}) - \widehat{\pi}(\mathbf{X})}{\widehat{\pi}(\mathbf{X})}\right\|_2 \le \varepsilon^{-1}\|\widehat{\pi} - \pi\|_2. \tag{92}$$

Since $\pi = \pi^t\pi^g$ and $\widehat{\pi} = \widehat{\pi}^t\widehat{\pi}^g$,

$$\widehat{\pi} - \pi = (\widehat{\pi}^t - \pi^t)\widehat{\pi}^g + \pi^t(\widehat{\pi}^g - \pi^g), \tag{93}$$

so by the triangle inequality and $0 \le \pi^t, \widehat{\pi}^g \le 1$ (discrete Z),

$$\|\widehat{\pi} - \pi\|_2 \le \|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2. \tag{94}$$

Therefore

$$\left\|\frac{\pi}{\widehat{\pi}} - 1\right\|_2 \le \varepsilon^{-1}\Big(\|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2\Big) \tag{95}$$

and it remains to bound the second factor. Since $b^-(z,\mathbf{X})$ is bounded away from 0, we have

$$\left\|\frac{\gamma_u(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_u(\mathbf{X})}{b^-(z,\mathbf{X})} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_l(\mathbf{X})}{b^+(z,\mathbf{X})}\right\|_2 \le \varepsilon^{-1}\Big(\|\widehat{\gamma}_u - \gamma_u(\widehat{Q}^+;\cdot)\|_2 + \|\widehat{\gamma}_l - \gamma_l(\widehat{Q}^+;\cdot)\|_2\Big), \tag{96}$$

by the triangle inequality. Combining with the previous inequality yields

$$\underbrace{\left\|\frac{\pi}{\widehat{\pi}} - 1\right\|_2}_{=O_p(r_{n,\pi})} \cdot \underbrace{\left\|\frac{\gamma_u(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_u(\mathbf{X})}{b^-} - \frac{\gamma_l(\widehat{Q}^+;\mathbf{X}) - \widehat{\gamma}_l(\mathbf{X})}{b^+}\right\|_2}_{=O_p(r_{n,\gamma})} = O_p(r_{n,\pi}r_{n,\gamma}), \tag{97}$$

where the second $O_p(r_{n,\gamma})$ is by definition of $r_{n,\gamma}$ (as the L_2 rate for estimating the conditional tail means at the cutoff used in the pseudo-outcome).

Step 3: Bounding the cutoff-induced term by $O_p(r_{n,Q}^2)$. We now need to control the term $\mu_{\widehat{Q}^+}(\mathbf{X}) - \mu^+(\mathbf{X})$. Fix (t, z) and $\mathbf{x}$, and define the scalar function

$$\mathcal{L}_{\mathbf{x}}(q) := q + \frac{1}{b^-(z,\mathbf{x})}\,\mathbb{E}[(Y-q)_+ \mid T=t, Z=z, \mathbf{X}=\mathbf{x}] - \frac{1}{b^+(z,\mathbf{x})}\,\mathbb{E}[(q-Y)_+ \mid T=t, Z=z, \mathbf{X}=\mathbf{x}]\,. \tag{98}$$

By construction,

$$\mu_{\widehat{Q}^+}(\mathbf{x}) = \mathcal{L}_{\mathbf{x}}(\widehat{Q}^+(t,z,\mathbf{x})), \qquad \mu^+(\mathbf{x}) = \mathcal{L}_{\mathbf{x}}(Q^+(t,z,\mathbf{x})), \tag{99}$$

where $Q^+(t,z,\mathbf{x})$ is the optimal cutoff from Theorem 4.2.

Assume (as is standard for quantile/CVaR-style expansions) the conditional CDF $F_{Y|t,z,\mathbf{x}}$ is differentiable in a neighborhood of $Q^+(t,z,\mathbf{x})$ with density $f_{Y|t,z,\mathbf{x}}$ bounded by $\bar{f} < \infty$. Then, $\mathcal{L}_{\mathbf{x}}$ is differentiable and

$$\mathcal{L}'_{\mathbf{x}}(q) = 1 - \frac{1 - F_{Y|t,z,\mathbf{x}}(q)}{b^-(z,\mathbf{x})} - \frac{F_{Y|t,z,\mathbf{x}}(q)}{b^+(z,\mathbf{x})}. \tag{100}$$

Moreover, $\mathcal{L}'_{\mathbf{x}}$ is Lipschitz with

$$|\mathcal{L}''_{\mathbf{x}}(q)| = \left|\left(\frac{1}{b^-(z,\mathbf{x})} - \frac{1}{b^+(z,\mathbf{x})}\right) f_{Y|t,z,\mathbf{x}}(q)\right| \le \bar{f}\left(\frac{1}{b^-(z,\mathbf{x})} + \frac{1}{b^+(z,\mathbf{x})}\right) \le L, \tag{101}$$

for a finite constant L (uniform in $\mathbf{x}$ by Assumption 4.6).

By optimality of $Q^+(t,z,\mathbf{x})$ for $\mathcal{L}_{\mathbf{x}}$, we have $\mathcal{L}'_{\mathbf{x}}(Q^+(t,z,\mathbf{x})) = 0$. Therefore, by the fundamental theorem of calculus,

$$\begin{aligned}
\left|\mathcal{L}_{\mathbf{x}}(\widehat{Q}^+(t,z,\mathbf{x})) - \mathcal{L}_{\mathbf{x}}(Q^+(t,z,\mathbf{x}))\right| &= \left|\int_{Q^+(t,z,\mathbf{x})}^{\widehat{Q}^+(t,z,\mathbf{x})} \left(\mathcal{L}'_{\mathbf{x}}(u) - \mathcal{L}'_{\mathbf{x}}(Q^+(t,z,\mathbf{x}))\right)\,\mathrm{d}u\right| \\
&\le \int_{Q^+(t,z,\mathbf{x})}^{\widehat{Q}^+(t,z,\mathbf{x})} L\,|u - Q^+(t,z,\mathbf{x})|\,\mathrm{d}u \\
&\le \frac{L}{2}\,|\widehat{Q}^+(t,z,\mathbf{x}) - Q^+(t,z,\mathbf{x})|^2.
\end{aligned}$$

Taking $L_2(P_{\mathbf{X}})$ norms yields

$$\|\mu_{\widehat{Q}^+} - \mu^+\|_2 = O_p\Big(\|\widehat{Q}^+ - Q^+\|_2^2\Big) = O_p(r_{n,Q}^2). \tag{102}$$

Conclusion. Combining Step 2 and Step 3 in the decomposition from Eq. (90) yields

$$\left\|\mathbb{E}\big[\phi^+_{t,z}(S;\widehat{\eta}) - \phi^+_{t,z}(S;\eta) \mid \mathbf{X}\big]\right\|_2 = O_p\big(r_{n,\pi}\,r_{n,\gamma} + r_{n,Q}^2\big)\,, \tag{103}$$

which is exactly Eq. (21). □

D.6. Proof of Corollary 4.10

Corollary 4.10 (Quasi-oracle rates and inference (discrete Z))**.** *Suppose Assumptions 4.6 and 4.8 hold, and let $r_{n,\pi}, r_{n,\gamma}, r_{n,Q}$ be as in Theorem 4.7.*

CAPO rates: The CAPO upper-bound estimator satisfies

$$\|\widehat{\mu}^+(t,z,\cdot) - \mu^+(t,z,\cdot)\|_2 = O_p\big(\delta_n + r_{n,\pi}\,r_{n,\gamma} + r_{n,Q}^2\big)\,.$$

In particular, if $r_{n,\pi}\,r_{n,\gamma} + r_{n,Q}^2 = o_p(\delta_n)$, then $\|\widehat{\mu}^+(t,z,\cdot) - \mu^+(t,z,\cdot)\|_2 = O_p(\delta_n)$.

APO rates: The APO upper-bound estimator $\widehat{\psi}^+(t,z) = \mathbb{E}_n[\widehat{\phi}^+_{t,z}]$ satisfies

$$\left|\widehat{\psi}^+(t,z) - \psi^+(t,z)\right| = O_p\left(n^{-1/2} + r_{n,\pi}\,r_{n,\gamma} + r_{n,Q}^2\right). \tag{23}$$

If moreover $r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2 = o_p(n^{-1/2})$*, then*

$$\sqrt{n}\left(\widehat{\psi}^+(t,z) - \psi^+(t,z)\right) \rightsquigarrow \mathcal{N}\big(0,\, V^+(t,z)\big)\,, \tag{24}$$

i.e., the APO bound estimator is asymptotically normal with variance $V^+(t,z) := \mathrm{Var}(\phi^+_{t,z}(S;\eta))$ *(efficiency bound).*

Proof. We prove the CAPO and APO statements for the upper bound; the lower-bound case follows by the same argument with the sign-swapped pseudo-outcome.

CAPO bound rate. Let $m^+_{t,z}(\mathbf{x}) := \mathbb{E}[\widehat{\phi}^+_{t,z} \mid \mathbf{X} = \mathbf{x}]$ denote the conditional mean of the (cross-fitted) pseudo-outcome. By Assumption 4.8,

$$\|\widehat{\mu}^+(t,z,\cdot) - m^+_{t,z}(\cdot)\|_2 = O_p(\delta_n). \tag{104}$$

By the triangle inequality,

$$\|\widehat{\mu}^+(t,z,\cdot) - \mu^+(t,z,\cdot)\|_2 \le \|\widehat{\mu}^+(t,z,\cdot) - m^+_{t,z}(\cdot)\|_2 + \|m^+_{t,z}(\cdot) - \mu^+(t,z,\cdot)\|_2. \tag{105}$$

The second term is precisely the conditional bias induced by nuisance estimation. Applying Theorem 4.7 yields

$$\|m^+_{t,z}(\cdot) - \mu^+(t,z,\cdot)\|_2 = O_p(r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}). \tag{106}$$

Combining the two bounds gives the stated CAPO rate.

APO rate and asymptotic normality. Recall $\widehat{\psi}^+(t,z) = \mathbb{E}_n[\widehat{\phi}^+_{t,z}]$. Decompose

$$\widehat{\psi}^+(t,z) - \psi^+(t,z) = (\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z}(S;\eta)] + \mathbb{E}\big[\phi^+_{t,z}(S;\widehat{\eta}) - \phi^+_{t,z}(S;\eta)\big] + R_n, \tag{107}$$

where $R_n := (\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z}(S;\widehat{\eta}) - \phi^+_{t,z}(S;\eta)]$ is an empirical-process term. Under Assumption 4.6(iii), $\phi^+_{t,z}(S;\cdot)$ is uniformly bounded, and with cross-fitting $R_n = o_p(n^{-1/2})$ by standard arguments (conditioning on training folds and applying Hoeffding/Bernstein inequalities).

The first term is $O_p(n^{-1/2})$ by the CLT. The second term is bounded by Theorem 4.7 (after integrating over $\mathbf{X}$), yielding

$$\big|\mathbb{E}\big[\phi^+_{t,z}(S;\widehat{\eta}) - \phi^+_{t,z}(S;\eta)\big]\big| = O_p(r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}). \tag{108}$$

This proves Eq. (23). If additionally $r_{n,\pi} r_{n,\gamma} + r^2_{n,Q} = o_p(n^{-1/2})$, then the nuisance-induced bias term and R_n are $o_p(n^{-1/2})$, hence

$$\sqrt{n}\big(\widehat{\psi}^+(t,z) - \psi^+(t,z)\big) = \sqrt{n}(\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z}(S;\eta)] + o_p(1) \rightsquigarrow \mathcal{N}(0, V^+(t,z)), \tag{109}$$

with $V^+(t,z) = \mathrm{Var}(\phi^+_{t,z}(S;\eta))$. □

D.7. Proof of Proposition 4.11

Proposition 4.11 (Consistency for sharp bounds (discrete Z))**.** *Assume the conditions of Corollary 4.10 hold. Suppose* $\delta_n = o_p(1)$ *and* $r_{n,Q} = o_p(1)$*, and, in addition, either* $r_{n,\pi} = o_p(1)$ *or* $r_{n,\gamma} = o_p(1)$*. Then,* $\|\widehat{\mu}^{\pm}(t,z,\cdot) - \mu^{\pm}(t,z,\cdot)\|_2 = o_p(1)$ *and* $|\widehat{\psi}^{\pm}(t,z) - \psi^{\pm}(t,z)| = o_p(1)$*, Consequently, the estimated CAPO and APO intervals converge to the* sharp *identified intervals.*

Proof. We show the claim for the CAPO upper bound; the other bounds (CAPO lower, APO upper/lower) follow similarly.

By Corollary 4.10,

$$\|\widehat{\mu}^+(t,z,\cdot) - \mu^+(t,z,\cdot)\|_2 = O_p\big(\delta_n + r_{n,\pi} r_{n,\gamma} + r^2_{n,Q}\big)\,. \tag{110}$$

Under the proposition assumptions, $\delta_n = o_p(1)$ and $r_{n,Q} = o_p(1)$. Moreover, if either $r_{n,\pi} = o_p(1)$ or $r_{n,\gamma} = o_p(1)$, then $r_{n,\pi} r_{n,\gamma} = o_p(1)$. Therefore, the right-hand side is $o_p(1)$, implying $\|\widehat{\mu}^+(t,z,\cdot) - \mu^+(t,z,\cdot)\|_2 = o_p(1)$.

For APOs, the corresponding statement follows from the APO rate in Corollary 4.10 and the same convergence of the remainder. Finally, repeating the same argument for the lower bound (using its analogous pseudo-outcome) establishes convergence of both endpoints and, hence, convergence of the estimated intervals to the sharp identified intervals. □

D.8. Proof of Corollary 4.12

Corollary 4.12 (Asymptotic validity under misspecified cutoffs (discrete Z))**.** *Assume the conditions of Corollary 4.10 hold. Let $\overline{Q}^{\pm}(t,z,\mathbf{x})$ be any measurable cut-off and define the induced (possibly non-sharp) bounds*

$$\begin{aligned}\overline{\mu}^{\pm}(t,z,\mathbf{x};\overline{Q}^{\pm}) &= \overline{Q}^{\pm}(t,z,\mathbf{x}) \\ &+ \frac{1}{b^{\mp}(z,\mathbf{x})}\mathbb{E}\big[(Y-\overline{Q}^{\pm}(t,z,\mathbf{x}))_+ \mid t,z,\mathbf{x}\big] \\ &- \frac{1}{b^{\pm}(z,\mathbf{x})}\mathbb{E}\big[(\overline{Q}^{\pm}(t,z,\mathbf{x})-Y)_+ \mid t,z,\mathbf{x}\big].\end{aligned} \tag{25}$$

(and analogously $\overline{\psi}^{\pm}(t,z) := \mathbb{E}[\overline{\mu}^{\pm}(t,z,\mathbf{X})]$). Then, $[\overline{\mu}^{-}(t,z,\mathbf{x}),\overline{\mu}^{+}(t,z,\mathbf{x})]$ is a valid (not necessarily sharp) CAPO interval, and likewise for $[\overline{\psi}^{-}(t,z),\overline{\psi}^{+}(t,z)]$.

Moreover, if $\widehat{Q}^{\pm} \to \overline{Q}^{\pm}$ in L_2 and either (i) $(\widehat{\pi}^t,\widehat{\pi}^g)$ is consistent, or (ii) $(\widehat{\gamma}_u^{\pm},\widehat{\gamma}_l^{\pm})$ is consistent for the tail-moment targets induced by $\overline{Q}^{\pm}$, then the resulting estimated (C)APO intervals converge to $[\overline{\mu}^{-},\overline{\mu}^{+}]$ and $[\overline{\psi}^{-},\overline{\psi}^{+}]$ and are asymptotically valid, though potentially conservative. If $\overline{Q}^{\pm} = Q^{\pm}$, the bounds coincide with the sharp bounds.

Proof. We prove the CAPO claim; the APO claim follows by taking expectations over $\mathbf{X}$.

Step 1: Any cutoff induces a valid (conservative) interval. Fix $(t,z,\mathbf{x})$ and define for any scalar cutoff q the upper and lower tail objectives

$$\mathcal{L}_{\mathbf{x}}^{+}(q) := q + \frac{1}{b^{-}(z,\mathbf{x})}\mathbb{E}[(Y-q)_+ \mid T=t,Z=z,\mathbf{X}=\mathbf{x}] - \frac{1}{b^{+}(z,\mathbf{x})}\mathbb{E}[(q-Y)_+ \mid T=t,Z=z,\mathbf{X}=\mathbf{x}], \tag{111}$$

$$\mathcal{L}_{\mathbf{x}}^{-}(q) := q + \frac{1}{b^{+}(z,\mathbf{x})}\mathbb{E}[(Y-q)_+ \mid T=t,Z=z,\mathbf{X}=\mathbf{x}] - \frac{1}{b^{-}(z,\mathbf{x})}\mathbb{E}[(q-Y)_+ \mid T=t,Z=z,\mathbf{X}=\mathbf{x}]. \tag{112}$$

By Theorem 4.2 (equivalently, the standard Rockafellar–Uryasev variational form),

$$\mu^{+}(t,z,\mathbf{x}) = \inf_q \mathcal{L}_{\mathbf{x}}^{+}(q), \qquad \mu^{-}(t,z,\mathbf{x}) = \sup_q \mathcal{L}_{\mathbf{x}}^{-}(q), \tag{113}$$

with optimizers $q = Q^{+}(t,z,\mathbf{x})$ and $q = Q^{-}(t,z,\mathbf{x})$. Hence, for any measurable $\overline{Q}^{+}(t,z,\mathbf{x})$ and $\overline{Q}^{-}(t,z,\mathbf{x})$,

$$\overline{\mu}^{+}(t,z,\mathbf{x};\overline{Q}^{+}) := \mathcal{L}_{\mathbf{x}}^{+}(\overline{Q}^{+}(t,z,\mathbf{x})) \;\geq\; \mu^{+}(t,z,\mathbf{x}), \qquad \overline{\mu}^{-}(t,z,\mathbf{x};\overline{Q}^{-}) := \mathcal{L}_{\mathbf{x}}^{-}(\overline{Q}^{-}(t,z,\mathbf{x})) \;\leq\; \mu^{-}(t,z,\mathbf{x}), \tag{114}$$

so $[\overline{\mu}^{-}(t,z,\mathbf{x}),\overline{\mu}^{+}(t,z,\mathbf{x})]$ contains the sharp CAPO interval and is therefore valid.

Step 2: Convergence to the induced bounds. Fix measurable cutoffs $\overline{Q}^{\pm}$ and define the induced hinge-mean targets

$$\overline{\gamma}_u^{\pm}(t,z,\mathbf{x}) := \mathbb{E}\Big[(Y-\overline{Q}^{\pm}(t,z,\mathbf{x}))_+ \mid T=t,Z=z,\mathbf{X}=\mathbf{x}\Big], \tag{115}$$

$$\overline{\gamma}_l^{\pm}(t,z,\mathbf{x}) := \mathbb{E}\Big[(\overline{Q}^{\pm}(t,z,\mathbf{x})-Y)_+ \mid T=t,Z=z,\mathbf{X}=\mathbf{x}\Big]. \tag{116}$$

Let $\overline{\eta}^{\pm} := (\pi^t,\pi^g,\overline{Q}^{\pm},\overline{\gamma}_u^{\pm},\overline{\gamma}_l^{\pm})$. By Theorem 4.4, the corresponding pseudo-outcome is conditionally unbiased: $\mathbb{E}[\phi_{t,z}^{\pm}(S;\overline{\eta}^{\pm}) \mid \mathbf{X}] = \overline{\mu}^{\pm}(t,z,\mathbf{X};\overline{Q}^{\pm})$.

Now consider the estimated pseudo-outcome $\phi_{t,z}^{\pm}(S;\widehat{\eta}^{\pm})$ and write $\widehat{Q} = \widehat{Q}^{\pm}$, $\overline{Q} = \overline{Q}^{\pm}$ for brevity. The same conditional-expectation algebra as in the proof of Theorem 4.7 yields the decomposition

$$\mathbb{E}[\phi_{t,z}^{\pm}(S;\widehat{\eta}^{\pm}) \mid \mathbf{X}] = \mu_{\widehat{Q}}^{\pm}(\mathbf{X}) + \left(\frac{\pi(\mathbf{X})}{\widehat{\pi}(\mathbf{X})} - 1\right)\Delta_{\widehat{Q}}^{\pm}(\mathbf{X}), \tag{117}$$

where $\mu^{\pm}_{\widehat{Q}}(\mathbf{X})$ is the induced bound functional evaluated at $\widehat{Q}$ (i.e., $\mathcal{L}^{\pm}_{\mathbf{X}}(\widehat{Q})$) and $\Delta^{\pm}_{\widehat{Q}}(\mathbf{X})$ collects the conditional-mean regression errors at cutoff $\widehat{Q}$.

First, since $(u)_+$ is 1-Lipschitz, for each $\mathbf{X}$,

$$|\gamma_u(\widehat{Q};\mathbf{X}) - \gamma_u(\overline{Q};\mathbf{X})| \le |\widehat{Q}(\mathbf{X}) - \overline{Q}(\mathbf{X})|, \qquad |\gamma_l(\widehat{Q};\mathbf{X}) - \gamma_l(\overline{Q};\mathbf{X})| \le |\widehat{Q}(\mathbf{X}) - \overline{Q}(\mathbf{X})|. \tag{118}$$

Using $b^{\pm}$ bounded away from 0, this implies $\|\mu^{\pm}_{\widehat{Q}} - \overline{\mu}^{\pm}(\cdot;\overline{Q})\|_2 \lesssim \|\widehat{Q} - \overline{Q}\|_2 = o_p(1)$ whenever $\widehat{Q} \to \overline{Q}$ in L_2.

Second, for the product term, Assumption 4.6 implies $\|\pi/\widehat{\pi} - 1\|_2$ is bounded, and, if $(\widehat{\pi}^t, \widehat{\pi}^g)$ is consistent, then $\|\pi/\widehat{\pi} - 1\|_2 = o_p(1)$. Moreover,

$$\|\gamma_u(\widehat{Q};\cdot) - \widehat{\gamma}^{\pm}_u\|_2 \le \|\overline{\gamma}^{\pm}_u - \widehat{\gamma}^{\pm}_u\|_2 + \|\gamma_u(\widehat{Q};\cdot) - \gamma_u(\overline{Q};\cdot)\|_2 \le \|\overline{\gamma}^{\pm}_u - \widehat{\gamma}^{\pm}_u\|_2 + \|\widehat{Q} - \overline{Q}\|_2, \tag{119}$$

and similarly for the lower hinge mean. Hence, if $(\widehat{\gamma}^{\pm}_u, \widehat{\gamma}^{\pm}_l)$ is consistent for the induced targets $(\overline{\gamma}^{\pm}_u, \overline{\gamma}^{\pm}_l)$ and $\widehat{Q} \to \overline{Q}$, then $\|\Delta^{\pm}_{\widehat{Q}}\|_2 = o_p(1)$, so the product term is $o_p(1)$ even if $(\widehat{\pi}^t, \widehat{\pi}^g)$ is misspecified (but bounded away from 0).

Combining the two parts gives

$$\|\mathbb{E}[\phi^{\pm}_{t,z}(S;\widehat{\eta}^{\pm}) \mid \mathbf{X}] - \overline{\mu}^{\pm}(t,z,\mathbf{X};\overline{Q}^{\pm})\|_2 = o_p(1). \tag{120}$$

Under Assumption 4.8 with $\delta_n = o_p(1)$, the final-stage regression therefore yields $\|\widehat{\mu}^{\pm}(t,z,\cdot) - \overline{\mu}^{\pm}(t,z,\cdot;\overline{Q}^{\pm})\|_2 = o_p(1)$, and the sample-average estimator gives $\widehat{\psi}^{\pm}(t,z) \to \overline{\psi}^{\pm}(t,z)$. Thus the estimated (C)APO intervals converge to the induced (conservative) intervals and are asymptotically valid. If $\overline{Q}^{\pm} = Q^{\pm}$, then $\overline{\mu}^{\pm} = \mu^{\pm}$ and the limits coincide with the sharp bounds. □

D.9. Proof of Theorem C.2

Theorem C.2 (Second-order nuisance error (continuous Z))**.** *Assume Z is continuous and Assumptions 4.6 and C.1 hold. Let $\widehat{\eta} = (\widehat{\pi}^t, \widehat{\pi}^g, \widehat{Q}^+, \widehat{\gamma}^+_u, \widehat{\gamma}^+_l)$ be the cross-fitted nuisances used in $\phi^+_{t,z,h}(S;\widehat{\eta})$ (Eq. (20) with $\omega_{z,h}(Z) = K_h(Z - z)$).*

Define nuisance error rates (in L_2 norms over the appropriate arguments) by

$$r_{n,\pi} := \|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2, \qquad r_{n,Q} := \|\widehat{Q}^+ - Q^+\|_2, \tag{41}$$

$$r_{n,\gamma} := \|\widehat{\gamma}^+_u - \gamma_u(\widehat{Q}^+;\cdot)\|_2 + \|\widehat{\gamma}^+_l - \gamma_l(\widehat{Q}^+;\cdot)\|_2, \tag{42}$$

where the norms are taken over the random variables that the corresponding nuisance is evaluated on (e.g., $(Z,\mathbf{X})$ for $\pi^g(Z \mid \mathbf{X})$, $Q^+(t,Z,\mathbf{X})$, and $\gamma^{\pm}(t,Z,\mathbf{X})$).

Then, the conditional bias induced by nuisance estimation satisfies

$$\left\| \mathbb{E}\Big[\phi^+_{t,z,h}(S;\widehat{\eta}) - \phi^+_{t,z,h}(S;\eta) \mid \mathbf{X}\Big] \right\|_2 = O_p\left(\frac{r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}}{\sqrt{h}} \right). \tag{43}$$

Proof. We mirror the proof of Theorem 4.7 and highlight only the changes required for continuous Z. Fix (t,z) and suppress (t,z) in the notation. As before, by cross-fitting, we may condition on the training folds and treat $\widehat{\eta}$ as fixed when taking expectations over the held-out fold.

Key modification. For continuous Z, define

$$A_h := \mathbf{1}_{[T=t]} K_h(Z - z), \qquad \pi(Z,\mathbf{X}) := \pi^t(\mathbf{X})\, \pi^g(Z \mid \mathbf{X}), \qquad \widehat{\pi}(Z,\mathbf{X}) := \widehat{\pi}^t(\mathbf{X})\, \widehat{\pi}^g(Z \mid \mathbf{X}), \tag{121}$$

so that the (true) localized selection weight is

$$\kappa_{t,z,h}(S) = \frac{A_h}{\pi(Z,\mathbf{X})}. \tag{122}$$

The discrete-Z algebra carries through with A replaced by A_h and $\pi(\mathbf{X})$ replaced by $\pi(Z,\mathbf{X})$. The only substantive difference is that L_2-norms of kernel-weighted terms pick up a factor $h^{-1/2}$ via $\int K_h(u)^2\, du = O(1/h)$.

A useful kernel moment bound. Under Assumptions 4.6 and C.1, there exists a constant $C < \infty$ such that, for any square-integrable measurable function $G(S)$,

$$\|\mathbb{E}[\kappa_{t,z,h}(S)\, G(S) \mid \mathbf{X}]\|_2 \le \frac{C}{\sqrt{h}}\, \|G(S)\|_2. \tag{123}$$

Indeed, by conditional Cauchy–Schwarz,

$$\big(\mathbb{E}[\kappa_{t,z,h} G \mid \mathbf{X}]\big)^2 \le \mathbb{E}[\kappa_{t,z,h}^2 \mid \mathbf{X}]\, \mathbb{E}[G^2 \mid \mathbf{X}], \tag{124}$$

and $\mathbb{E}[\kappa_{t,z,h}^2 \mid \mathbf{X}]$ is of order $1/h$ because K_h^2 integrates to $O(1/h)$ and $\pi^t(\mathbf{X}), \pi^g(\cdot \mid \mathbf{X})$ are bounded away from 0 (overlap).

Step 1: Conditional expectation decomposition. Write $\phi_h(S;\cdot)$ for Eq. (20) with $\omega_{z,h}(Z) = K_h(Z - z)$. As in the discrete proof, take conditional expectations given $\mathbf{X}$ and use iterated expectations to replace the in-sample hinge terms by their corresponding conditional-mean targets (evaluated at the cutoff used in the pseudo-outcome). This yields a decomposition of the form

$$\mathbb{E}[\phi_h(S;\widehat{\eta}) - \phi_h(S;\eta) \mid \mathbf{X}] = \underbrace{\big(\mu_{h,\widehat{Q}^+}(\mathbf{X}) - \mu_h^+(\mathbf{X})\big)}_{\text{cutoff-induced error}} + \underbrace{\mathbb{E}\left[\left(\frac{\pi(Z,\mathbf{X})}{\widehat{\pi}(Z,\mathbf{X})} - 1\right) \kappa_{t,z,h}(S)\, \Delta_{\widehat{Q}^+}(S) \;\middle|\; \mathbf{X}\right]}_{\text{propensity} \times \text{regression product term}}, \tag{125}$$

where $\mu_{h,\widehat{Q}^+}(\mathbf{X})$ denotes the bound functional induced by the cutoff $\widehat{Q}^+$ (holding the remaining targets at their population values for that cutoff), and $\Delta_{\widehat{Q}^+}(S)$ collects the hinge-mean regression discrepancies at cutoff $\widehat{Q}^+$ (the continuous-Z analogue of the bracketed term in Eq. (90) of the discrete proof).

Step 2: Bounding the product term. By overlap, $\widehat{\pi}(Z,\mathbf{X})$ is bounded away from 0; hence

$$\left\| \frac{\pi(Z,\mathbf{X})}{\widehat{\pi}(Z,\mathbf{X})} - 1 \right\|_2 \lesssim \|\widehat{\pi}^t - \pi^t\|_2 + \|\widehat{\pi}^g - \pi^g\|_2 = O_p(r_{n,\pi}), \tag{126}$$

where norms are taken over the arguments on which the nuisances are evaluated (here $(Z,\mathbf{X})$ for π^g). Moreover, by definition of $r_{n,\gamma}$, $\|\Delta_{\widehat{Q}^+}(S)\|_2 = O_p(r_{n,\gamma})$. Applying Eq. (123) with $G(S) := \left(\frac{\pi}{\widehat{\pi}} - 1\right) \Delta_{\widehat{Q}^+}(S)$ gives

$$\left\| \mathbb{E}\left[\left(\frac{\pi}{\widehat{\pi}} - 1\right) \kappa_{t,z,h} \Delta_{\widehat{Q}^+} \mid \mathbf{X} \right] \right\|_2 \le \frac{C}{\sqrt{h}} \left\| \left(\frac{\pi}{\widehat{\pi}} - 1\right) \Delta_{\widehat{Q}^+} \right\|_2 = O_p\left(\frac{r_{n,\pi}\, r_{n,\gamma}}{\sqrt{h}} \right). \tag{127}$$

Step 3: Bounding the cutoff-induced term. The discrete proof bounds the cutoff-induced term using the envelope/FOC property of the cutoff and a second-order Taylor expansion, yielding a quadratic dependence on $\widehat{Q}^+ - Q^+$. The same argument applies here pointwise in the arguments of the cutoff (the cutoff remains an optimizer of the same tail objective, now for the localized target), so

$$|\mu_{h,\widehat{Q}^+}(\mathbf{X}) - \mu_h^+(\mathbf{X})| \lesssim \mathbb{E}\Big[\kappa_{t,z,h}(S) \left|\widehat{Q}^+(t,Z,\mathbf{X}) - Q^+(t,Z,\mathbf{X})\right|^2 \;\Big|\; \mathbf{X} \Big]. \tag{128}$$

Applying Eq. (123) with $G(S) := \left|\widehat{Q}^+(t,Z,\mathbf{X}) - Q^+(t,Z,\mathbf{X})\right|^2$ and using the same bounded-moment simplification as in the discrete proof gives

$$\|\mu_{h,\widehat{Q}^+} - \mu_h^+\|_2 = O_p\left(\frac{r_{n,Q}^2}{\sqrt{h}} \right). \tag{129}$$

Conclusion. Combining Steps 2–3 in Eq. (125) yields

$$\|\mathbb{E}[\phi_h(S;\widehat{\eta}) - \phi_h(S;\eta) \mid \mathbf{X}]\|_2 = O_p\left(\frac{r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2}{\sqrt{h}} \right), \tag{130}$$

which is exactly the claim. □

D.10. Proof of Corollary C.3

Corollary C.3 (Quasi-oracle rates and inference (continuous Z))**.** *Assume the conditions of Theorem C.2 and that the second-stage regression learner $\widehat{\mathbb{E}}_n[\cdot \mid \mathbf{X} = \mathbf{x}]$ satisfies Assumption 4.8 with rate δ_n when regressing $\phi^+_{t,z,h}(S;\eta)$ on $\mathbf{X}$.*

Then:

CAPO rates: The CAPO upper-bound estimator satisfies

$$\|\widehat{\mu}^+_h(t,z,\cdot) - \mu^+_h(t,z,\cdot)\|_2 = O_p\left(\delta_n + \frac{r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}}{\sqrt{h}}\right). \tag{44}$$

APO rates: The APO upper-bound estimator $\widehat{\psi}^+(t,z) = \mathbb{E}_n[\widehat{\phi}^+_{t,z,h}]$ satisfies

$$|\widehat{\psi}^+_h(t,z) - \psi^+_h(t,z)| = O_p\left(\frac{1}{\sqrt{nh}} + \frac{r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q}}{\sqrt{h}}\right). \tag{45}$$

$\sqrt{nh}$-CLT (central limit theorem) for the (smoothed) APO. If $r_{n,\pi}\, r_{n,\gamma} + r^2_{n,Q} = o_p(n^{-1/2})$, then

$$\sqrt{nh}\left(\widehat{\psi}^+_h(t,z) - \psi^+_h(t,z)\right) \rightsquigarrow \mathcal{N}\big(0,\ V^+_h(t,z)\big), \tag{46}$$

where one valid asymptotic variance target is $V^+_h(t,z) := \mathrm{Var}\big(\sqrt{h}\,\phi^+_{t,z,h}(S;\eta)\big)$.

Finally, if the smoothing bias satisfies $|\psi^+_h(t,z) - \psi^+(t,z)| = o((nh)^{-1/2})$ (e.g., via undersmoothing under z-smoothness), then the same CLT holds with $\psi^+(t,z)$ in place of $\psi^+_h(t,z)$.

Proof. We follow the proof of Corollary 4.10, replacing Theorem 4.7 by Theorem C.2 and tracking the kernel-induced scaling.

CAPO rate. Let $m^+_{t,z,h}(\mathbf{x}) := \mathbb{E}[\widehat{\phi}^+_{t,z,h} \mid \mathbf{X} = \mathbf{x}]$ denote the conditional mean of the (cross-fitted) localized pseudo-outcome. By Assumption 4.8,

$$\|\widehat{\mu}^+_h(t,z,\cdot) - m^+_{t,z,h}(\cdot)\|_2 = O_p(\delta_n). \tag{131}$$

By the triangle inequality,

$$\|\widehat{\mu}^+_h(t,z,\cdot) - \mu^+_h(t,z,\cdot)\|_2 \le \|\widehat{\mu}^+_h(t,z,\cdot) - m^+_{t,z,h}(\cdot)\|_2 + \|m^+_{t,z,h}(\cdot) - \mu^+_h(t,z,\cdot)\|_2. \tag{132}$$

The second term is exactly the conditional nuisance-induced bias controlled by Theorem C.2, giving the stated CAPO rate.

APO rate and $\sqrt{nh}$ asymptotic normality. Recall $\widehat{\psi}^+_h(t,z) = \mathbb{E}_n[\widehat{\phi}^+_{t,z,h}]$. Decompose

$$\widehat{\psi}^+_h(t,z) - \psi^+_h(t,z) = (\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z,h}(S;\eta)] + \mathbb{E}\Big[\phi^+_{t,z,h}(S;\widehat{\eta}) - \phi^+_{t,z,h}(S;\eta)\Big] + R_{n,h}, \tag{133}$$

where $R_{n,h} := (\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z,h}(S;\widehat{\eta}) - \phi^+_{t,z,h}(S;\eta)]$.

Under Assumption C.1 and overlap, $\mathrm{Var}(\phi^+_{t,z,h}(S;\eta)) = O(1/h)$, so $(\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z,h}(S;\eta)] = O_p((nh)^{-1/2})$ by the CLT. With cross-fitting and the same conditioning argument as in the discrete proof, $R_{n,h} = o_p((nh)^{-1/2})$.

The bias term is controlled by Theorem C.2 after integrating over $\mathbf{X}$, yielding the stated APO rate. If additionally $r_{n,\pi} r_{n,\gamma} + r^2_{n,Q} = o_p(n^{-1/2})$, then the bias term and $R_{n,h}$ are $o_p((nh)^{-1/2})$, implying

$$\sqrt{nh}\big(\widehat{\psi}^+_h(t,z) - \psi^+_h(t,z)\big) = \sqrt{nh}(\mathbb{E}_n - \mathbb{E})[\phi^+_{t,z,h}(S;\eta)] + o_p(1) \rightsquigarrow \mathcal{N}(0, V^+_h(t,z)), \tag{134}$$

with $V^+_h(t,z) = \mathrm{Var}(\sqrt{h}\,\phi^+_{t,z,h}(S;\eta))$. The final undersmoothing statement follows by adding/subtracting $\psi^+(t,z)$ and using the assumed bias condition. □

D.11. Proof of Proposition C.4

Proposition C.4 (Consistency for sharp bounds (continuous Z))**.** *Assume the conditions of Corollary C.3 and consider the corresponding lower-bound estimator $\widehat{\mu}_h^-(t,z,\cdot)$ constructed from the lower-bound pseudo-outcome (defined analogously to Eq. (20)). Suppose $\delta_n = o_p(1)$ and*

$$\frac{r_{n,\pi}\, r_{n,\gamma} + r_{n,Q}^2}{\sqrt{h}} = o_p(1). \tag{47}$$

Then,

$$\|\widehat{\mu}_h^{\pm}(t,z,\cdot) - \mu_h^{\pm}(t,z,\cdot)\|_2 = o_p(1), \qquad |\widehat{\psi}_h^{\pm}(t,z) - \psi_h^{\pm}(t,z)| = o_p(1). \tag{48}$$

Consequently, the estimated CAPO and APO intervals converge to the sharp *kernel-localized identified intervals for the bandwidth-indexed targets.*

Moreover, if the smoothing bias vanishes at the appropriate rate (e.g., $|\psi_h^{\pm}(t,z) - \psi^{\pm}(t,z)| = o((nh)^{-1/2})$), then the estimated intervals are asymptotically sharp for the original pointwise bounds as $h \downarrow 0$.

Proof. The argument is identical to the proof of Proposition 4.11, replacing Corollary 4.10 by Corollary C.3. Under the stated assumptions, $\delta_n = o_p(1)$ and $\frac{r_{n,\pi} r_{n,\gamma} + r_{n,Q}^2}{\sqrt{h}} = o_p(1)$, hence both CAPO endpoints converge in L_2 to the sharp kernel-localized endpoints $\mu_h^{\pm}(t,z,\cdot)$. The APO convergence follows from the APO rate in Corollary C.3. Finally, if the smoothing bias vanishes at the stated rate, the same conclusion holds for the pointwise (unsmoothed) bounds. □

D.12. Proof of Corollary C.5

Corollary C.5 (Asymptotic validity under misspecified cutoffs (continuous Z))**.** *Fix measurable cutoffs $\overline{Q}^{\pm}(t,z,\mathbf{x})$ (not necessarily equal to the sharp cut-offs) and let $\overline{\mu}_h^{\pm}(t,z,\mathbf{x};\overline{Q}^{\pm})$ and $\overline{\psi}_h^{\pm}(t,z;\overline{Q}^{\pm})$ denote the resulting (possibly non-sharp) kernel-localized bound functionals induced by these cutoffs (i.e., the targets obtained by replacing $Q^{\pm}$ in the pseudo-outcomes and taking the conditional/unconditional expectations as in Eq. (40)). Then, the induced intervals*

$$\big[\overline{\mu}_h^-(t,z,\mathbf{x};\overline{Q}^-),\ \overline{\mu}_h^+(t,z,\mathbf{x};\overline{Q}^+)\big] \quad \textit{and} \quad \big[\overline{\psi}_h^-(t,z;\overline{Q}^-),\ \overline{\psi}_h^+(t,z;\overline{Q}^+)\big] \tag{49}$$

are (not necessarily sharp) valid CAPO and APO intervals for the kernel-localized targets.

Moreover, if $\widehat{Q}^{\pm} \to \overline{Q}^{\pm}$ in L_2 and either

(i) $(\widehat{\pi}^t, \widehat{\pi}^g)$ is consistent, or

(ii) the corresponding tail-moment regressions $(\widehat{\gamma}_u^{\pm}, \widehat{\gamma}_l^{\pm})$ are consistent for the targets induced by $\overline{Q}^{\pm}$,

then the estimated endpoints converge to the induced (conservative) targets and the resulting (C)APO intervals remain asymptotically valid, though potentially conservative. If $\overline{Q}^{\pm}$ equals the sharp cut-offs, then the induced bounds coincide with the sharp bounds, and the intervals are asymptotically sharp as well.

Proof. We adapt the proof of Corollary 4.12 and indicate only the continuous-Z differences.

Step 1: Any cutoffs induce a valid (conservative) localized interval. Fix $(t,z,\mathbf{x})$. For each exposure level u, the discrete-Z proof shows (via the same Rockafellar-Uryasev tail objectives) that evaluating the tail objective at an arbitrary cutoff $\overline{Q}^{\pm}(t,u,\mathbf{x})$ yields conservative endpoints $\overline{\mu}^{\pm}(t,u,\mathbf{x};\overline{Q}^{\pm})$ that contain the sharp pointwise endpoints $\mu^{\pm}(t,u,\mathbf{x})$.

Kernel localization preserves this ordering because $K_h(\cdot) \geq 0$ and integrates to 1. Indeed, the continuous-Z localized targets are obtained by the same conditional-expectation construction as in Eq. (40), and the weight $\kappa_{t,z,h}(S)$ transports pointwise statements in u into their localized analogues around z. Therefore,

$$\overline{\mu}_h^-(t,z,\mathbf{x};\overline{Q}^-) \ \leq\ \mu_h^-(t,z,\mathbf{x}) \ \leq\ \mu_h^+(t,z,\mathbf{x}) \ \leq\ \overline{\mu}_h^+(t,z,\mathbf{x};\overline{Q}^+), \tag{135}$$

such that the CAPO interval is valid (though not necessarily sharp). The APO claim follows by taking expectations over $\mathbf{X}$.

Step 2: Convergence to the induced (conservative) localized bounds. The convergence argument follows Step 2 of the discrete-Z proof, with the single change that kernel-weighted terms are controlled using the bound in Eq. (123) (hence the extra factor $h^{-1/2}$ in intermediate inequalities). Under $\widehat{Q}^{\pm} \to \overline{Q}^{\pm}$ in L_2 and either (i) $(\widehat{\pi}^t, \widehat{\pi}^g)$ consistent or (ii) $(\widehat{\gamma}_u^{\pm}, \widehat{\gamma}_l^{\pm})$ consistent for the targets induced by $\overline{Q}^{\pm}$, the same decomposition yields

$$\left\| \mathbb{E}\left[\phi_{t,z,h}^{\pm}(S; \widehat{\eta}^{\pm}) \mid \mathbf{X} \right] - \overline{\mu}_h^{\pm}(t, z, \mathbf{X}; \overline{Q}^{\pm}) \right\|_2 = o_p(1). \tag{136}$$

With Assumption 4.8 and $\delta_n = o_p(1)$, the second-stage regression therefore implies $\|\widehat{\mu}_h^{\pm}(t, z, \cdot) - \overline{\mu}_h^{\pm}(t, z, \cdot; \overline{Q}^{\pm})\|_2 = o_p(1)$, and the sample-average estimator yields $\widehat{\psi}_h^{\pm}(t, z) \to \overline{\psi}_h^{\pm}(t, z; \overline{Q}^{\pm})$. Thus the estimated intervals converge to the induced (conservative) localized intervals and remain asymptotically valid. If $\overline{Q}^{\pm} = Q^{\pm}$, these limits coincide with the sharp localized bounds. □

E. Practical considerations

E.1. Applications of exposure mappings

Our framework accommodates a range of *exposure mappings* $g(\cdot)$ that reduce the (typically high-dimensional) vector of neighbors' treatments into a low-dimensional exposure variable for unit i. The choice of g should be guided by substantive knowledge about how interference operates, and by what is measured in the data. Below, we summarize common mappings and settings where they arise.

(1) Applications of the weighted-mean exposure

A common exposure mapping is the number of treated neighbors $g(T_{\mathcal{N}_i}) = \sum_{j\in\mathcal{N}_i} T_j$. This is natural when spillovers scale approximately with the number of treated contacts: repeated encouragement from multiple peers can increase salience, multiple treated farmers can raise local demonstration intensity, and multiple trained coworkers can increase adoption of a workflow tool. In spatial policy applications, a count of nearby treated sites (e.g., treated intersections or corridors) can proxy local exposure intensity to infrastructure changes.

A more realistic mapping often weights neighbors by interaction intensity: $g(T_{\mathcal{N}_i}) = \sum_{j\in\mathcal{N}_i} w_{ij}T_j$, $w_{ij} \geq 0$, where w_{ij} encodes geographic distance decay, communication frequency, tie strength, or mobility flows. This is common in epidemiology (contact rates), spatial economics (commuting flows), and online platforms (interaction networks). In transport and infrastructure settings, weights based on travel time or distance can encode that nearer interventions plausibly matter more.

When neighborhood size varies, a normalized mapping captures saturation: $g(T_{\mathcal{N}_i}) = \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} T_j$. This is widely used in education (fraction of classmates treated), workplaces (share of coworkers trained), and community programs (share of households reached by a campaign). Proportion-based mappings are also natural in information environments (online or offline) where the fraction of one's social neighborhood exposed to an intervention (e.g., a correction or informational nudge) shapes beliefs or behavior.

(2) Applications of the threshold exposure

A simple mapping is $g(T_{\mathcal{N}_i}) = \mathbf{1}\Big\{\sum_{j\in\mathcal{N}_i} T_j \geq 1\Big\}$, i.e., whether unit i has at least one treated neighbor. This is appropriate when spillovers plausibly operate through *presence* rather than intensity: diffusion of a new practice or technology can start once a single close contact adopts it (e.g., demonstration effects or peer-to-peer referrals). In public health, the presence of a vaccinated (or otherwise treated) contact may affect risk via reduced transmission in small networks (e.g., household or close-contact structures), where the key margin is whether at least one relevant contact is treated.

A thresholded saturation mapping, $g(T_{\mathcal{N}_i}) = \mathbf{1}\Big\{\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} T_j \geq c\Big\}$, is appropriate when spillovers exhibit *nonlinearities* such as coordination, social norms, or capacity constraints. Examples include collective action (private incentives change after a critical mass), technology standards (compatibility benefits after adoption passes a threshold), and community compliance contexts (program effects emerge only once participation exceeds a minimum level). In behavioral climate interventions, a threshold can represent a social-norm mechanism: behavior changes once individuals perceive that "most peers" act.

(3) High-order spillover effects

If interference operates through longer paths than captured by $\mathcal{N}_i$ (e.g., two-hop network effects, market-level general equilibrium, or broader media spillovers), then a strictly local exposure definition may be inadequate. Enlarging $\mathcal{N}_i$ or using weighted kernels can address this conceptually, but typically worsens overlap and can further widen bounds.

E.2. Limitations of our partial-identification bounds

Our partial-identification results yield identification-robust statements under interference and limited structure. That said, the bounds come with concrete limitations.

First of all, our bounds can be wide when the data are weakly informative. Partial identification is conservative by construction. When counterfactual information is scarce, e.g., rare high-saturation neighborhoods, limited overlap across exposure regimes, or strongly clustered assignment, the bounds may be wide. This reflects genuine lack of information under the maintained assumptions, not an estimation failure. While our approach targets robustness, finite samples can be sensitive

to rare exposure levels and to the quality of nuisance estimation (e.g., outcome regression and any assignment/exposure models). Heavy-tailed outcomes or highly imbalanced exposures can amplify instability, motivating careful overlap diagnostics, effective-sample-size reporting by exposure regime, and sensitivity checks across learners. Furthermore, network data is often incomplete (missing ties, mismeasured distances, unknown interaction strengths). Although we assume a fully observed network in our paper, errors in T_j, $\mathcal{N}_i$, or weights w_{ij} propagate directly into $g(T_{\mathcal{N}_i})$. Because the bounds are defined with respect to observed exposures, measurement error can attenuate or distort the effective exposure regimes and complicate substantive interpretation.

F. Implementation details

F.1. Data generation

We study the finite-sample performance of our proposed bound estimators in a network interference setting with potentially misspecified exposure mappings. The data-generating process is designed to closely mirror the structural assumptions of Section 4 while allowing for controlled violations of the exposure mapping.

Units, covariates, and network. We observe N units indexed by $i = 1, \dots, n$. Each unit is endowed with a d-dimensional covariate vector $\mathbf{X}_i \sim \mathcal{U}(-1, 1)$, drawn independently across units. Units are connected through a known undirected network $G = (V, E)$, where neighborhoods are defined as $\mathcal{N}_i = \{j : (i, j) \in E\}$ and node-specific degrees are denoted by $n_i = |\mathcal{N}_i|$.

Overall, we conduct six experiments. We generate each three networks with $N = 1000$ nodes and a 1-dimensional covariate (small datasets) and three networks with $N = 6000$ nodes and 6-dimensional covariates (large datasets). Each network is chosen to demonstrate the theoretical properties of our framework in dealing with different forms of exposure misspecification: we employ a random Erdős–Rényi network for testing exposure mapping (1), a scale-free Barabási–Albert network for exposure mapping (2), and community-structured stochastic block model exposure mappings (3). We elaborate on the design in Subsection F.2.

Treatment assignment. Each unit receives a binary treatment $T_i \in \{0, 1\}$. Treatments are assigned independently conditional on covariates according to a logistic propensity score model $\pi^t(\mathbf{x}) := \mathbb{P}(T_i = 1 \mid \mathbf{X}_i = \mathbf{x}) = \text{logit}^{-1}(\beta_T^\top \mathbf{x})$, where $\beta_T \in \mathbb{R}^d$ is fixed across simulations.

Potential outcomes and observed data. Potential outcomes depend on individual treatment, exposure, and covariates according to

$$Y_i(t, z) = m(t, z, \mathbf{X}_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{137}$$

with

$$m(t, z, \mathbf{x}) = \tau\, t + \delta\, z + \gamma\, tz + f(\mathbf{x}), \tag{138}$$

where τ captures the direct effect of treatment, δ the spillover effect of exposure, η allows for treatment–exposure interaction, and $f(\mathbf{x})$ is a nonlinear baseline function of covariates. The observed outcome is given by $Y_i = Y_i(T_i, Z_i^*)$, so that interference operates through the true exposure mapping g^*.

Target estimands. Our primary targets are the conditional average potential outcomes (CAPOs)

$$\mu(t, z, \mathbf{x}) := \mathbb{E}[Y(t, z) \mid \mathbf{X} = \mathbf{x}], \tag{139}$$

and their induced causal effects. In particular, we consider (i) conditional direct effects $\mu(1, z, \mathbf{x}) - \mu(0, z, \mathbf{x})$, (ii) conditional spillover effects $\mu(t, z_1, \mathbf{x}) - \mu(t, z_0, \mathbf{x})$, as well as their averaged (APO) counterparts.

Experimental variations. Across simulation scenarios, we vary the network density, the degree of exposure-mapping misspecification through ε, the discreteness or continuity of Z, and the sample size n. All reported results are averaged over multiple Monte Carlo repetitions.

F.2. Choice of network structure

We deliberately vary the underlying network structure across simulation scenarios to ensure that each form of exposure mapping misspecification is evaluated in a setting where it is substantively meaningful. Different misspecifications interact with distinct structural properties of networks, and using a single network model throughout would mask or attenuate these effects.

Specifically, we employ the following networks: (1) *Weighted versus unweighted mean exposure:* Erdős–Rényi (ER) networks, whose homogeneous degree distribution isolates the effect of heterogeneous influence weights from confounding degree heterogeneity; (2) *Threshold-based exposure mappings:* scale-free networks generated by a Barabási–Albert (BA) process, where heavy-tailed degree distributions imply that small shifts in the threshold can lead to substantial misclassification of exposure, particularly for highly connected units; (3) *Higher order spillovers:* stochastic block models

Table 2. Simulation design and parameter specifications.

Component	**Specification / Values**
Units (nodes)	$N = 3000/6000$
Covariate dimension	$d = 1/6$
Treatment propensity	$\beta_T = 0.8$
Direct effect	$\tau = 0.8$
Spillover effect	$\delta = 0.6$
Interaction	$\gamma = 0.2$
Outcome model nonlinearity	$0.6\tanh(X) + 0.4\sin(X) - 0.2X^2$
Noise	$\xi_i \sim \mathcal{N}(0, 1)$
Kernel bandwidth	$h = 0.1$
Mean misspecification	$\varepsilon = 0.03$
Threshold misspecifition	$c^* = 0.45$
Cross-fitting folds	$K = 5$
Runs	20

(SBMs) with pronounced community structure, where spillovers naturally propagate beyond direct neighbors through short multi-step paths, making truncation of the exposure radius consequential.

Across all scenarios, the network choice is therefore tailored to highlight the specific failure mode induced by the corresponding exposure mapping misspecification, rather than to optimize estimator performance.

F.3. Implementation

We implement our experiments in Python. All code for replication is available in our GitHub repository at `https://github.com/m-schroder/ExposureMisspecification`.

We estimate nuisance components via cross-fitted learners: the treatment propensity is fit with gradient-boosted classifiers; the exposure density uses a Gaussian-mixture conditional model for continuous exposures (or multinomial/gradient boosting for discrete cases). We fit our quantile regression and the conditional expectations γ with XGBoost models. For the second-stage regression, we as well employ XGBoost with cross-fitting. We select hyperparameters through cross-validation (log-loss for propensity/exposure models, pinball loss for quantile models, and MSE for regression).