Scalable Sensitivity and Uncertainty Analyses for Causal-Effect Estimates of Continuous-Valued Interventions

Andrew Jesson*
OATML
Department of Computer Science
University of Oxford

Alyson Douglas
AOPP
Department of Physics
University of Oxford

Peter Manshausen
AOPP
Department of Physics
University of Oxford

Maëlys Solal
Department of Computer Science
University of Oxford

Nicolai Meinshausen
Seminar for Statistics
Department of Mathematics
ETH Zurich

Philip Stier
AOPP
Department of Physics
University of Oxford

Yarin Gal
OATML
Department of Computer Science
University of Oxford

Uri Shalit
Machine Learning and Causal Inference in Healthcare Lab
Technion – Israel Institute of Technology

Abstract

Estimating the effects of continuous-valued interventions from observational data is a critically important task for climate science, healthcare, and economics. Recent work focuses on designing neural network architectures and regularization functions to allow for scalable estimation of average and individual-level dose-response curves from high-dimensional, large-sample data. Such methodologies assume ignorability (observation of all confounding variables) and positivity (observation of all treatment levels for every covariate value describing a set of units), assumptions problematic in the continuous treatment regime. Scalable sensitivity and uncertainty analyses to understand the ignorance induced in causal estimates when these assumptions are relaxed are less studied. Here, we develop a continuous treatment-effect marginal sensitivity model (CMSM) and derive bounds that agree with the observed data and a researcher-defined level of hidden confounding. We introduce a scalable algorithm and uncertainty-aware deep models to derive and estimate these bounds for high-dimensional, large-sample observational data. We work in concert with climate scientists interested in the climatological impacts of human emissions on cloud properties using satellite observations from the past 15 years. This problem is known to be complicated by many unobserved confounders.

1 Introduction

Understanding the causal effect of a continuous variable (termed "treatment") on individual units and subgroups is crucial across many fields. In economics, we might like to know the effect of price on demand from different customer demographics. In healthcare, we might like to know the effect of medication dosage on health outcomes for patients of various ages and comorbidities. And in

*Correspondence to `andrew.jesson@cs.ox.ac.uk`

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

climate science, we might like to know the effects of anthropogenic emissions on cloud formation and lifetimes under variable atmospheric conditions. In many cases, these effects must be estimated from observational data as experiments are often costly, unethical, or otherwise impossible to conduct.

Estimating causal effects from observational data can only be done under certain conditions, some of which are not testable from data. The most prominent are the common assumptions that all confounders between treatment and outcome are measured ("no hidden confounders"), and any level of treatment could occur for any observable covariate vector ("positivity"). These assumptions and their possible violations introduce uncertainty when estimating treatment effects. Estimating this uncertainty is crucial for decision-making and scientific understanding. For example, understanding how unmeasured confounding can change estimates about the impact of emissions on cloud properties can help to modify global warming projection models to account for the uncertainty it induces.

We present a novel marginal sensitivity model for continuous treatment effects. This model is used to develop a method that gives the user a corresponding interval representing the "ignorance region" of the possible treatment outcomes per covariate and treatment level [D'A19] for a specified level of violation of the no-hidden confounding assumption. We adapt prior work [Tan06, KMZ19, JMGS21] to the technical challenge presented by continuous treatments. Specifically, we modify the existing model to work with propensity score densities instead of propensity score probabilities (see Section 3 below) and propose a method to relate ignorability violations to the unexplained range of outcomes. Further, we derive bootstrapped uncertainty intervals for the estimated ignorance regions and show how to efficiently compute the intervals, thus providing a method for quantifying the uncertainty presented by finite data and possible violations of the positivity assumption. We validate our methods on synthetic data and provide an application on real-world satellite observations of the effects of anthropogenic emissions on cloud properties. For this application, we develop a new neural network architecture for estimating continuous treatment effects that can take into account spatiotemporal covariates. We find that the model accurately captures known patterns of cloud deepening in response to anthropogenic emission loading with realistic intervals of uncertainty due to unmodeled confounders in the satellite data.

2 Problem Setting

Let the random variable $\mathbf{X} \in \mathcal{X}$ model observable covariates. For clarity, we will assume that $\mathcal{X}$ is a d-dimensional continuous space: $\mathcal{X} \subseteq \mathbb{R}^d$, but this does not preclude more diverse spaces. Instances of $\mathbf{X}$ are denoted by $\mathbf{x}$. The observable continuous treatment variable is modeled as the random variable $\mathrm{T} \in \mathcal{T} \subseteq \mathbb{R}$. Instances of T are denoted by t. Let the random variable $\mathrm{Y} \in \mathcal{Y} \subseteq \mathbb{R}$ model the observable continuous outcome variable. Instances of Y are denoted by y. Using the Neyman-Rubin potential outcomes framework [Ney23, Rub74, Sek08], we model the potential outcome of a treatment level t by the random variable $\mathrm{Y_t} \in \mathcal{Y}$. Instances of $\mathrm{Y_t}$ are denoted by $\mathrm{y_t}$. We assume that the observational data, $\mathcal{D}_n$, consists of n realizations of the random variables, $\mathcal{D}_n = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)\}_{i=1}^n$. We let the observed outcome be the potential outcome of the assigned treatment level, $\mathrm{y}_i = \mathrm{y}_{\mathrm{t}_i}$, thus assuming non-interference and consistency [Rub80]. Moreover, we assume that the tuple $(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)$ are i.i.d. samples from the joint distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y_T})$, where $\mathrm{Y_T} = \{\mathrm{Y_t} : \mathrm{t} \in \mathcal{T}\}$.

We are interested in the **conditional average potential outcome (CAPO)** function, $\mu(\mathbf{x}, \mathrm{t})$, and the **average potential outcome (APO)** — or dose-response function — $\mu(\mathrm{t})$, for continuous valued treatments. These functions are defined by the expectations:

$$\mu(\mathbf{x}, \mathrm{t}) \coloneqq \mathbb{E}\left[\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}\right] \quad (1) \qquad \mu(\mathrm{t}) \coloneqq \mathbb{E}\left[\mu(\mathbf{X}, \mathrm{t})\right]. \quad (2)$$

Under the assumptions of ignorability, $\mathrm{Y_T} \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$, and positivity, $p(\mathrm{t} \mid \mathbf{X} = \mathbf{x}) > 0 : \forall \mathrm{t} \in \mathcal{T}, \forall \mathbf{x} \in \mathcal{X}$ — jointly known as *strong ignorability* [RR83] — the CAPO and APO are identifiable from the observational distribution $P(\mathbf{X}, \mathrm{T}, \mathrm{Y_T})$ as:

$$\widetilde{\mu}(\mathbf{x}, \mathrm{t}) = \mathbb{E}\left[\mathrm{Y} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}\right] \quad (3) \qquad \widetilde{\mu}(\mathrm{t}) = \mathbb{E}\left[\widetilde{\mu}(\mathbf{X}, \mathrm{t})\right]. \quad (4)$$

In practice, however, these assumptions rarely hold. For example, there will almost always be unobserved confounding variables, thus violating the ignorability (also known as unconfoundedness or exogeneity) assumption, $\mathrm{Y_T} \not\perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$. Moreover, due to both the finite sample of observed data, $\mathcal{D}$, and also the continuity of treatment T, there will most certainly be values, $\mathrm{T} = \mathrm{t}$, that are unobserved for a given covariate measurement, $\mathbf{X} = \mathbf{x}$, leading to violations or near violations of the positivity assumption (also known as overlap).

3 Methods

We propose the continuous marginal sensitivity model (CMSM) as a new marginal sensitivity model (MSM [Tan06]) for continuous treatment variables. The set of 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 data, $\mathcal{D}$. But, the set of marginal distributions of the potential outcomes, $\{P(\mathrm{Y_t} \mid, \mathbf{X} = \mathbf{x}) : \mathrm{t} \in \mathcal{T}\}$, each given as a continuous mixture,

$$P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = \int_{\mathcal{T}} p(\mathrm{t}' \mid \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}) d\mathrm{t}',$$

are not. This is due to the general unidentifiability of the component distributions, $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$, where $\mathrm{Y_t}$ cannot be observed for units under treatment level $\mathrm{T} = \mathrm{t}'$ for $\mathrm{t}' \neq \mathrm{t}$: the well-known "fundamental problem of causal inference" [Hol86]. Yet, under the ignorability assumption, the factual $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and counterfactual $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x})$ are equal for all $\mathrm{t}' \in \mathcal{T}$. Thus, $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ and $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ are identical, and any divergence between them is indicative of hidden confounding. But, such divergence is not observable in practice.

The CMSM supposes a degree of divergence between the unidentifiable $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ and the identifiable $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ by assuming that the rate of change of $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ with respect to $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ is bounded by some value greater than or equal to 1. The Radon-Nikodym derivative formulates the divergence, $\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) = \frac{dP(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})}{dP(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})}$, under the assumption that $P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x})$ is absolutely continuous with respect to $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$, $\forall \mathrm{t} \in \mathcal{T}$.

Proposition 1. *Under the additional assumption that $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$ and the Lebesgue measure are mutually absolutely continuous, the Radon-Nikodym derivative above is equal to the ratio between the unidentifiable "complete" propensity density for treatment $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$ and the identifiable "nominal" propensity density for treatment $p(\mathrm{t} \mid \mathbf{x})$,*

$$\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t}) = \frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}, \tag{5}$$

Proof (Appendix A.3) and an analysis of this proposition are given in Appendix A.

The value $\lambda(\mathrm{y_t}; \mathbf{x}, \mathrm{t})$ cannot be identified from the observational data alone; the merit of the CMSM is that enables a domain expert to express their belief in what is a plausible degree hidden confounding through the parameter $\Lambda \geq 1$. Where, $\Lambda^{-1} \leq p(\mathrm{t} \mid \mathbf{x}) / p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x}) \leq \Lambda$, reflects a hypothesis that the "complete", unidentifiable propensity density for subjects with covariates $\mathbf{X} = \mathbf{x}$ can be different from the observable "nominal" propensity density by at most a factor of Λ. These inequalities allow for the specification of user hypothesized complete propensity density functions, $p(\mathrm{t} \mid \mathrm{y}, \mathbf{x})$, and we define the CMSM as the set of such functions that agree with the inequalities.

Definition 1. *Continuous Marginal Sensitivity Model (CMSM)*

$$\mathcal{P}(\Lambda) \coloneqq \left\{ p(\mathrm{t} \mid \mathrm{y}, \mathbf{x}) : \frac{1}{\Lambda} \leq \frac{p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} \leq \Lambda, \forall \mathrm{y} \in \mathbb{R}, \forall \mathbf{x} \in \mathcal{X} \right\} \tag{6}$$

Remark. Note that the CMSM is defined in terms of a *density ratio*, $p(\mathrm{t} \mid \mathbf{x}) / p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$, whereas the MSM for binary-valued treatments is defined in terms of an *odds ratio*, $\frac{P(\mathrm{t} \mid \mathbf{x})}{(1 - P(\mathrm{t} \mid \mathbf{x}))} / \frac{P(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})}{(1 - P(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x}))}$. Importantly, naively substituting densities into the MSM for binary-treatments would violate the condition that $\lambda > 0$ as the densities $p(\mathrm{t} \mid \mathbf{x})$ or $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$ can each be greater than one, which would result in a negative $1 - p(\mathrm{t} \mid \cdot)$. The odds ratio is familiar to practitioners. The density ratio is less so. We offer a transformation of the sensitivity analysis parameter Λ in terms of the unexplained range of the outcome later.

3.1 Continuous Treatment Effect Bounds Without Ignorability

The CAPO and APO (dose-response) functions cannot be point identified from observational data without ignorability. Under the CMSM with a given Λ, we can only identify a set of CAPO and APO functions jointly consistent with the observational data $\mathcal{D}$ and the continuous marginal sensitivity model. All of the functions in this set are possible from the point of view of the observational data

alone. So to cover the range of all possible functional values, we seek an interval function that maps covariate values, $\mathbf{X} = \mathbf{x}$, to the upper and lower bounds of this set for every treatment value, t.

For $\mathrm{t} \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{X}$, let $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})$ denote the density of the distribution $P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x})$. As a reminder, this distribution is identifiable from observational data, but without further assumptions the CAPO, $\mu(\mathbf{x}, \mathrm{t}) = \mathbb{E}\left[\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}\right]$, is not. We can express the CAPO in terms of its identifiable and unidentifiable components as

$$\begin{aligned}\mu(\mathbf{x}, \mathrm{t}) &= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t}|\mathrm{t},\mathbf{x})}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t}|\mathrm{t},\mathbf{x})}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})} d\mathrm{y_t}} = \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y}, \mathbf{x})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}, \\ &\equiv \mu(w(\mathrm{y}, \mathbf{x}); \mathbf{x}, \mathrm{t}, \Lambda)\end{aligned} \tag{7}$$

where, by a one-to-one change of variables, $\frac{1}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})} = \frac{1}{\Lambda p(\mathrm{t}|\mathbf{x})} + w(\mathrm{y}, \mathbf{x})(\frac{\Lambda}{p(\mathrm{t}|\mathbf{x})} - \frac{1}{\Lambda p(\mathrm{t}|\mathbf{x})})$ with $w : \mathcal{Y} \times \mathcal{X} \to [0, 1]$. Both [KMZ19] and later [JMGS21] provide analogous expressions for the CAPO in the discrete treatment regime under the MSM, and we provide our derivation in Lemma 1.

The uncertainty set that includes all possible values of $w(\mathrm{y}, \mathbf{x})$ that agree with the CMSM, *i.e.*, the set of functions that violate ignorability by no more than Λ, can now be expressed as $\mathcal{W} = \{w : w(\mathrm{y}, \mathbf{x}) \in [0, 1] \quad \forall \mathrm{y} \in \mathcal{Y}, \forall \mathbf{x} \in \mathcal{X}\}$.

With this set of functions, we can now define the CAPO and APO bounds under the CMSM. The CAPO lower, $\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)$, and upper, $\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)$, bounds under the CMSM with parameter Λ are:

$$\begin{aligned}\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) &\coloneqq \inf_{w \in \mathcal{W}} \mu(w(\mathrm{y}, \mathbf{x}); \mathbf{x}, \mathrm{t}, \Lambda) \\ &= \inf_{w \in \mathcal{W}^H_{\text{ni}}} \mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda)\end{aligned} \tag{8}$$

$$\begin{aligned}\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) &\coloneqq \sup_{w \in \mathcal{W}} \mu(w(\mathrm{y}, \mathbf{x}); \mathbf{x}, \mathrm{t}, \Lambda) \\ &= \sup_{w \in \mathcal{W}^H_{\text{nd}}} \mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda)\end{aligned} \tag{9}$$

Where the sets $\mathcal{W}^H_{\text{ni}} = \{w : w(y) = H(\mathrm{y}_H - \mathrm{y})\}_{\mathrm{y}_H \in \mathcal{Y}}$, and $\mathcal{W}^H_{\text{nd}} = \{w : w(y) = H(\mathrm{y} - \mathrm{y}_H)\}_{\mathrm{y}_H \in \mathcal{Y}}$, and $H(\cdot)$ is the Heaviside step function. Lemma 2 in appendix D proves the equivalence in eq. (9) for bounded Y. The equivalence in eq. (8) can be proved analogously.

The APO lower, $\underline{\mu}(\mathrm{t}; \Lambda)$, and upper, $\overline{\mu}(\mathrm{t}; \Lambda)$, bounds under the CMSM with parameter Λ are:

$$\underline{\mu}(\mathrm{t}; \Lambda) \coloneqq \mathbb{E}\left[\underline{\mu}(\mathbf{X}, \mathrm{t}; \Lambda)\right] \tag{10}$$

$$\overline{\mu}(\mathrm{t}; \Lambda) \coloneqq \mathbb{E}\left[\overline{\mu}(\mathbf{X}, \mathrm{t}; \Lambda)\right] \tag{11}$$

Remark. It is worth pausing here and breaking down Equation (7) to get an intuitive sense of how the specification of Λ in the CMSM affects the bounds on the causal estimands. When $\Lambda \to 1$, then the $(\Lambda^2 - 1)^{-1}$ term (and thus the denominator) in Equation (7) tends to infinity. As a result, the CAPO under Λ converges to the empirical estimate of the CAPO — $\mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda \to 1) \to \widetilde{\mu}(\mathbf{x}, \mathrm{t})$ — as expected. Thus, the supremum and infimum in Equations (8) and (9) become independent of w, and the ignorance intervals concentrate on point estimates. Next, consider complete relaxation of the ignorability assumption, $\Lambda \to \infty$. Then, the $(\Lambda^2 - 1)^{-1}$ term tends to zero, and we are left with,

$$\mu(w; \cdot, \Lambda \to \infty) \to \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}))p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y}}{\int_{\mathcal{Y}} w(\mathrm{y})p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})d\mathrm{y},} = \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \mathop{\mathbb{E}}_{p(w(\mathrm{y})|\mathbf{x},\mathrm{t})}[\mathrm{Y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t})],$$

where, $p(w(\mathrm{y}) \mid \mathbf{x}, \mathrm{t}) \equiv \frac{w(\mathrm{y})p(\mathrm{y}|\mathrm{t},\mathbf{x})}{\int_{\mathcal{Y}} w(\mathrm{y}')p(\mathrm{y}'|\mathrm{t},\mathbf{x})d\mathrm{y}'}$, a distribution over Y given $\mathbf{X} = x$ and $\mathrm{T} = \mathrm{t}$. Thus, when we *relax* the ignorability assumption entirely, the CAPO can be anywhere in the range of Y.

The parameter Λ relates to the proportion of unexplained range in Y assumed to come from unobserved confounders after observing $\mathbf{x}$ and t. When a user sets Λ to 1, they assume that the entire unexplained range of Y comes from unknown mechanisms independent of T. As the user increases Λ, they attribute some of the unexplained range of Y to mechanisms causally connected to T. For bounded $\mathrm{Y_t}$, this proportion can be calculated as:

$$\rho(\mathbf{x}, \mathrm{t}; \Lambda) \coloneqq \frac{\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) - \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)}{\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda \to \infty) - \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda \to \infty)} = \frac{\overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda) - \underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda)}{\mathrm{y}_{\max} - \mathrm{y}_{\min} \mid \mathbf{X} = x, \mathrm{T} = \mathrm{t}}.$$

The user can sweep over a set of Λ values and report the bounds corresponding to a ρ value they deem tolerable (e.g., $\rho = 0.5$ yields bounds for the assumption that half the unexplained range in Y is due

to unobserved confounders). For unbounded outcomes, the limits can be estimated empirically by increasing Λ to a large value. Refer to Figure 10 in the appendix for a comparison between ρ and Λ.

For another way to interpret Λ, in Appendix A.3.1 we Λ can be presented as a bound on the Kullback–Leibler divergence between the nominal and complete propensity scores through the relationship: $|\log(\Lambda)| \geq D_{\mathrm{KL}}\left(P(\mathrm{Y_t} \mid \mathrm{T}=\mathrm{t}, \mathbf{X}=\mathbf{x}) || P(\mathrm{Y_t} \mid \mathbf{X}=\mathbf{x})\right)$.

3.2 Semi-Parametric Interval Estimator

Following [JMGS21], we develop a semi-parametric estimator of the bounds in eqs. (8) to (11). Under assumption Λ, the bounds on the expected potential outcome over $\mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda)$ are completely defined in terms of identifiable quantities: namely, the conditional density of the outcome given the assigned treatment and measured covariates, $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})$; and the conditional expected outcome $\widetilde{\mu}(\mathbf{x}, \mathrm{t})$. Thus, we define a density estimator, $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$, and estimator, $\mu(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})$, parameterized by instances $\boldsymbol{\theta}$ of the random variable $\boldsymbol{\Theta}$. The choice of density estimator is ultimately up to the user and will depend on the scale of the problem examined and the distribution of the outcome variable Y. In Section 3.5, we will outline how to define appropriate density estimators for high-dimensional, large-sample, continuous-valued treatment problems. Next, we need an estimator of the integrals in $\mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta})$, eq. (7). We use Monte-Carlo (MC) integration to estimate the expectation of arbitrary functions $h(y)$ with respect to the parametric density estimate $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$: $I(h(\mathrm{y})) := \frac{1}{m}\sum_{i=1}^{m} h(\mathrm{y}_i), \quad \mathrm{y}_i \sim p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$. We outline how the Gauss-Hermite quadrature rule is an alternate estimator of these expectations in Appendix C. The integral estimators allow for the semi-parametric estimators for the CAPO and APO bounds under the CMSM to be defined.

Algorithm 1 Grid 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 $\{\mathrm{y} \sim p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})\}$.

```
function GRIDSEARCH(x, t, Λ, θ, Ŷ)
    μ̄ ← −∞, ȳ ← 0
    μ̲ ← ∞, y̲ ← 0
    for y_H ∈ Ŷ do
        κ̄ ← μ(H(y − y_H); x, t, Λ, θ)
        κ̲ ← μ(H(y_H − y); x, t, Λ, θ)
        if κ̄ > μ̄ then
            μ̄ ← κ̄, ȳ ← y_H
        if κ̲ < μ̲ then
            μ̲ ← κ̲, y̲ ← y_H
    return y̲, ȳ
```

The semi-parametric CAPO bound estimators under the CMSM with sensitivity parameter Λ are:

$$\underline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) := \inf_{w \in \mathcal{W}_{\mathrm{ni}}^{H}} \mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}) \quad (12) \qquad \overline{\mu}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) := \sup_{w \in \mathcal{W}_{\mathrm{nd}}^{H}} \mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}) \quad (13)$$

where,

$$\mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}) \equiv \widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) + \frac{I\left(w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}))\right)}{(\Lambda^2 - 1)^{-1} + I\left(w(\mathrm{y})\right)}.$$

The semi-parametric APO bound estimators under the CMSM with sensitivity parameter Λ are:

$$\underline{\mu}(\mathrm{t}; \Lambda, \boldsymbol{\theta}) := \mathbb{E}\left[\underline{\mu}(\mathbf{X}, \mathrm{t}; \Lambda, \boldsymbol{\theta})\right] \quad (14) \qquad \overline{\mu}(\mathrm{t}; \Lambda, \boldsymbol{\theta}) := \mathbb{E}\left[\overline{\mu}(\mathbf{X}, \mathrm{t}; \Lambda, \boldsymbol{\theta})\right] \quad (15)$$

Theorem 1. *In the limit of data ($n \to \infty$) and MC samples ($m \to \infty$), for observed $(\mathbf{X}=\mathbf{x}, \mathrm{T}=\mathrm{t}) \in \mathcal{D}_n$, we assume that $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$ converges in measure to $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})$, $\widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta})$ is a consistent estimator of $\widetilde{\mu}(\mathbf{x}, \mathrm{t})$, and $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})$ is bounded away from 0 uniformly for all $\mathrm{y_t} \in \mathcal{Y}$. 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 in Appendix E.*

3.3 Solving for w

We are interested in a scalable algorithm to compute the intervals on the CAPO function, eqs. (12) and (13), and the APO (dose-response) function, eqs. (14) and (15). The need for scalability stems not only from dataset size. The intervals also need to be evaluated for arbitrarily many values of the continuous treatment variable, t, and the sensitivity parameter Λ. The bounds on the CAPO function can be calculated independently for each instance $\mathbf{x}$, and the limits on the APO are an expectation over the CAPO function bounds.

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

$$\widehat{\underline{\mu}}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) \coloneqq \mu(H(\underline{\mathrm{y}} - \mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}),$$
$$\widehat{\overline{\mu}}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}) \coloneqq \mu(H(\mathrm{y} - \overline{\mathrm{y}}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}),$$

where $\underline{\mathrm{y}}$ and $\overline{\mathrm{y}}$ are found using Algorithm 1. See Algorithm 2 and Appendix F for optional methods.

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

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

3.4 Uncertainty about the Continuous Treatment Effect Interval

Following [ZSB19], [DG21], and [CCN+21], we construct $(1-\alpha)$ statistical confidence intervals for the upper and lower bounds under the CMSM using the percentile bootstrap estimator. [JMSG20] and [JMGS21] have shown that statistical uncertainty is appropriately high for regions with poor overlap. Let $P_{\mathcal{D}}$ be the empirical distribution of the observed data sample, $\mathcal{D} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^n = \{\mathbf{S}_i\}_{i=1}^n$. Let $\widehat{P}_{\mathcal{D}} = \{\widehat{\mathcal{D}}_k\}_{k=1}^{n_b}$ be the bootstrap distribution over n_b datasets, $\widehat{\mathcal{D}}_k = \{\widehat{\mathbf{S}}_i\}_{i=1}^n$, sampled with replacement from the empirical distribution, $P_{\mathcal{D}}$. Let Q_α be the α-quantile of $\mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta})$ in the bootstrap resampling distribution: $Q_\alpha \coloneqq \inf_{\mu^*}\left\{\widehat{P}_{\mathcal{D}}(\mu(w(\mathrm{y}); \mathbf{x}, \mathrm{t}, \Lambda, \boldsymbol{\theta}) \leq \mu^*) \geq \alpha\right\}$. Finally, let $\boldsymbol{\theta}_k$ be the parameters of the model of the k-th bootstrap sample of the data. Then, the bootstrap confidence interval of the upper and lower bounds of the CAPO function under the CMSM is given by: $\mathrm{CI}_b\left(\mu(\mathbf{x}, \mathrm{t}; \Lambda, \alpha)\right) \coloneqq \left[\underline{\mu}_b(\mathbf{x}, \mathrm{t}; \Lambda, \alpha), \overline{\mu}_b(\mathbf{x}, \mathrm{t}; \Lambda, \alpha)\right]$, where,

$$\underline{\mu}_b(\mathbf{x}, \mathrm{t}; \Lambda, \alpha) = Q_{\alpha/2}\left(\left\{\widehat{\underline{\mu}}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}_k)\right\}_{k=1}^b\right), \ \overline{\mu}_b(\mathbf{x}, \mathrm{t}; \Lambda, \alpha) = Q_{1-\alpha/2}\left(\left\{\widehat{\overline{\mu}}(\mathbf{x}, \mathrm{t}; \Lambda, \boldsymbol{\theta}_k)\right\}_{k=1}^b\right).$$

Furthermore, the bootstrap confidence interval of the upper and lower bounds of the APO (dose-response) function under the CMSM are given by: $\mathrm{CI}_b\left(\mu(\mathrm{t}; \Lambda, \alpha)\right) \coloneqq \left[\underline{\mu}_b(\mathrm{t}; \Lambda, \alpha), \overline{\mu}_b(\mathrm{t}; \Lambda, \alpha)\right]$, where,

$$\underline{\mu}_b(\mathrm{t}; \Lambda, \alpha) = Q_{\alpha/2}\left(\left\{\widehat{\underline{\mu}}(\mathrm{t}; \Lambda, \boldsymbol{\theta}_k)\right\}_{k=1}^b\right), \qquad \overline{\mu}_b(\mathrm{t}; \Lambda, \alpha) = Q_{1-\alpha/2}\left(\left\{\widehat{\overline{\mu}}(\mathrm{t}; \Lambda, \boldsymbol{\theta}_k)\right\}_{k=1}^b\right).$$

3.5 Scalable Continuous Treatment Effect Estimation

Following [SJS17], [SLB+20], and [NYLN21], we propose using neural-network architectures with two basic components: a feature extractor, $\boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\theta})$ ($\boldsymbol{\phi}$, for short) and a conditional outcome prediction block $f(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta})$. The feature extractor design will be problem and data specific. In Section 5, we look at using both a simple feed-forward neural network, and also a transformer [VSP+17]. For the conditional outcome block, we depart from more complex structures ([SLB+20, NYLN21]) and simply focus on a residual [HZRS16], feed-forward, S-learner [KSBY19] structure. For the final piece of the puzzle, we follow [JMGS21] and propose a n_{y} component Gaussian mixture density:

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

and $\widetilde{\mu}(\mathbf{x}, \mathrm{t}; \boldsymbol{\theta}) = \sum_{j=1}^{n_{\mathrm{y}}} \widetilde{\pi}_j(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta}) \widetilde{\mu}_j(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta})$ [Bis94]. Models are optimized by maximizing the log-likelihood of $p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}, \boldsymbol{\theta})$.

4 Related Works

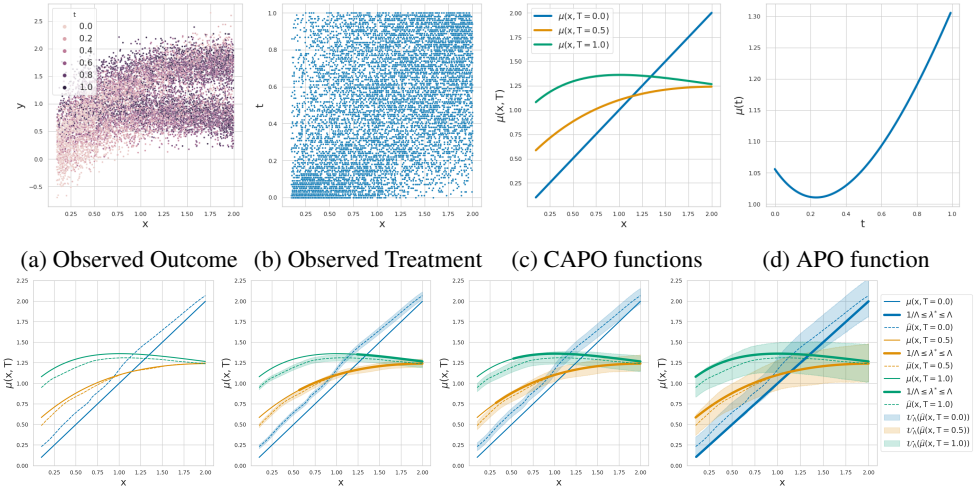
Scalable Continuous Treatment Effect Estimation. Using neural networks to provide scalable solutions for estimating the effects of continuous-valued interventions has received significant

attention in recent years. [BJvdS20] provide a Generative Adversarial Network (GAN) approach. The dose-response network (DRNet) [SLB+20] provides a more direct adaptation of the TarNet [SJS17] architecture for continuous treatments. The varying coefficient network VCNet [NYLN21] generalizes the DRNet approach and provides a formal result for incorporating the target regularization technique presented by [SBV19]. The RieszNet [CCN+21] provides an alternative approach for targeted regularization. Adaptation of each method is straightforward for use in our sensitivity analysis framework by replacing the outcome prediction head of the model with a suitable density estimator.

Sensitivity and Uncertainty Analyses for Continuous Treatment Effects. The prior literature for continuous-valued treatments has focused largely on parametric methods assuming linear treatment/outcome, hidden-confounder/treatment, and hidden-confounder/outcome relationships [CHH16, DHCH16, MSDH16, Ost19, CH20a, CH20b]. In addition to linearity, these parametric methods need to assume the structure and distribution of the unobserved confounding variable(s). [CKC+19] allows for sensitivity analysis for arbitrary structural causal models under the linearity assumption. The MSM relaxes both the distributional and linearity assumptions, as does our CMSM extension. A two-parameter sensitivity model based on Riesz-Frechet representations of the target functionals, here the APO and CAPO, is proposed by [CCN+21] as a way to incorporate confidence intervals and sensitivity bounds. In contrast, we use the theoretical background of the marginal sensitivity model to derive a one-parameter sensitivity model. [DBSC21] purport to quantify the bias induced by unobserved confounding in the effects of continuous-valued interventions, but they do not present a formal sensitivity analysis. Simultaneously and independently of this work, [MVSG] are deriving a sensitivity model that bounds the partial derivative of the log density ratio between complete and nominal propensity densities. Bounding the effects of continuous valued interventions has also been explored using instrumental variable models [KKS20, HWZW21, PZW+22].

5 Experiments

Here we empirically validate our method. First, we consider a synthetic structural causal model (SCM) to demonstrate the validity of our method. Next, we show the scalability of our methods by applying them to a real-world climate-science-inspired problem. Implementation details (appendix H), datasets (appendix G), and code are provided at `https://github.com/oatml/overcast`.



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

(e) Λ=1.0 (f) Λ=1.1 (g) Λ=1.2 (h) Λ=1.6

Figure 1: Figures 1a to 1d: Synthetic data and ground truth functions. Figures 1e to 1h Causal uncertainty under hypothesized Λ values. Solid lines are ground truth; thick solid lines where the true λ^* is within the range of hypothesized Λ, thin solid lines otherwise. The dotted lines are the estimated CAPO. Shaded regions are estimated CMSM intervals.

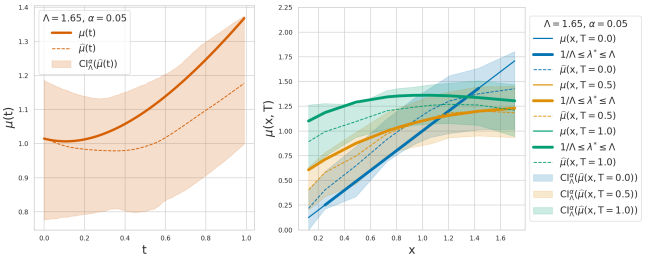
5.1 Synthetic

Figure 1 presents the synthetic dataset (additional details about the SCM are given in Appendix G.1). Figure 1a plots the observed outcomes, y, against the observed confounding covariate, x. Each

datapoint is colored by the magnitude of the observed treatment, t. The binary unobserved confounder, u, induces a bi-modal distribution in the outcome variable, y, at each measured value, x. Figure 1b plots the assigned treatment, t, against the observed confounding covariate, x. We can see that the coverage of observed treatments, t, varies for each value of x. For example, there is uniform coverage at $\mathrm{X} = 1$, but low coverage for high treatment values at $\mathrm{X} = 0.1$, and low coverage for low treatment values at $\mathrm{X} = 2.0$. Figure 1c plots the true CAPO function over the domain of observed confounding variable, X, for several values of treatment ($\mathrm{T} = 0.0$, $\mathrm{T} = 0.5$, and $\mathrm{T} = 1.0$). For lower magnitude treatments, t, the CAPO function becomes more linear, and for higher values, we see more effect heterogeneity and attenuation of the effect size as seen from the slope of the CAPO curve for $\mathrm{T} = 0.5$ and $\mathrm{T} = 1.0$. Figure 1d plots the the APO function over the domain of the treatment variable T.

Causal Uncertainty We want to show that in the limit of large samples (we set n to $100k$), the bounds on the CAPO and APO functions under the CMSM include the ground truth when the CMSM is correctly specified. That is, when $1/\Lambda \leq \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) \leq \Lambda$, for user specified parameter Λ, the estimated intervals should cover the true CAPO or APO. This is somewhat challenging to demonstrate as the true density ratio $\lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u})$ (eq. (50)), varies with t, x, and u. Figures 1e to 1h work towards communicating this. In Figure 1e, we see that each predicted CAPO function (dashed lines) is biased away from the true CAPO functions (solid lines). We use thick solid lines to indicate cases where $1/\Lambda \leq \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) \leq \Lambda$, and thin solid lines otherwise. Therefore thick solid lines indicate areas where we expect the causal intervals to cover the true functions. Under the erroneous assumption of ignorability ($\Lambda = 1$), the CMSM bounds have no width. In Figure 1f, we see that as we relax our ignorability assumption ($\Lambda = 1.1$) the intervals (shaded regions) start to grow. Note the thicker orange line: this indicates that for observed data described by $\mathrm{X} > 0.5$ and $\mathrm{T} = 0.5$, the actual density ratio is in the bounds of the CMSM with parameter $\Lambda = 0.5$. We see that our predicted bounds cover the actual CAPO function for these values. We see our bounds grow again in Figure 1g when we increase Λ to 1.2. We see that more data points have λ^* values that lie in the CMSM range and that our bounds cover the actual CAPO function for these values. In Figure 1h we again increase the parameter of the CMSM. We see that the bounds grow again, and cover the true CAPO functions for all of the data that satisfy $1/\Lambda \leq \lambda^*(\mathrm{t}, \mathrm{x}, \mathrm{u}) \leq \Lambda$.

Statistical Uncertainty Now we relax the infinite data assumption and set $n = 1000$. This decrease in data will increase the estimator error for the CAPO and APO functions. So the estimated functions will not only be biased due to hidden confounding, but they may also be erroneous due to finite sample variance. We show this in Figure 2b where the blue dashed line deviates from the actual blue solid line as $\mathbf{x}$ increases beyond 1.0. However, Figure 2b shows that under the correct CMSM, the uncertainty aware confidence intervals, section 3.4, cover the actual CAPO functions for the range of treatments considered. Figure 2a demonstrates that this holds for the APO function as well.

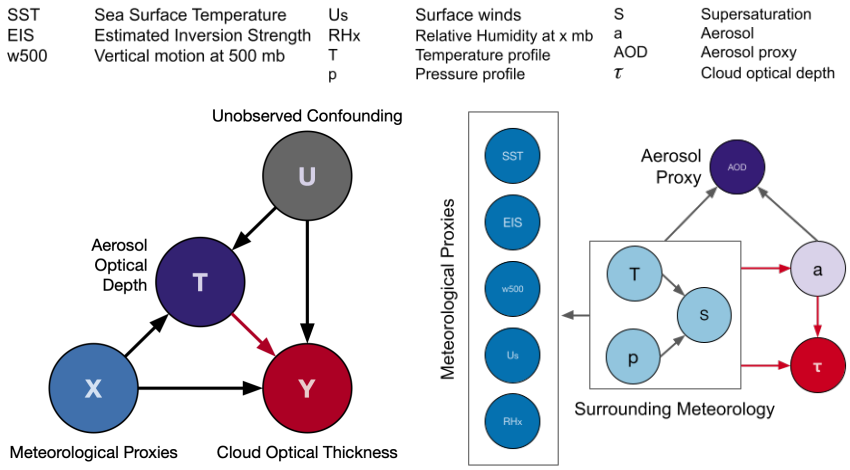


(a) APO Function (b) CAPO Functions

Figure 2: Statistical and causal uncertainty, α is statistical significance level for the bootstrap. see Figure 1 for other details.

5.2 Estimating Aerosol-Cloud-Climate Effects from Satellite Data

Background The development of the model above, and the inclusion of treatment as a continuous variable with multiple, unknown confounders, is motivated by a real-life use case for a prime topic in climate science. Aerosol-cloud interactions (ACI) occur when anthropogenic emissions in the form of aerosol enter a cloud and act as cloud condensation nuclei (CCN). An increase in the number of CCN results in a shift in the cloud droplets to smaller sizes which increases the brightness of a cloud and delays precipitation, increasing the cloud's lifetime, extent, and possibly thickness [Two77, Alb89, TCGB17]. However, the magnitude and sign of these effects are heavily dependent on the environmental conditions surrounding the cloud [DL20]. Clouds remain the largest source of uncertainty in our future climate projections [MDZP+21]; it is pivotal to understand how human emissions may be altering their ability to cool. Our current climate models fail to accurately emulate



(a) (b)

Figure 3: Causal diagrams. Figure 3a, a simplified causal diagram representing what we are reporting within; aerosol optical depth (AOD, regarded as the treatment T) modulates cloud optical depth (τ, Y), which itself is affected by hidden confounders (U) and the meteorological proxies (X). Figure 3b, an expanded causal diagram of ACI. The aerosol (a) and aerosol proxy (AOD), the true confounders (light blue), their proxies (dark blue), and the cloud optical depth (red).

ACI, leading to uncertainty bounds that could offset global warming completely or double the effects of rising CO_2 [BRA+13].

Defining the Causal Relationships Clouds are integral to multiple components of the climate system, as they produce precipitation, reflect incoming sunlight, and can trap outgoing heat [SF09]. Unfortunately, their interconnectedness often leads to hidden sources of confounding when trying to address how anthropogenic emissions alter cloud properties.

Ideally, we would like to understand the effect of aerosols (a) on the cloud optical thickness, denoted τ. However, this is currently impossible. Aerosols come in varying concentrations, chemical compositions, and sizes [SGW+16] and we cannot measure these variables directly. Therefore, we use aerosol optical depth (AOD) as a continuous, 1-dimensional proxy for aerosols. Figure 3b accounts for the known fact that AOD is an imperfect proxy impacted by its surrounding meteorological environment [CNP+17]. The meteorological environment is also a confounder that impacts cloud thickness τ and aerosol concentration a. Additionally, we depend on simulations of the current environment in the form of reanalysis to serve as its proxy.

Here we report AOD as a continuous treatment and the environmental variables as covariates. However, aerosol is the actual treatment, and AOD is only a confounded, imperfect proxy (Figure 3a). This model cannot accurately capture all causal effects and uncertainty due to known and unknown confounding variables. We use this simplified model as a test-bed for the methods developed within this paper and as a demonstration that they can scale to the underlying problem. Future work will tackle the more challenging and realistic causal model shown in Figure 3b, noting that the treatment of interest a is multi-dimensional and cannot be measured directly.

Model We use daily observed $1^\circ \times 1^\circ$ means of clouds, aerosol, and the environment from sources shown in Table 1 of Appendix G. To model the spatial correlations between the covariates on a given day, we use multi-headed attention [VSP+17] to define a transformer-based feature extractor. Modeling the spatial dependencies between meteorological variables is motivated by confounding that may be latent in the relationships between neighboring variables. These dependencies are unobserved from the perspective of a single location. This architectural change respects both the assumed causal graph (fig. 3a) and some of the underlying physical causal structure. We see in Figure 4 (Left) that modeling *context* with the transformer architecture significantly increases the predictive accuracy of the model when compared to a simple feed-forward neural network (*no context*). **Discussion & Results** The results for the APO of cloud optical depth (τ) as the "treatment", AOD, increases are shown in Figure 4. As the assumed strength of confounding increases ($\Lambda > 1$), the range of uncertainty

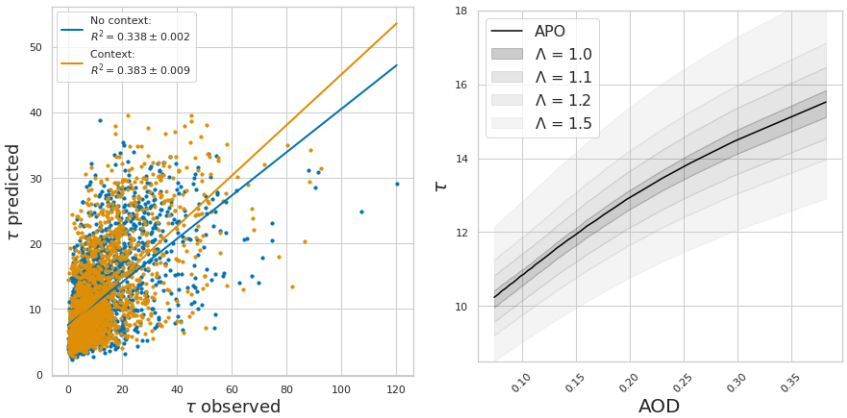


Figure 4: Left: The values of the observed, true τ against the modeled τ. Right: The curve for continuous treatment outcome of our aerosol proxy (AOD) on cloud optical depth (τ). The darkest shaded region (Λ = 1) represents the uncertainty in the treatment outcome from the ensemble due to finite data. As the strength of confounders increases ($\Lambda > 1.0$), the range of uncertainty in the treatment outcome increases.

in the treatment outcome increases. Within this range of confounding, the modeled outcomes agree with two conflicting hypotheses. First, that aerosol acts to invigorate the cloud, inducing a large response that would follow a maximum curve within this uncertainty range [CS11, DL21]. And second, that aerosol has little impact on cloud depth, and the actual response is a minimal, flat line [GGS+19]. We further find the reported dose-response curves in agreement with multiple estimates of aerosol-cloud interactions using satellite observations [BTG02, MSJ+07, TCQB19]. The upper bound for $\log \Lambda = .2$ agrees with measurements of the in-cloud environment and aerosol-cloud interactions from aircraft-mounted sensors [PZ13], this may indicate the need for additional control variables when using satellite data.

The resolution of the satellite observations ($1^\circ \times 1^\circ$ daily means) could be averaging various cloud types and obscuring the signal. Future work will investigate how higher resolution (20km $\times$ 20km) data with constraints on cloud type may resolve some confounding influences. However, even our more detailed causal model (Figure 3b) cannot account for all confounders; we expected, and have seen, imperfections in our model of this complex effect. The model's results require further expert validation to interpret the outcomes and uncertainty.

Societal Impact Geoengineering of clouds by aerosol seeding could offset some amount of warming due to climate change, but also have disastrous global impacts on weather patterns [DGL+22]. Given the uncertainties involved in understanding aerosol-cloud interactions, it is paramount that policy makers are presented with projected outcomes if a proposals assumptions are wrong or relaxed.

Acknowledgments and Disclosure of Funding

We would like to thank Angela Zhou for introducing us to the works of [ZSB19] and [DG21]. These works use the percentile bootstrap for finite sample uncertainty estimation within their sensitivity analysis methods. We would also like to thank Lewis Smith for helping us understand the Marginal Sensitivity Model of [Tan06] in detail. Finally, we would like to thank Clare Lyle and all anonymous reviewers for their valuable feedback.

This research was supported by the European Research Council (ERC) project constRaining the EffeCts of Aerosols on Precipitation (RECAP) under the European Union's Horizon 2020 research and innovation program with grant agreement no. 724602 and from the European Union's Horizon 2020 research and innovation program project Constrained aerosol forcing for improved climate projections (FORCeS) under grant agreement No 821205. and Marie Skłodowska-Curie grant agreement No 860100 (iMIRACLI). This work used JASMIN, the UK's collaborative data analysis environment (http://jasmin.ac.uk). U.S. was partially supported by the Israel Science Foundation (grant No. 1950/19).

References

[Alb89] Bruce A Albrecht. Aerosols, cloud microphysics, and fractional cloudiness. *Science*, 245(4923):1227–1230, 1989.

[BAC+15] Michael G Bosilovich, Santha Akella, Lawrence Coy, Richard Cullather, Clara Draper, Ronald Gelaro, Robin Kovach, Qing Liu, Andrea Molod, Peter Norris, et al. Merra-2: Initial evaluation of the climate. 2015.

[Bis94] Christopher M. Bishop. Mixture density networks. 1994.

[BJvdS20] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *arXiv preprint arXiv:2002.12326*, 2020.

[BP06] Bryan A Baum and Steven Platnick. Introduction to modis cloud products. In *Earth science satellite remote sensing*, pages 74–91. Springer, 2006.

[BRA+13] Olivier Boucher, David Randall, Paulo Artaxo, Christopher Bretherton, Gragam Feingold, Piers Forster, V-M Kerminen, Yutaka Kondo, Hong Liao, Ulrike Lohmann, et al. Clouds and aerosols. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 571–657. Cambridge University Press, 2013.

[BTG02] Francois-Marie Bréon, Didier Tanré, and Sylvia Generoso. Aerosol effect on cloud droplet size monitored from satellite. *Science*, 295(5556):834–838, 2002.

[CCN+21] Victor Chernozhukov, Carlos Cinelli, Whitney Newey, Amit Sharma, and Vasilis Syrgkanis. Omitted variable bias in machine learned causal models. *arXiv preprint arXiv:2112.13398*, 2021.

[CH20a] Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67, 2020.

[CH20b] Carlos Cinelli and Chad Hazlett. An omitted variable bias framework for sensitivity analysis of instrumental variables. *Work. Pap*, 2020.

[CHH16] Nicole Bohme Carnegie, Masataka Harada, and Jennifer L Hill. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *Journal of Research on Educational Effectiveness*, 9(3):395–420, 2016.

[CKC+19] Carlos Cinelli, Daniel Kumor, Bryant Chen, Judea Pearl, and Elias Bareinboim. Sensitivity analysis of linear structural causal models. In *International conference on machine learning*, pages 1252–1261. PMLR, 2019.

[CNP+17] Matthew W Christensen, David Neubauer, Caroline A Poulsen, Gareth E Thomas, Gregory R McGarragh, Adam C Povey, Simon R Proud, and Roy G Grainger. Unveiling aerosol–cloud interactions–part 1: Cloud contamination in satellite products enhances the aerosol indirect forcing estimate. *Atmospheric Chemistry and Physics*, 17(21):13151–13164, 2017.

[CS11] Matthew W Christensen and Graeme L Stephens. Microphysical and macrophysical responses of marine stratocumulus polluted by underlying ships: Evidence of cloud deepening. *Journal of Geophysical Research: Atmospheres*, 116(D3), 2011.

[D'A19] Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3478–3486. PMLR, 2019.

[DBSC21] Gianluca Detommaso, Michael Brückner, Philip Schulz, and Victor Chernozhukov. Causal bias quantification for continuous treatment. *arXiv preprint arXiv:2106.09762*, 2021.

[DG21] Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *arXiv preprint arXiv:2102.04543*, 2021.

[DGL+22] Michael S Diamond, Andrew Gettelman, Matthew D Lebsock, Allison McComiskey, Lynn M Russell, Robert Wood, and Graham Feingold. Opinion: To assess marine cloud brightening's technical feasibility, we need to know what to study—and when to stop. *Proceedings of the National Academy of Sciences*, 119(4), 2022.

[DHCH16] Vincent Dorie, Masataka Harada, Nicole Bohme Carnegie, and Jennifer Hill. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470, 2016.

[DL20] Alyson Douglas and Tristan L'Ecuyer. Quantifying cloud adjustments and the radiative forcing due to aerosol–cloud interactions in satellite observations of warm marine clouds. *Atmospheric Chemistry and Physics*, 20(10):6225–6241, 2020.

[DL21] Alyson Douglas and Tristan L'Ecuyer. Global evidence of aerosol-induced invigoration in marine cumulus clouds. *Atmospheric Chemistry and Physics*, 21(19):15103–15114, 2021.

[FKH17] Stefan Falkner, Aaron Klein, and Frank Hutter. Combining hyperband and bayesian optimization. In *NIPS 2017 Bayesian Optimization Workshop (Dec 2017)*, 2017.

[GGS+19] Edward Gryspeerdt, Tom Goren, Odran Sourdeval, Johannes Quaas, Johannes Mülmenstädt, Sudhakar Dipu, Claudia Unglaub, Andrew Gettelman, and Matthew Christensen. Constraining the aerosol influence on cloud liquid water path. *Atmospheric Chemistry and Physics*, 19(8):5331–5347, 2019.

[GMS+17] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (merra-2). *Journal of climate*, 30(14):5419–5454, 2017.

[HI04] Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

[Hol86] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[HWZW21] Yaowei Hu, Yongkai Wu, Lu Zhang, and Xintao Wu. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12104–12112, 2021.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JMGS21] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4829–4838. PMLR, 18–24 Jul 2021.

[JMSG20] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33:11637–11649, 2020.

[KKS20] Niki Kilbertus, Matt J Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20108–20119. Curran Associates, Inc., 2020.

[KMZ19] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.

[KSBY19] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

[LLN+18] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[MDZP+21] V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, , and B. Zhou. Ipcc, 2021: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. 2021.

[MNW+18] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A distributed framework for emerging ai applications, 2018.

[MSDH16] Joel A Middleton, Marc A Scott, Ronli Diakow, and Jennifer L Hill. Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323, 2016.

[MSJ+07] Gunnar Myhre, Frode Stordal, M Johnsrud, YJ Kaufman, D Rosenfeld, Trude Storelvmo, Jon Egill Kristjansson, Terje Koren Berntsen, A Myhre, and Ivar SA Isaksen. Aerosol-cloud interaction inferred from modis satellite data and global aerosol models. *Atmospheric Chemistry and Physics*, 7(12):3081–3101, 2007.

[MVSG] Myrl G Marmarelis, Greg Ver Steeg, and Aram Galstyan. Bounding the effects of continuous treatments for hidden confounders.

[Ney23] Jerzy Neyman. edited and translated by dorota m. dabrowska and terrence p. speed (1990). on the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923.

[NYLN21] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

[Ost19] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

[PGM+19] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[PZ13] D Painemal and P Zuidema. The first aerosol indirect effect quantified through airborne remote sensing during vocals-rex. *Atmospheric Chemistry and Physics*, 13(2):917–931, 2013.

[PZW+22] Kirtan Padh, Jakob Zeitler, David Watson, Matt Kusner, Ricardo Silva, and Niki Kilbertus. Stochastic causal programming for bounding treatment effects. *arXiv preprint arXiv:2202.10806*, 2022.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[RR83] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[Rub74] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[Rub80] Donald B Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[SBV19] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 2507–2517, 2019.

[Sek08] Jasjeet S Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–32, 2008.

[SF09] Bjorn Stevens and Graham Feingold. Untangling aerosol effects on clouds and precipitation in a buffered system. *Nature*, 461(7264):607–613, 2009.

[SGW+16] Nick AJ Schutgens, Edward Gryspeerdt, Natalie Weigum, Svetlana Tsyro, Daisuke Goto, Michael Schulz, and Philip Stier. Will a perfect model agree with perfect observations? the impact of spatial sampling. *Atmospheric Chemistry and Physics*, 16(10):6335–6353, 2016.

[SJS17] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[SLB+20] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M Buhmann, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.

[Tan06] Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

[TCGB17] Velle Toll, Matthew Christensen, Santiago Gassó, and Nicolas Bellouin. Volcano and ship tracks indicate excessive aerosol-induced cloud water increases in a climate model. *Geophysical research letters*, 44(24):12–492, 2017.

[TCQB19] Velle Toll, Matthew Christensen, Johannes Quaas, and Nicolas Bellouin. Weak average liquid-cloud-water response to anthropogenic aerosols. *Nature*, 572(7767):51–55, 2019.

[Two77] Sean Twomey. The influence of pollution on the shortwave albedo of clouds. *Journal of the atmospheric sciences*, 34(7):1149–1152, 1977.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[ZSB19] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
      i. we claim to introduce a novel marginal sensitivity model for continuous valued treatment effect (the CMSM). See Section 3.
      ii. we claim to derive bounds for the CAPO and APO functions that agree with the CMSMS and observed data. See Section 3.1.
      iii. we claim to provide tractable estimators of the CAPO and APO bounds. See Sections 3.2 and 3.3.
      iv. we claim to provide bounds that account for finite-sample (statistical) uncertainty. See Section 3.4.
      v. we claim to provide a novel architecture for scalable estimation of the effects of continuous valued interventions. See Section 3.5.
      vi. we claim that the bounds cover the true ignorance interval for well specified Λ. See Figures 1e to 1h and Theorem 1.
      vii. we claim that this model scales to real-world, large-sample, high-dimensional data. See Section 5.2
   (b) Did you describe the limitations of your work? [Yes]
      i. We have discussed the major limitation of sensitivity analysis methods, that unobserved confounding is not identifiable from data alone. We have tried to be honest and transparent that our method provides users with a way to communicate the uncertainty induced when relaxing the ignorability assumption. We do not claim that lambda is in any way identifiable without further assumptions.
      ii. In Section 5.2, we have clearly discussed the limitations of analyses of aerosol-cloud interactions using satellite data where we only see underlying causal mechanisms through proxy variables. We hope this paper serves as a stepping stone for work that specifically addresses those issues.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We have five theoretical results. Proposition 1, Equation (7), Equation (8), Equation (9), and Theorem 1. All assumptions are stated for each.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proof of Proposition 1 is given in Appendix A.3. The proof of Equation (7) is given in Lemma 1. The proofs for Equation (8) and Equation (9) are given in Lemma 2. The proof for Theorem 1 is given in Appendix E.
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code, data, and instructions are provided in the suppleemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify these details in Appendix G and Appendix H as well as in the provided code.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Both random seeds and random bootstrapped sampling of the training data.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] this is outlined in Appendix H

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We use existing satellite data and open source code libraries that we have cited.

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] we provide a new synthetic dataset and code base

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

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 $\lambda_B(\mathrm{Y_t}, \mathbf{x}; \mathrm{t}) : \mathcal{Y} \to [0, \inf)$ such that,

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

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

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

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 $e(\mathbf{x}, \mathrm{y_t}) \equiv Pr(\mathrm{T} = 1 \mid \mathbf{X} = \mathbf{x}, \mathrm{Y_t} = \mathrm{y_t})$ 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}' \mid \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{\cancel{r(\mathbf{x}, \mathrm{t}')}}}{\frac{p(\mathrm{T}=\mathrm{t} \mid \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} \mid \mathrm{y_t}, \mathbf{x}) \cancel{p(\mathrm{y_t})}}{\cancel{r(\mathbf{x}, \mathrm{t})}}}{\frac{p(\mathrm{T}=\mathrm{t} \mid \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,

$$\begin{aligned}\{ & \\ & P(\mathrm{Y_t} \mid \mathbf{X} = \mathbf{x}) = \\ & \quad \int_{\mathcal{T}} p(\mathrm{t}' \mid \mathbf{x}) P(\mathrm{Y_t} \mid \mathrm{T} = \mathrm{t}', \mathbf{X} = \mathbf{x}) d\mathrm{t}' \\ & : \mathrm{t} \in \mathcal{T} \\ \} & \end{aligned} \tag{26}$$

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 aslo 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 (6)

Lemma 1.

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

Proof. Recall that the expected 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,

$$\begin{aligned} \mu(\mathbf{x}, \mathrm{t}) &= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \\ &= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})p(\mathrm{t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \\ &= \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{y_t} \mid \mathrm{t}, \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}, \end{aligned}$$

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:

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

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

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

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

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

Following [KMZ19] again, we reparameterize the inverse complete propensity density as, $\frac{1}{p(\mathrm{t}|\mathrm{y_t},\mathbf{x})} = \alpha(\mathbf{x}; \mathrm{t}, \Lambda) + w(\mathrm{y}, \mathbf{x})(\beta(\mathbf{x}; \mathrm{t}, \Lambda) - \alpha(\mathbf{x}; \mathrm{t}, \Lambda))$ 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,

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

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

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

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

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

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

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

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

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

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

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: $\int_{-\infty}^{\infty} \exp\left(-\mathrm{y}^2\right) f(\mathrm{y}) d\mathrm{y}$. In this case,

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

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 $\mu(\mathbf{x}, \mathrm{t}) = \frac{\int_{\mathcal{Y}} \mathrm{y_t} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y_t} \mid \mathbf{x})}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} d\mathrm{y_t}}$, and applying a one-to-one change of variables, $\frac{1}{p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x})} = \alpha(\mathbf{x}; \mathrm{t}, \Lambda) + w(\mathrm{y})(\beta(\mathbf{x}; \mathrm{t}, \Lambda) - \alpha(\mathbf{x}; \mathrm{t}, \Lambda))$ 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 $\mathcal{W}^H_{\text{nd}}$. 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 $\mathcal{W}^H_{\text{ni}}$.

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

$$\text{maximize} \quad \frac{a \langle \mathrm{y}, w(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + c}{b \langle 1, w(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + d} \tag{44a}$$

$$\text{subject to} \quad 0 \leq w(\mathrm{y}) \leq 1 : \forall \mathrm{y} \in \mathcal{Y}, \tag{44b}$$

where $\langle \cdot, \cdot \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})}$ 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:

$$\text{maximize} \quad a \langle \mathrm{y}, \widetilde{w}(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + c \widetilde{v}(\mathbf{x}) \tag{45a}$$

$$\text{subject to} \quad \widetilde{w}(\mathrm{y}) \leq \widetilde{v}(\mathbf{x}) : \forall \mathrm{y} \in \mathcal{Y} \tag{45b}$$

$$-\widetilde{w}(\mathrm{y}) \leq 0 : \forall \mathrm{y} \in \mathcal{Y} \tag{45c}$$

$$b \langle 1, \widetilde{w}(\mathrm{y}) \rangle_{p(\mathrm{y} \mid \mathrm{t}, \mathbf{x})} + d \widetilde{v}(\mathbf{x}) = 1 \tag{45d}$$

$$\widetilde{v}(\mathbf{x}) \geq 0, \tag{45e}$$

where

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

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:

$$\text{minimize} \quad \gamma \tag{46a}$$

$$\text{subject to} \quad \rho(\mathrm{y}) - \eta(\mathrm{y}) + \gamma b p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) = a \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) : \forall \mathrm{y} \in \mathcal{Y} \tag{46b}$$

$$-\langle 1, \rho(\mathrm{y}) \rangle + \gamma d \geq c \tag{46c}$$

$$\rho(\mathrm{y}) \in \mathbb{R}_+, \eta(\mathrm{y}) \in \mathbb{R}_+, \gamma \in \mathbb{R} \tag{46d}$$

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

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

[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 $\mathcal{W}_{\mathrm{nd}}^H$.

This result holds under

$$\mu(\mathbf{x}, \mathrm{t}) = \frac{\int_{\mathcal{Y}} \mathrm{y} p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y} + (\lambda^2 - 1) \int_{\mathcal{Y}} \mathrm{y} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{1 + (\lambda^2 - 1) \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, \tag{48a}$$

$$= \frac{\int_{\mathcal{Y}} \mathrm{y}_{\mathrm{t}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} | \mathbf{x})}{p(\mathrm{t} | \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}{\int_{\mathcal{Y}} \frac{p(\mathrm{t}, \mathrm{y}_{\mathrm{t}} | \mathbf{x})}{p(\mathrm{t} | \mathrm{y}_{\mathrm{t}}, \mathbf{x})} d\mathrm{y}_{\mathrm{t}}}, \tag{48b}$$

$$= \widetilde{\mu}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathcal{Y}} w(\mathrm{y})(\mathrm{y} - \widetilde{\mu}(\mathbf{x}, \mathrm{t})) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}{(\Lambda^2 - 1)^{-1} + \int_{\mathcal{Y}} w(\mathrm{y}) p(\mathrm{y} \mid \mathrm{t}, \mathbf{x}) d\mathrm{y}}, \tag{48c}$$

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

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

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

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

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(x*, t*, Λ, θ, Ŷ)
    μ̄ ← −∞, κ̄ ← ∞
    μ̲ ← ∞, κ̲ ← −∞
    δ̲ ← True, δ̄ ← True
    while δ̲ do
        y* ← Pop(Ŷ_c)                          ▷ Ŷ_c a copy of Ŷ
        κ̲ ← κ̂_θ(x, t; Λ, H(y* − y))
        if κ̲ < μ̲ then
            μ̲ ← κ̲
        else
            δ̲ ← False
    while δ̄ do
        y* ← Pop(Ŷ_c)                          ▷ Ŷ_c a copy of Ŷ
        κ̄ ← κ̂_θ(x, t; Λ, H(y − y*))
        if κ̄ > μ̄ then
            μ̄ ← κ̄
        else
            δ̄ ← False
    return μ̲, μ̄
```

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. $p(\mathrm{t} \mid \mathrm{y_t}, \mathbf{x}) > 0, \forall \mathrm{y_t} \in \mathcal{Y}$.

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,

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

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

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

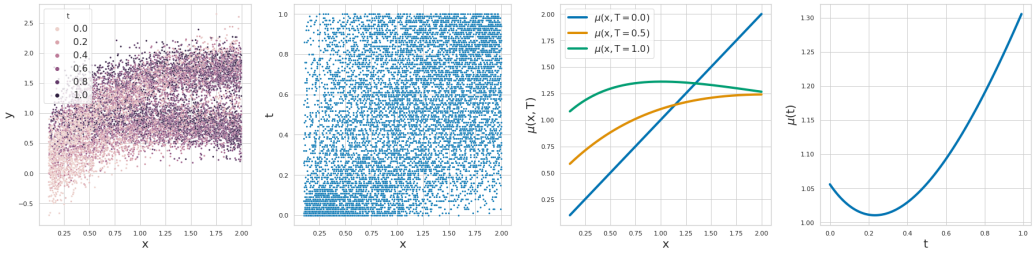
where, $N_{\mathrm{u}} \sim p(\mathrm{u}) := \text{Bern}(\mathrm{u} \mid 0.5)$, $N_{\mathrm{x}} \sim p(\mathrm{x}) := \text{Unif}[\mathrm{x} \mid 0.1, 2.0]$, $N_{\mathrm{t}} \sim p(\mathrm{t} \mid \mathrm{x}, \mathrm{u}) := \text{Beta-Binomial}(\mathrm{t} \mid n = 100, \alpha = \mathrm{x} + \gamma_{\mathrm{t}}\mathrm{u}, \beta = 1)$, 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,

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

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



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

Figure 5: Synthetic data with hidden confounding

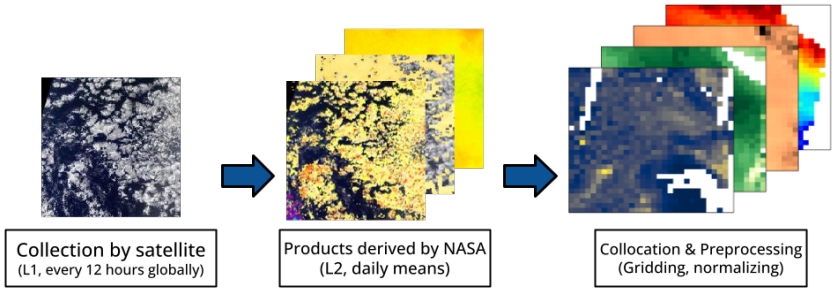


Figure 6: Workflow of observed clouds from satellite to ingestion by model.

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.

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

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

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

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

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, $\boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\theta})$ ($\boldsymbol{\phi}$, for short), and a conditional outcome prediction block $f(\boldsymbol{\phi}, \mathrm{t}; \boldsymbol{\theta})$, 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**) 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.

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 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.

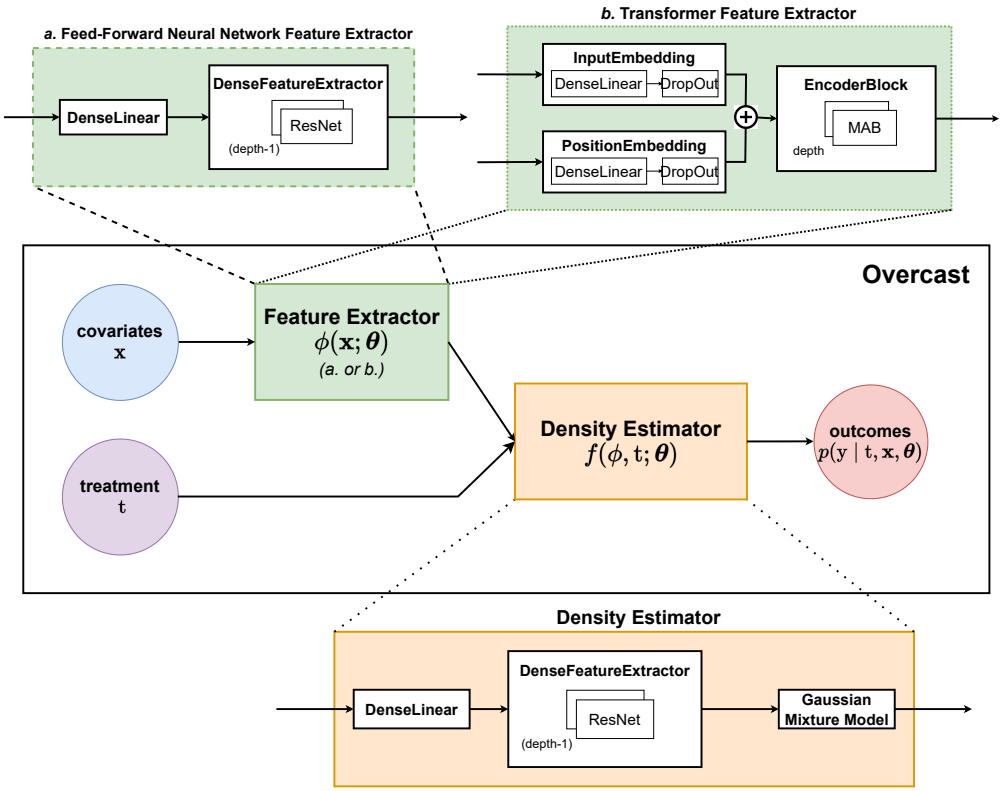


Figure 7: **Overcast model architecture.** The inputs are represented by circles, in **blue** the covariates, and in **purple** the treatment. In the **red** circle is the output of the model, the outcomes distribution. The model has different feature extractors (in green) for the feed-forward neural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

H.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

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

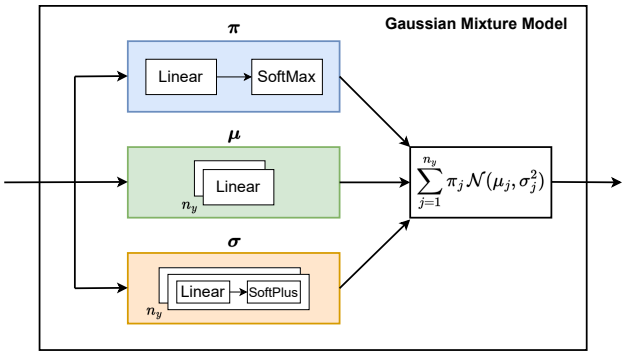
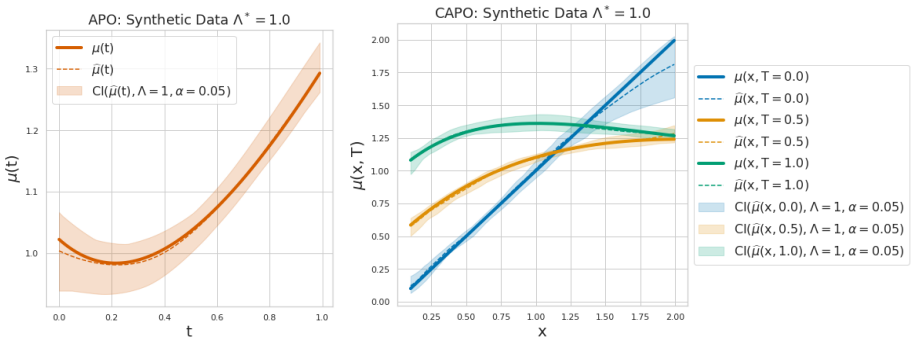


Figure 8: **Overcast Gaussian mixture model.** The mixing coefficients π are estimated with a linear layer and a SoftMax layer, to obtain $\tilde{\pi}$, represented in **blue** in the figure. The vector of means of the Gaussian kernels $\tilde{\mu}$ is obtained by n_y linear layers (in **green** in the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



(a) APO Functions

(b) CAPO Functions

Figure 9: Investigating statistical uncertainty using unconfounded synthetic data.

we let just a small amount of the unexplained range of 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).

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).

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.

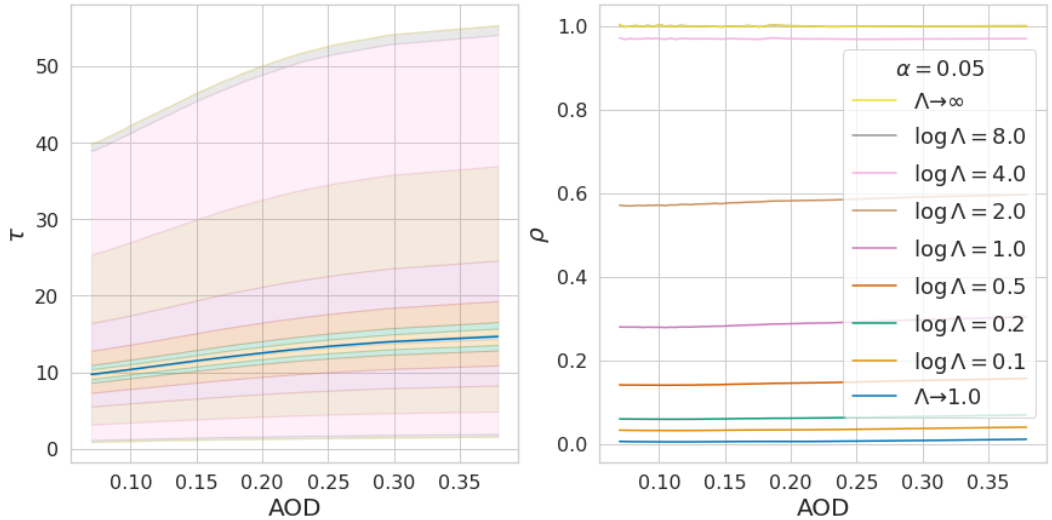
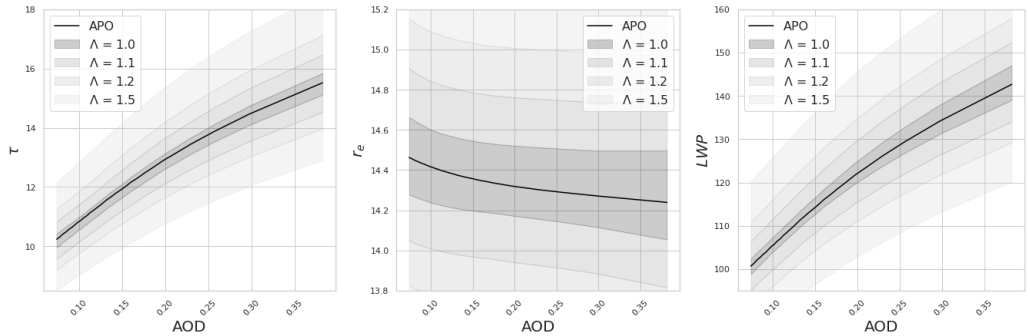


Figure 10: Interpreting Λ as a proportion (ρ) of the unexplained range of Y_t attributed to unobserved confounding variables.

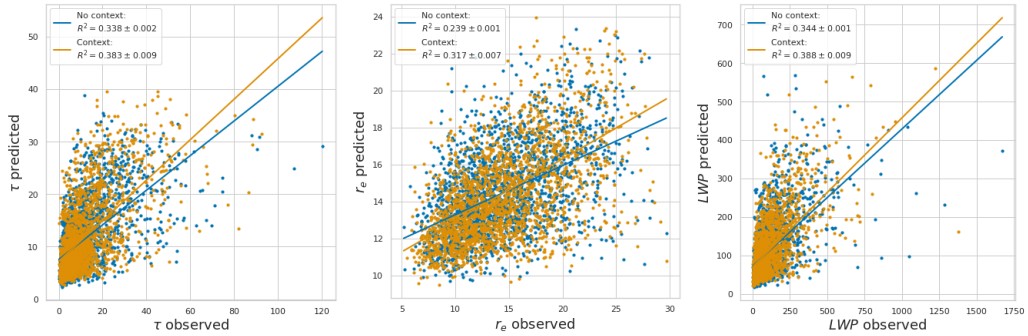


(a) (b) (c)

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

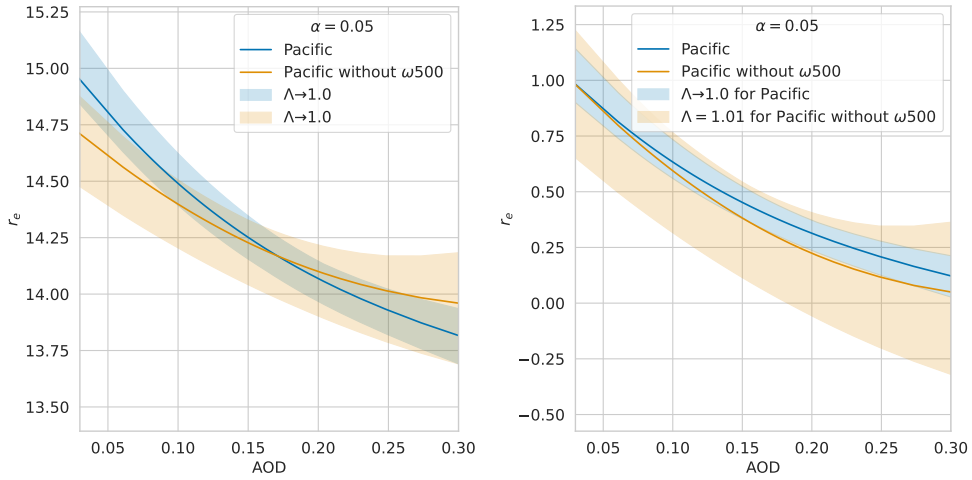
In Figure 13 we compare the same region with different covariates to identify an appropriate Λ. We fit one model on data from the Pacific (blue) and one model from the Pacific omitting $\omega 500$ from the covariates (orange). The shaded bounds in blue are the ignorance region for $\Lambda \to 1$ for 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) (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.



(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.