Quantifying Ignorance in Individual-Level Causal-Effect Estimates under Hidden Confounding

Andrew Jesson [1] **Sören Mindermann** [1] **Yarin Gal** [1] **Uri Shalit** [2]

Abstract

We study the problem of learning conditional average treatment effects (CATE) from high-dimensional, observational data with unobserved confounders. Unobserved confounders introduce ignorance—a level of unidentifiability—about an individual's response to treatment by inducing bias in CATE estimates. We present a new parametric interval estimator suited for high-dimensional data, that estimates a range of possible CATE values when given a predefined bound on the level of hidden confounding. Further, previous interval estimators do not account for ignorance about the CATE associated with samples that may be underrepresented in the original study, or samples that violate the overlap assumption. Our interval estimator also incorporates model uncertainty so that practitioners can be made aware of such out-of-distribution data. We prove that our estimator converges to tight bounds on CATE when there may be unobserved confounding and assess it using semi-synthetic, high-dimensional datasets.

1. Introduction

How will a patient's health be affected by taking a given medication (Perez, 2019)? How will a job seeker's employment be affected by participating in a training program? Making effective personalized recommendations depends on being able to answer such questions. Answering such questions requires knowledge about the causal effect that a treatment or intervention (medication, training program) has on a person. And knowing the effect of the treatment requires knowledge about the individual.

Randomized controlled trials (RCTs) are the gold standard for discovering population-level causal effects of such treatments. However, in many cases, RCTs are prohibitively expensive or unethical. For example, researchers cannot randomly prescribe smoking to assess health risks. Observational data, often with larger sample sizes, lower costs, and more relevance to the target population, offer an alternative way to learn about individual-level causal effects. The price paid for using observational data, however, is lower certainty in the estimated causal effects.

When there is sufficient knowledge about both the population and the individual, inferring the individual's response to treatment is possible, and corresponding recommendations can be made with relative certainty. A widely used quantity expressing an individual's response to treatment is the Conditional Average Treatment Effect (CATE), which is defined in the next section.

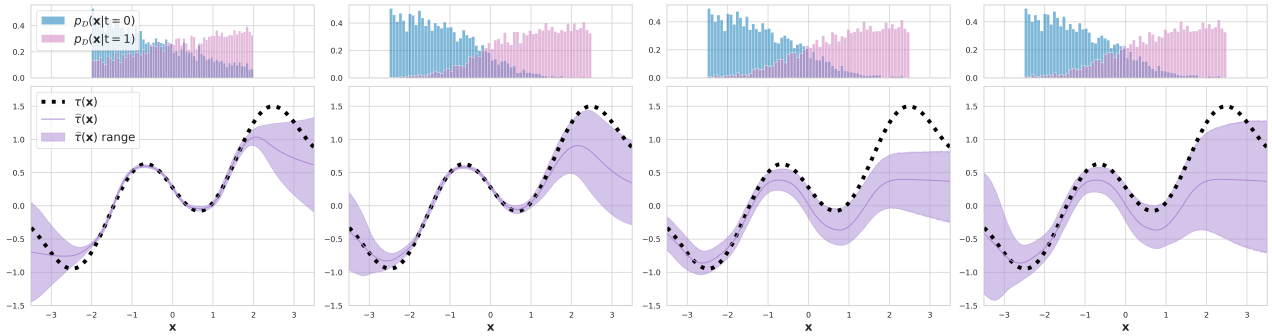
There are, however, many reasons why we would not know enough about someone to make an informed recommendation. For example, there may be **insufficient similarity**: when an individual is unrepresented in the study population, which can be the case if the data comes from a small study or just one hospital. There may also be **insufficient overlap** (ubiquitous, especially for high-dimensional data (D'Amour et al., 2020)): when an individual lacks representation in either the treatment or control group, which can be the case if there are socio-economic barriers to accessing treatment. Finally, there may be **insufficient context**: when there are unobserved factors (confounders) that influence both an individual's odds of receiving treatment, as well as their outcome.

When confronted by such ignorance about a person's response to treatment, recommending treatments based on a model's point estimate of the CATE can be dangerous - doubly so in high-stakes domains such as health care. Instead, it may be preferable to *defer* the recommendation when the CATE estimate is uncertain: this might entail consulting with a domain expert, using a safe default treatment, collecting additional data on subjects similar to the one in question, or broadening the context of the study by incorporating additional confounding covariates.

In this paper, we provide a measure of ignorance that unifies

[1]OAMTL, University of Oxford [2]Machine Learning and Causal Inference in Healthcare Lab, Technion – Israel Institute of Technology. Correspondence to: Andrew Jesson <andrew.jesson@cs.ox.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).



(a) similarity (b) overlap (c) context (d) unified

Figure 1. The purple shaded areas in the lower panes depict regions of ignorance about a unit's response to treatment. The training data density for the untreated and treated groups are shown in the upper panes. (1a) For ignorance due to **insufficient similarity**, the region should get wider as the distance between $\mathbf{x}$ and the training data increases. (1b) For ignorance due to **insufficient overlap** the region should get wider as $P(\mathrm{T}=0 \mid \mathbf{x})$ or $P(\mathrm{T}=1 \mid \mathbf{x}) \to 1$. (1c) Ignorance due to **insufficient context** results in an arbitrarily biased CATE estimator $\widehat{\tau}(\mathbf{x})$, hence the discrepancy between the blue solid line and the black dashed line. Therefore, the ignorance region should include the true CATE $\tau(\mathbf{x})$ on the training data manifold where overlap is satisfied. (1d) All sources of ignorance jointly.

all three of the above sources (similarity, overlap, and context), which is expressed as an interval of possible CATE values for each individual. The width of the interval increases as the assumptions underlying each source are challenged more severely. When CATE estimates are used to recommend treatment (e.g. "treat if and only if CATE is positive"), we envision the ignorance interval as being used to *defer* the decision: it might be better not to give a recommendation at all rather than give a highly uncertain one.

We take Bayesian deep learning as a starting point: such methods provide multiple functions to explain observed data (illustrated in Figure 2), functions that tend to agree with one another for well-represented data but disagree with one another where data is under or unrepresented. Thus, Bayesian methods can be used in quantifying ignorance due to **insufficient similarity and overlap** by measuring *epistemic uncertainty* (the disagreement between functional predictions of the outcome), which has been used in the context of CATE estimates by Jesson et al. (2020).

That leaves us with ignorance due to **insufficient context**, also known as unobserved confounding. Unobserved confounding manifests as unexplained variance in the estimates of both the outcome and the individual's propensity for treatment and induces a bias in the estimates of causal effects. Standard Bayesian methods account for unexplained variance in the outcome, known as *aleatoric uncertainty*; however, without further assumptions, it is in general impossible to identify which part of this uncertainty is due to confounding (Pearl et al., 2009). Therefore, we turn to causal sensitivity analysis to quantify the ignorance in causal-effect estimates due to the bias induced by hidden confounding. Causal sensitivity analysis includes a diverse family of frameworks, whose common goal is to give bounds on the treatment-effect under the assumption of some "level" of unobserved confounding, either at a population level (Rosenbaum & Rubin, 1983; Robins et al., 2000b; Imbens, 2003; Rosenbaum, 2014; Dorie et al., 2016; Franks et al., 2019; Veitch & Zaveri, 2020) or at the level of individuals (Yadlowsky et al., 2018; Kallus et al., 2019).

Specifically, we build on recent work by Kallus et al. (2019) and introduce a novel method that can scale to large-sample, high-dimensional data, and convey information about *all three sources of ignorance* mentioned above. In section 3.2 we present a new functional interval estimator that predicts a range of possible CATE values when given a bound on the influence of hidden confounding. We prove that our estimator converges to tight bounds on CATE for a given bound on hidden confounding. In section 3.5 we present a CATE interval estimator integrating all sources of uncertainty mentioned above. In section 4 we demonstrate that our new method scales to high-dimensional data by evaluating it on existing benchmarks and introducing a new high-dimensional dataset.

2. Sources of Ignorance in Causal Inference

In this section we formalize the idea of being ignorant about an individual and their response to treatment by framing it as a violation of one or more of the requisite assumptions needed to identify treatment-effects.

The individual's response to treatment is formally known as the *individual treatment effect* or ITE. The ITE of a binary treatment $\mathrm{T} \in \{0, 1\}$ on an individual i is the difference in potential outcomes $\mathrm{Y}_i^1 - \mathrm{Y}_i^0$. The potential outcome Y_i^1 describes the outcome were the individual i treated, whereas the potential outcome Y_i^0 describes the outcome were they

not treated. The ITE is a fundamentally unobservable quantity since it is only possible to measure one potential outcome for a given individual. However, when individuals are described by a set of covariates $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$, then we can model the Conditional Average Treatment Effect (CATE) (Abrevaya et al., 2015), $\tau(\mathbf{x}) = \mathbb{E}[\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}[\mathrm{Y}^1 \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[\mathrm{Y}^0 \mid \mathbf{X} = \mathbf{x}]$, which is the expected difference in potential outcomes over units (possibly individuals) who share the same measured covariates $\mathbf{X} = \mathbf{x}$.

The estimation of $\tau(\mathbf{x})$ relies on an observational dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i) : i = 1, \dots n\}$. From such data, the expected potential outcome $\mathbb{E}[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{X} = \mathbf{x}]$ is identifiable as the conditional expectation over observed outcomes $\mu(\mathbf{x}, \mathrm{t}) \equiv \mathbb{E}[\mathrm{Y} | \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}]$ (Rubin, 1974) under the following assumptions:

1. $(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)$ are i.i.d. draws from the same population $P_{\mathcal{D}}(\mathbf{X}, \mathrm{T}, \mathrm{Y}^0, \mathrm{Y}^1)$.
2. Overlap (Positivity): $e_{\mathrm{t}}(\mathbf{x}) \equiv P(\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}) > 0 : \mathrm{t} \in \{0, 1\}$.
3. Unconfoundedness (Exchaneability, Sufficiency, Exogeneity): $\{(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T}\} \mid \mathbf{X}$.

A further assumption which is not our focus here is the stable unit treatment value assumption which briefly stated means that each unit's observed outcome corresponds exactly and only to its treatment assignment. That is, for an individual i we observe the outcome $y_i = t_i \mathrm{Y}_i^1 + (1 - t_i)\mathrm{Y}_i^0$. When these assumptions hold, the CATE for individuals sharing the same measured covariates $\mathbf{X} = \mathbf{x}$ is given by

$$\tau(\mathbf{x}) = \mu(\mathbf{x}, 1) - \mu(\mathbf{x}, 0). \tag{1}$$

In practice, an estimator $\widehat{\tau}(\mathbf{x})$ for $\tau(\mathbf{x})$ is learned from a finite dataset, and ignorance about an individual's response to treatment is due to both observational data being finite and possible violations of the above assumptions.

First, the dataset $\mathcal{D}$ is a finite sample from $P_{\mathcal{D}}(\mathbf{X}, \mathrm{T}, \mathrm{Y}^0, \mathrm{Y}^1)$ of size n, so there is **limited similarity** – for a test point $\mathbf{x}^*$ there might not be any similar train points $\mathbf{x}$. Furthermore, test samples might come from a different marginal distribution $P_{\mathcal{D}'}(\mathbf{X})$ than the one the training dataset is drawn from, i.e. covariate shift, a scenario which violates Assumption 1. Figure 1a illustrates such violations of Assumption 1. The range of $\widehat{\tau}(\mathbf{x})$ (purple shaded areas) should be tight around values of $\mathbf{x}$ that are observed in $\mathcal{D}$ and get wider for individuals described by $\mathbf{x}$ that are not.

Second, the treatment assignment may be such that for units described by covariates $\mathbf{X} = \mathbf{x}$, the observed treatment indicator T is all 0 or all 1, so there is **limited overlap** (D'Amour et al., 2020). For example, a given test point $\mathbf{x}^*$ may have similar points in the train set with treatment assignment $T = 0$ but none with $T = 1$. Therefore, we cannot accurately estimate $\mathbf{x}^*$'s response under $T = 1$. Such violations of the overlap assumption are especially common for high-dimensional covariates which likely contain ample information to predict the treatment (Assumption 2). Figure 1b illustrates such violations of the overlap assumption. Here, overlap is not satisfied at the left and right edges of the data. Therefore, the uncertainty for $\widehat{\tau}(\mathbf{x})$ should be tight around values of $\mathbf{x}$ for which there are both treated and untreated examples (darker area in top pane, $-2 \leq \mathbf{x} \leq 1.5$) and get wider around values of $\mathbf{x}$ where there are only either treated ($P(\mathrm{T} = 1 \mid \mathbf{x}) \to 1 \colon \mathbf{x} > 1.5$) or untreated examples ($P(\mathrm{T} = 0 \mid \mathbf{x}) \to 1 \colon \mathbf{x} < 2$).

Third, there is **limited context** about the individual ($\mathcal{X}$ is only d-dimensional). For a point $\mathbf{x}^*$ we might not have enough context to correctly estimate its true response under one or both treatments T. This is especially important if treatment in the train set was assigned based on an unobserved factor which also affects the outcome Y, which is a violation of Assumption 3. Figure 1c illustrates such violations of the unconfoundedness assumption. Such violations result in $\widehat{\tau}(\mathbf{x})$ (blue solid line) being a biased estimator of the true CATE (black dotted line). The bias is induced here by having the probability of treatment and the outcome be affected by a confounding variable u, which is not included in the set of covariates $\mathbf{x}$ given to the estimator $\widehat{\tau}(\mathbf{x})$.

A *unified* measure of uncertainty would correspond to the width of the range of CATE values that accounts for all of the above sources of ignorance in the estimate of $\widehat{\tau}(\mathbf{x})$, for all values of $\mathbf{x}$, as illustrated in Figure 1d.

3. Proposed Method

We first introduce the ideas which are needed to develop our approach: how to evaluate epistemic uncertainty for CATE using Bayesian deep learning (Jesson et al., 2020), and a method for expressing violations of unconfoundedness assumption (Kallus et al., 2019) in the context of CATE estimation. We then develop our novel proposed estimator.

3.1. Preliminaries

3.1.1. Quantifying Ignorance due to Insufficient Similarity and Overlap

The expectations in Equation 1 are typically expressed using parametric (Robins et al., 2000a; Tian et al., 2014; Shalit et al., 2017) or non-parametric models (Hill, 2011; Xie et al., 2012; Alaa & van der Schaar, 2017; Gao & Han, 2020). Parametric models assume predictions are generated from $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$, the conditional distribution over outcomes Y

given covariates $\mathbf{x}$, treatment t, and parameters $\boldsymbol{\omega} \in \mathcal{W}$. A common choice for continuous Y is a Gaussian distribution with density,

$$f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) = \mathcal{N}\left(\mathrm{y} \mid \mu_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \sigma^2_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})\right), \qquad (2)$$

which assumes that y is given by a deterministic function $\mu_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$ with additive Gaussian noise scaled by $\sigma_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$. For large, high dimensional datasets, neural networks yield suitable functional estimators $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$ and $\widehat{\sigma}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$. The mean function is then used to define a parametric CATE estimator, $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}) = \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0)$.

Standard neural network optimization often seeks a single set of parameters $\boldsymbol{\omega}_{\mathrm{ML}}$ that maximize the likelihood of the observed data $\mathcal{D}$ under the model. Therefore, it yields one prediction for novel observations $\mathbf{x}^*$, even when an $\mathbf{x}^*$ lies outside of those observed in $\mathcal{D}$, and so there is no way to discern whether $\mathbf{x}^*$ is in-distribution or out-of-distribution.

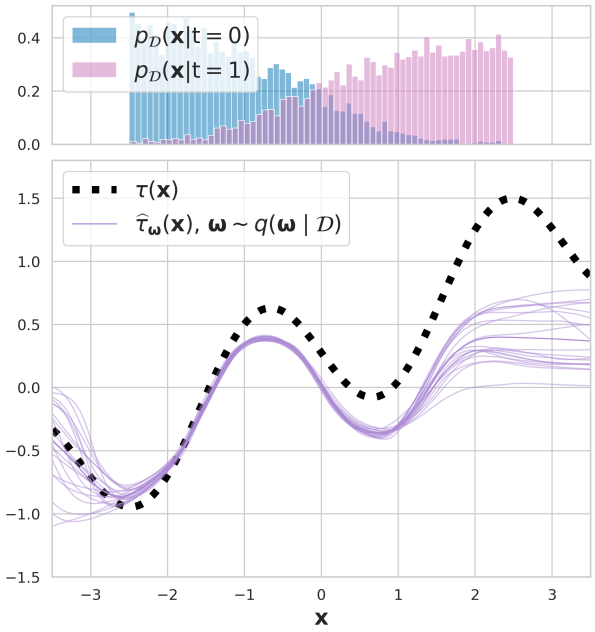


Figure 2. Samples from posterior over functions agree on the training data, but disagree off the training support. However the disagreement does not account for the bias induced by hidden confounding, hence the discrepancy between the purple samples from the model, and the true CATE $\tau(x)$ in the black dashed line.

Bayesian Deep Learning (BDL), instead aims to generate samples from the posterior distribution of the parameters given the observed data $p(\boldsymbol{\Omega} \mid \mathcal{D})$, *e.g.* from a variational approximation of the posterior $q(\boldsymbol{\Omega} \mid \mathcal{D})$ (MacKay, 1992; Hinton & Van Camp, 1993; Barber & Bishop, 1998; Gal & Ghahramani, 2016). Ideally, each sample $\boldsymbol{\omega} \sim q(\boldsymbol{\Omega} \mid \mathcal{D})$ induces a unique functional explanation that, given sufficient flexibility of the neural network, should predict $\mathrm{y} \in \mathcal{D}$. When these models work as intended, then for points $\mathbf{x}^*$ far away from the training set $\mathcal{D}$ the function values $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}^*, \mathrm{t})$ will have high variance, hence by the law of total variance, so will $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}^*)$. Figure 2 illustrates how induced functions for the CATE $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x})$ for different samples of $\boldsymbol{\omega}$ agree with one another on the training data, but disagree away from the training data. Indeed, in recent work Jesson et al. (2020) show that with high-dimensional data, BDL methods are effective at quantifying the uncertainty in CATE estimates arising from insufficient similarity and insufficient overlap.

Non-parametric methods, such as Bayesian Additive Regression Trees (BART) (Hill, 2011) or Gaussian Processes (GPs) (Alaa & van der Schaar, 2017) are also capable of expressing such uncertainty, but do not always scale well to big or high-dimensional data.

While existing Bayesian methods are well suited to account for ignorance due to insufficient similarity and overlap, the approaches above were developed under Assumption 3 (unconfoundedness) and so cannot easily account for the bias in $\widehat{\tau}(\mathbf{x})$ induced by insufficient context (hidden confounding). This is also illustrated in Figure 2. Specifically, note that even though the functions induced by sampled parameters agree with one another close to the training data, they are still biased away from the true CATE function. In order to relax Assumption 3, such ignorance must be accounted for by some other means, as we now discuss.

3.1.2. Quantifying Ignorance due to Insufficient Context

When there is insufficient context, the unconfoundedness assumption $(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$ does not necessarily hold. The challenge in this case is to quantitatively express the degree of violation of this conditional independence. We follow in the footsteps of recent work by Yadlowsky et al. (2018) and Kallus et al. (2019) who use the Marginal Sensitivity Model (MSM) proposed by Tan (2006) for this purpose.

Let $e_{\mathrm{t}}(\mathbf{x}) = P(\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x})$ be the *nominal* propensity score, and $e_{\mathrm{t}}(\mathbf{x}, \mathrm{y}) = P(\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}, \mathrm{Y}^{\mathrm{t}} = \mathrm{y})$ be the *complete* propensity score. The complete propensity, being conditioned on the potential outcome, is by construction both unconfounded and unobserved. The MSM supposes that the odds of receiving treatment under the complete propensity $\frac{e_{\mathrm{t}}(\mathbf{x},\mathrm{y})}{(1-e_{\mathrm{t}}(\mathbf{x},\mathrm{y}))}$ for individuals described by $\mathbf{x}$ differs from the odds of receiving treatment under the nominal propensity $\frac{e_{\mathrm{t}}(\mathbf{x})}{(1-e_{\mathrm{t}}(\mathbf{x}))}$ by at most a factor of Γ. That is,

$$\Gamma^{-1} \leq \frac{(1 - e_{\mathrm{t}}(\mathbf{x}))e_{\mathrm{t}}(\mathbf{x}, \mathrm{y})}{e_{\mathrm{t}}(\mathbf{x})(1 - e_{\mathrm{t}}(\mathbf{x}, \mathrm{y}))} \leq \Gamma.$$

As such $\Gamma > 1$ can be interpreted as a degree of supposed hidden confounding, whereas $\Gamma = 1$ is equivalent to the unconfoundedness assumption.

In order to incorporate the MSM into a CATE bound, Kallus et al. (2019) propose using the following factorization for

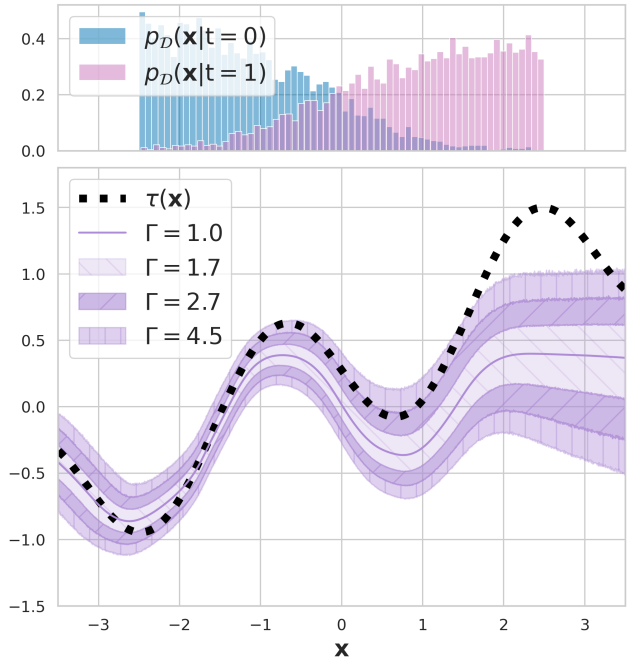


Figure 3. Varying Γ for Marginal Sensitivity Model. Ground truth $\Gamma^* = 2.7$. While the bounds follow the true CATE $\tau(x)$ on the support of $p_{\mathcal{D}}(\mathbf{x})$, they become nonsensical for out-of-distribution data ($\mathbf{x} < -2.5$ and $\mathbf{x} > 2.5$) and when there is a lack of overlap.

the expectation of the potential outcome Y^{t}:

$$\mathbb{E}[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{X} = \mathbf{x}] = \mu(w_{\mathrm{t}}; \mathbf{x}, \mathrm{t}) = \frac{\int \mathrm{y} w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x}) e_{\mathrm{t}}(\mathbf{x}) f(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}{\int w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x}) e_{\mathrm{t}}(\mathbf{x}) f(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}. \quad (3)$$

Equation (3) expresses the unbiased conditional expectation of the potential outcome in terms of the unidentifiable inverse complete propensity $w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x}) = 1/e_{\mathrm{t}}(\mathbf{x}, \mathrm{y})$ and the identifiable nominal propensity $e_{\mathrm{t}}(\mathbf{x})$ and conditional density $f(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$ of the outcome.

The MSM can then be used to define an ignorance set that includes all possible values of $w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x})$ that would violate unconfoundedness by no more than Γ, that is

$$\mathcal{W}_{\mathrm{t}}(\mathbf{x}, \Gamma) = \{w_{\mathrm{t}} : w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x}) \in [\alpha_{\mathrm{t}}(\mathbf{x}, \Gamma), \beta_{\mathrm{t}}(\mathbf{x}, \Gamma)] \forall \mathrm{y}\},$$

where $\alpha_{\mathrm{t}}(\mathbf{x}; \Gamma) = \frac{1}{\Gamma e_{\mathrm{t}}(\mathbf{x})} + 1 - \frac{1}{\Gamma}$, and $\beta_{\mathrm{t}}(\mathbf{x}; \Gamma) = \frac{\Gamma}{e_{\mathrm{t}}(\mathbf{x})} + 1 - \Gamma$. Given the set $\mathcal{W}_{\mathrm{t}}(\mathbf{x}, \Gamma)$ expressing bounded violations of unconfoundedness, (Kallus et al., 2019) suggest upper and lower bounds on the CATE as follows: $\overline{\tau}(\mathbf{x}; \Gamma) = \overline{\mu}(\mathbf{x}, 1; \Gamma) - \underline{\mu}(\mathbf{x}, 0; \Gamma)$, and $\underline{\tau}(\mathbf{x}; \Gamma) = \underline{\mu}(\mathbf{x}, 1; \Gamma) - \overline{\mu}(\mathbf{x}, 0; \Gamma)$, where

$$\underline{\mu}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \inf_{w_{\mathrm{t}} \in \mathcal{W}_{\mathrm{t}}(\mathbf{x}; \Gamma)} \mu(w_{\mathrm{t}}; \mathbf{x}, \mathrm{t}). \quad (4a)$$

$$\overline{\mu}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \sup_{w_{\mathrm{t}} \in \mathcal{W}_{\mathrm{t}}(\mathbf{x}; \Gamma)} \mu(w_{\mathrm{t}}; \mathbf{x}, \mathrm{t}) \quad (4b)$$

Taken together this gives an ignorance interval

$$\mathcal{T}(\mathbf{x}, \Gamma) = [\underline{\tau}(\mathbf{x}; \Gamma), \overline{\tau}(\mathbf{x}; \Gamma)]. \quad (5)$$

The ignorance interval $\mathcal{T}(\mathbf{x}, \Gamma)$ is completely defined with respect to identifiable estimands. For example, the likelihood in equation (2) can be used to model the density $f(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$ and a parametric model with Bernoulli likelihood, $p(\mathrm{t} \mid \mathbf{x}, \boldsymbol{\omega}) = \mathrm{Bern}(\mathrm{t} \mid \widehat{e}_{\boldsymbol{\omega}}(\mathbf{x}))$, can be used to model the identifiable nominal propensity for treatment $e_{\mathrm{t}}(\mathbf{x})$.

Kallus et al. (2019) uses a non-parameteric kernel based method and discrete line search to learn a function that maps x to the identifiable CATE intervals: $\mathcal{T}(\mathbf{x}, \Gamma)$. Figure 3 illustrated the bounds given by such a model for given assumptions on Γ.

For average treatment effects, there are two approaches for interpreting the bounds on $\tau(\mathbf{x})$ (Tan, 2006). One approach seeks the smallest value Γ_s such that the interval $[\underline{\tau}(\mathbf{x}; \Gamma_s), \overline{\tau}(\mathbf{x}; \Gamma_s)]$ crosses 0. This approach then reports that the CATE becomes sensitive to hidden confounding at Γ_s. The other approach sets a cutoff Γ_c and examines how the CATE changes for plausible Γ values below Γ_c.

There are two main limitations of the approach of (Kallus et al., 2019) that this paper seeks to address. First, as is evident in the regions of $\mathbf{x}$ that lie out of distribution ($\mathbf{x} < -2.5$ or $\mathbf{x} > 2.5$), the bounds become nonsensical (as expected), and there is no way to identify that a measurement $\mathbf{x}$ is actually out of distribution; more generally, it does not account for sources of ignorance other than unconfoundedness. Second, the method does not scale well computationally to large sample sizes, and does not scale well statistically to high-dimensional datasets as it relies on weighted kernel regression to estimate the outcome. We will now propose a method for incorporating parametric models (including BDL models) instead of the non-parametric one proposed in (Kallus et al., 2019), thus enabling better scaling to high-dimensional and large-sample setting, while at the same time also accounting for all sources of ignorance.

3.2. Estimating Bounds on $\widehat{\tau}(\mathbf{x})$ for a Fixed Degree of Hidden Confounding, Γ

We start by developing a parametric interval estimator for $\mathcal{T}(\mathbf{x}, \Gamma)$ as defined in Eq. (5). Our parametric estimator for $\mathbb{E}[\mathrm{Y}^{\mathrm{t}} \mid \mathbf{X} = \mathbf{x}]$ under hidden confounding is based off of the following equivalent expression for Equation (3)

$$\mu(w_{\mathrm{t}}; \mathbf{x}, \mathrm{t}) = \mu(\mathbf{x}, \mathrm{t}) + \frac{\int \mathrm{r}_{\mathrm{t}}(\mathrm{y}) w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x}) f(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}{\int w_{\mathrm{t}}(\mathrm{y} \mid \mathbf{x}) f(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}},$$

where the residual is given by $\mathrm{r_t(y)} = (\mathrm{y} - \mu(\mathbf{x}, \mathrm{t}))$ (see Lemma 2 in the Appendix for proof). This expression is still given in terms of both identifiable and unidentifiable quantities.

Building off the derivation in Lemma 1 of Kallus et al. (2019), we can then express the infimum and supremum in (4) as

$$\underline{\mu}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \inf_{\mathrm{y}^* \in \mathcal{Y}} \mu(\mathbf{x}, \mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} \mathrm{r_t(y)} f(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma} + \mathrm{P}(\mathrm{Y} \leq \mathrm{y}^* \mid \mathbf{x}, \mathrm{t})},$$

$$\overline{\mu}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \sup_{\mathrm{y}^* \in \mathcal{Y}} \mu(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathrm{y}^*}^{\infty} \mathrm{r_t(y)} f(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}{\alpha_{\mathrm{t}}'^{\Gamma} + \mathrm{P}(\mathrm{Y} > \mathrm{y}^* \mid \mathbf{x}, \mathrm{t})},$$

where $\alpha_{\mathrm{t}}'^{\Gamma} = \frac{\alpha_{\mathrm{t}}(\mathbf{x};\Gamma)}{\beta_{\mathrm{t}}(\mathbf{x};\Gamma) - \alpha_{\mathrm{t}}(\mathbf{x};\Gamma)}$ and $\mathcal{Y}$ is the space of outcomes as before (see Lemma 3 in the Appendix for proof). As such, the bounds on $\mathbb{E}[\mathrm{Y^t} \mid \mathbf{X} = \mathbf{x}]$ given Γ are now completely defined in terms of identifiable quantities; namely, the nominal propensity for treatment $e_{\mathrm{t}}(\mathbf{x})$, the conditional distribution of the outcome $p(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$, and its density function $f(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$, from which $\alpha_{\mathrm{t}}'^{\Gamma}$, $\mu(\mathbf{x}, \mathrm{t})$, and $\mathrm{P}(\cdot \mid \mathbf{x}, \mathrm{t})$ are straightforwardly derived.

Where Kallus et al. (2019) use a kernel-based estimator, we instead model the identifiable $p(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ directly. Thus, a generative model $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ from which to sample y and evaluate $\hat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$, and a propensity score estimator $\widehat{e}_{\mathrm{t},\boldsymbol{\omega}}(\mathbf{x})$ to evaluate $\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x}, \mathrm{t})$ are needed. Because hidden confounders induce multi-modal distributions over Y, we model $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ with a Gaussian Mixture denisty over J mixture components, noting that with a sufficient number of mixture components it can approximate any continuous distribution (Titterington et al., 1985). Thus, our density function becomes

$$f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) = \sum_{j=1}^{J} \widehat{\pi}_{\boldsymbol{\omega}}^{j}(\mathbf{x}, \mathrm{t}) \mathcal{N}\left(\mathrm{y} \mid \widehat{\mu}_{\boldsymbol{\omega}}^{j}(\mathbf{x}, \mathrm{t}), \widehat{\sigma}_{\boldsymbol{\omega}}^{j2}(\mathbf{x}, \mathrm{t})\right),$$

and $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) = \sum_{j=1}^{J} \widehat{\pi}_{\boldsymbol{\omega}}^{j}(\mathbf{x}, \mathrm{t}) \widehat{\mu}_{\boldsymbol{\omega}}^{j}(\mathbf{x}, \mathrm{t})$ (Bishop, 1994). We expand on this in the Appendix. $\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x}, \mathrm{t})$ is calculated in the same manner as $\alpha_{\mathrm{t}}'^{\Gamma}$, where $e_{\mathrm{t}}(\mathbf{x})$ is replaced by $\widehat{e}_{\mathrm{t},\boldsymbol{\omega}}(\mathbf{x})$ in the terms for $\alpha_{\mathrm{t}}(\mathbf{x}, \Gamma)$ and $\beta_{\mathrm{t}}(\mathbf{x}, \Gamma)$.

Given these models, we can now define the parametric interval CATE estimator, $\widehat{\mathcal{T}}_{\boldsymbol{\omega}}(\mathbf{x}, \Gamma) = [\underline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma), \overline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma)]$:

$$\underline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma) = \underline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, 1) - \overline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, 0), \tag{7a}$$

$$\overline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma) = \overline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, 1) - \underline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, 0), \tag{7b}$$

where

$$\underline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \inf_{\mathrm{y}^* \in \mathcal{Y}} \underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t}), \tag{8a}$$

$$\overline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \sup_{\mathrm{y}^* \in \mathcal{Y}} \overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t}), \tag{8b}$$

and for $\widehat{\mathrm{r}}_{\boldsymbol{\omega}}(\mathrm{y}, \mathrm{t}) = \mathrm{y} - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$:

$$\underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t}) = \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) + \frac{\int_{-\infty}^{\mathrm{y}^*} \widehat{\mathrm{r}}_{\boldsymbol{\omega}}(\mathrm{y}, \mathrm{t}) f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}{\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x}, \mathrm{t}) + \int_{-\infty}^{\mathrm{y}^*} f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}},$$

$$\overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t}) = \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) + \frac{\int_{\mathrm{y}^*}^{\infty} \widehat{\mathrm{r}}_{\boldsymbol{\omega}}(\mathrm{y}, \mathrm{t}) f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}{\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x}, \mathrm{t}) + \int_{\mathrm{y}^*}^{\infty} f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t}) d\mathrm{y}}.$$

3.3. Computing the Interval Estimator

Where Kallus et al. (2019) define their interval estimator as an optimization problem over n weight variables, where n is the size of the training set, we instead characterize ours as an optimization problem over m samples of y from the modeled conditional distribution $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$. Because $\underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t})$ is convex and $\overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t})$ is concave with increasing y^*, we can employ a similar discrete line search as Kallus et al. (2019) to solve this optimization problem. However, where their search has $\mathcal{O}(n)$ time complexity, ours is independent of the dataset size and has $\mathcal{O}(m)$ time complexity. m is a user-defined parameter that controls the stability of predicted $\underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k; \mathbf{x}, \mathrm{t})$ and $\overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k; \mathbf{x}, \mathrm{t})$, defined below.

This solution uses Monte-Carlo integration to estimate $\underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t})$ and $\overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(\mathrm{y}^*; \mathbf{x}, \mathrm{t})$, so as m increases the Monte-Carlo estimates converge to the integral. One could use other methods to evaluate the integrals, such as Bayesian Quadrature.

The algorithm proceeds by reordering the samples of y such that $\mathrm{y}_1 \leq \mathrm{y}_2 \leq \ldots \mathrm{y}_m$ and defining the following terms for $k \in \{1, \ldots, m\}$, $\mathbf{x} \in \mathcal{X}$, and $\Gamma \geq 1$:

$$\underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k; \mathbf{x}, \mathrm{t}) = \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) + \frac{\frac{1}{m}\sum_{i=1}^{k} \widehat{\mathrm{r}}_{\boldsymbol{\omega}}(\mathrm{y}, \mathrm{t})}{\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x}, \mathrm{t}) + \frac{k}{m}},$$

$$\overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k; \mathbf{x}, \mathrm{t}) = \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) + \frac{\frac{1}{m}\sum_{i=k+1}^{m} \widehat{\mathrm{r}}_{\boldsymbol{\omega}}(\mathrm{y}, \mathrm{t})}{\alpha_{\boldsymbol{\omega}}'^{\Gamma}(\mathbf{x}, \mathrm{t}) + 1 - \frac{k}{m}}.$$

Then, $\underline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k^L; \mathbf{x}, \mathrm{t})$, and $\overline{\widehat{\mu}}_{\boldsymbol{\omega}}^{\Gamma}(\mathbf{x}, \mathrm{t}) = \overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k^H; \mathbf{x}, \mathrm{t})$, with

$$k^L = \inf\{k = 1, \ldots, m : \underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k; \mathbf{x}, \mathrm{t}) \leq \underline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k+1; \mathbf{x}, \mathrm{t})\}$$

$$k^H = \inf\{k = 1, \ldots, m : \overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k; \mathbf{x}, \mathrm{t}) \geq \overline{\widehat{\lambda}}_{\boldsymbol{\omega}}^{\Gamma}(k+1; \mathbf{x}, \mathrm{t})\}.$$

3.4. Tightness of Bounds

Theorem 1. *Suppose that*

i $n \to \infty$, *and* $\mathbf{x} \in \mathcal{D}$.

ii Y *is a bounded random variable.*

iii $f_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$ *converges in measure to* $f(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$.

iv $\widehat{e}_{\mathrm{t},\boldsymbol{\omega}}(\mathbf{x})$ *and* $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x})$ *are consistent estimators of* $\mathbb{E}[\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x}]$ *and* $\mathbb{E}[\mathrm{Y} \mid \mathbf{X} = \mathbf{x}, \mathrm{T} = \mathrm{t}]$.

v $e_{\mathrm{t}}(\mathbf{x}, \mathrm{y})$ *is bounded away from 0 and 1 uniformly over* $\mathbf{x} \in \mathcal{X}$, $\mathrm{y} \in \mathcal{Y}$, *and* $\mathrm{t} \in \{0, 1\}$ *(overlap assumption).*

Then, for $\mathrm{t} \in \{0, 1\}$, $\underline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) \xrightarrow{p} \underline{\mu}^{\Gamma}(\mathbf{x}, \mathrm{t})$, $\overline{\widehat{\mu}}^{\Gamma}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) \xrightarrow{p} \overline{\mu}^{\Gamma}(\mathbf{x}, \mathrm{t})$, *which imply that both* $\underline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma) \xrightarrow{p} \underline{\tau}(\mathbf{x}; \Gamma)$, *and* $\overline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma) \xrightarrow{p} \overline{\tau}(\mathbf{x}; \Gamma)$.

Proof for Theorem 1 is given in Appendix B.

3.5. Model Uncertainty for $\widehat{\tau}(\mathbf{x})$ Interval Estimates

In order to account for uncertainty that arises from both a lack of similarity and a lack of overlap, it is necessary to propagate the uncertainty in $\boldsymbol{\omega}$ to the estimates of the lower and upper bounds on $\widehat{\tau}(\mathbf{x})$. The CATE bounds above are for a given parameterization $\boldsymbol{\omega} \sim q(\boldsymbol{\Omega} \mid \mathcal{D})$. By taking the expectation over $\boldsymbol{\omega}$, we arrive at

$$\begin{aligned} \underline{\widehat{\tau}}(\mathbf{x}; \Gamma) &= \mathop{\mathbb{E}}_{\boldsymbol{\omega}}[\underline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma)] - 2 \cdot \sqrt{\mathop{\mathrm{Var}}_{\boldsymbol{\omega}}[\underline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma)]} \\ \overline{\widehat{\tau}}(\mathbf{x}; \Gamma) &= \mathop{\mathbb{E}}_{\boldsymbol{\omega}}[\overline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma)] + 2 \cdot \sqrt{\mathop{\mathrm{Var}}_{\boldsymbol{\omega}}\left[\overline{\widehat{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma)\right]} \end{aligned} \tag{11}$$

$$\widehat{\mathcal{T}}(\mathbf{x}, \Gamma) = \left[\underline{\widehat{\tau}}(\mathbf{x}; \Gamma), \overline{\widehat{\tau}}(\mathbf{x}; \Gamma)\right],$$

which we name the predictive interval (using 2 standard deviations as the Bayesian confidence level here). The expectations and variances in eq. (11) can then be evaluated via Monte Carlo integration.

4. Experiments

In this section we evaluate our methods using synthetic and semi-synthetic datasets. To assess our method on high-dimensional data, we introduce a new benchmark dataset, HC-MNIST. To illustrate how our uncertainty aware bounds can be used for deferring treatment, we introduce a hidden confounding variant of the IHDP dataset (Hill, 2011). Details about the data generating processes including dataset links, code links, and validation splitting procedures are given in Appendix C.

The sampling procedure outlined in subsections 3.3-3.5 for the estimator in eq. (11) requires models for $p(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ and the nominal propensity $e_{\mathrm{t}}(\mathbf{x})$. We use a mixture density network for $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$ and a standard neural network with categorical likelihood for $\mathrm{e}_{\mathrm{t},\boldsymbol{\omega}}(\mathbf{x})$. Deep Ensembles (Lakshminarayanan et al., 2017) are used to approximate sampling $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D})$. In general, modelling $p(\boldsymbol{\Omega} \mid \mathcal{D})$ is a choice to be made by the practitioner, for example, by using Bayesian Neural Networks or simpler Bayesian models for $p_{\boldsymbol{\omega}}(\mathrm{Y} \mid \mathbf{x}, \mathrm{t})$. Details for each experiment, including architectures, hyper-parameter tuning, training procedures, and compute infrastructure are detailed in Appendix D.

4.1. Simulated Data

We first consider the one-dimensional example introduced by Kallus et al. (2019) C.1. Figure 3, generated with $n = 10000$ and $\log \Gamma^* = 1$, illustrates the nonlinear CATE function of these data. This is a useful example because both the CATE and the bias induced by hidden confounding are heterogeneous in $\mathbf{x}$. Further, Figure 3 shows that our estimator, outlined in sections 3.2 and 3.3, converges to tight bounds on the CATE interval for varying choices of Γ, achieving coverage when the assumed Γ matches the true value Γ^* used to generate the data. For this experiment and the next we assume that the outcomes correspond to costs, so that we aim to treat when $\tau(\mathbf{x}) \leq 0$.

For a quantitative evaluation, we use the same minimax-optimal policy as Kallus et al. (2019), namely, $\pi^*(\mathbf{x}; \Gamma) = \mathbb{I}(\overline{\tau}(\mathbf{x}; \Gamma) \leq 0) + \pi_0(\mathbf{x})\mathbb{I}(\underline{\tau}(\mathbf{x}; \Gamma) < 0 < \overline{\tau}(\mathbf{x}; \Gamma))$. This says that the optimal policy always treats when $\overline{\tau}(\mathbf{x}; \Gamma) \leq 0$ and otherwise reverts to the default policy $\pi_0(\mathbf{x})$. Setting $\pi_0(\mathbf{x}) = 0$, *do not treat*, our approximation to the optimal policy is given by $\widehat{\pi}(\mathbf{x}; \Gamma) = \mathbb{I}(\widehat{\overline{\tau}}(\mathbf{x}; \Gamma) \leq 0)$. The risk associated with a given policy is defined as $V(\pi; \tau) = \mathbb{E}[\pi(\mathbf{x})\mathrm{Y}^1 + (1 - \pi(\mathbf{x}))\mathrm{Y}^0]$. Intuitively, policy risk will be minimized when $\widehat{\overline{\tau}}(\mathbf{x})$ is aligned exactly with the true CATE $\tau(\mathbf{x})$, and any deviations between $\widehat{\overline{\tau}}(\mathbf{x})$ and $\tau(\mathbf{x})$ will result in a higher policy risk score. To compare different methods on a finite sample, we report the *Policy Risk Error* as the mean squared error between the risk of the optimal treatment policy $\mathbb{I}(\tau(x) < 0)$, and the policy risk of a given policy π.

In Table 1, we compare the Policy Risk Error of our method to the one proposed by Kallus et al. (2019). The average and 95% confidence intervals over 50 random realizations of training ($n = 1000$), validation ($n = 100$), and test ($n = 1000$) datasets are reported. On the diagonals we assess each policy and method with a "well-specified" $\Gamma = \Gamma^*$. These results show empirical evidence for the tightness of our interval estimator's bounds, and improved accuracy w.r.t. Kallus et al. (2019) on this low-dimension problem.

4.2. HC-MNIST: Hidden Confounding with High-dimensional Data

For this experiment, we adopt the one-dimensional simulated setting into a high-dimensional setting C.2. Specifically, we assign to each image of the MNIST dataset (LeCun,

Table 1. **Simulated Data:** Policy risk errors for various policies under data generating processes with different Γ^*. Average test-set policy risk errors and 95% confidence intervals over 50 randomly generated datasets are reported. Statistically significant improvements for "well-specified" $\Gamma = \Gamma^*$, as determined by a paired t-test (1% threshold), shown in green. Policy risk errors are multiplied by 100 for readability.

$n = 1000$	Proposed Method			Kallus et al. (2019)		
$\log \Gamma^*$	$\hat{\pi}(\mathbf{x}; \exp(0.5))$	$\hat{\pi}(\mathbf{x}; \exp(1.0))$	$\hat{\pi}(\mathbf{x}; \exp(1.5))$	$\hat{\pi}(\mathbf{x}; \exp(0.5))$	$\hat{\pi}(\mathbf{x}; \exp(1.0))$	$\hat{\pi}(\mathbf{x}; \exp(1.5))$
0.5	$\mathbf{0.07 \pm 0.03}$	0.28 ± 0.03	0.38 ± 0.04	$\mathbf{0.10 \pm 0.04}$	0.44 ± 0.09	0.99 ± 0.38
1.0	0.71 ± 0.20	$\mathbf{0.10 \pm 0.04}$	0.31 ± 0.03	0.48 ± 0.19	$\mathbf{0.25 \pm 0.11}$	0.81 ± 0.39
1.5	3.99 ± 0.59	0.75 ± 0.18	$\mathbf{0.13 \pm 0.04}$	3.33 ± 0.61	0.52 ± 0.19	$\mathbf{0.52 \pm 0.40}$

Table 2. **HC-MNIST:** Policy risk for various policies under data generating processes with different Γ^*. The proposed method approaches the ideal policy value of -1.41 under optimal policy given the true CATE. Average test-set policy risk errors and 95% confidence intervals over 20 randomly generated datasets are reported. This shows that our method scales well to large-sample, high-dimensional datasets.

	Proposed Method		
$\log \Gamma^*$	$\hat{\pi}(\mathbf{x}; \exp(0.5))$	$\hat{\pi}(\mathbf{x}; \exp(1.0))$	$\hat{\pi}(\mathbf{x}; \exp(1.5))$
0.5	$\mathbf{-1.40 \pm 0.01}$	-1.36 ± 0.01	-1.35 ± 0.01
1.0	-1.32 ± 0.02	$\mathbf{-1.40 \pm 0.01}$	-1.36 ± 0.01
1.5	-1.98 ± 0.02	-1.30 ± 0.02	$\mathbf{-1.38 \pm 0.01}$

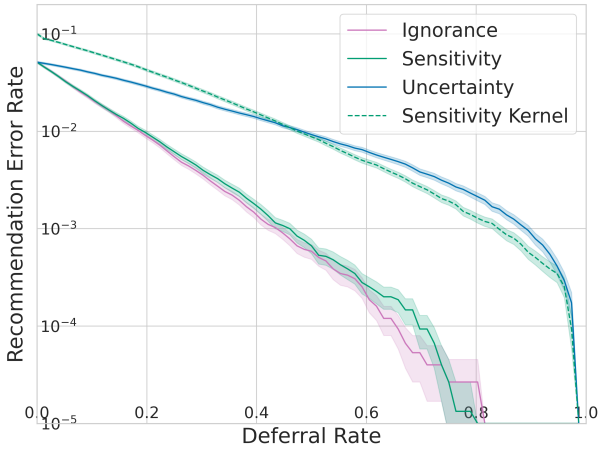


Figure 4. **IHDP Hidden Confounding:** Error rate as we sweep over the percentage of deferred points. We propose that recommendations should be deferred when there is ignorance. On the x-axis we vary the share of recommendations deferred, simulating various levels of practitioner caution. *Ignorance* (ours) accounts for all lack of knowledge. *Uncertainty* (Jesson et al., 2020) accounts only for insufficient similarity and overlap. *Sensitivity* only accounts for hidden confounding, without accounting for insufficient similarity and overlap; implemented by omitting the variance term in Eq. (11). *Sensitivity Kernel* is the kernel method of Kallus et al. (2019), which does not account for other sources of ignorance. Results show that all sources of ignorance are important on IHDP with one hidden confounder.

1998) a latent feature $\phi \in [-2, 2]$ as follows: all images of the digits 0 are assigned a $\phi \in [-2, -1.6]$, all images 1 have $\phi \in [-1.6, -1.2]$, and so on up to the digit 9. The images of every digit are sorted by brightness and ordered equally within the interval of ϕ values assigned to images of that digit. Finally, these one-dimensional hidden values ϕ are used as the inputs to the same model of hidden confounding introduced by Kallus et al. (2019) and used in the simulated data experiments above. We report the results of our method in Table 2, showing it achieves near optimal policy risk under the true level of hidden confounding. We do not to report results for Kallus et al. (2019) here as their kernel based method did not scale well to the full dataset size of MNIST, and it did not give sensible results when training only on a subset of the dataset.

4.3. IHDP Hidden Confounding

In this section we demonstrate how our uncertainty-aware interval estimator can be used to inform deferral policies for treatment recommendations. To this end we use the IHDP dataset (Hill, 2011) as Jesson et al. (2020) show that low overlap and/or similarity are problems for IHDP. For insufficient context, we induce hidden confounding by hiding covariate x_9 during model training and CATE estimation; however, it is still used for the generation of synthetic observed outcomes as per the response surface B described by Hill (2011) C.3.

In contrast to the above experiments, treatment $T = 1$ is recommended if and only if $\tau(\mathbf{x}) > 0$; we propose a deferral policy that simulates deferral to an expert and withholds a recommendation if the predicted CATE interval intersects 0. We select Γ_s such that the uncertainty aware CATE interval $[\widehat{\underline{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma_s), \widehat{\overline{\tau}}_{\boldsymbol{\omega}}(\mathbf{x}; \Gamma_s)]$ crosses 0. We then defer predictions with the lowest Γ_s value; these are predictions the model is least sure about. We compare using the same policy for the Kallus et al. (2019) method, and to the epistemic uncertainty based method proposed by Jesson et al. (2020). We report the error rate between recommendations given by $\mathbb{I}(\tau(\mathbf{x}) > 0)$ and $\mathbb{I}(\widehat{\underline{\tau}}(\mathbf{x}) > 0)$ on the remaining recommendations that were not deferred.

In Figure 4, we see that the epistemic *uncertainty* policy (blue solid line) has a moderate decrease in error rate as the rate of deferral increases. The green solid *sensitivity*

line shows that the error rate decreases as we defer recommendations based only on levels of hidden confounding. We should see the same behavior for the sensitivity method (green dashed line) proposed by Kallus et al. (2019), but it appears to struggle for higher dimensional covariates. The purple solid *ignorance* line shows that using the uncertainty aware CATE interval further improves results, showing that our method can account for all sources of ignorance discussed.

5. Conclusion

In this paper we aim to create a framework for jointly expressing the multiple sources of uncertainty, or ignorance, in individual-level causal inference. This includes uncertainty due to finite samples and due to possible violations of the standard causal inference assumptions of overlap and no-hidden confounding, as well as uncertainty due to out-of-distribution data. The novel interval estimator we present can scale to large samples and high-dimensional data, and performs well on semi-synthetic, high-dimensional datasets. We hope this work leads to further interest in research encompassing the varied possible sources of uncertainty in statistical machine learning models.

6. Acknowledgements

We would like to thank Joost van Amersfoort, Jan Brauner, OATML group members, and all anonymous reviewers for sharing their valuable feedback and insights. A.J. would like to thank Lisa, Milad, Joost, Luisa, Lewis, Tim, and Andreas for their friendship and support over a very challenging year. U.S. was partially supported by the Israel Science Foundation (grant No. 1950/19).

References

Abrevaya, J., Hsu, Y.-C., and Lieli, R. P. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015. doi: 10.1080/07350015.2014.975555. URL https://doi.org/10.1080/07350015.2014.975555.

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 3424–3432, 2017.

Barber, D. and Bishop, C. M. Ensemble learning in bayesian neural networks. *Nato ASI Series F Computer and Systems Sciences*, 168:215–238, 1998.

Bergstra, J., Yamins, D., and Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/bergstra13.html.

Bishop, C. M. Mixture density networks. 1994. URL http://publications.aston.ac.uk/id/eprint/373/.

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35 (20):3453–3470, 2016.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 2020.

Franks, A., D'Amour, A., and Feller, A. Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

Gao, Z. and Han, Y. Minimax optimal nonparametric estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2002.06471*, 2020.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hinton, G. E. and Van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Imbens, G. W. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2): 126–132, May 2003. doi: 10.1257/000282803321946921. URL https://www.aeaweb.org/articles?id=10.1257/000282803321946921.

Jesson, A., Mindermann, S., Shalit, U., and Gal, Y. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

Kallus, N., Mao, X., and Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2281–2290. PMLR, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I., and Stoica, I. Ray: A distributed framework for emerging ai applications, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pearl, J. et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

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

Perez, C. C. *Invisible women: Exposing data bias in a world designed for men*. Random House, 2019.

Robins, J. M., Hernán, M. A., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):551, 2000a.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pp. 1–94. Springer, 2000b.

Rosenbaum, P. R. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

Rosenbaum, P. R. and Rubin, D. B. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(2):212–218, 1983. ISSN 00359246. URL http://www.jstor.org/stable/2345524.

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

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Tan, Z. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006. doi: 10.1198/016214506000000023. URL https://doi.org/10.1198/016214506000000023.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Titterington, D. M., Afm, S., Smith, A. F., Makov, U., et al. *Statistical analysis of finite mixture distributions*, volume 198. John Wiley & Sons Incorporated, 1985.

Veitch, V. and Zaveri, A. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding, 2020.

Xie, Y., Brand, J. E., and Jann, B. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

Yadlowsky, S., Namkoong, H., Basu, S., Duchi, J., and Tian, L. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.