Bounding the Effects of Continuous Treatments for Hidden Confounders

Myrl G. Marmarelis  
USC Information Sciences Institute  
4676 Admiralty Way  
Marina del Rey, CA 90292  
myrlm@isi.edu

Greg Ver Steeg  
USC Information Sciences Institute  
4676 Admiralty Way  
Marina del Rey, CA 90292  
gregv@isi.edu

Neda Jahanshad  
USC Stevens Neuroimaging and Informatics Institute  
4676 Admiralty Way  
Marina del Rey, CA 90292  
neda.jahanshad@ini.usc.edu

Aram Galstyan  
USC Information Sciences Institute  
4676 Admiralty Way  
Marina del Rey, CA 90292  
galstyan@isi.edu

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 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 causal inference is to separate the effect of one treatment variable from the influence of many related but irrelevant “confounding” variables. Physical interventions accomplish this most effectively, but practical barriers often force researchers to rely on observational studies and clever statistics. 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. Usually, the covariates are inadequate to account for all the hidden paths between the treatment and the outcome, and propensity-based approaches may struggle to discern the real effect. The scientific community at large continues to vacillate on the health implications and ideal consumption levels of coffee [Atroszko 2019], alcohol [Ystrom et al. 2022], and cheese [Godos et al. 2020], to name a few substances.

Most sensitivity analyses to hidden confounding require the treatment categories to be binary or at least discrete. This weakens the empirical studies that would have been better specified by dose-response curves [Calabrese and Baldwin 2001] from a continuous treatment variable, for example.
Estimated dose responses are indeed vulnerable in the presence of hidden confounders. 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, Tübingen, 2022], health sciences [Vegetable et al., 2021], and machine learning [Ghassami et al., 2021, Colon, 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.

A number of quite recent attempts have been made to handle unobserved confounding with continuous treatments. Some optimize the treatment-effect bounds through generative models [Padh et al., 2022, Hu et al., 2021], rather than a sensitivity model. Another employs instrumental variables [Kilbertus et al., 2020]. One 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 additional structure to the data-generating (observed outcome, treatment, covariates) process. Proximal causal learning [Tchetgen et al., 2020, Mastouri et al., 2021] requires additional proxy structures. Chen et al. [2022] rely on multiple large dataset partitions.

1.2 Contributions

Our first contribution is to propose a unique sensitivity model (§2.1) that extends the MSM to a treatment continuum. Next, we derive general (§3) and specialized (§3.2) formulas. We devise an efficient algorithm (§C) to compute ignorance bounds over dose-response curves, following up with experiments on real (§4) datasets.

2 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 heterogeneously 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
The first two assumptions involved in Rubin’s framework are that observations of outcome, assigned treatment, and covariates \( \{ Y(t) , T(t) , X(t) \} \) are i.i.d. draws from the same population and that all treatments have a chance to occur for each covariate vector: \( p_{T|X}(t | x) > 0 \) (overlap/positivity) for all \( t , x \in [0,1] \times X \), 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:

\[
\{ (Y_t)_{t \in T} \neq \tau \} \ | \ X.
\]

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

\[
p(y_t | x) = \int_0^1 p(y_t | \tau , x) p(\tau | x) \ d\tau,
\]

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 \( \tau \) 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.

**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, \( p(\tau | x) \) means \( \frac{d}{d\tau} \mathbb{P}[T \leq \tau | X = x] \), and \( p(y_t | \tau , x) \) means \( \frac{d}{d\tau} \mathbb{P}[Y_t \leq u | T = \tau , X = x] \) at \( u = y_t \).

**Interpretation.** How would one interpret \( p(y_t | \tau , x) \)? The potential-outcomes vector \( \{ Y_t \}_{t \in T} \) 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 \( \int p(y_t | z) p(z | \tau , x) \ dz \) because \( z \) subsumes both \( x \) and \( \tau \). 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 \( \tau \).

**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 \( \tau \) 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 \( \tau \) on the potential outcomes.
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** (Fully Observed Confounding at No Treatment). The utter lack of treatment is not informed by potential outcomes: $p(\tau = 0 \mid y_t, x) = p(\tau = 0 \mid x)$ 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 $\delta$MSM 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 §4.

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.

### 3.1 Dealing with an unreliable approximation

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

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

Denote with $\tilde{p}(y|\tau, x)$ an approximation of first or second order as laid out above. We will encounter that even $\partial_\tau p(y|\tau, x)|_{\tau = t}$ is intractable. Thankfully, it can be bounded using the $\delta$MSM machinery. Let us quantify the reliability of this approximation by a set of weights $w_i(t) \leq 1$, where typically (but need not necessarily) $w_i(t) = 1$. Decompose the integral of Equation 1—

$$p(y|t) = \int_0^1 w_i(\tau)p(y|\tau, x)p(\tau|x) \, d\tau + \int_0^1 (1 - w_i(\tau))p(y|\tau, x)p(\tau|x) \, d\tau \quad \text{(A)}$$

This separation into recoverable (A) and entirely unknown (B), demarcated by the weights, ensures that the inaccurate regimes of the approximation vanish (as $w_i(\tau) \to 0$ away from $t$) and are replaced with the ignorant quantity. We simplify part B of Equation 3 first, into $p(y|t|x)[1 - \int_0^1 w_i(\tau)p(\tau|y_i, x) \, d\tau]$. We witness already that $p(y|t|x)$ shall take the form of

$$p(y|t|x) \approx \frac{\int_0^1 w_i(\tau)p(y_i|\tau, x)p(\tau|x) \, d\tau}{\int_0^1 w_i(\tau)p(\tau|y_i, x) \, d\tau}. \quad (4)$$

How the approximation error of Equation 2 carries into Equation 4 depends on the peakedness of the weight function. To proceed further demands reflecting on Assumptions 1 & 2 as we do in §A.

**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.

### 3.2 Tractable weight combinations

In addition to developing the general framework above, we derive analytical forms for a specific parametrization 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

$$(T | X = x) \sim \text{Beta}(\alpha(x), \beta(x)), \quad \text{for arbitrary } \alpha(x), \beta(x), \quad (5)$$

$$w_i(\tau) = \frac{\tau^{a_i-1}(1 - \tau)^{b_i-1}}{c_i} = \frac{\tau^{rt}(1 - \tau)^{r(1-t)}}{c_i}, \quad a_i + b_i = r + 2, \quad r > 0. \quad (6)$$

We designed the reliability weights to mirror the propensity’s form by rescaling a Beta density. We assert that $w_i(\tau)$ peaks at $\tau = t$, and that $w_i(\tau) = 1$. We find that $c_i \triangleq t^{r(1-t)}r^{(1-t)}$, even though the solution is irrelevant for our purposes. The mode is fixed: $(a_i - 1)/(a_i + b_i - 2) = t$. See §B for details on the solution.
4 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 cumulative density function.

Semi-synthetic Model Sensitivity Trade-offs

Figure 3. Ignorance (horizontal) versus recall (vertical) in four replications of a counterfactual model estimated on real Biobank covariates with a synthetic binary outcome, and known dose response.

We obtained real observational data from the UK Biobank and performed one semi-synthetic and one fully empirical experiment. In the latter, we relied on discarding covariates to induce greater hidden confounding. See §D for details.

The objective. We chose to investigate the coverage of $E[Y_t]$ from ignorance intervals on counterfactual models trained with inadequate covariates. In the semi-synthetic case (Figure 3), the covariates were real but the outcome was simulated, and the treatment effect was known to be linear after a logit transformation. Coverage in the empirical case (Figure 4) was more difficult in the absence of a ground truth. There, we trained an uncensored model on an expanded set of covariates to act as an approximate target for the smaller model’s ignorance intervals. In both cases, the recall was expressed as the portion of the dose-response curve that satisfied the relevant objective; ignorance was, (a) the average width of the intervals in logits for the semi-synthetic, and (b) normalized to the 95%-confidence intervals of the uncensored model for the empirical study.

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 $\Gamma$-scaling the Algorithm 1 weights without any propensity.

Empirical Model Sensitivity Trade-offs

Figure 4. Ignorance (up to 2.25) versus recall for the model estimated on the real censored dataset. Four disjoint sections of the covariates were censored for the different panels. Curves begin at $\Gamma := 1$. 
5 Discussion

The utility of our framework is evident in the above showcased results. We demonstrated that, in the presence of a semi-synthetic ground truth, the $\delta$MSM covers the full dose-response curve most efficiently. In the empirical study, we showed that discrepancies in the potential-outcomes continuum between a censored model and a full model are most efficiently bridged under the $\delta$MSM assumption.

Moving forward, we plan on generalizing the permissible anchoring points that were restricted by Assumption 2 at $T = 0$. Even further, the same $\delta$MSM can be applied to treatment variables in $(-\infty, +\infty)$ or $[0, +\infty)$, inviting new propensity/weight parametrizations. We shall also investigate the superficial similarity of the $\delta$MSM to the Riesz representers for the average causal derivative under the general framework of [Chernozhukov et al. 2021].

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. The method also bears utility in the context of fairness in machine learning. It can help decision makers identify subpopulations for whom the sample is too biased to reliably draw conclusions. Nevertheless, researchers must be careful to maintain a healthy degree of skepticism towards observational results even after properly calibrating the partially identified effects.

Acknowledgments and Disclosure of Funding

This material is based upon work supported, in part, by the Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002.

References


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]

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] We introduce a marginal sensitivity model for continuous treatments.
   (b) Did you describe the limitations of your work? [Yes] Most notably, Assumption 2.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We mentioned in §5 how this line of work can help to guard against erroneous conclusions from observational studies. However, there is no replacement for an actual interventional study, and we must be careful to maintain a healthy degree of skepticism on observational results even after performing the sensitivity analysis.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumptions 1–3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See §A–§C.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See §D and the attached source code. A link to the Julia package on Github will be made available upon publication.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See §D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We studied bootstrapped 95% confidence intervals for all our empirical results, namely in §4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] These results were not bottlenecked by computational resources. A couple of Nvidia 2080 Ti GPUs were used for training, and a Macbook Pro for the rest of the computations.

4. If you are using existing assets (e.g., code, data, models) or curating/releaseing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] All owners of the datasets used herein were credited in §D. Besides common Julia packages and frameworks, all novel functionality was implemented by the first author.
   (b) Did you mention the license of the assets? [Yes] See §D.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The source code will be included as a Github link in the de-anonymized version.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A] Our datasets were acquired from institutions with clear guidelines.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] This study only reported broad aggregations upon the data.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
A Additional derivations

We will expand the denominator of Equation 4 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 §3.2. Without loss of generality, consider

\[
\frac{\partial_\tau \log p(\tau | y_t, x)}{\partial_\tau \log p(\tau | y_t, x)} = \frac{\partial_\tau \log p(\tau | x) + \gamma(\tau | y_t, x)}{\partial_\tau \log p(\tau | x) + \gamma(\tau | y_t, x)} \leq \log \Gamma. \tag{7}
\]

We may attempt to integrate both sides;

\[
\int_0^{t^\prime} \partial_\tau \log p(\tau | y_t, x) \, d\tau = \int_0^{t^\prime} \partial_\tau \log p(\tau | x) \, d\tau + \int_0^{t^\prime} \gamma(\tau | y_t, x) \, d\tau
\]

\[
\iff \log p(\tau = t^\prime | y_t, x) - \log p(\tau = 0 | y_t, x) = \log p(\tau = t^\prime | x) - \log p(\tau = 0 | x) + \lambda(t^\prime | y_t, x),
\]

\[
\log p(\tau | y_t, x) = \log p(\tau | x) + \lambda(\tau | y_t, x) \quad \text{(by Assumption 2)}.
\]

Clearly \( |\lambda(\tau | y_t, x)\| \leq \tau \log \Gamma \) because it integrates \( \gamma \), bounded by \( \pm \log \Gamma\), over a support with length \( \tau \). Subsequently \( \Lambda(\tau | y_t, x) \) is bounded by \( \Gamma^\pm \). 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 \( w_i(\gamma) \).

Consider Equation [5];

\[
\int_0^1 w_t(\gamma) \bar{p}(y_t | \tau)p(\tau | x) \, d\tau = p(y_t | t, x) \int_0^1 w_t(\gamma)p(\tau | x) \, d\tau
\]

\[
+ g_1(y_t | t, x) \int_0^1 w_t(\gamma)(\tau - t)p(\tau | x) \, d\tau + g_2(y_t | t, x) \int_0^1 w_t(\gamma)(\tau - t)^2 p(\tau | x) \, d\tau,
\]

where \( g_k(y_t | t, x) \equiv \partial^k \bar{p}(y_t | \tau, x) | \tau = t \tag{9} \)

Lightening the notation with a shorthand for the weighted expectations, \( \langle \cdot \rangle_{\tau} \equiv \int_0^1 w_t(\gamma)(\cdot)p(\tau | x) \, d\tau \), 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

\[
p(y_t | x) = \frac{p(\tau | y_t, x)p(y_t | x)}{p(\tau | x)} \iff \partial_\tau p(y_t | x) = p(y_t | x) \cdot \frac{\partial_\tau p(\tau | y_t, x)}{p(\tau | x)}.
\]

The \( p(y_t | x) \) will be moved to the other side of the equation as needed; by Equation [8]

\[
\frac{\partial_\tau p(\tau | y_t, x)}{p(\tau | x)} = \frac{\partial_\gamma}{\partial_\tau} \Lambda(\tau | y_t, x).
\]

Expanding,

\[
\begin{aligned}
\frac{\partial}{\partial_\tau} \exp\left\{ \int_0^\tau \gamma(\tau | y_t, x) \, d\tau \right\} &= \gamma(\tau | y_t, x) \exp\left\{ \int_0^\tau \gamma(\tau | y_t, x) \, d\tau \right\} \\
&= (\gamma \Lambda)(\tau | y_t, x).
\end{aligned}
\]

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

\[
\tilde{g}_k(y_t | t, x) \triangleq p(y_t | x)^{-1} g_k(y_t | t, x) = \left( \frac{\partial_\tau}{\partial_\tau} \right)^k \frac{p(\tau | y_t, x)}{p(\tau | x)} \bigg|_{\tau = t}.
\]

11
The second derivative may be calculated in terms of the ignorance quantities \( \gamma, \Lambda \):
\[
\tilde{g}_2(y_t|t, x) = \partial_r^2 \gamma (\tau|y_t, x) \Lambda (\tau|y_t, x) \\
= (\gamma^2 + \dot{\gamma}) \Lambda (\tau|y_t, x) \\
= (\gamma^2 + \dot{\gamma}) \Lambda (\tau|y_t, x).
\]

And finally we address \( \tilde{p}(y_t|x) \). Carrying over the components of Equation \[9\] into Equation \[10\]
\[
\tilde{p}(y_t|x) = \frac{p(y_t|x)}{(\Lambda (\tau|y_t, x))_\tau - \tilde{g}_1(y_t|t, x)((\tau - t)\Lambda (\tau|y_t, x))^2}
\]
where these expectations \( \mathbb{E}_\tau[] \) 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 \( \gamma(\tau|y_t, x) \). To make use of this formula, one first procures the set of admissible \( \bar{d}(\tau|y_t, x) \in [\tilde{d}(\tau|y_t, x), \bar{d}(\tau|y_t, x)] \) that violate ignorability up to a factor \( \Gamma \) according to the \( \delta \)-MSM. 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.

### B Analytical solutions for the Beta parametrization

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. The expectations found in Equation \[10\] are now available in closed form, and can be bounded in terms of just two extra free parameters, \( \Gamma \) and \( r \). Guidance on setting the violation factor \( \Gamma \) is discussed elsewhere, e.g. [4].

Shapes are symmetrical about \( t \) and \( \Gamma \) decreases with \( r \).

![Beta weight schemes](image)

**Figure 5.** 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 findings.** We pose a third and final assumption, which enables us to state Proposition \[4\]. 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 \( \gamma^2 \) dominates, and consequently \( |\dot{\gamma} + \gamma^2| \leq |\gamma^2| + \varepsilon \) for small \( \varepsilon \geq 0 \).

**Proposition 1** (Beta Parametrizations). The formulations in Equations \[5\] & \[6\] admit analytical solutions to the ignorance denominator in Equation \[10\]. With \( \alpha, \beta \) implicitly referring to \( \alpha(x), \beta(x) \),
\[
\mathbb{E}_\tau [\Lambda(\tau|y_t, x)] \in {}_1F_1 (\alpha + a_t - 1; \alpha + \beta + r; \pm \log \Gamma),
\]
where again the operator \( \mathbb{E}_\tau \) is employed as in Equation \[10\] and \( {}_1F_1 \) denotes Kummer’s confluent hypergeometric function [Mathews Jr. et al. 2021]. In addition, \( \mathbb{E}_\tau [\tau - t] \) and \( \mathbb{E}_\tau [(\tau - t)^2] \) can be readily computed by means of the first two moments of the Beta distribution.

### C Computing the ignorance intervals

After deriving \( \tilde{p}(y_t|x) \) in Equation \[10\] and a specific solution with Proposition \[4\] we must find a way to bound the partially identified expectations with respect to this distribution. Concretely, we seek to characterize the Individual Treatment Effect (ITE) \( \mathbb{E}[f(Y_i)|X = x] \) or Average Treatment Effect (ATE) \( \mathbb{E}[F(Y_t)] \) for any task-specific \( f(y) \). This is accomplished with a Monte Carlo importance sampler of \( n \) outcome realizations \( y_t \), drawn from proposal \( q(y) \):
\[
\tilde{\mathbb{E}} [f(Y_t)|X = x] = \frac{\sum_{i=1}^{n} f(y_t) \tilde{p}(y_t = y_t|x)/q(y_t)}{\sum_{i=1}^{n} \tilde{p}(y_t = y_t|x)/q(y_t)}.
\]
For the ATE, one additionally averages over covariates with sample size $m$:

$$
\E[f(Y)] = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} f(y_i) \tilde{p}(y_i = y_j | x_j) / q(y_i)}{\sum_{i=1}^{n} \sum_{j=1}^{m} \tilde{p}(y_i = y_j | x_j) / q(y_i)}.
$$

(12)

Even though $\tilde{p}(y_i | x)$ is a normalized probability density, it contains partially identified quantities. It is untenable to constrain a search along the candidate values for each $d(t | y_i = y_i, x)$ to even approximately ensure $\int_{x} \tilde{p}(y_i = y_i | x) \, dx = 1$. For this reason the bias of an estimator without the corrective denominator of Equation (11) would be uncontrollable [Tokdar and Kass, 2010]. A greedy algorithm may be deployed to maximize $\E[f(Y)] | X = x$ in the form above by optimizing weights $w_i$ attached to each $f(y_i)$, within the range

$$
\frac{p(y_i | t, x)}{d(t | y_i, x) q(y_i)}, \quad \frac{p(y_i | t, x)}{d(t | y_i, x) q(y_i)}.
$$

The minimum may be achieved by a trivial adaptation. Maximizing and minimizing $\E[f(Y)] | X = x$ with respect to the bounding quantities $(\gamma, \Lambda)$ enables the resolution of ignorance bounds on the basis of $\Gamma$ from Definition 2. Our Algorithm 1 adapts the method of Jesson et al. [2021] to heterogeneous weight bounds $[w_i, \bar{w}_i]$ per draw $i$.

**Algorithm 1:** The expectation maximizer, with $O(n)$ runtime if intermediate $\Delta_j$ are memoized.

**D  Experimental details**

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].

**Semi-synthetic evaluation.** In our first experiment, the synthetic binary outcome was generated by linearly combining the covariates and treatment and then applying a logistic curve for a Bernoulli probability. As the logit-transformed ATE was known to be linear, “recall” was evaluated as the portion of the ignorance intervals that permitted the actual linear effect along each section of the dose response.

**Semi-synthetic dataset.** The 148 MRI fields were normalized such that values floored at each variable’s 25% quantile and ceiled at the 95% quantile were mapped to the range $[0, 1]$. In each of the four replications, a random quarter of the covariates were assigned a nonzero, normally distributed coefficient, and one of them was deemed the treatment variable with unit coefficient. After computing the outcomes, we randomly resampled a third of the feature values, with replacement, in order to introduce noise to the covariates while preserving the marginals.

**Semi-synthetic estimators.** Both the predictor and the propensity model, censored and uncensored, were trained in ensembles of 32 artificial neural networks with one 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.
**Empirical evaluation.** 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 bounds. These competing quantities are a form of recall and (the opposite of) precision, respectively.

Denote \((y_c, \underline{y})\) the partially identified bounds of the censored model and \((y, \overline{y})\) the full model’s bounds for \( \hat{E}[Y_t] \), both at 95% confidence from percentile bootstrapping. For some \( \Gamma := s \) and a \( t \in [0, 1] \) grid, of length 17 in our case, ignorance is \( \sum_t [\overline{y}(s,t) - y_c(s,t)]/\sum_t [\overline{y}(t) - y(t)] \), and recall is the normalized intersection between the bounds:

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

**Empirical dataset.** 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.

**Empirical 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 6.
Figure 6. **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.