Bounding the Effects of Continuous Treatments for Hidden Confounders

Myrl G. Marmarelis[1] **Greg Ver Steeg**[1] **Neda Jahanshad**[2] **Aram Galstyan**[1]

[1]USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292
[2]USC Stevens Neuroimaging and Informatics Institute, 4676 Admiralty Way, Marina del Rey, CA 90292

Abstract

Observational studies often seek to infer the causal effect of a treatment even though both the assigned treatment and the outcome depend on other confounding variables. An effective strategy for dealing with confounders is to estimate a propensity model that corrects for the relationship between covariates and assigned treatment. Unfortunately, the confounding variables themselves are not always observed, in which case we can only bound the propensity, and therefore bound the magnitude of causal effects. In many important cases, like administering a dose of some medicine, the possible treatments belong to a continuum. Sensitivity models, which are required to tie the true propensity to something that can be estimated, have been explored for binary treatments. We propose one for *continuous treatments*. We develop a framework to compute *ignorance intervals* on the partially identified dose-response curves, enabling us to quantify the susceptibility of an inference to hidden confounders. We show with simulations and three real-world observational studies that our approach can give non-trivial bounds on causal effects from continuous treatments in the presence of hidden confounders.

1 INTRODUCTION

The goal of machine learning is often to build a good predictor. A separate task that is crucial to any scientific endeavor is to identify the causal drivers. This entails recording as many covariates as possible and adjusting for any confounding before drawing conclusions. Sometimes an interventional study is prohibitive due to cost or ethics, like forcing someone to smoke. In those cases, one must rely on observational studies and remove confounding *post hoc*. A plethora

of methods operate on the basis of a learned propensity model for the assigned treatment conditioned on covariates, for instance to reweigh the sample and remove any visible biases. Quite frequently, the covariates are inadequate to account for all the hidden paths between the treatment and the outcome, and propensity-based approaches cannot discern the real effect. Routine discoveries of dietary and lifestyle confounders lead researchers to vascillate on the health implications of e.g. coffee [Atroszko, 2019], alcohol [Ystrom et al., 2022], and cheese [Godos et al., 2020].
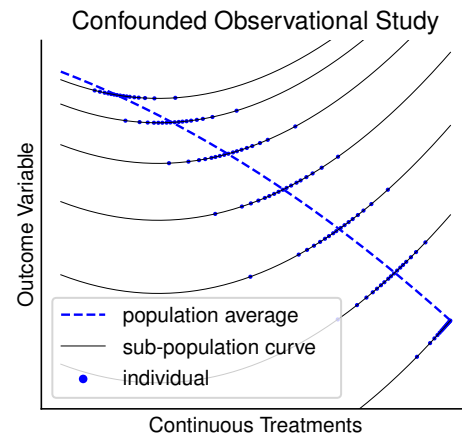

Figure 1. When a hidden confounder is distorting the assigned treatments in sub-populations, the overall population-level trend may appear flipped in comparison to each sub-population's dose response. We deal with the cases where the variables that separate the sub-populations are unobserved. Refer to §D for details on the illustration.

We shall highlight the pitfalls of failing to accommodate possible confounders on estimated dose responses. The widely noted J-shaped curves are common in physiology [Calabrese and Baldwin, 2001]. Their characteristic dynamic occurs wherever a substance in moderate concentration exhibits vastly different effects than either extremes of high or low doses. A simulated example in Figure 1 demonstrates how a

J-shaped treatment effect can appear flipped in observational data due to confounding. The phenomenon is an example of Simpson's paradox [Simpson, 1951, Yule, 1903].

1.1 RELATED WORKS.

There is growing interest in causal methodology for treatments (or exposures, interventions) that take on specific values within a continuum, especially in the fields of econometrics [e.g. Huang et al., 2021, Tübbicke, 2022], health sciences [Vegetabile et al., 2021], and machine learning [Ghassami et al., 2021, Colangelo and Lee, 2021, Kallus and Santacatterina, 2019]. So far, much scrutiny on partially identified potential outcomes has focused on the case of binary treatments, the simplest setting [e.g. Rosenbaum and Rubin, 1983, Louizos et al., 2017, Lim et al., 2021]. A number of creative approaches were exhibited in the past few years to make strides in this binary setting. Most of them relied on a *sensitivity model* for bounding the extent of possible unobserved confounding, to which downstream tasks may be adapted by noting which treatment effects degrade the quickest.

Quite recently, attempts were made to handle unobserved confounding with continuous treatments by optimizing the treatment effect bounds with generative models [Padh et al., 2022, Hu et al., 2021], rather than a sensitivity model. One employs instrumental variables [Kilbertus et al., 2020]. Another, with a sensitivity model, was developed in parallel to the present work [Jesson et al., 2022].

Regarding binary treatments, the so-called Marginal Sensitivity Model (MSM) due to Tan [2006] continues to be studied extensively [Zhao et al., 2019, Veitch and Zaveri, 2020, Yin et al., 2021]. Variations thereof include Rosenbaum's earlier sensitivity model [2002] that enjoys ties to regression coefficients [Yadlowsky et al., 2020]. Other groups have borrowed strategies from deep learning [Wu and Fukumizu, 2022] rather than opting for the MSM. Another active line of work constructs bounds not due to ignorance on confounding but instead viewed from the lens of robustness [Guo et al., 2022, Makar et al., 2020, Johansson et al., 2020]. The MSM is highly interpretable with its single free parameter, and applicable to a wide swath of models.

Other approaches require a more complex data-generating process than the simple observed outcome, treatment, covariates tuple (Y, T, X). For instance, proximal causal learning [Tchetgen et al., 2020, Mastouri et al., 2021] loosens confounding requirements with additional proxy variables. Chen et al. [2022] rely on multiple large dataset partitions.

1.2 CONTRIBUTIONS.

The MSM is incompatible with continuous treatments. Our first contribution is to propose a unique sensitivity model

(§1.4) that extends the MSM to a treatment continuum. Next, we derive general formulas (§2) and specialize them to versatile closed forms (§2.2) culminating in Theorem 1. We devise an efficient algorithm (§3) to compute ignorance bounds over dose-response curves, following up with experiments on simulated (§4) and real (§5) datasets.

1.3 POTENTIAL OUTCOMES.

Causal inference is often cast in the nomenclature of potential outcomes, due to Rubin [1974]. The broad goal is to measure a treatment's effect on an individual, marked by a set of covariates, while accounting for all the confounding between the covariates and the treatment variable. Effects could manifest heterogenously across individuals. In non-interventional settings, observed covariates may not entirely overlap across treatment regimens. It is typical to estimate two models, (1) the outcome predictor and (2) a model for the propensity of treatment conditioned on the covariates. The latter may help account for biases.

The first two assumptions involved in Rubin's framework are that observations of outcome, assigned treatment, and covariates $\{Y^{(i)}, T^{(i)}, X^{(i)}\}$ are i.i.d draws from the same population and that all treatments have a chance to occur for each covariate vector: $p_{T|X}(t \mid x) > 0$ (overlap/positivity) for all $t, x \in [0, 1] \times \mathcal{X}$, specifically in our context of continuous treatments. The third and most challenging of these fundamental assumptions is that of *ignorability*, or sufficiency. Our study is concerned with the scenarios where that assumption is violated: when there exists a dependency, not blocked by the covariates, between the assigned treatment and true potential outcomes. Let $p(y_t|x)$ denote the probability density function of *potential* outcome $Y_t = y_t$ from a treatment $t \in [0, 1]$, given covariates $X = x$. Formally, we have a violation of ignorability:

$$\{(Y_t)_{t \in \mathcal{T}} \not\perp\!\!\!\perp T\} \mid X.$$

It is only realistic to observe samples of $Y|T = t, X = x$ with density $p(y_t|t, x)$. However, to account for possible hidden confounding, we also require a $p(y_t|\tau \neq t, x)$ for quantifying treatment effects of the general form $\mathbb{E}[f(Y_t)|X]$, involving the density

$$p(y_t|x) = \int_0^1 p(y_t|\tau, x)p(\tau|x) \, \mathrm{d}\tau, \qquad (1)$$

where $p(y_t|\tau, x)$ is the distribution of potential outcomes conditioned on actual treatment $T = \tau \in [0, 1]$ that may differ from the potential outcome's index t. Throughout this study, y_t will indicate the value of the potential outcome at treatment t, and to disambiguate with *assigned* treatment τ will be used for events where $T = \tau$. For instance, we may care about the counterfactual of a smoker's ($\tau = 1$) health outcome had they not smoked ($y_{t=0}$), where $T = 0$

signifies no smoking and $T = 1$ is "full" smoking. We aim to develop some intuition before introducing the novelties.
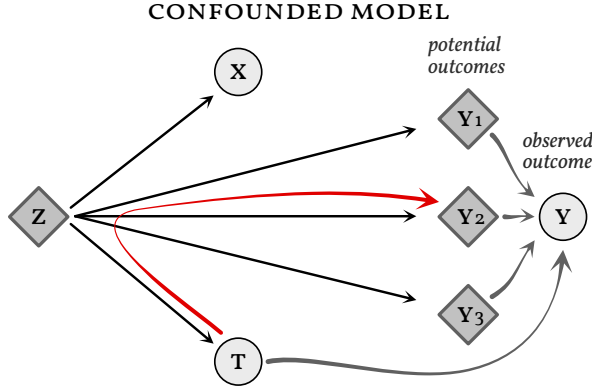

CONFOUNDED MODEL

Figure 2. A simplified setting with three potential outcomes Y_1, Y_2, Y_3 arising from treatment assignments $T = 1, 2, 3$. Black arrows represent the ground-truth process. The red arrow shows what could happen when confounder Z is hidden and covariate X is inadequate to block all links between T and $Y_{1,2,3}$ through Z, leading Z to "leak" into the estimated treatment effect. The probability density $p(y_t|\tau, x)$, found in the integrand of Equation 1, diverges from $p(y_t|x)$ due to this leakage.

On notation. We will use the shorthand $p(\cdots)$ with lowercase variables whenever working with probability densities of the corresponding stochastic variables in uppercase. In other words,

$$p(\tau|x) \text{ means } \frac{\mathrm{d}}{\mathrm{d}\tau}\mathbb{P}[T \leq \tau|X = x], \quad \text{and}$$

$$p(y_t|\tau, x) \text{ means } \frac{\mathrm{d}}{\mathrm{d}u}\mathbb{P}[Y_t \leq u|T = \tau, \ X = x]\Big|_{u=y_t}.$$

Interpretation. How would one interpret $p(y_t|\tau, x)$? The potential-outcomes vector $(Y_t)_{t\in[0,1]}$ of infinite dimensionality is *intrinsic* to each individual with true confounder Z, for which X is a noisy proxy. By "true" confounder we refer to any set of variables that suffice to block all backdoor paths between Y_t and T. The potential-outcomes vector would only change from knowledge of assigned treatment $T = \tau$ if it betrayed additional information about Z, absent in X, that further informed any Y_t. We may express $p(y_t|\tau, x)$ explicitly in terms of hypothetical true confounders as $\int p(y_t|z)p(z|\tau, x)\,\mathrm{d}z$ because z subsumes both x and τ. This way, $p(y_t|z)$ is the true potential outcome and $p(z|\tau, x)$ acts as a filter for how parts of the true confounder mix together into the proxy x and the assigned treatment τ.

Propensities. The probability density $p(\tau|x)$ is termed the *nominal propensity*. A quantity often examined is the *complete propensity*, specifically referring to $p(\tau|y_t, x)$ in our realm. The complete propensity can differ from $p(\tau|x)$

because of hidden confounders. In that instance, conditioning on potential outcome y_t modulates the distribution. Similarly, by connection through Bayes' rule, conditioning the potential outcomes $p(y_t|x)$ on assigned treatment τ modulates those distributions. Absent any unobserved confounding, $p(y_t|\tau, x) = p(y_t|x)$ and Equation 1 trivializes. See Figure 2 for a graphical illustration on the runaway influence of τ on the potential outcomes.

Sensitivity. Explored by Kallus et al. [2019] and Jesson et al. [2021] among many others, the Marginal Sensitivity Model (MSM) serves to bound the extent of (putative) hidden confounding in the regime of binary treatments $T' \in \{0, 1\}$. Specifically, it couples the odds of treatment under the nominal propensity to the odds of treatment under complete propensity, limiting the discrepancy:

Definition 1 (The Marginal Sensitivity Model). For binary treatment $t' \in \{0, 1\}$ and violation factor $\Gamma \geq 1$, the following ratio is bounded:

$$\Gamma^{-1} \leq \left[\frac{p(t'|x)}{1 - p(t'|x)}\right]^{-1}\left[\frac{p(t'|y_{t'}, x)}{1 - p(t'|y_{t'}, x)}\right] \leq \Gamma.$$

Restricting ourselves to binary treatments affords us a number of conveniences. For instance, one probability value is sufficient to describe the whole propensity landscape on a set of conditions, $p(1-t'|\cdots) = 1-p(t'|\cdots)$. As we transfer to the separate context of $t \in [0, 1]$, we must contend with infinite treatments and infinite potential outcomes.

1.4 TOWARDS CONTINUOUS SENSITIVITY.

We require a constraint on the fundamentally unknowable quantity $p(\tau|y_t, x)$ for any treatment $T = \tau \in [0, 1]$ and potential outcome y_t, for possibly contrary treatment assignments $\tau \neq t$. As with the MSM, our target is to associate $p(\tau|y_t, x)$ to the knowable $p(\tau|x)$. In other words, we seek to constrain the knowledge conferred on propensity by a single potential outcome y_t. It is not necessary for the functions pertaining to $(y_t)_{t\in[0,1]}$ to exhibit any degree of smoothness in t. The potential-outcome variables are treated as entries in an infinitely long vector. However, we do impose that the propensity probability densities $p(\tau|\dots)$ are at least once differentiable in τ. What sort of analogue exists for the notion of "odds" in the MSM?

Contrast treatment τ versus $\tau + \delta$ locally, for some infinitesimal δ, at any part of the curve. A translation of the MSM might appear as

$$\left[\frac{p(\tau + \delta|x)}{p(\tau|x)}\right]^{-1}\left[\frac{p(\tau + \delta|y_t, x)}{p(\tau|y_t, x)}\right].$$

Let us peer into one of those ratios. In logarithms,

$$\delta^{-1} \log \frac{p(\tau + \delta | x)}{p(\tau | x)} = \frac{\log p(\tau + \delta | x) - \log p(\tau | x)}{\delta}$$

$$\xrightarrow[\delta \to 0]{} \frac{\partial \log p(\tau | x)}{\partial \tau} \triangleq \partial_\tau \log p(\tau | x).$$

Hence, we introduce the notion of an infinitesimal MSM (δMSM), tying $\partial_\tau \log p(\tau | y_t, x)$ to $\partial_\tau \log p(\tau | x)$.

Definition 2 (The Infinitesimal Marginal Sensitivity Model). For treatments in the closed unit interval, $t \in [0, 1]$, and violation factor $\Gamma \geq 1$, the following holds everywhere:

$$\left| \partial_\tau \log \frac{p(\tau | y_t, x)}{p(\tau | x)} \right| \leq \log \Gamma.$$

We crafted the δMSM with the intention of functionally mirroring the MSM—locally, on a treatment continuum. Whereas Definition 2 is stated in logarithms, Definition 1 is not; the difference is merely cosmetic and hyperparameter Γ plays an equivalent role in both structures. Nevertheless, the emergent properties are vastly different.

2 THE FRAMEWORK

We list the core assumptions surrounding our problem.

Assumption 1 (Bounded Hidden Confounding). Invoking Definition 2, the violation of ignorability is constrained by a δMSM with some $\Gamma \geq 1$.

Assumption 2 (Observed Confounding at No Treatment). The utter lack of treatment is not informed by potential outcomes: $p(\tau = 0 | y_t, x) = p(\tau = 0 | x)$ for all t and y_t.

Assumption 2 states that we look for sensitivity to hidden confounders outside the control group at $T = 0$. The restriction is reasonable in situations like the following: we seek to estimate the effect of a prescription drug, and some clinics prescribe different dosages. Our $T = 0$ group would be individuals who have not received any such prescription, and $T > 0$ would place patients on a scale depending on prescription dosage. We expect a dramatically lessened vulnerability to hidden confounders for the well-represented—in observed *and* unobserved attributes—control group. From a technical perspective, Assumption 2 is necessary for our derivations, and should be interpreted as a blind spot in the sensitivity model rather than a requirement for the underlying process. There is no additional constraint, besides the δMSM itself, on how much the complete propensity function may fluctuate around any $T > 0$. We motivate and validate this assumption in the real world with §5.

Next, we proceed with derivations. The key to cracking open Equation 1 is to carve out a region inside the domain of

integration where an approximation can be trusted. This will extrapolate from the singular point $\tau = t$ where estimation is feasible.

2.1 DEALING WITH AN UNRELIABLE APPROXIMATION.

We expand $p(y_t | \tau, x)$ around $\tau = t$, where $p(y_t | t, x) = p(y | t, x)$ is learnable from data. Suppose that $p(y_t | \tau, x)$ is twice differentiable almost everywhere in τ. Construct a Taylor expansion

$$p(y_t | \tau, x) = p(y_t | t, x) + (\tau - t)\partial_\tau p(y_t | \tau, x)|_{\tau=t}$$
$$+ \frac{(\tau - t)^2}{2} \partial_\tau^2 p(y_t | \tau, x)|_{\tau=t} + \mathcal{O}(\tau - t)^3. \quad (2)$$

Denote with $\tilde{p}(y_t | \tau, x)$ an approximation of first or second order as laid out above. We will encounter that even $\partial_\tau p(y_t | \tau, x)|_{\tau=t}$ is intractable. Thankfully, it can be bounded using the δMSM machinery. Let us quantify the reliability of this approximation by a set of weights $0 \leq w_t(\tau) \leq 1$, where typically (but need not necessarily) $w_t(t) = 1$. Split Equation 1 along these weights by applying the following identity, and then approximate the first part only:

$$p(y_t | x) = \int_0^1 w_t(\tau) p(y_t | \tau, x) p(\tau | x) \, \mathrm{d}\tau$$
$$+ \int_0^1 [1 - w_t(\tau)] p(y_t | \tau, x) p(\tau | x) \, \mathrm{d}\tau$$
$$\approx \int_0^1 \underbrace{w_t(\tau) \tilde{p}(y_t | \tau, x) p(\tau | x) \, \mathrm{d}\tau}_{(A) \text{ the approximated quantity}} \quad (3)$$
$$+ \int_0^1 \underbrace{[1 - w_t(\tau)] p(\tau | y_t, x) p(y_t | x) \, \mathrm{d}\tau}_{(B) \text{ by Bayes' rule}}.$$

This separation into recoverable (A) and entirely unknown (B), demarcated by the weights, ensures that the inaccurate regimes of the approximation vanish (as $w_t(\tau) \to 0$ away from t) and are replaced with the ignorant quantity. We simplify part B of Equation 3 first:

$$\int_0^1 [1 - w_t(\tau)] p(\tau | y_t, x) p(y_t | x) \, \mathrm{d}\tau$$
$$= p(y_t | x) \left[1 - \int_0^1 w_t(\tau) p(\tau | y_t, x) \, \mathrm{d}\tau \right].$$

We witness already that $p(y_t | x)$ shall take the form of

$$p(y_t | x) \approx \frac{\int_0^1 w_t(\tau) \tilde{p}(y_t | \tau, x) p(\tau | x) \, \mathrm{d}\tau}{\int_0^1 w_t(\tau) p(\tau | y_t, x) \, \mathrm{d}\tau}. \quad (4)$$

How the approximation error of Equation 2 carries into Equation 4 is addressed by Proposition 2, for the cases considered in this study. To proceed further demands reflecting

on Assumptions 1 & 2. We will expand the denominator first, and then repurpose the results for the derivatives of Equation 2 that appear in the numerator. This part shows how the assumed model serves to characterize the unknown quantities, but the impatient reader may skip to §2.2. Without loss of generality, consider

$$\partial_\tau \log p(\tau|y_t, x) = \partial_\tau \log p(\tau|x) + \gamma(\tau|y_t, x),$$
$$|\gamma(\tau|y_t, x)| \le \log \Gamma. \quad (5)$$

We may attempt to integrate both sides;

$$\int_0^{t'} \partial_\tau \log p(\tau|y_t, x)\, \mathrm{d}\tau = \int_0^{t'} \partial_\tau \log p(\tau|x)\, \mathrm{d}\tau$$
$$+ \underbrace{\int_0^{t'} \gamma(\tau|y_t, x)\, \mathrm{d}\tau}_{\triangleq \lambda(t'|y_t, x)}$$
$$\iff \log p(\tau = t'|y_t, x) - \log p(\tau = 0 \,|y_t, x)$$
$$= \log p(\tau = t'|x) - \log p(\tau = 0 \,|x) + \lambda(t'|y_t, x),$$
$$\log p(\tau|y_t, x) = \log p(\tau|x) + \lambda(\tau|y_t, x)$$
$$\text{(by Assumption 2)}.$$

$$\therefore \quad p(\tau|y_t, x) = p(\tau|x)\Lambda(\tau|y_t, x), \quad \Lambda \triangleq \exp\{\lambda\}. \quad (6)$$

Clearly $|\lambda(\tau|y_t, x)| \le \tau \log \Gamma$ because it integrates γ, bounded by $\pm \log \Gamma$, over a support with length τ. Subsequently $\Lambda(\tau|y, t)$ is bounded by $\Gamma^{\pm\tau}$. We are now equipped with the requisite tools to properly bound $p(y_t|x)$—or an approximation thereof, erring on ignorance via reliability weights $w_t(\tau)$. The full derivation may be found in §A.

Predicting potential outcomes. The recovery of a fully normalized probability density $\tilde{p}(y_t|x)$ via Equation 4 is laid out below. It may be approximated with Monte Carlo or solved in closed form with specific formulations for the weights and propensity. Generally, it takes on the form $\tilde{p}(y_t|x) = d(t|y_t, x)^{-1}p(y_t|t, x)$, where

$$d(t|y_t, x) \triangleq \mathbb{E}_\tau[\Lambda(\tau|y_t, x)] - (\gamma\Lambda)(t|y_t, x)\, \mathbb{E}_\tau[\tau - t]$$
$$- \frac{1}{2}((\dot{\gamma} + \gamma^2)\Lambda)(t|y_t, x)\, \mathbb{E}_\tau[(\tau - t)^2], \quad (7)$$

and said expectations, $\mathbb{E}_\tau[\cdot]$, are with respect to the implicit distribution $q(\tau|t, x) \propto w_t(\tau)p(\tau|x)$. The notation $\dot{\gamma}$ denotes a derivative in the first argument of $\gamma(t|y_t, x)$.

To make use of this formula, one first procures the set of admissible $d(t|y_t, x) \in [\underline{d}(t|y_t, x), \overline{d}(t|y_t, x)]$ that violate ignorability up to a factor Γ according to the δMSM. Then, considering their reciprocals as importance weights [Tokdar and Kass, 2010], tight bounds on the partially identified expectations over $\tilde{p}(y_t|x)$ may be optimized.

A note on ensemble uncertainty. One should quantify empirical uncertainties [Jesson et al., 2020] alongside sensitivity to hidden confounding. In our experiments we learn both the predictor and the propensity model as ensembles from bootstrapped resampled [Lo, 1987] data. Then $\tilde{p}(y_t|x)$ can also be resampled for confidence intervals via its component ensembles.

2.2 TRACTABLE WEIGHT COMBINATIONS.

In addition to developing the general framework above, we derive analytical forms for a specific paramametrization to the weighting function and propensity distribution. Here, we look to the Beta function and its associated probability density for a natural solution. Suppose that

$$(T|X = x) \sim \text{Beta}(\alpha(x), \beta(x)),$$
$$\text{for arbitrary } \alpha(x), \beta(x), \quad (8)$$

$$w_t(\tau) = \frac{\tau^{a_t - 1}(1 - \tau)^{b_t - 1}}{c_t} = \frac{\tau^{rt}(1 - \tau)^{r(1-t)}}{c_t},$$
$$a_t + b_t = r + 2, \quad r > 0. \quad (9)$$

We designed the reliability weights to mirror the propensity's form by rescaling a Beta density. We assert that $w_t(\tau)$ peaks at $\tau = t$, and that $w_t(\tau) = 1$. We find that $c_t \triangleq t^{rt}(1 - t)^{r(1-t)}$, even though the solution is irrelevant for our purposes. The mode is fixed:

$$\frac{a_t - 1}{a_t + b_t - 2} = t.$$

The remaining degree of freedom disappears by a precision constraint $a_t + b_t - 2 = r$ for some $r > 0$. Constraining a more complex dispersion statistic like variance is much more difficult.
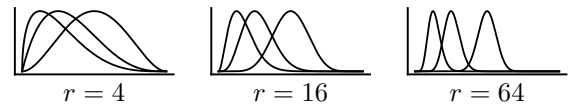
Figure 3. Beta weight schemes $w_t(\tau)$ in the unit square, plotted for centers $t = 0.125, 0.25, 0.5$. Shapes are symmetrical about $t = 0.5$. Trust declines with r.

The expectations found in Equation 7 are now available in closed form, and can be bounded in terms of just two extra free parameters, Γ and r. Guidance on setting the violation factor Γ is discussed elsewhere, e.g. §5; as for the class of weights, high r conveys poor trust in the Equation 2 approximation, as characterized by Proposition 2 below.

The findings. We pose a third and final assumption, which enables us to state Theorem 1. The main insight to unlocking

those expectations is that each one of them involves an integral with the product $w_t(\tau)p(\tau|x)$ over its normalization constant, yielding the moments of a Beta distribution.

Assumption 3 (Second-order Simplification). The quantity $\dot{\gamma}(\tau|y_t, x)$ cannot be characterized as-is. We grant that γ^2 dominates over the former, and consequently

$$\left| (\dot{\gamma} + \gamma^2)\Lambda \right| \leq \left| \gamma^2 \Lambda \right| + \varepsilon, \quad \text{for small } \varepsilon \geq 0.$$

Theorem 1 (Beta Parametrizations). *The formulations in Equations 8 & 9 admit analytical solutions to the ignorance denominator in Equation 7.*

$$\mathbb{E}_\tau[\Lambda(\tau|y_t, x)] \in \\ {}_1F_1(\alpha(x) + a_t - 1; \; \alpha(x) + \beta(x) + r; \; \pm \log \Gamma),$$

where again the operator $\mathbb{E}_\tau$ is employed as in Equation 7, and $_1F_1$ denotes Kummer's confluent hypergeometric function [Mathews Jr. et al., 2021]. In addition, with α, β implicitly referring to $\alpha(x), \beta(x)$,

$$\mathbb{E}_\tau[\tau - t] = \frac{(1-t)\alpha - t\beta}{\alpha + \beta + r},$$

$$\mathbb{E}_\tau[(\tau - t)^2] = \Big[(\alpha + \beta + r)(\alpha + \beta + r + 1)\Big]^{-1} \\ \cdot \Big[(\alpha^2 + \beta^2 + \alpha + \beta - r)t^2 \\ - (2\alpha^2 + 2\alpha\beta + 2\alpha - r)t + (\alpha + 1)\alpha\Big].$$

To take these findings one step further, we compare any weighting function $w_t(\tau)$ to a baseline or phantom Beta scheme with relatively low r.

Proposition 2 (Absolute Accuracy). *Suppose that*

1. *$\partial_\tau^2 p(y_t|\tau, x)$ is Q-Lipschitz, and that*

2. *the compound propensity-weights $w_t(\tau)p(\tau|x)$ _and_ $w_t(\tau)p(\tau|y_t, x)$ are bounded on both sides by a phantom Beta scheme, of the form in Equation 9 with narrowness $r^\star$, up to a margin of $c^\star \geq 1$.*

More specifically by the second item,

$$w_t^\star(\tau)/c^\star \leq \{w_t(\tau)p(\tau|x), w_t(\tau)p(\tau|y_t, x)\} \leq c^\star w_t^\star(\tau).$$

Then the outcome approximator $\tilde{p}(y_t|x)$ of second order (Equation 7) has an absolute error stemming from the Taylor remainder on $p(y_t|\tau, x)$, stated in terms of these constants:

$$|p(y_t|x) - \tilde{p}(y_t|x)| \leq (c^\star)^2 (r^\star + 1) k(r^\star) Q.$$

The unitless $k(r^\star)$, pertaining to the numerator in Equation 4, was solved numerically for certain values of $r^\star$;

$r^\star$	0.5	1	2	4
$(r^\star + 1) k(r^\star)$	0.03	0.02	0.01	0.005

Pertinent to the above result is the tendency for compound propensity-weights $w_t(\tau)p(\tau|x)$ to be narrower than the weights $w_t(\tau)$ themselves. This holds especially in the Beta family of propensities, per the discussion surrounding Equation 8. See §B for more details on the proposition.

3 COMPUTING THE IGNORANCE INTERVALS

After deriving $\tilde{p}(y_t|x)$ in Equation 7 and a specific solution with Theorem 1, we must find a way to bound the partially identified expectations with respect to this distribution. Concretely, we seek to characterize $\mathbb{E}[f(Y_t)|X = x]$ for any task-specific $f(y)$. This is accomplished with a Monte Carlo importance sampler of n realizations y_i drawn from proposal $q(y)$:

$$\tilde{\mathbb{E}}[f(Y_t)|X = x] = \frac{\sum_{i=1}^n f(y_i)\tilde{p}(y_t = y_i|x)/q(y_i)}{\sum_{i=1}^n \tilde{p}(y_t = y_i|x)/q(y_i)}. \tag{10}$$

Even though $\tilde{p}(y_t|x)$ is a normalized probability density, it contains partially identified quantities. It is untenable to constrain a search along the candidate values for each $d(t|y_t = y_i, x)$ to even approximately ensure $\int_{\mathcal{Y}} \tilde{p}(y_t = y|x) \, \mathrm{d}y = 1$. For this reason the bias of an estimator without the corrective denominator of Equation 10 would be uncontrollable [Tokdar and Kass, 2010]. A greedy algorithm may be deployed to maximize $\tilde{\mathbb{E}}[f(Y_t)|X = x]$ in this form by using weights

$$\underline{w}_i := \frac{p(y_i|t, x)}{\overline{d}(t|y_i, x)q(y_i)}, \qquad \overline{w}_i := \frac{p(y_i|t, x)}{\underline{d}(t|y_i, x)q(y_i)}.$$

The minimum may be achieved by a trivial extension. Maximizing and minimizing $\tilde{\mathbb{E}}[f(Y_t)|X = x]$ with respect to the bounding quantities (γ, Λ) enables the resolution of ignorance bounds on the basis of Γ from Definition 2.

Our algorithm, displayed below, adapts the method of Jesson et al. [2021] to heterogeneous weight bounds $[\underline{w}_i, \overline{w}_i]$ per draw i. View a proof of correctness in §C.

4 A SYNTHETIC EXPERIMENT

We first studied an idealized scenario with one hidden confounder and no covariate shift. In our design we sought to induce a complicated enough dose-response function in (T, X), and a true propensity $T|Z$ that deliberately obeys Assumption 2. We constructed a piecewise uniform distribution for $T|Z$ with variable categorical probabilities assigned to each "bucket," always residing in the probability simplex. The coefficients (see §E) were concocted prior to experimentation. For covariates, true confounder z_1 passed through entirely and z_2 remained hidden.

Crucially, we placed an intentional relationship between the single covariate $X \in [0, 1]$ and the amount of unobserved

input : $\{(\underline{w}_i, \overline{w}_i, f_i)\}_{i=1}^n$ ordered by ascending f_i.
output : $\max_w \mathbb{E}[f(X)]$ estimated by importance
sampling with n draws.

Initialize $w_i \leftarrow \overline{w}_i$ for all $i = 1, 2, \ldots n$;
for $j = 1, 2, \ldots n$ **do**
  Compute $\Delta_j \triangleq \sum_{i=1}^n w_i(f_j - f_i)$;
  if $\Delta_j < 0$ **then**
    $w_j \leftarrow \underline{w}_j$;
  else
    break;
  end
end
Return $\sum_i w_i f_i / \sum_i w_i$;

Algorithm 1: The expectation maximizer, with $\mathcal{O}(n)$ runtime if intermediate Δ_j results are memoized.

confounding. The model suffered more near $X = 0$ than $X = 1$. We challenged our δMSM to discern this and grow the ignorance intervals more for lower X, as it should have been informed by the propensity estimates. And it was. We pinpoint this differential growth in Figure 4.

4.1 A GROUNDED IGNORANCE MEASURE.

In reality, an unconfounded ground truth—like the black line in Figure 4—would stay unobserved. This instance, however, lends us special tools for model evaluation. Uncertainty of an ignorance interval $[\underline{y}, \overline{y}]$ with respect to a true Bernoulli($y^\star$) outcome was measured in an information-theoretic way. We calculated for every (t, x) the expression $\int_{\underline{y}}^{\overline{y}} \mathrm{KL}_B(y^\star \| y) \, \mathrm{d}y$, where KL_B is shorthand for the Kullback-Leibler divergence between two Bernoulli variables with the given probabilities, true $y^\star$ and inferred y. This way, we grounded the ignorance intervals on their average approximation cost.

5 RESULTS FROM AN OBSERVATIONAL STUDY

The most pertinent application for the framework laid out above is an observational study with incomplete or noisy covariates and a continuous treatment variable. More concretely, the treatment variable should be transformed and scaled into the unit interval such that $T = 0$ signifies a control with a complete lack of treatment. Every *kind* of individual should be about equally likely to fall in the $(T = 0)$ cohort (Assumption 2.) As for shaping the $(T > 0)$ regime, the domain should inform whether a linear scale is employed, versus an empirical or parametric (i.e. standard normal) cumulative density function.

The data. We conducted multiple experiments based on observational studies, showcased in Table 1. Refer to §F for particulars on the dataset construction, model design, and test/train learning curves. A quarter of the data were reserved randomly for the test sets.

Dataset	dim $\mathcal{X}$	dim $\mathcal{X}_c$	Sample	Outcome
Vitamin	14	3	253	Bernoulli
Energy	14	5	5,158	Gamma
Brain (1–4)	100	(73, 58, 86, 93)	14,551	Gamma

Table 1. Exposition of the datasets constructed for our experiments. Columns display the dimensionality of the full covariate vectors X, the censored versions, sample sizes, and outcome parametrizations. For Brain, the outcome is 74 variables; the rest are univariate.

The estimators. In both experiments, the predictor and ensemble were trained as ensembles of artificial neural networks with two inner residual layers. We chose the "swish" activation function for its well-behaved gradients [Ramachandran et al., 2017]. The flagship ADAM optimizer was employed for stochastic gradient descent with a learning rate of 10^{-3} and no mini-batching. Our analysis involved bootstrapping the ensemble to garner empirical uncertainties. Outcomes were modeled as in the last column of Table 1.

Test-set log-likelihoods do not degrade much between the full and censored datasets. One cannot look to predictive performance for how well the fully unconfounded dose-response curves are captured, since any hidden confounding would persist. Synthetic ground-truth potential outcomes always carry the risk of presenting an unrealistic scenario [Curth et al., 2021].

The objective. We chose to investigate the coverage of $\mathbb{E}[Y_t]$ from ignorance intervals on censored covariates. The notion of coverage [McCandless et al., 2007] is vital for downstream decision-making. The larger dataset Brain afforded us the ability to approximate a pseudo-population $\tilde{X}$ by reweighing X such that $\tilde{X} \perp\!\!\!\perp T$, in a manner akin to inverse propensity weighting [Reiffel, 2020]. This allowed us to scrutinize the less overlapping parts of the population. We took an unweighted expectation for Vitamin and Energy.

Our test relies on approximating an unconfounded model by collecting a large set of covariates, and then learning another model on a heavily censored version. Our reasoning is that the censored model would suffer from a greater degree of hidden confounding. The censored model could then be assessed along all potential-outcome predictions, by pretending that the full model represented the real dose-response curves. A pertinent metric would be how much a sensitivity model with $\Gamma \geq 1$ swallows the "real" dose responses, as a trade-off against the sheer area of the ignorance

Synthetic Outcomes at Individual Covariates
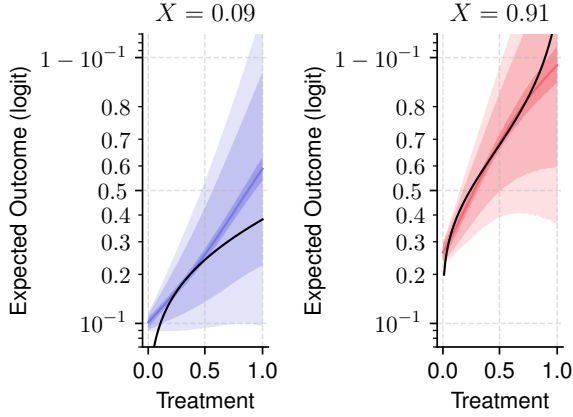
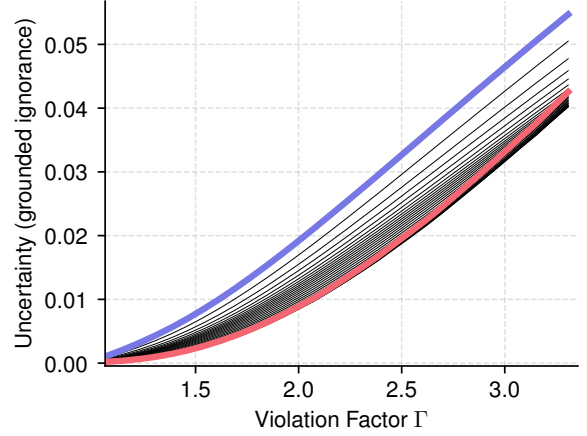Bounds Growth at Individual Covariates

Figure 4. Greater extent of hidden confounding at the lower $X = 0.09$ translates to wider ignorance intervals than at the higher $X = 0.91$. *Left*: graded ignorance intervals at $\Gamma := 1$, then $\Gamma := 2.125$, $\Gamma := 3.25$, with $r = 128$ and 95% ensemble uncertainty. The **black line** is the actual $\mathbb{E}[Y_t|X]$, from integrating the ground truth over the hidden $p(z_2)$. Bernoulli expectations are plotted in a logit transformation because the significance of a perturbation is proportional to $y(1-y)$. *Right*: the selective growth of ignorance intervals across X, computed using the divergence measure described in §4.1.

intervals. These competing quantities can be viewed as a form of recall and (the opposite of) precision, respectively.

Denote $(\underline{y_c}, \overline{y_c})$ the partially identified bounds of the censored model and $(\underline{y}, \overline{y})$ the full model's 95% confidence interval for $\tilde{\mathbb{E}}[Y_t]$, from percentile bootstrapping. The partial identification was bootstrapped as well. For some $\Gamma := s$ and a $t \in [0, 1]$ grid, of length 17 in our case, **ignorance** is $\sum_t [\overline{y_c}(s, t) - \underline{y_c}(s, t)]/\sum_t [\overline{y}(t) - \underline{y}(t)]$, and **recall** is the normalized intersection between the bounds:

$$\frac{\sum_t \max\{0, \min\{\overline{y_c}(s, t), \overline{y}(t)\} - \max\{\underline{y_c}(s, t), \underline{y}(t)\}\}}{\sum_t [\overline{y}(t) - \underline{y}(t)]}.$$

The comparisons. We compared our δMSM with $r = 32$ throughout (**solid** in the figures on this page) to other sensitivity models: namely, (**dotted**) an analogue developed independently and in parallel to the present work, with just one free parameter [Jesson et al., 2022]; (**dashed**) the product of shoehorning a continuous model into the binary MSM by triggering a binary treatment at $T > 0.5$ and discretizing the propensity at the threshold; and (**dot/dash**) a baseline sensitivity model that emerges from Γ-scaling the Algorithm 1 weights without any propensity.

6 DISCUSSION

The utility of our framework is evident in the above show-cased results. We demonstrated that, in the presence of a solid ground truth (§4), the procedure discriminates between more and less confounded outcomes by selectively growing the ignorance bounds (Figure 4).

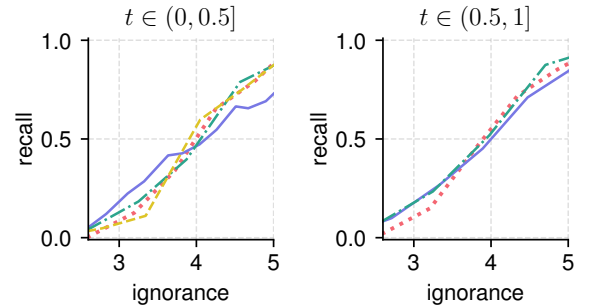Vitamin Model Sensitivity Trade-offs
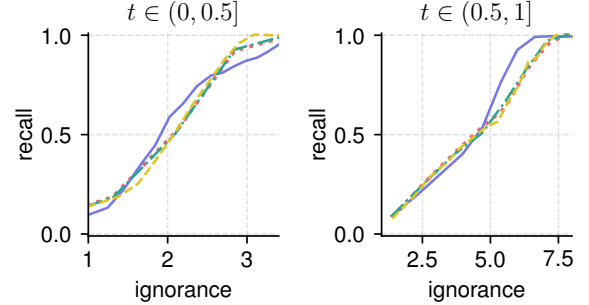
Energy Model Sensitivity Trade-offs

Figure 5. Fraction of the full model's response captured by the censored model (vert.), judged in terms of bounds overlap, versus the relative area of the partial identification (horiz.). The response grid was partitioned into equal halves. Ignorance window was selected to cover the range of recall.
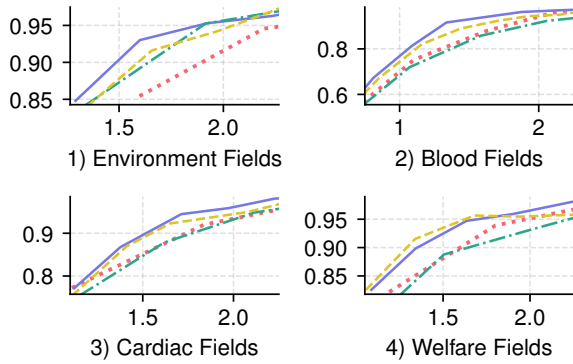
Figure 6. Ignorance (up to 2.25) versus recall, as in Figure 5, but for Brain. Four disjoint sections of the covariates were censored for the different panels. Curves begin at $\Gamma := 1$.

With the real datasets (§5), we analyzed how the censored model recovered the full model's continuum of potential outcomes. The novel δMSM consistently outperformed the other contenders for the Brain dataset, in Figure 6. We noticed via Figure 5 that the δMSM's behavior is markedly different from the alternatives, with clear improvement in the latter half of the response for Vitamin and for small ignorances in Energy. These findings suggest higher coverage efficiency for the δMSM in some realistic scenarios.

Ethical implications. Sensitivity models for hidden confounders can help to guard against erroneous conclusions from observational studies. We generalized this line of analysis to the regime of continuous treatments, thereby increasing its practical applicability. However, there is no replacement for an actual interventional study, and researchers must be careful to maintain a healthy degree of skepticism towards observational results even after properly calibrating the partially identified effects.

Future work. Our scope solely considered the univariate distributions $p(y_t|x)$ for each potential outcome. As such, it was only necessary to grapple with the complete propensity *of first order*, $p(\tau \mid y_t, x)$. In theory there could exist more exhaustive "complete" propensities such as $p(\tau \mid y_{t_1}, y_{t_2}, x)$ that we term *of second order*. These high-order propensities only become relevant alongside joint distributions, like $p(y_{t_1}, y_{t_2}|x)$. It may be fruitful to study these.

7 CONCLUSION

We successfully bridged a hidden-confounding sensitivity model to continuous treatments. Our extension is parsimonious in that it imposes minimal additional requirements on the data and the predictor model. Not only did we derive the analytical forms needed to work with Beta-parametrized

propensities, but we also provided the ingredients for a more general sampling-based approach. We carried these findings into a complete pipeline for quantifying ignorance intervals on dose-response curves. Namely, we bounded the approximation error in Proposition 2 and detailed an efficient optimization algorithm. Our observational experimentation validated the method's core assumptions.

REFERENCES

Paweł A Atroszko. Is a high workload an unaccounted confounding factor in the relation between heavy coffee consumption and cardiovascular disease risk? *The American Journal of Clinical Nutrition*, 110(5):1257–1258, 2019.

Edward J Calabrese and Linda A Baldwin. U-shaped dose-responses in biology, toxicology, and public health. *Annual Review of Public Health*, 22(1):15–33, 2001. doi: 10.1146/annurev.publhealth.22.1.15. PMID: 11274508.

You-Lin Chen, Lenon Minorics, and Dominik Janzing. Correcting confounding via random selection of background variables. *arXiv preprint arXiv:2202.02150*, 2022.

Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2021.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Amiel A Dror, Nicole Morozov, Amani Daoud, Yoav Namir, Orly Yakir, Yair Shachar, Mark Lifshitz, Ella Segal, Lior Fisher, Matti Mizrachi, et al. Pre-infection 25-hydroxyvitamin d3 levels and association with severity of covid-19 illness. *Plos one*, 17(2):e0263069, 2022.

ENTSO-E. Transparency platform. https://transparency.entsoe.eu.

AmirEmad Ghassami, Numair Sani, Yizhen Xu, and Ilya Shpitser. Multiply robust causal mediation analysis with continuous treatments. *arXiv preprint arXiv:2105.09254*, 2021.

Justyna Godos, Maria Tieri, Francesca Ghelfi, Lucilla Titta, Stefano Marventano, Alessandra Lafranconi, Angelo Gambera, Elena Alonzo, Salvatore Sciacca, Silvio Buscemi, et al. Dairy foods and health: an umbrella review of observational studies. *International Journal of Food Sciences and Nutrition*, 71(2):138–151, 2020.

Wenshuo Guo, Mingzhang Yin, Yixin Wang, and Michael Jordan. Partial identification with noisy covariates: A robust optimization approach. In *First Conference on Causal Learning and Reasoning*, 2022.

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.

Wei Huang, Oliver Linton, and Zheng Zhang. A unified framework for specification tests of continuous treatment effect models. *Journal of Business & Economic Statistics*, 0(0):1–14, 2021. doi: 10.1080/07350015.2021.1981915.

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.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *ICML*, 2021.

Andrew Jesson, Alyson Douglas, Peter Manshausen, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*, 2022.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Nathan Kallus and Michele Santacatterina. Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments. *arXiv preprint arXiv:1910.11972*, 2019.

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.

Niki Kilbertus, Matt J Kusner, and Ricardo Silva. A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems*, 33: 20108–20119, 2020.

Justin Lim, Christina X Ji, Michael Oberst, Saul Blecker, Leora Horwitz, and David Sontag. Finding regions of heterogeneity in decision-making via expected conditional covariance. *Advances in Neural Information Processing Systems*, 34:15328–15343, 2021.

Albert Y. Lo. A large sample study of the bayesian bootstrap. *The Annals of Statistics*, 15(1):360–375, 1987.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Maggie Makar, Fredrik Johansson, John Guttag, and David Sontag. Estimation of bounds on potential outcomes for decision making. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6661–6671. PMLR, 13–18 Jul 2020.

Afsaneh Mastouri, Yuchen Zhu, Limor Gultchin, Anna Korba, Ricardo Silva, Matt Kusner, Arthur Gretton, and Krikamol Muandet. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*, pages 7512–7523. PMLR, 2021.

Wesley N. Mathews Jr., Mark A. Esrick, ZuYao Teoh, and James K. Freericks. A physicist's guide to the solution of kummer's equation and confluent hypergeometric functions. *arXiv preprint arXiv:2111.04852*, 2021.

Lawrence C. McCandless, Paul Gustafson, and Adrian Levy. Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statist Med*, 26:2331–2347, 2007.

Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N. Sotiropoulos, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W. Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M. Matthews, and Stephen M. Smith. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nat Neurosci*, 19:1523–1536, 2016.

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.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

James A Reiffel. Propensity score matching: The 'devil is in the details' where more may be hidden than you know. *The American Journal of Medicine*, 133(2):178–181, 2020.

P. R. Rosenbaum. *Observational Studies*. Springer, 2002.

P. R. Rosenbaum and D. B. Rubin. 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.

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

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

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

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Surya T. Tokdar and Robert E. Kass. Importance sampling: A review. *WIREs Computational Statistics*, 2(1):54–60, 2010.

Stefan Tübbicke. Entropy balancing for continuous treatments. *J Econ Methods*, 11(1):71–89, 2022.

Brian G Vegetabile, Beth Ann Griffin, Donna L Coffman, Matthew Cefalu, Michael W Robbins, and Daniel F McCaffrey. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology*, 21(1):69–110, 2021.

Victor Veitch and Anisha Zaveri. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10999–11009. Curran Associates, Inc., 2020.

Pengzhou Abel Wu and Kenji Fukumizu. β-intact-VAE: Identifying and estimating causal effects under limited overlap. In *International Conference on Learning Representations*, 2022.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2020.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M. Blei. Conformal sensitivity analysis for individual treatment effects. *arXiv preprint arXiv:2112.03493v2*, 2021.

Eivind Ystrom, Eirik Degerud, Martin Tesli, Anne Høye, Ted Reichborn-Kjennerud, and Øyvind Næss. Alcohol consumption and lower risk of cardiovascular and all-cause mortality: the impact of accounting for familial factors in twins. *Psychological Medicine*, pages 1–9, 2022.

G. Undy Yule. NOTES ON THE THEORY OF ASSOCIATION OF ATTRIBUTES IN STATISTICS. *Biometrika*, 2(2):121–134, 02 1903. ISSN 0006-3444. doi: 10.1093/biomet/2.2.121. URL https://doi.org/10.1093/biomet/2.2.121.

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)*, 81(4):735–761, 2019.

A ADDITIONAL DERIVATIONS

Consider Equation 3.A:

$$\int_0^1 w_t(\tau)\tilde{p}(y_t|\tau,x)p(\tau|x)\,\mathrm{d}\tau = \underbrace{p(y_t|t,x)\int_0^1 w_t(\tau)p(\tau|x)\,\mathrm{d}\tau}_{(A.0)}$$

$$+ \underbrace{g_1(y_t|t,x)\int_0^1 w_t(\tau)(\tau-t)p(\tau|x)\,\mathrm{d}\tau}_{(A.1)} + \underbrace{g_2(y_t|t,x)\int_0^1 w_t(\tau)\frac{(\tau-t)^2}{2}p(\tau|x)\,\mathrm{d}\tau}_{(A.2)},$$

$$\text{where}\quad g_k(y_t|t,x) \triangleq \partial_\tau^k p(y_t|\tau,x)|_{\tau=t}. \quad (11)$$

Lightening the notation with a shorthand for the weighted expectations, $\langle\cdot\rangle_\tau \triangleq \int_0^1 w_t(\tau)(\cdot)p(\tau|x)\,\mathrm{d}\tau$, it becomes apparent that we must grapple with the pseudo-moments $\langle 1\rangle_\tau$, $\langle \tau-t\rangle_\tau$, and $\langle(\tau-t)^2\rangle_\tau$. Note that t should not be mistaken for a "mean" value.

Furthermore, we have yet to fully characterize $g_k(y_t|t,x)$. Observe that

$$p(y_t|\tau,x) = \frac{p(\tau|y_t,x)p(y_t|x)}{p(\tau|x)} \quad\Longleftrightarrow\quad \partial_\tau p(y_t|\tau,x) = p(y_t|x)\cdot\frac{\partial}{\partial\tau}\frac{p(\tau|y_t,x)}{p(\tau|x)}.$$

The $p(y_t|x)$ will be moved to the other side of the equation as needed; by Equation 6,

$$\frac{\partial}{\partial\tau}\frac{p(\tau|y_t,x)}{p(\tau|x)} = \frac{\partial}{\partial\tau}\Lambda(\tau|y_t,x).$$

Expanding,

$$= \frac{\partial}{\partial\tau}\exp\left\{\int_0^\tau \gamma(\tau|y_t,x)\,\mathrm{d}\tau\right\} = \gamma(\tau|y_t,x)\exp\left\{\int_0^\tau \gamma(\tau|y_t,x)\,\mathrm{d}\tau\right\}$$

$$= (\gamma\Lambda)(\tau|y_t,x).$$

Appropriate bounds will be calculated for $g_2(y_t|t,x)$ next, utilizing the finding above as their main ingredient. Let

$$\tilde{g}_k(y_t|t,x) \triangleq p(y_t|x)^{-1}g_k(y_t|t,x) = \left(\frac{\partial}{\partial\tau}\right)^k \frac{p(\tau|y_t,x)}{p(\tau|x)}\Bigg|_{\tau=t}.$$

The second derivative may be calculated in terms of the ignorance quantities γ,Λ:

$$\tilde{g}_2(y_t|t,x) = \partial_\tau\gamma(\tau|y_t,x)\Lambda(\tau|y_t,x)$$
$$= \gamma(\tau|y_t,x)^2\Lambda(\tau|y_t,x) + \dot{\gamma}(\tau|y_t,x)\Lambda(\tau|y_t,x)$$
$$= (\gamma^2 + \dot{\gamma})\Lambda(\tau|y_t,x).$$

And finally we address $\tilde{p}(y_t|x)$. Carrying over the components of Equation 11 into Equation 3,

$$\tilde{p}(y_t|x) = \frac{p(y_t|t,x)\langle 1\rangle_\tau}{\langle\Lambda(\tau|y_t,x)\rangle_\tau - \tilde{g}_1(y_t|t,x)\langle\tau-t\rangle_\tau - \tilde{g}_2(y_t|t,x)\langle(\tau-t)^2\rangle_\tau}$$

$$= \frac{p(y_t|t,x)}{\mathbb{E}_\tau[\Lambda(\tau|y_t,x)] - (\gamma\Lambda)(t|y_t,x)\,\mathbb{E}_\tau[\tau-t] - \frac{1}{2}((\dot{\gamma}+\gamma^2)\Lambda)(t|y_t,x)\,\mathbb{E}_\tau[(\tau-t)^2]}, \quad (12)$$

where these expectations $\mathbb{E}_\tau[\cdot]$ are with respect to the implicit distribution $q(\tau|t,x) \propto w_t(\tau)p(\tau|x)$. The first term in the denominator, $\mathbb{E}_\tau[\Lambda(\tau|y_t,x)]$, may be approximately bounded by the same Algorithm 1.

B DETAILS ON PROPOSITION 2

The numerics were accomplished via Wolfram Mathematica 12. Our steps for calculating key quantities are listed below.

Firstly, since the second derivative employed in the Taylor expansion is Q-Lipschitz, we have $\left|\partial_\tau^3 p(y_t|\tau, x)\right| \leq Q$. Let us denote the Taylor remainder in this scope as $\rho(y_t|\tau, x)$. By Taylor's theorem,

$$|\rho(y_t|\tau, x)| \leq \frac{|\tau - t|^3}{6} Q.$$

We carry over the remainder into the formulation of $\tilde{p}(y_t|x)$ as in Equation 4, reproduced more explicitly below.

$$\tilde{p}(y_t|x) = \frac{\int_0^1 w_t(\tau)\left[p(y_t|\tau, x) + \rho(y_t|\tau, x)\right]p(\tau|x)\, \mathrm{d}\tau}{\int_0^1 w_t(\tau)p(\tau|y_t, x)\, \mathrm{d}\tau}$$

By acknowledging that an exact approximation requires $\rho(y_t|\tau, x) = 0$, we express the absolute error in our imperfect approximator as

$$|\tilde{p}(y_t|x) - p(y_t|x)| = \left|\frac{\int_0^1 w_t(\tau)\rho(y_t|\tau, x)p(\tau|x)\, \mathrm{d}\tau}{\int_0^1 w_t(\tau)p(\tau|y_t, x)\, \mathrm{d}\tau}\right| \leq \frac{\left|\int_0^1 c^\star w_t^\star(\tau)\rho(y_t|\tau, x)\, \mathrm{d}\tau\right|}{\int_0^1 w_t(\tau)p(\tau|y_t, x)\, \mathrm{d}\tau}.$$

We analytically solved the definite integral $\int_0^1 c^\star w_t^\star(\tau)\frac{1}{6}Q|\tau - t|^3\, \mathrm{d}\tau \geq \left|\int_0^1 c^\star w_t^\star(\tau)\rho(y_t|\tau, x)\, \mathrm{d}\tau\right|$, bounding the numerator. We omitted the full expansion for brevity. Then, for specific values of $r^\star$ as displayed in the table of the theorem statement, we maximized the quantity numerically over the support $0 \leq t \leq 1$ in order to obtain $k(r^\star)$.

Next, on reducing the denominator, we employed the opposite side of our bound:

$$\int_0^1 w_t(\tau)p(\tau|y_t, x)\, \mathrm{d}\tau \;\geq\; \frac{1}{c^\star}\int_0^1 w_t^\star(\tau)\, \mathrm{d}\tau \;=\; \frac{B(tr^\star + 1, (1 - t)r^\star + 1)}{c^\star\, t^{tr^\star}(1 - t)^{(1-t)r^\star}}.$$

The resultant fraction is symmetric about $t = \frac{1}{2}$. We posit—for it was too unweildy to prove—that the expression is concave in $t \in [0, 1]$ for all $r^\star > 0$. We validated this claim by testing the non-positivity of the second derivative at relevant values of $r^\star$ across the domain in t. It follows that the fraction achieves its minimum at the boundaries $t = 0, 1$. In either of those cases it may be evaluated by means of the identity $B(r^\star + 1, 1) = (r^\star + 1)^{-1}$. Hence, we assert that the above fraction is no lesser than $[c^\star(r^\star + 1)]^{-1}$.

Regardless of the validity of the above conjecture, the quantities displayed in the table of Proposition 2 are correct because the relation surely holds at those points in $r^\star$.

C CORRECTNESS OF ALGORITHM 1

The algorithm functions by incrementally reallocating mass (relative, in the weights) to the righthand side, from a cursor beginning on the lefthand side of the "tape".

Proof. Firstly we characterize the indicator quantity Δ_j. Differentiate the quantity to be maximized with respect to w_j;

$$
\frac{\partial}{\partial w_j} \frac{\sum_i w_i f_i}{\sum_i w_i} = \frac{f_j}{\sum_i w_i} - \frac{\sum_i w_i f_i}{\left(\sum_i w_i\right)^2}
$$

$$
= \frac{f_j \sum_i w_i - \sum_i w_i f_i}{\left(\sum_i w_i\right)^2}
$$

$$
\propto \underbrace{\sum_i w_i (f_j - f_i)}_{\triangleq \Delta_j} \quad \text{up to some positive factor.}
$$

Hence, Δ_j captures the sign of the derivative.

We shall proceed with induction. Begin with the first iteration, $j = 1$. No weights have been altered since initialization yet. Therefore we have

$$
\Delta_1 = \sum_i \overline{w}_i (f_1 - f_i).
$$

Since $\forall i$, $f_1 \leq f_i$ due to the prior sorting, Δ_1 is either negative or zero. If zero, trivially terminate the procedure as all function values are identical.

Now assume that by the time the algorithm reaches some $j > 1$, all $w_k = \underline{w}_k$ for $1 \leq k < j$. In other words,

$$
\Delta_j = \sum_{i<j} \underline{w}_i \underbrace{(f_j - f_i)}_{(+)} + \sum_{i>j} \overline{w}_i \underbrace{(f_j - f_i)}_{(-)}.
$$

Per the algorithm, we would flip the weight $w_j \leftarrow \underline{w}_j$ only if $\Delta_j < 0$. In that case,

$$
\sum_{i<j} \underline{w}_i (f_j - f_i) < \sum_{i>j} \overline{w}_i (f_i - f_j), \quad \text{where both sides are non-negative.}
$$

Notice that the above is not affected by the current value of w_j. This update can only increase the current estimate because the derivative remains negative and the weight at j is non-increasing. We *must* verify that the derivatives for the previous weights, indexed at $k < j$, remain negative. Otherwise, the procedure would need to backtrack to possibly flip some weights back up.

More generally, with every decision for weight assignment, we seek to ensure that the condition detailed above is not violated for any weights that have been finalized. That includes the weights before j, and those after j at the point of termination. Returning from this digression, at $k < j$ after updating w_j,

$$
\Delta_k = \sum_{i \leq j} \underline{w}_i (f_k - f_i) + \sum_{i>j} \overline{w}_i (f_k - f_i).
$$

To glean the sign of this, we refer to a quantity that we know.

$$
\sum_{i<j} \underline{w}_i (f_j - f_i) < \sum_{i>j} \overline{w}_i (f_i - f_j)
$$

$$
\iff \sum_{i \leq j} \underline{w}_i (f_k - f_i) < \sum_{i>j} \overline{w}_i (f_i - f_j) + \sum_{i \leq j} \underline{w}_i (f_k - f_j)
$$

$$
\iff \underbrace{\sum_{i \leq j} \underline{w}_i (f_k - f_i) + \sum_{i>j} \overline{w}_i (f_k - f_i)}_{\Delta_k} < \underbrace{\sum_{i>j} \overline{w}_i (f_k - f_j) + \sum_{i \leq j} \underline{w}_i (f_k - f_j)}_{\text{negative.}}
$$

The remaining fact to be demonstrated is that upon termination, when $\Delta_j \geq 0$, no other pseudo-derivatives $\Delta_{j'}$, $j' > j$ are negative. This must be the case simply because $f_{j'} \geq f_j$. $\qquad\square$

D ON THE INTRODUCTORY ILLUSTRATION

The setting in Figure 1 lacks covariates. Its conditional expected outcome $\mathbb{E}[Y \mid Z = z, T = t]$ is defined as $3(t - \frac{1}{4})^2 - 4z^2$. To produce the dashed line, observations are fully confounded by the treatment regime $p_{T|Z}(t \mid z) = \delta(t - z)$. The amalgamation of individual curves leads to a population-level average pointing in the wrong direction. Here, the overlap assumption for inverse propensity weighting is violated here. Still, "true" potential outcomes are revealed by forcing independence between T and Z.

The solid line is the result of taking an expectation over $Z \sim \text{Uniform}[0, 1]$. The blue dots represent observations from a more realistic, partially confounded scenario where $(T \mid Z = z) \sim \text{Beta}(48z + 1, 48(1 - z) + 1)$, i.e. treatments are centered at z with a narrow spread.

E SYNTHETIC SCENARIO DETAILS

Setup. 1,500 training and 500 testing examples were generated. We trained ensembles of eight neural networks with two hidden layers of 16 units, for the Bernoulli predictor and Beta propensity each. On constructing the idealized world,

$$
\begin{aligned}
Z = (Z_1, Z_2) \quad &\sim \quad \text{i.i.d Unif}(0, 1) \\
(T | Z = z) \quad &\sim \quad
\begin{cases}
U(0, 0.2) & \text{w.p} & 0.1 \\
U(0.2, 0.5) & \text{w.p} & 0.4 - 0.2z_2 \\
U(0.5, 0.7) & \text{w.p} & 0.3 + 0.3z_2 \\
U(0.7, 0.8) & \text{w.p} & 0.1 - 0.1z_2 \\
U(0.8, 1) & \text{w.p} & 0.1
\end{cases} \\
(X | Z = z) \quad &\sim \quad \text{Dirac}(z_1) \\
(Y | T = t, Z = z) \quad &\sim \quad \text{Bernoulli}(z_1^{1 - z_2} t^{z_2})
\end{aligned}
\tag{13}
$$

F EXPERIMENTAL DETAILS

Brain dataset. Data from the UK Biobank were accessed under application 11559. From the brain Magnetic Resonance Imaging (MRI) data we extracted the 74 fields corresponding to parcelized cortical volumes on the left and the right hemispheres each [Miller et al., 2016]. Then we summed each left/right pair to arrive at 74 positively valued outcomes. Six groups of semantically related fields composed the long covariate vector:

- 6 basic details: age, weight, sex, standing height, seated height, and month of birth.
- 3 reported activity measurements: weekly minutes spent walking, engaged in moderate activity, and vigorous activity.

- 27 *environmental* variables surveying the pollution and greenery surrounding the person's life.
- 42 *blood* measurements from cell counts to calcium concentration.
- 15 *cardiac* measurements including ECG and PWA modalities.
- 8 *welfare* indices for English citizens assessing the following: deprivation, income, employment, health, education, housing, crime, and living environment.

Listwise deletion was employed to handle any missing value. To censor the covariates, the four largest sectors (*italicized*) encompassing various confounding variables were omitted, one at a time. The treatment variable was the walking field taken from the triad of activity measurements, scaled to the unit interval such that any recording of at least two hours per day was set to $T = 0$ and any lesser amount had T increase up to 1 according to an empirical CDF.

Brain estimators. Both the predictor and the propensity model, censored and uncensored, were trained in ensembles of 32 artificial neural networks with four inner residual layers of 32 activation units each. A dropout of 0.05 was imposed on these layers. Additionally, an L^2-regularization on the inner layers with weight 10^{-3} was applied. All predictors were trained for 10,000 epochs and propensities for 5,000 epochs. The outcome predictor parametrized a Gamma distribution, and the propensity model parametrized a Beta distribution. See Figure 7.

Vitamin dataset. We studied individuals from a recently published record of COVID-19 mortality from Israel, with a focus on pre-infection Vitamin D levels. The data acquired from Dror et al. [2022] included 253 individuals, with 38 deaths attributed to COVID-19. The full set of recorded covariates were gender (two categories,) age (three levels,) body-mass index (two levels,) religion (five categories,) and six binary-labeled comorbidities. Vitamin D levels, the treatment variable, were scaled to the unit interval such that "healthy" levels of at least 40 ng/mL were allocated to $T = 0$ and T increased to 1 linearly as vitamin levels tended to zero. The censored dataset omitted the religion and comorbidity categories.

Vitamin estimators. The singular inner predictor- and propensity-model layers comprised four activation units each, owing to the smaller set of covariates. The dropout was increased to 0.1. The smaller models were trained in ensembles of 64 in this case.

Energy dataset. Hourly time series of electricity generation in Spain were gathered for the period from April 1st, 2021 to November 1st, 2021 [ENTSO-E]. The outcome was the wholesale price of electricity. The market price of

natural gas spiked in Europe during that time frame, especially within the testing set that was temporally ordered after the training set—*not* randomly assigned. Hence, we observed substantial non-stationarity between the model estimates and the out-of-sample covariates used in the ignorance/recall evaluation. The treatment variable was the carbon price dictated by the EU Emissions Trading System (ETS), linearly scaled between 0 and 1.

Energy estimators. The predictor models, censored and uncensored, each had one inner layer of eight activation units. The propensity models had four inner units instead. The dropout was increased to 0.2 here, and the models were trained in ensembles of 16.
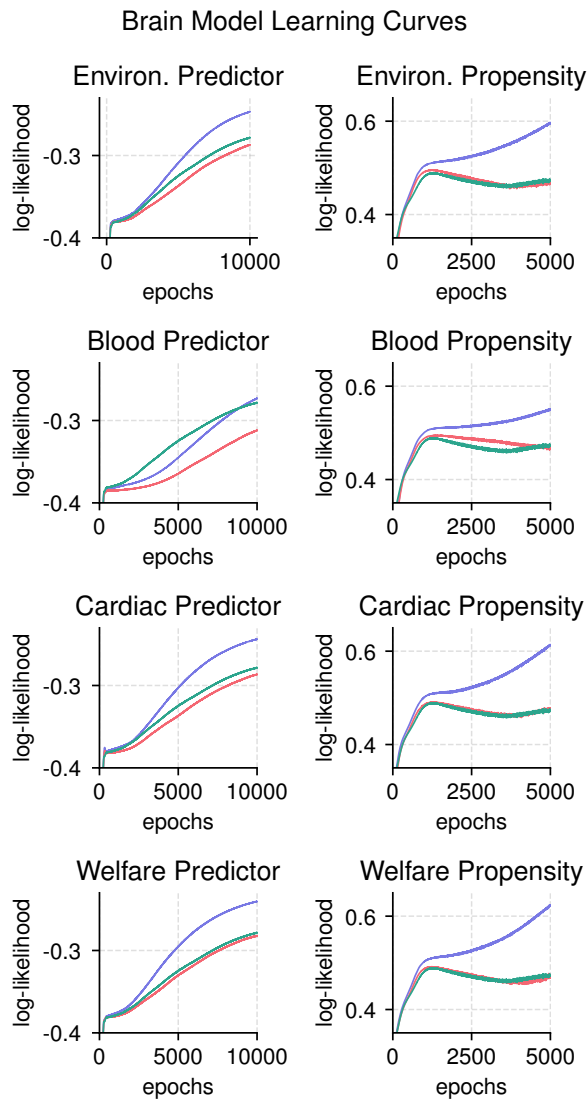

Figure 7. **Blue**: censored-model likelihood in the train set; **red**: censored-model likelihood in the test set; and **green**: full-model likelihood in the test set.

G   SOURCE-CODE AVAILABILITY

Please visit `https://github.com/marmarelis/TreatmentCurves.jl`.