A Breaking down the Continuous Treatment Marginal Sensitivity Model

Let's go deeper into the Continuous Treatment Marginal Sensitivity Model (CMSM).

A.1 MSM for binary treatment values

This section details the Marginal Sensitivity Model of [Tan06]. For binary treatments, $\mathcal{T}_B = \{0, 1\}$, the (nominal) propensity score, $e(\mathbf{x}) \equiv Pr(T = 1 \mid \mathbf{X} = \mathbf{x})$, states how the treatment status, t, depends on the covariates, $\mathbf{x}$, and is identifiable from observational data. The potential outcomes, Y_0 and Y_1, conditioned on the covariates, $\mathbf{x}$, are distributed as $P(\mathrm{Y}_0 \mid \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathbf{X} = \mathbf{x})$. Each of these conditional distributions can be written as mixtures with weights based on the propensity score:

$$\begin{aligned} P(\mathrm{Y}_0 \mid \mathbf{X} = \mathbf{x}) &= (1 - e(\mathbf{x}))P(\mathrm{Y}_0 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}) + e(\mathbf{x})P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}), \\ P(\mathrm{Y}_1 \mid \mathbf{X} = \mathbf{x}) &= (1 - e(\mathbf{x}))P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}) + e(\mathbf{x})P(\mathrm{Y}_1 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}). \end{aligned} \tag{16}$$

The conditional distributions of each potential outcome given the observed treatment, $P(\mathrm{Y}_0 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$, are identifiable from observational data, whereas the conditional distributions of each potential outcome given the counterfactual treatment, $P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$ are not. Under ignorability, $\{\mathrm{Y}_0, \mathrm{Y}_1\} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X} = \mathbf{x}$, $P(\mathrm{Y}_0 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y}_1 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$. Therefore, any deviation from these equalities will be indicative of hidden confounding. However, because the distributions $P(\mathrm{Y}_0 \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_1 \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x})$ are unidentifiable, the MSM postulates a relationship between each pair of identifiable and unidentifiable components.

The MSM assumes that $P(\mathrm{Y}_t \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x})$ is absolutely continuous with respect to $P(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})$ for all $\mathrm{t} \in \mathcal{T}_B$. Therefore, given that $P(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y}_t \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x})$ are σ-finite measures, by the Radon-Nikodym theorem, there exists a function $\lambda_B(\mathrm{Y}_\mathrm{t}, \mathbf{x}; \mathrm{t}) : \mathcal{Y} \to [0, \inf)$ such that,

$$P(\mathrm{Y}_t \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x}) = \int_{\mathcal{Y}} \lambda_B(\mathrm{Y}_\mathrm{t}, \mathbf{x}; \mathrm{t}) dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x}). \tag{17}$$

Rearranging terms, $\lambda_B(\mathrm{Y}_\mathrm{t}, \mathbf{x}; \mathrm{t})$ is expressed as the Radon-Nikodym derivative or ratio of densities,

$$\begin{aligned} \lambda_B(\mathrm{Y}_\mathrm{t}, \mathbf{x}; \mathrm{t}) &= \frac{dP(\mathrm{Y}_t \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}, \\ &= \frac{p(\mathrm{y}_t \mid \mathrm{T} = 1 - t, \mathbf{X} = \mathbf{x})}{p(\mathrm{y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}. \end{aligned} \tag{18}$$

By Bayes's rule, $\lambda(\mathrm{Y}_0, \mathbf{x}; 0)$ and $\lambda(\mathrm{Y}_1, \mathbf{x}; 1)$ are expressed as odds ratios,

$$\begin{aligned} \lambda_B(\mathrm{Y}_0, \mathbf{x}; 0) &= \frac{1 - e(\mathbf{x})}{e(\mathbf{x})} \Big/ \frac{1 - e(\mathbf{x}, \mathrm{y}_0)}{e(\mathbf{x}, \mathrm{y}_0)}, \\ \lambda_B(\mathrm{Y}_1, \mathbf{x}; 1) &= \frac{e(\mathbf{x})}{1 - e(\mathbf{x})} \Big/ \frac{e(\mathbf{x}, \mathrm{y}_1)}{1 - e(\mathbf{x}, \mathrm{y}_1)}, \end{aligned} \tag{19}$$

where $e(\mathbf{x}, \mathrm{y}_\mathrm{t}) \equiv Pr(\mathrm{T} = 1 \mid \mathbf{X} = \mathbf{x}, \mathrm{Y}_\mathrm{t} = \mathrm{y}_\mathrm{t})$ is the unidentifiable complete propensity for treatment.

Finally, the MSM further postulates that the odds of receiving the treatment $\mathrm{T} = 1$ for subjects with covariates $\mathbf{X} = \mathbf{x}$ can only differ from $e(\mathbf{x})/(1 - e(\mathbf{x}))$ by at most a factor of Λ,

$$\Lambda^{-1} \leq \lambda_B(\mathrm{Y}_t, \mathbf{x}; t) \leq \Lambda. \tag{20}$$

$$\alpha(e(\mathbf{x}, \mathrm{t}), \Lambda) = \frac{1}{\Lambda e(\mathbf{x}, \mathrm{t})} + 1 - \frac{1}{\Lambda} \leq \frac{1}{e(\mathbf{x}, \mathrm{t}, \mathrm{y}_t)} \leq \frac{\Lambda}{e(\mathbf{x}, \mathrm{t})} + 1 - \Lambda = \beta(e(\mathbf{x}, \mathrm{t}), \Lambda) \tag{21}$$

A.2 Modifying the MSM for categorical treatment values

For categorical treatments, $\mathcal{T}_C = \{\mathrm{t}_i\}_{i=1}^{n_c}$, the (nominal) generalized propensity score [HI04], $r(\mathbf{x}, \mathrm{t}) \equiv Pr(T = \mathrm{t} \mid \mathbf{X} = \mathbf{x})$, states how the treatment status, t, depends on the covariates, $\mathbf{x}$, and is identifiable from observational data. The potential outcomes, $\{\mathrm{Y_t} : \mathrm{t} \in \mathcal{T}_C\}$, conditioned on the covariates, $\mathbf{x}$, are distributed as $\{P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) : \mathrm{t} \in \mathcal{T}_C\}$. Again, each of these conditional distributions can be written as mixtures with weights based on the propensity density, yielding the following set of mixture distributions:

$$\left\{ P(\mathrm{Y}_t \mid \mathbf{X} = \mathbf{x}) = \sum_{\mathrm{t}' \in \mathcal{T}_C} r(\mathbf{x}, \mathrm{t}') P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}) \right\}. \tag{22}$$

Each conditional distribution of the potential outcome given the observed treatment, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, is identifiable from observational data, but each conditional distribution of the potential outcome given the counterfactual treatment, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$, and therefore each mixture $P(\mathrm{Y}_t \mid \mathbf{X} = \mathbf{x})$, is not. Under the ignorability assumption, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$ for all $\mathrm{t}' \in \mathcal{T}_C$.

In order to recover the form of the binary treatment MSM, we can postulate a relationship between the unidentifiable $P(\mathrm{Y}_t \mid \mathbf{X} = \mathbf{x}) - r(\mathbf{x}, \mathrm{t}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and the identifiable $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) - r(\mathbf{x}, \mathrm{t}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. Under the assumption that $P(\mathrm{Y}_t \mid \mathbf{X} = \mathbf{x}) - r(\mathbf{x}, \mathrm{t}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ is absolutely continuous with respect to $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) - r(\mathbf{x}, \mathrm{t}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, we define the Radon-Nikodym derivative

$$\begin{aligned}
\lambda_C(\mathrm{Y_t}, \mathbf{x}; \mathrm{t}) &= \frac{d(P(\mathrm{Y_t} \mid, \mathbf{X} = \mathbf{x}) - r(\mathbf{x}, \mathrm{t}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}))}{d(1 - r(\mathbf{x}, \mathrm{t})) P(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}, \\
&= \frac{1}{1 - r(\mathbf{x}, \mathrm{t})} \left(\frac{dP(\mathrm{Y_t} \mid, \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})} - \frac{r(\mathbf{x}, \mathrm{t}) dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})} \right), \\
&= \frac{1}{1 - r(\mathbf{x}, \mathrm{t})} \left(\frac{\sum_{\mathrm{t}' \in \mathcal{T}_C} r(\mathbf{x}, \mathrm{t}') dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})} - \frac{r(\mathbf{x}, \mathrm{t}) dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})} \right), \\
&= \frac{1}{1 - r(\mathbf{x}, \mathrm{t})} \left(\frac{\sum_{\mathrm{t}' \in \mathcal{T}_C} r(\mathbf{x}, \mathrm{t}') p(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})}{p(\mathrm{y}_t \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})} - \frac{r(\mathbf{x}, \mathrm{t}) p(\mathrm{y_t} \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})}{p(\mathrm{y_t} \mid \mathrm{T} = t, \mathbf{X} = \mathbf{x})} \right), \\
&= \frac{1}{1 - r(\mathbf{x}, \mathrm{t})} \left(\frac{\sum_{\mathrm{t}' \in \mathcal{T}_C} \cancel{r(\mathbf{x}, \mathrm{t}')} \frac{p(\mathrm{T} = \mathrm{t}' | \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{\cancel{r(\mathbf{x}, \mathrm{t}')}}}{\frac{p(\mathrm{T} = \mathrm{t} | \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{r(\mathbf{x}, \mathrm{t})}} - \frac{r(\mathbf{x}, \mathrm{t}) \frac{p(\mathrm{T} = \mathrm{t} | \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{\cancel{r(\mathbf{x}, \mathrm{t})}}}{\frac{p(\mathrm{T} = \mathrm{t} | \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{\cancel{r(\mathbf{x}, \mathrm{t})}}} \right), \\
&= \frac{r(\mathbf{x}, \mathrm{t})}{1 - r(\mathbf{x}, \mathrm{t})} \frac{1 - p(\mathrm{T} = \mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}{p(\mathrm{T} = \mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}, \\
&= \frac{r(\mathbf{x}, \mathrm{t})}{1 - r(\mathbf{x}, \mathrm{t})} \Big/ \frac{r(\mathbf{x}, \mathrm{t}, \mathrm{y_t})}{1 - r(\mathbf{x}, \mathrm{t}, \mathrm{y_t})},
\end{aligned} \tag{23}$$

where, $r(\mathbf{x}, \mathrm{t}, \mathrm{y_t}) \equiv p(\mathrm{T} = \mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$ is the unidentifiable complete propensity density for treatment.

Finally, the categorical MSM further postulates that the odds of receiving the treatment $\mathrm{T} = \mathrm{t}$ for subjects with covariates $\mathbf{X} = \mathbf{x}$ can only differ from $r(\mathbf{x}, \mathrm{t})/(1 - r(\mathbf{x}, \mathrm{t}))$ by at most a factor of Λ,

$$\Lambda^{-1} \leq \lambda_C(\mathrm{Y}_t, \mathbf{x}; t) \leq \Lambda. \tag{24}$$

$$\alpha(r(\mathbf{x}, \mathrm{t}), \Lambda) = \frac{1}{\Lambda r(\mathbf{x}, \mathrm{t})} + 1 - \frac{1}{\Lambda} \leq \frac{1}{r(\mathbf{x}, \mathrm{t}, \mathrm{y}_t)} \leq \frac{\Lambda}{r(\mathbf{x}, \mathrm{t})} + 1 - \Lambda = \beta(r(\mathbf{x}, \mathrm{t}), \Lambda) \tag{25}$$

A.3 Defining the Continuous MSM (CMSM) in terms of densities for continuous-valued interventions

The conditional distributions of the potential outcomes given the observed treatment assigned,

$$\{P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) : \mathrm{t} \in \mathcal{T}\},$$

are identifiable from observational data. However, the marginal distributions of the potential outcomes over all possible treatments,

$$\left\{ P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = \int_{\mathcal{T}} p(\mathrm{t}' \mid \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}) d\mathrm{t}' : \mathrm{t} \in \mathcal{T} \right\} \tag{26}$$

are not. This is because the component distributions, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$, are not identifiable when $\mathrm{t}' \neq \mathrm{t}$ as $\mathrm{Y_t}$ cannot be observed for units under treatment level $\mathrm{T} = \mathrm{t}'$. Under the ignorability assumption, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) = P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$ for all $\mathrm{t}' \in \mathcal{T}$, and so $P(\mathrm{Y_t} \mid, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ are identical. Therefore, any divergence between $P(\mathrm{Y_t} \mid, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ will be indicative of hidden confounding.

Where in the binary setting the MSM postulates a relationship between the unidentifiable $P(\mathrm{Y_t} \mid \mathrm{T} = 1 - \mathrm{t}, \mathbf{X} = \mathbf{x})$ and identifiable $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, our CMSM postulates a relationship between the unidentifiable $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ and the identifiable $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$.

> The Radon-Nikodym theorem involves a measurable space (X, Σ) on which two σ-finite measures are defined, μ and ν."
> – Wikipedia

In our setting, the measurable space is $(\mathbb{R}, \Sigma)$, and our σ-finite measures are, $\mu = P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and $\nu = P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$: $\mathrm{Y_t} \in \mathcal{Y} \subseteq \mathbb{R}$.

> If ν is absolutely continuous with respect to μ (written $\nu \ll \mu$), then there exists a Σ-measurable function $f : X \to [0, \infty)$, such that $\nu(A) = \int_A f d\mu$ for any measurable set $A \subseteq X$.
> – Wikipedia

We then need to assume that $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) \ll P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, that is $P(A \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) = 0$ implies $P(A \mid \mathbf{X} = \mathbf{x}) = 0$ for any measurable set A.

This leads us to a proof for Proposition 1

Proof. Further, in our setting we have $f = \lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t})$, therefore

$$P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = \int_{\mathcal{Y}} \lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}). \tag{27}$$

Let the range of $\mathrm{Y_t}$ be the measurable space $(\mathcal{Y}, \mathcal{A})$, and $\nu(A)$ denote the Lebesgue measure for any measurable $A \in \mathcal{A}$. Then,

$$\begin{aligned}
\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) &= \frac{dP(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})} && \text{(28a)} \\
&= \frac{dP(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})}{d\nu} \frac{d\nu}{dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})} && \text{(28b)} \\
&= \frac{dP(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})}{d\nu} \left(\frac{dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}{d\nu} \right)^{-1} && \text{(28c)} \\
&= \frac{d}{d\nu} \int_A p(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x}) d\nu \left(\frac{d}{d\nu} \int_A p(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) d\nu \right)^{-1} && \text{(28d)} \\
&= \frac{p(\mathrm{y_t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})} && \text{(28e)} \\
&= \frac{p(\mathrm{t} \mid \mathbf{X} = \mathbf{x})}{p(\mathrm{t} \mid \mathrm{Y_t} = \mathrm{y_t}, \mathbf{X} = \mathbf{x})} && \text{(28f)}
\end{aligned}$$

Equation (28a) by the Radon-Nikodym derivative. Equation (28a)-Equation (28c) hold $\nu-$almost everywhere under the assumption $P(\mathrm{Y_t} \in A \mid \mathbf{x}) \ll \nu(A) \sim P(\mathrm{Y_t} \in A \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. Equation (28c)-Equation (28d) by the Radon-Nikodym theorem. Equation (28d)-Equation (28e) by the fundamental theorem of calculus under the assumption that $p(\mathrm{y_t} \mid \mathbf{x})$ and $p(\mathrm{y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ be continuous for $\mathrm{y_t} \in \mathcal{Y}$. Equation (28e)-Equation (28f) by Bayes's Rule. □

The sensitivity analysis parameter Λ then bounds the ratio, which leads to our bounds for the inverse complete propensity density:

$$\begin{aligned} \frac{1}{\Lambda} \leq & \frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} \leq \Lambda, \\ \frac{1}{\Lambda p(\mathrm{t} \mid \mathbf{x})} \leq & \frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} \leq \frac{\Lambda}{p(\mathrm{t} \mid \mathbf{x})} \\ \alpha(p(\mathrm{t} \mid \mathbf{x}), \Lambda) \leq & \frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} \leq \beta(p(\mathrm{t} \mid \mathbf{x}), \Lambda) \end{aligned} \tag{29}$$

A.3.1 KL Divergence

The bounds on the density ratio can also be expressed as bounds on the Kullback–Leibler divergence between $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$.

$$\Lambda^{-1} \leq \frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} \leq \Lambda, \tag{30}$$

$$\log\left(\Lambda^{-1}\right) \leq \log\left(\frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}\right) \leq \log\left(\Lambda\right) \tag{31}$$

$$\mathop{\mathbb{E}}_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} \log\left(\Lambda^{-1}\right) \leq \mathop{\mathbb{E}}_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} \log\left(\frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}\right) \leq \mathop{\mathbb{E}}_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} \log\left(\Lambda\right) \tag{32}$$

$$\log\left(\Lambda^{-1}\right) \leq \mathop{\mathbb{E}}_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} \log\left(\frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}\right) \leq \log\left(\Lambda\right) \tag{33}$$

$$\log\left(\Lambda^{-1}\right) \leq \int_{\mathcal{Y}} \log\left(\frac{dP(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}\right) dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) \leq \log\left(\Lambda\right) \tag{34}$$

$$\log\left(\Lambda^{-1}\right) \leq -D_{\mathrm{KL}}\left(P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) || P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})\right) \leq \log\left(\Lambda\right) \tag{35}$$

$$\left|\log\left(\Lambda\right)\right| \geq D_{\mathrm{KL}}\left(P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}) || P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})\right) \tag{36}$$

B Derivation of Equation (7)

Lemma 1.

$$\mu(\mathbf{x}, \mathrm{t}) = \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})dy}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})dy} \tag{37}$$

Proof. Recall that the conditional average potential outcome, $\mu(\mathbf{x}, \mathrm{t}) = \mathbb{E}[\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}]$, is unidentifiable without further assumptions. Following [KMZ19], we start from,

$$\begin{aligned}
\mu(\mathbf{x}, \mathrm{t}) &= \mathbb{E}\left[\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}\right], \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} p(\mathrm{y_t} \mid \mathbf{x}) d\mathrm{y_t}}{\int_{\mathcal{Y}} p(\mathrm{y_t} \mid \mathbf{x}) d\mathrm{y_t}}, \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x}) p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x}) p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \\
&= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}},
\end{aligned}$$

which is convenient as it decomposes $\mu(\mathbf{x}, \mathrm{t})$ into it's identifiable, $p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})$, and unidentifiable, $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$, parts.

Now, following [JMGS21], we add and subtract the empirical conditional outcome $\widetilde{\mu}(\mathbf{x}, \mathrm{t}) = \mathbb{E}[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}]$ from the right-hand-side above:

$$\mu(\mathbf{x}, \mathrm{t}) = \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \tag{39a}$$

$$= \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}), \tag{39b}$$

$$= \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}) \frac{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \tag{39c}$$

$$= \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}} - \frac{\int_{\mathcal{Y}} \widetilde{\mu}(\mathbf{x}, \mathrm{t}) \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \tag{39d}$$

$$= \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} (\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t})) \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}. \tag{39e}$$

Following [KMZ19] again, we reparameterize the inverse complete propensity density as, $\frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} = \alpha(\mathbf{x}; \mathrm{t}, \Lambda) + w(\mathrm{y}, \mathbf{x})(\beta(\mathbf{x}; \mathrm{t}, \Lambda) - \alpha(\mathbf{x}; \mathrm{t}, \Lambda))$ with $w : \mathcal{Y} \times \mathcal{X} \to [0, 1]$. We will shorten this

expression to $\frac{1}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})} = \alpha + w(\mathrm{y},\mathbf{x})(\beta-\alpha)$ below. This gives,

$$\mu(\mathbf{x},\mathrm{t}) = \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))\frac{p(\mathrm{y_t}|\mathrm{t},\mathbf{x})}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})}d\mathrm{y_t}}{\int_{\mathcal{Y}}\frac{p(\mathrm{y_t}|\mathrm{t},\mathbf{x})}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})}d\mathrm{y_t}}, \tag{40a}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\int_{\mathcal{Y}}(\alpha+w(\mathrm{y},\mathbf{x})(\beta-\alpha))(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\int_{\mathcal{Y}}(\alpha+w(\mathrm{y},\mathbf{x})(\beta-\alpha))p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40b}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\alpha\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t} + (\beta-\alpha)\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\alpha\int_{\mathcal{Y}}p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t} + (\beta-\alpha)\int_{\mathcal{Y}}w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40c}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\alpha\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t} + (\beta-\alpha)\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\alpha + (\beta-\alpha)\int_{\mathcal{Y}}w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40d}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{(\beta-\alpha)\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\alpha + (\beta-\alpha)\int_{\mathcal{Y}}w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40e}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\frac{\alpha}{\beta-\alpha} + \int_{\mathcal{Y}}w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40f}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\frac{1/(\Lambda p(\mathrm{t}|\mathbf{x}))}{\Lambda/p(\mathrm{t}|\mathbf{x})-1/(\Lambda p(\mathrm{t}|\mathbf{x}))} + \int_{\mathcal{Y}}w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40g}$$

$$= \widetilde{\mu}(\mathbf{x},\mathrm{t}) + \frac{\int_{\mathcal{Y}}(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t}))w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}{\frac{1}{\Lambda^2-1} + \int_{\mathcal{Y}}w(\mathrm{y},\mathbf{x})p(\mathrm{y_t}\mid\mathrm{t},\mathbf{x})d\mathrm{y_t}}, \tag{40h}$$

which concludes the proof. □

C Approximating integrals using Gauss-Hermite quadrature

Gauss-Hermite quadrature is a numerical method to approximate indefinite integrals of the following form: $\int_{-\infty}^{\infty}\exp(-\mathrm{y}^2)f(\mathrm{y})d\mathrm{y}$. In this case,

$$\int_{-\infty}^{\infty}\exp(-\mathrm{y}^2)f(\mathrm{y})d\mathrm{y} \approx \sum_{i=1}^{m} g_i f(\mathrm{y}),$$

where m is the number of samples drawn. The y_i are the roots of the physicists Hermite polynomial $H_m^*(\mathrm{y})(i=1,2,\ldots,m)$ and the weights are given by

$$g_i = \frac{2^{m-1}m!\sqrt{\pi}}{m^2[H_{m-1}^*(\mathrm{y}_k)]^2}$$

This method can be used to calculate the expectation of a function, $h(y)$, with respect to a Gaussian distributed outcome $p(\mathrm{y}) = \mathcal{N}(\mathrm{y}\mid\mu,\sigma^2)$ through a change of variables, such that,

$$\begin{aligned}\mathop{\mathbb{E}}_{p(\mathrm{y})}[h(\mathrm{y})] &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{\pi}}\exp\left(-\mathrm{y}^2\right)h\left(\sqrt{2}\sigma\mathrm{y}+\mu\right)d\mathrm{y} \\ &\approx \frac{1}{\sqrt{\pi}}\sum_{i=1}^{m} g_i h\left(\sqrt{2}\sigma\mathrm{y}+\mu\right).\end{aligned} \tag{41}$$

Definition 2. Gauss-Hermite quadrature integral estimator when $p(\mathrm{y}|\mathrm{t},\mathbf{x},\boldsymbol{\theta})$ is a parametric Gaussian density estimator, $\mathcal{N}(\mathrm{y}\mid\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}),\widetilde{\sigma}^2(\mathbf{x},\mathrm{t};\boldsymbol{\theta}))$*:*

$$I_G(h(\mathrm{y})) \coloneqq \frac{1}{\sqrt{\pi}}\sum_{i=1}^{m} g_i h\left(\sqrt{2}\widetilde{\sigma}^2(\mathbf{x},\mathrm{t};\boldsymbol{\theta})\mathrm{y}+\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta})\right)$$

Alternatively, when the density of the outcome is modelled using a n_{y} component Gaussian mixture, $p(\mathrm{y}) = \sum_{j=1}^{n_{\mathrm{y}}} \pi_j \mathcal{N}(\mathrm{y} \mid \mu_j, \sigma_j^2)$

$$\begin{aligned}\mathop{\mathbb{E}}_{p(\mathrm{y})}[h(\mathrm{y})] &= \frac{1}{\sqrt{\pi}} \sum_{j=1}^{n_{\mathrm{y}}} \pi_j \int_{-\infty}^{\infty} \exp\left(-\mathrm{y}^2\right) h\left(\sqrt{2}\sigma_j \mathrm{y} + \mu_j\right) d\mathrm{y}, \\ &\approx \frac{1}{\sqrt{\pi}} \sum_{j=1}^{n_{\mathrm{y}}} \pi_j \sum_{i=1}^{m} g_i h\left(\sqrt{2}\sigma_j \mathrm{y} + \mu_j\right).\end{aligned}$$

Definition 3. Gauss-Hermite quadrature integral estimator for expectations when $p(\mathrm{y}|\mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$ is a parametric Gaussian Mixture Density, $\sum_{j=1}^{n_{\mathrm{y}}} \widetilde{\pi}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) \mathcal{N}\left(\mathrm{y} \mid \widetilde{\mu}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}), \widetilde{\sigma}_j^2(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\right)$:

$$I_{GM}(h(\mathrm{y})) \coloneqq \frac{1}{\sqrt{\pi}} \sum_{j=1}^{n_{\mathrm{t}}} \widetilde{\pi}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) \sum_{i=1}^{m} g_i h\left(\sqrt{2}\widetilde{\sigma}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\mathrm{y} + \widetilde{\mu}_j(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})\right)$$

D Optimization over step functions

Lemma 2. *The sensitivity bounds given in Equations (8) and (9) have the following equivalent expressions:*

$$\begin{aligned}\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) &= \sup_{w(\mathrm{y}) \in \mathcal{W}_{nd}^H} \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}, \\ \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) &= \inf_{w(\mathrm{y}) \in \mathcal{W}_{ni}^H} \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}},\end{aligned}$$

where $\widetilde{\mu}(\mathbf{x}, \mathrm{t}) = \mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$, $\mathcal{W}_{nd}^H = \{w : H(\mathrm{y} - \mathrm{y}_H)\}_{\mathrm{y}_H \in \mathcal{Y}}$, $\mathcal{W}_{ni}^H = \{w : H(\mathrm{y}_H - \mathrm{y})\}_{\mathrm{y}_H \in \mathcal{Y}}$, *and*

$$H(\mathrm{z}) \coloneqq \begin{cases} 1, \mathrm{z} \geq 0 \\ 0, \mathrm{z} < 0 \end{cases},$$

Proof. We follow the argument of [KMZ19] and show that our alternative formulations of $\alpha(\cdot, \Lambda)$ and $\beta(\cdot, \Lambda)$ do not change the conclusions of their linear program solution. Starting from $\mu(\mathbf{x}, \mathrm{t}) = \frac{\int_{\mathcal{Y}} \mathrm{y}_{\mathrm{t}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}$, and applying a one-to-one change of variables, $\frac{1}{p(\mathrm{t} \mid \mathrm{y}_{\mathrm{t}}, \mathbf{x})} = \alpha(\mathbf{x}; \mathrm{t}, \Lambda) + w(\mathrm{y})(\beta(\mathbf{x}; \mathrm{t}, \Lambda) - \alpha(\mathbf{x}; \mathrm{t}, \Lambda))$ with $w : \mathcal{Y} \to [0, 1]$, $\alpha(\mathbf{x}; \mathrm{t}, \Lambda) = 1/\Lambda p(\mathrm{t} \mid \mathbf{x})$, $\beta(\mathbf{x}; \mathrm{t}, \Lambda) = \Lambda / p(\mathrm{t} \mid \mathbf{x})$, we arrive at:

$$\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) = \sup_{w: \mathcal{Y} \to [0,1]} \frac{\int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}, \tag{42}$$

and

$$\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) = \inf_{w: \mathcal{Y} \to [0,1]} \frac{\int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}, \tag{43}$$

after some cancellations. Duality can be used to prove that the $w^*(\mathrm{y})$ which achieves the supremum in Equation (42) belongs to the set of step functions $\mathcal{W}_{\mathrm{nd}}^H$. An analogous proof for Equation (43) would show that the $w^*(\mathrm{y})$ which achieves the infimum in Equation (43) belongs to the set of step functions $\mathcal{W}_{\mathrm{ni}}^H$.

The optimization problem in Equation (42) can be rewritten as a linear-fractional program:

$$\begin{aligned}\text{maximize} \quad & \frac{a\langle \mathrm{y}, w(\mathrm{y}) \rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} + c}{b\langle 1, w(\mathrm{y}) \rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} + d} && \text{(44a)} \\ \text{subject to} \quad & 0 \leq w(\mathrm{y}) \leq 1 : \forall \mathrm{y} \in \mathcal{Y}, && \text{(44b)}\end{aligned}$$

where $\langle \cdot, \cdot \rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})}$ is the inner product with respect to $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})$, $a = b = \lambda^2 - 1$, $c = \int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}$, and $d = \int_{\mathcal{Y}} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}$.

The linear-fractional program of Equation (44) is equivalent to the following linear program:

$$\begin{aligned}
\text{maximize} \quad & a\langle \mathrm{y}, \widetilde{w}(\mathrm{y})\rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} + c\widetilde{v}(\mathbf{x}) && \text{(45a)}\\
\text{subject to} \quad & \widetilde{w}(\mathrm{y}) \leq \widetilde{v}(\mathbf{x}) : \forall \mathrm{y} \in \mathcal{Y} && \text{(45b)}\\
& -\widetilde{w}(\mathrm{y}) \leq 0 : \forall \mathrm{y} \in \mathcal{Y} && \text{(45c)}\\
& b\langle 1, \widetilde{w}(\mathrm{y})\rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} + d\widetilde{v}(\mathbf{x}) = 1 && \text{(45d)}\\
& \widetilde{v}(\mathbf{x}) \geq 0, && \text{(45e)}
\end{aligned}$$

where

$$\widetilde{w}(\mathrm{y}) = \frac{w(\mathrm{y})}{b\langle 1, w(\mathrm{y})\rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} + d} \quad \text{and} \quad \widetilde{v}(\mathbf{x}) = \frac{1}{b\langle 1, w(\mathrm{y})\rangle_{p(\mathrm{y}|\mathrm{t},\mathbf{x})} + d}$$

by the Charnes-Cooper transformation.

Let the dual function $\rho(\mathrm{y})$ be associated with the primal constraint eq. (45b), the dual function $\eta(\mathrm{y})$ be associated with the primal constraint eq. (45c), and γ be the dual variable associated with the primal constraint eq. (45d). The dual program is then:

$$\begin{aligned}
\text{minimize} \quad & \gamma && \text{(46a)}\\
\text{subject to} \quad & \rho(\mathrm{y}) - \eta(\mathrm{y}) + \gamma b p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) = a\mathrm{y}p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) : \forall \mathrm{y} \in \mathcal{Y} && \text{(46b)}\\
& -\langle 1, \rho(\mathrm{y})\rangle + \gamma d \geq c && \text{(46c)}\\
& \rho(\mathrm{y}) \in \mathbb{R}_+, \eta(\mathrm{y}) \in \mathbb{R}_+, \gamma \in \mathbb{R} && \text{(46d)}
\end{aligned}$$

At most one of $\rho(\mathrm{y})$ or $\eta(\mathrm{y})$ is non-zero by complementary slackness; therefore, condition eq. (46b) implies that

$$\begin{aligned}
\rho(\mathrm{y}) &= (\lambda^2 - 1)p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) \max\{\mathrm{y} - \gamma, 0\} \text{ when } \eta = 0,\\
\eta(\mathrm{y}) &= (\lambda^2 - 1)p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) \max\{\gamma - \mathrm{y}, 0\} \text{ when } \rho = 0.
\end{aligned}$$

[KMZ19] argue that constraint eq. (46c) ought to be tight (an equivalence) at optimality, otherwise there would exist a smaller, feasible γ that satisfies the linear program. Therefore,

$$\begin{aligned}
-\langle 1, \rho(\mathrm{y})\rangle + \gamma d &= c,\\
-\int_{\mathcal{Y}} (\lambda^2 - 1)p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) \max\{\mathrm{y} - \gamma, 0\} d\mathrm{y} + \gamma \int_{\mathcal{Y}} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} &= \int_{\mathcal{Y}} \mathrm{y}p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y},\\
(\lambda^2 - 1) \int_{\mathcal{Y}} \max\{\mathrm{y} - \gamma, 0\} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} &= \int_{\mathcal{Y}} (\gamma - \mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}.
\end{aligned} \tag{47}$$

Letting $C_{\mathrm{Y}} > 0$ such that $|\mathcal{Y}| \leq C_{\mathrm{Y}}$, it is impossible that either $\gamma > C_{\mathrm{Y}}$ (the r.h.s. would be 0 and the l.h.s. would be > 0) or $\gamma < -C_{\mathrm{Y}}$ (the r.h.s. would be > 0 and the l.h.s. would be < 0). Thus, $\exists \mathrm{y}^* \in [-C_{\mathrm{Y}}, C_{\mathrm{Y}}]$ such that when $\mathrm{y} < \mathrm{y}^*$, $\eta > 0$ so $w = 0$ and when $\mathrm{y} \geq \mathrm{y}^*$, $\rho > 0$ so $w = 1$. Therefore, the optimal $w^*(\mathrm{y})$ that achieves the supremum in Equation (42) is in $\mathcal{W}^H_{\text{nd}}$.

This result holds under

$$\begin{aligned}
\mu(\mathbf{x}, \mathrm{t}) &= \frac{\int_{\mathcal{Y}} \mathrm{y}p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y}w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, && \text{(48a)}\\
&= \frac{\int_{\mathcal{Y}} \mathrm{y}_{\mathrm{t}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} | \mathbf{x})}{p(\mathrm{t} | \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} | \mathbf{x})}{p(\mathrm{t} | \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}, && \text{(48b)}\\
&= \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, && \text{(48c)}
\end{aligned}$$

thus concluding the proof (eq. (48b)-eq. (48c) by Lemma 1). □

D.0.1 Discrete search approaches

Let $\widehat{\mathcal{Y}} = \{\mathrm{y}_i \in \mathcal{Y}\}_{i=1}^k$ be a set of k values of y, then

$$\begin{aligned}
\underline{\mu}^H_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}) &= \min_{\mathrm{y}^*} \left\{ \widehat{\kappa}_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}; \Lambda, H(\mathrm{y}^* - \mathrm{y})) : \mathrm{y}^* \in \widehat{\mathcal{Y}} \right\},\\
\overline{\mu}^H_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}) &= \max_{\mathrm{y}^*} \left\{ \widehat{\kappa}_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}; \Lambda, H(\mathrm{y} - \mathrm{y}^*)) : \mathrm{y}^* \in \widehat{\mathcal{Y}} \right\}.
\end{aligned}$$

$$H(\mathrm{y}) \coloneqq \begin{cases} 1, \mathrm{y} > 0 \\ 0, \mathrm{y} \leq 0 \end{cases}$$

Algorithm 2 Line Search Interval Optimizer

Require: $\mathbf{x}^*$ is an instance of $\mathbf{X}$, t^* is a treatment level to evaluate, Λ is a belief in the amount of hidden confounding, $\boldsymbol{\theta}$ are optimized model parameters, $\widehat{\mathcal{Y}}$ is a set of unique values $\mathrm{y} \in \mathcal{Y}$ *sorted in ascending order*.

function LINESEARCH($\mathbf{x}^*$, t^*, Λ, $\boldsymbol{\theta}$, $\widehat{\mathcal{Y}}$)
  $\overline{\mu} \leftarrow -\infty, \overline{\kappa} \leftarrow \infty$
  $\underline{\mu} \leftarrow \infty, \underline{\kappa} \leftarrow -\infty$
  $\underline{\delta} \leftarrow$ True, $\overline{\delta} \leftarrow$ True
  while $\underline{\delta}$ **do**
    $\mathrm{y}^* \leftarrow \textsc{Pop}(\widehat{\mathcal{Y}}_c)$ ▷ $\widehat{\mathcal{Y}}_c$ a copy of $\widehat{\mathcal{Y}}$
    $\underline{\kappa} \leftarrow \widehat{\kappa}_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}; \Lambda, H(\mathrm{y}^* - \mathrm{y}))$
    if $\underline{\kappa} < \underline{\mu}$ **then**
      $\underline{\mu} \leftarrow \underline{\kappa}$
    else
      $\underline{\delta} \leftarrow$ False
  while $\overline{\delta}$ **do**
    $\mathrm{y}^* \leftarrow \textsc{Pop}(\widehat{\mathcal{Y}}_c)$ ▷ $\widehat{\mathcal{Y}}_c$ a copy of $\widehat{\mathcal{Y}}$
    $\overline{\kappa} \leftarrow \widehat{\kappa}_{\boldsymbol{\theta}}(\mathbf{x}, \mathrm{t}; \Lambda, H(\mathrm{y} - \mathrm{y}^*))$
    if $\overline{\kappa} > \overline{\mu}$ **then**
      $\overline{\mu} \leftarrow \overline{\kappa}$
    else
      $\overline{\delta} \leftarrow$ False
  return $\underline{\mu}, \overline{\mu}$

E Theorem 1

Assume that

1. $m \to \infty$,
2. $n \to \infty$,
3. $(\mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}) \in \mathcal{D}_n$,
4. $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$ converges in measure to $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})$,
5. $\widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})$ is a consistent estimator of $\widetilde{\mu}(\mathbf{x}, \mathrm{t})$,
6. $p(\mathrm{t} \mid \mathrm{y}_\mathrm{t}, \mathbf{x}) > 0, \forall \mathrm{y}_\mathrm{t} \in \mathcal{Y}$.

Then, $\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) \xrightarrow{p} \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)$ and $\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) \xrightarrow{p} \overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)$.

Proof. We prove that $\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) \xrightarrow{p} \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)$, from which $\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) \xrightarrow{p} \overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)$ can be proved analogously. Note that $\xrightarrow{p}$ denotes "convergence in probability". We need to show that $\lim_n P(|\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}_n) - \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)| \geq \epsilon) = 0$, for all $\epsilon > 0$. Where $\boldsymbol{\theta}_n$ are the model parameters corresponding to a dataset $\mathcal{D}_n$ of n observations. Recall that,

$$\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) \coloneqq \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \inf_{w \in \mathcal{W}_{\text{ni}}^H} \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}},$$

and

$$\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}_n) \coloneqq \widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}_n) + \inf_{w \in \mathcal{W}_{\text{ni}}^H} \frac{I_m\left(w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}_n))\right)}{(\Lambda^2 - 1)^{-1} + I_m\left(w(\mathrm{y})\right)},$$

where

$$I_m\left(w(\mathrm{y})(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}_n))\right)=\frac{1}{m}\sum_{i=1}^{m}w(\mathrm{y}_i)(\mathrm{y}_i-\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}_n)),$$

and

$$I_m\left(w(\mathrm{y})\right)=\frac{1}{m}\sum_{i=1}^{m}w(\mathrm{y}_i),$$

with $\mathrm{y}_i \sim p(\mathrm{y} \mid \mathrm{t},\mathbf{x},\boldsymbol{\theta}_n)$.

First, by Item 1 and the law of large numbers, both

$$\lim_{m\to\infty} I_m\left(w(\mathrm{y})(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}_n))\right)=\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}_n))p(\mathrm{y} \mid \mathrm{t},\mathbf{x};\boldsymbol{\theta}_n)d\mathrm{y},$$

and

$$\lim_{m\to\infty} I_m\left(w(\mathrm{y})\right)=\int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t},\mathbf{x};\boldsymbol{\theta}_n)d\mathrm{y}.$$

Therefore,

$$\lim_{m\to\infty}\underline{\mu}(\mathbf{x},\mathrm{t};\Lambda,\boldsymbol{\theta}_n)=\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}_n)+\inf_{w\in\mathcal{W}_{\mathrm{ni}}^{H}}\frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}_n))p(\mathrm{y} \mid \mathrm{t},\mathbf{x};\boldsymbol{\theta}_n)d\mathrm{y}}{(\Lambda^2-1)^{-1}+\int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t},\mathbf{x};\boldsymbol{\theta}_n)d\mathrm{y}}.$$

Note that this step was missed by [JMGS21].

From here, the proof for Theorem 1 from [JMGS21] can be followed, substituting in $(\Lambda^2-1)^{-1}$ where they write $\alpha_{\boldsymbol{\omega}}^{`}$ and $\alpha^{`}$.

□

F Optimization over continuous functions

Second, we need a functional estimator for $w(\mathrm{y},\mathbf{x})$. We use a neural network, $w(\mathrm{y},\mathbf{x};\boldsymbol{\omega})$, parameterized by $\boldsymbol{\omega}$ with sigmoid non-linearity on the output layer to satisfy the $w:\mathcal{Y}\times\mathcal{X}\to[0,1]$ constraint.

For each (Λ,t) pair, we then need to solve the following optimization problems:

$$\underline{\boldsymbol{\omega}}=\arg\min_{\boldsymbol{\omega}}\frac{1}{n}\sum_{i=1}^{n}\mu(w(\mathrm{y},\cdot;\boldsymbol{\omega});\mathbf{x}_i,\mathrm{t},\Lambda,\boldsymbol{\theta}),\quad \mathbf{x}_i\in\mathcal{D},$$

and

$$\overline{\boldsymbol{\omega}}=\arg\min_{\boldsymbol{\omega}}\frac{1}{n}\sum_{i=1}^{n}-\mu(w(\mathrm{y},\cdot;\boldsymbol{\omega});\mathbf{x}_i,\mathrm{t},\Lambda,\boldsymbol{\theta}),\quad \mathbf{x}_i\in\mathcal{D},$$

where

$$\begin{aligned}&\mu(w(\mathrm{y},\cdot;\boldsymbol{\omega});\mathbf{x},\mathrm{t},\Lambda,\boldsymbol{\theta})\\&\quad:=\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta})+\frac{I\left(w(\mathrm{y},\mathbf{x};\boldsymbol{\omega})(\mathrm{y}-\widetilde{\mu}(\mathbf{x},\mathrm{t};\boldsymbol{\theta}))\right)}{(\Lambda^2-1)^{-1}+I(w(\mathrm{y},\mathbf{x};\boldsymbol{\omega}))}.\end{aligned}$$

Each of these problems can then be optimized using stochastic gradient descent [Rud16] and error back-propogation [RHW86]. Since the optimization over $\boldsymbol{\omega}$ is non-convex, guarantees on this strategy finding the optimal solution have yet to be established. As an alternative, the line-search algorithm presented in [JMGS21] can also be used with small modifications. Under the assumptions of Theorem 1 in [JMGS21], with the additional assumption that T is a bounded random variable, we inherit their guarantees on the bound of the conditional average potential outcome.

The upper and lower bounds for the CAPO function under treatment $\mathrm{T}=\mathrm{t}$ and sensitivity parameter Λ can be estimated for any observed covariate value, $\mathbf{X}=\mathbf{x}$, as

$$\underline{\mu}(\mathbf{x},\mathrm{t};\Lambda,\boldsymbol{\theta})=\mu(w(\mathrm{y},\cdot;\underline{\boldsymbol{\omega}});\mathbf{x},\mathrm{t},\Lambda,\boldsymbol{\theta}),$$

and

$$\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) = \mu(w(\mathrm{y}, \cdot; \overline{\boldsymbol{\omega}}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}).$$

The upper and lower bounds for the APO (dose-response) function under treatment $\mathrm{T} = \mathrm{t}$ and sensitivity parameter Λ can be estimated over any set of observed covariates $\mathcal{D}_{\mathbf{x}} = \{\mathbf{x}_i\}_{i=1}^n$, as

$$\underline{\mu}(\mathrm{t}; \Lambda, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \underline{\mu}(\mathbf{x}_i, \mathrm{t}; \Lambda, \boldsymbol{\theta}), \quad \mathbf{x}_i \in \mathcal{D}_{\mathbf{x}},$$

$$\overline{\mu}(\mathrm{t}; \Lambda, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \overline{\mu}(\mathbf{x}_i, \mathrm{t}; \Lambda, \boldsymbol{\theta}), \quad \mathbf{x}_i \in \mathcal{D}_{\mathbf{x}}.$$

G Datasets

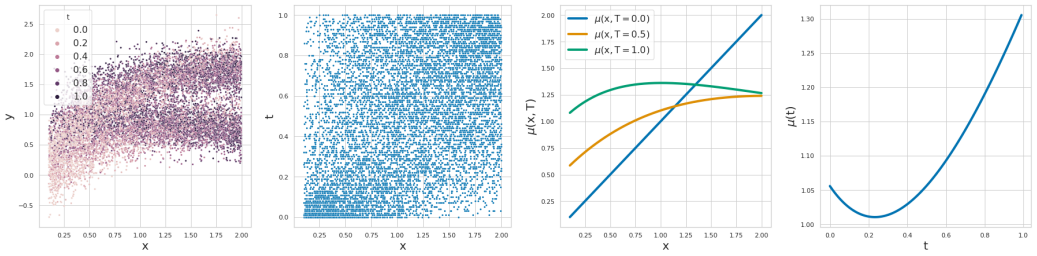
G.1 Synthetic

$$\begin{aligned} \mathrm{u} &\coloneqq N_{\mathrm{u}}, \\ \mathrm{x} &\coloneqq N_{\mathrm{x}}, \\ \mathrm{t} &\coloneqq N_{\mathrm{t}}, \\ \mathrm{y_t} &\coloneqq \mathrm{t} + \mathbf{x} \exp(-\mathrm{tx}) - \gamma_{\mathrm{y}}(\mathrm{u} - 0.5) * (0.5 * \mathrm{x} + 1) + N_{\mathrm{y}}, \end{aligned} \tag{49}$$

where, $N_{\mathrm{u}} \sim p(\mathrm{u}) \coloneqq \text{Bern}(\mathrm{u} \mid 0.5)$, $N_{\mathrm{x}} \sim p(\mathrm{x}) \coloneqq \text{Unif}[\mathrm{x} \mid 0.1, 2.0]$, $N_{\mathrm{t}} \sim p(\mathrm{t} \mid \mathrm{x}, \mathrm{u}) \coloneqq \text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}, \beta = 1)$, and $N_{\mathrm{y}} \sim \mathcal{N}(0, 0.04)$. For the results in this paper $\gamma_t = 0.3$ and $\gamma_y = 0.5$.

The ground truth ratio, $\lambda = \frac{p(\mathrm{t}|\mathrm{x})}{p(\mathrm{t}|\mathrm{x},\mathrm{u})}$, is then given by,

$$\begin{aligned} \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) &= \frac{\mathbb{E}_{p(\mathrm{u})}[p(\mathrm{t} \mid \mathrm{x}, \mathrm{u})]}{p(\mathrm{t} \mid \mathrm{x}, \mathrm{u})} \\ &= \frac{\sum_{\mathrm{u}'=0}^{1} 0.5 * \text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}', \beta = 1)}{\text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}, \beta = 1)} \end{aligned} \tag{50}$$



(a) Observed Outcome (b) Observed Treatment (c) CAPO function (d) APO function

Figure 5: Synthetic data with hidden confounding

G.2 Observations of clouds and aerosol

The Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Aqua satellite observes the Earth twice daily at ∼1 km x 1 km resolution native resolution (Level 1) [BP06]. We used the daily mean, 1° x 1° gridded version (Level 2) in order to somewhat homogenize our observations of clouds and the atmosphere confined to a region off the coast of South America in the Pacific basin. MODIS observations are fed into the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) real-time model in order to emulate the atmosphere and it's components, such as aerosol [GMS+17]. Aerosol optical depth at 550nm from MERRA-2 is derived from MODIS observations of aerosol from multiple satellites (Terra, Aqua, Suomi-NPP), with corrections for sun glint and near-cloud optical effects [BAC+15]. We collocated all gridded observations of clouds and reanalysis aerosol with our meteorological proxies of the environment (EIS, SST, w500, RH700, RH850), then normalized our features before feeding them into the model.

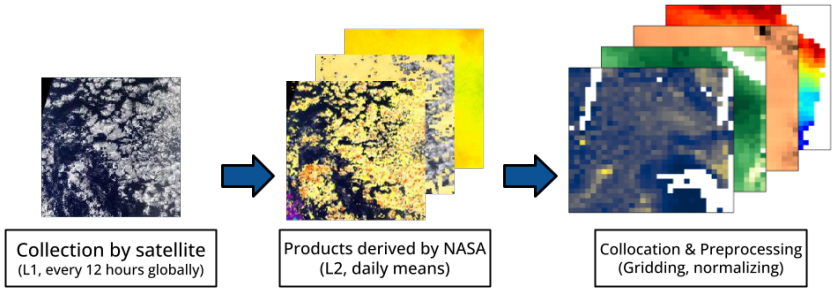


Figure 6: Workflow of observed clouds from satellite to ingestion by model.

Table 1: Sources of satellite observations.

Product name	Description
Cloud optical depth τ	MODIS (1.6, 2.1, 3.7 μm)
Precipitation	NOAA CMORPH
Sea Surface Temperature	NOAA WHOI
Vertical Motion	MERRA-2
Estimated Inversion Strength	MERRA-2
Relative Humidity	MERRA-2
Aerosol Optical Depth	MERRA-2

H Implementation Details

Experiments were run using a single NVIDIA GeForce GTX 1080 ti, an Intel(R) Core(TM) i7-8700K, on a desktop computer with 16GB of RAM. Code is written in python. Packages used include PyTorch [PGM+19], scikit-learn [PVG+11], Ray [MNW+18], NumPy, SciPy, and Matplotlib. We use ray tune [LLN+18] with HyperBand Bayesian Optimization [FKH17] search algorithm to optimize our network hyper-parameters. The hyper-parameters we consider are accounted for in Table 2. The final hyper-parameters used are given in Table 3. The hyper-parameter optimization objective is the batch-wise Pearson correlation averaged across all outcomes of the validation data for a single dataset realization with random seed 1331. All experiments reported can be completed in 30 hours using this setup.

Hyper-parameter	Search Space
hidden units	tune.qlograndint(32, 512, 32)
network depth	tune.randint(2, 5)
gmm components	tune.randint(1, 32)
attention heads	tune.randint(1, 8)
negative slope	tune.quniform(0.0, 0.5, 0.01)
dropout rate	tune.quniform(0.0, 0.5, 0.01)
layer norm	tune.choice([True, False])
batch size	tune.qlograndint(32, 256, 32)
learning rate	tune.quniform(1e-4, 1e-3, 1e-4)

Table 2: Hyper-parameter search space

H.1 Model Architecture

The general model architecture is shown in Figure 7. The models are neural-network architectures with two basic components: a feature extractor, $\phi(\mathbf{x};\boldsymbol{\theta})$ (ϕ, for short), and a conditional outcome prediction block $f(\phi, \mathrm{t};\boldsymbol{\theta})$, or density estimator. The covariates $\mathbf{x}$ (represented in **blue**) are given as input to the feature extractor, whose output is concatenated with the treatment t (represented in **purple**)

Hyper-parameter	Synthetic	ACCE NN	ACCE Transformer
hidden units	96	256	256
network depth	4	3	3
gmm components	24	24	24
attention heads	NA	NA	4
negative slope	0.05	0.04	0.01
dropout rate	0.04	0.2	0.5
layer norm	False	False	False
batch size	32	2048	32
learning rate	0.0015	1e-4	2e-4

Table 3: Final hyper-parameters for each dataset/model

and given as input to the density estimator which outputs a Gaussian mixture density $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$ from which we can sample to obtain samples of the outcomes (represented in **red**). Models are optimized by maximizing the log-likelihood, $\log p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$, using mini-batch stochastic gradient descent.

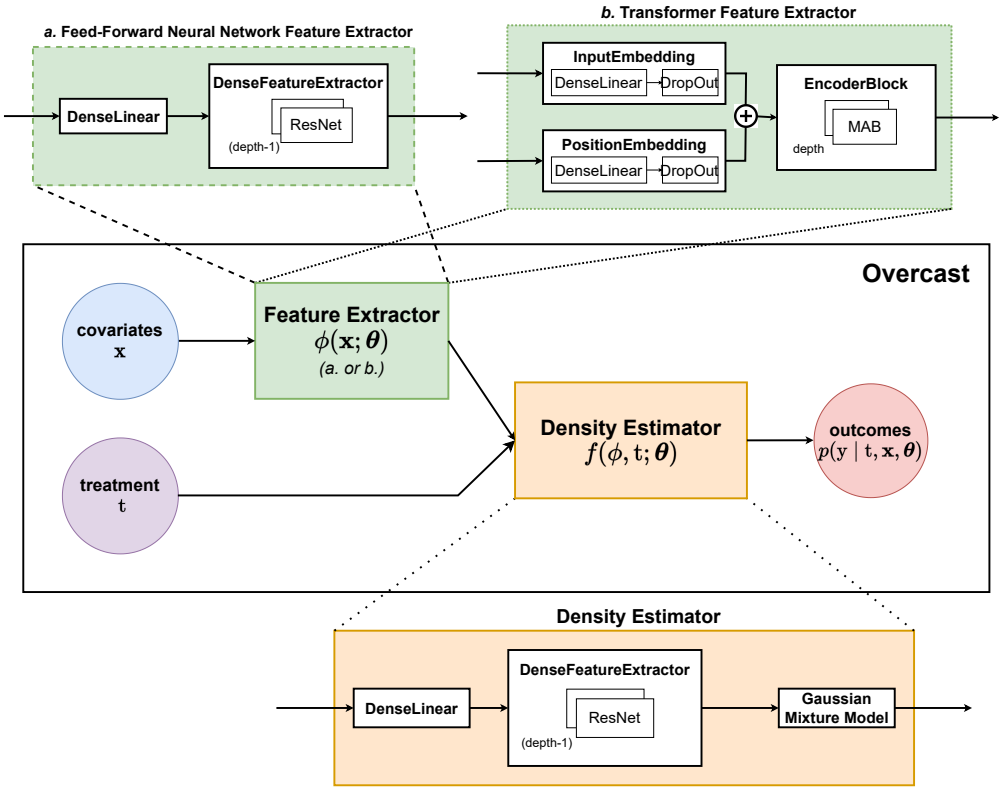


Figure 7: **Overcast model architecture.** The inputs are represented by circles, in **blue** the covariates, and in **purple** the treatment. In the **red** circle is the output of the model, the outcomes distribution. The model has different feature extractors (in green) for the feed-forward neural network and the transformer. It has a single density estimator (in orange).

H.1.1 Feature extractor

The feature extractor design is problem and data specific. In our case, we look at using both a simple feed-forward neural network and also a transformer. The transformer has the advantage of allowing us to model the spatio-temporal correlations between the covariates on a given day using the geographical coordinates of the observations as positional encoding. This is interesting when

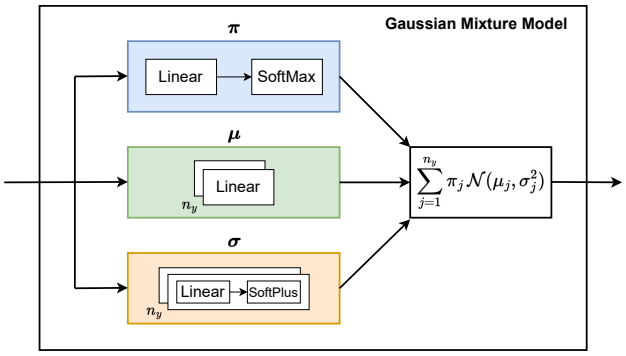


Figure 8: **Overcast Gaussian mixture model.** The mixing coefficients π are estimated with a linear layer and a SoftMax layer, to obtain $\tilde{\pi}$, represented in **blue** in the figure. The vector of means of the Gaussian kernels $\tilde{\mu}$ is obtained by n_y linear layers (in **green** in the diagram), whilst the vector of variances $\tilde{\sigma}$ is obtained by n_y blocks of linear layers and SoftPlus layers (in **orange** in the diagram).

studying ACI because confounding may be latent in the relationships between neighboring variables. Typically, environmental processes (which is one source of confounding) are dependent upon the spatial distribution of clouds, humidity and aerosol, and this feature extractor may capture these confounding effects better.

H.2 Density Estimator

The conditional outcome prediction block, relies on a n_{y} component Gaussian mixture density represented in Figure 8. It outputs:

$$p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{n_{\mathrm{y}}} \widetilde{\pi}_j(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta}) \mathcal{N}\left(\mathrm{y} \mid \widetilde{\mu}_j(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta}), \widetilde{\sigma}_j^2(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta})\right),$$

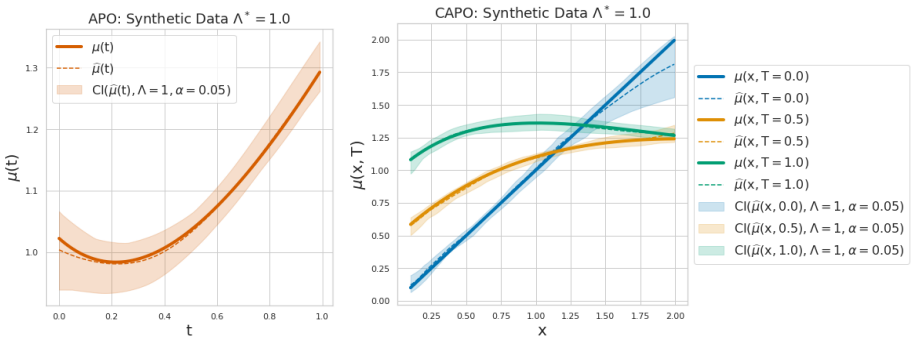
and

$$\widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) = \sum_{j=1}^{n_{\mathrm{y}}} \widetilde{\pi}_j(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta}) \widetilde{\mu}_j(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta}),$$

where $\mathcal{N}(\cdot \mid \mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

I Additional Results

I.1 Synthetic



(a) APO Functions (b) CAPO Functions

Figure 9: Investigating statistical uncertainty using unconfounded synthetic data.

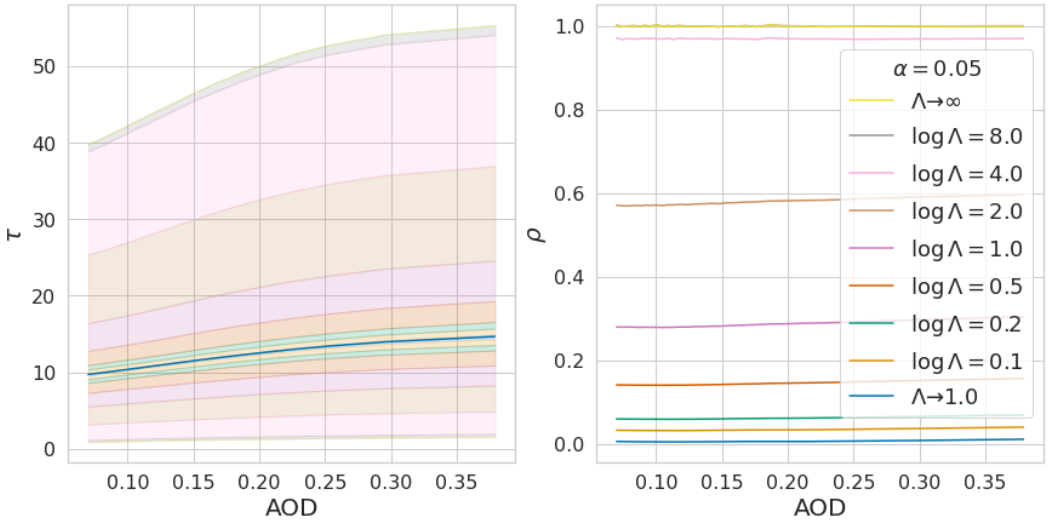
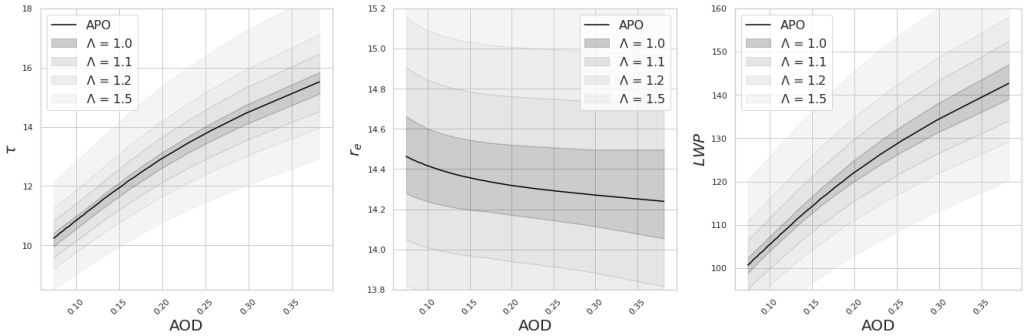


Figure 10: Interpreting Λ as a proportion (ρ) of the unexplained range of $\mathrm{Y_t}$ attributed to unobserved confounding variables.

I.2 Aerosol-Cloud-Climate Effects

In Figure 10 we show how Λ can be interpreted as the proportion, ρ, of the unexplained range of $\mathrm{Y_t}$ attributed to unobserved confounding variables. In the left figure, we plot the corresponding bounds for increasing values of Λ of the predicted AOD-τ dose-response curves. In the right figure we plot the ρ value for each Λ at each value of t. For the curves reported in Section 5.2: we find that $\Lambda = 1.1$ leads to $\rho \approx 0.04$, $\Lambda = 1.2$ leads to $\rho \approx 0.07$, and $\Lambda = 1.6$ leads to $\rho \approx 0.15$. This shows that when we let just a small amount of the unexplained range of $\mathrm{Y_t}$ be attributed to unobserved confounding, the range of the predicted APO curves become quite wide. If we were to completely relax the no-hidden-confounding assumption, the entire range seen in Figure 10 Left would be plausible for the APO function. This range dwarfs the predicted APO curve. These results highlight the importance of reporting such sensitivity analyses.

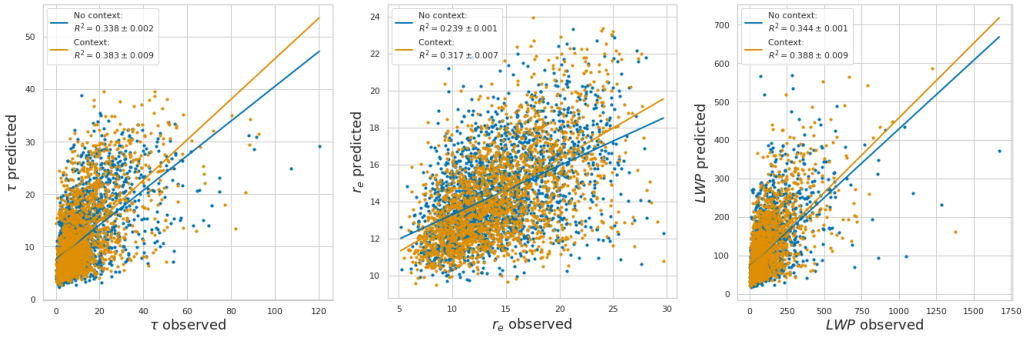
In Figure 11 we show additional dose response curves for cloud optical thickness (τ), water droplet effective radius (r_e), and liquid water path (LWP).



(a) (b) (c)

Figure 11: Average dose-response curves for other cloud properties. a) Cloud optical depth. b) Water droplet effective radius. c) Liquid water path.

In Figure 12 we show additional scatter plots comparing the neural network and transformer models for cloud optical thickness (τ), water droplet effective radius (r_e), and liquid water path (LWP).



(a) (b) (c)

Figure 12: Comparing transformer to feed-forward feature extractor at predicting cloud properties given covariates and AOD. a) Cloud optical depth. b) Water droplet effective radius. c) Liquid water path. We see a significant improvement in pearson correlation (R^2) in each case.

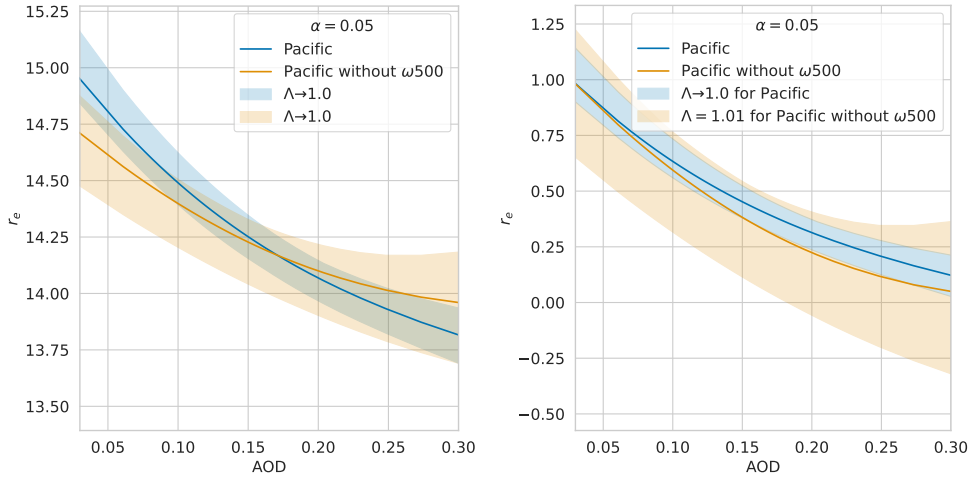
I.3 ω_{500} experiment

The Overcast models make use of expert knowledge about ACI to select the covariates. Ideally, they would include pressure profiles, temperature profiles and supersaturation since these are directly involved in cloud processes and impact the quality of AOD measurements as a proxy for aerosol concentration. Unfortunately, they are impossible to retrieve from satellite data, so we rely on meteorological proxies like relative humidity, sea surface temperature, inversion strengths, and vertical motion. Relying on these proxies however results in ignorability violations, which must be accounted for in the parameter Λ in order to derive appropriate plausible ranges of outcomes.

In the experiment that follows, we are removing a confounding variable from the model, therefore inducing hidden confounding. The covariate we remove is vertical motion at 500 mb, denoted by $\omega 500$. This experiment helps us gain some intuition about the influence of the parameter Λ and how it relates to the inclusion of confounding variables in the model.

In Figure 13 we compare the same region with different covariates to identify an appropriate Λ. We fit one model on data from the Pacific (blue) and one model from the Pacific omitting $\omega 500$ from the covariates (orange). The shaded bounds in blue are the ignorance region for $\Lambda \to 1$ for the Pacific. We then find the Λ that results in an ignorance interval around the Pacific omitting $\omega 500$ that covers the Pacific model prediction. From this, we can infer how the parameter Λ relates to the inclusion of covariates in the model. We show that we need to set $\Lambda = 1.01$ to account for the fact that we are omitting $\omega 500$ from our list of covariates. We also note that the slopes of the dose-response curves are slightly different, with worse predictions when omitting $\omega 500$ from the covariates, as expected.

This work attempts to set a new methodology for setting Λ which can be summarised as followed. Working with two datasets, which vary in only aspect, we train two different models: (i), the control model, and (ii), the experimental model. After training both models, we plot the dose-response curves for (i) and (ii) on the same plot. We can compare the shape and slope of these curves as well as their uncertainty bounds under the unconfoundedness assumption by plotting the ignorance region for $\Lambda \to 1$ for both models. Then, we are interested in setting Λ for model (ii) such that the uncertainty bounds cover the entire ignorance region of model (i) under the unconfoundedness assumption. For this, we are interested in comparing the slopes and thus min-max scale both curves.



(a) Unscaled

(b) Scaled by min-max with appropriate Λ

Figure 13: Dose-response curves with or without vertical motion at 500 mb ($\omega 500$) as a covariate.