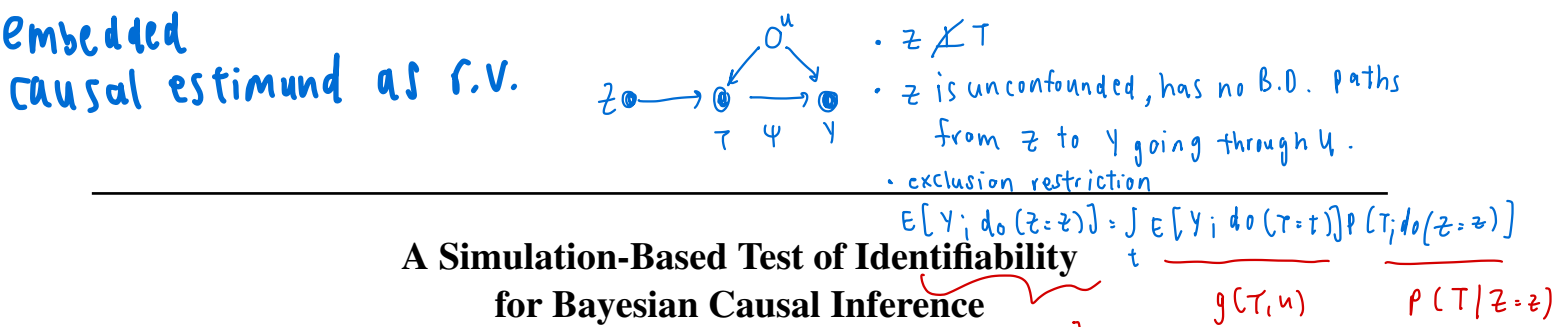


A Simulation-Based Test of Identifiability for Bayesian Causal Inference

Sam Witty [1] **David Jensen** [1] **Vikash Mansinghka** [2]

Abstract

This paper introduces a procedure for testing the identifiability of Bayesian models for causal inference. Although the do-calculus is sound and complete given a causal graph, many practical assumptions cannot be expressed in terms of graph structure alone, such as the assumptions required by instrumental variable designs, regression discontinuity designs, and within-subjects designs. We present simulation-based identifiability (SBI), a fully automated identification test based on a particle optimization scheme with simulated observations. This approach expresses causal assumptions as priors over functions in a structural causal model, including flexible priors using Gaussian processes. We prove that SBI is asymptotically sound and complete, and produces practical finite-sample bounds. We also show empirically that SBI agrees with known results in graph-based identification as well as with widely-held intuitions for designs in which graph-based methods are inconclusive.

1. Introduction

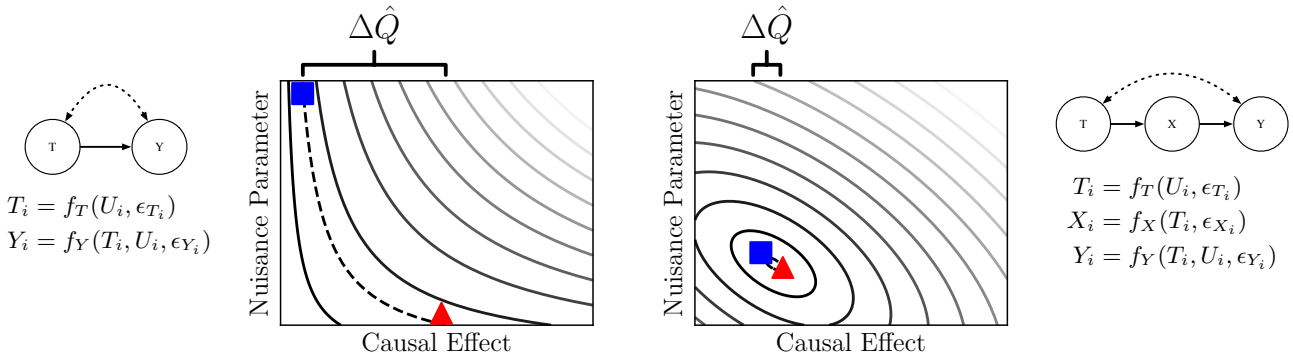
Given a causal graph, the do-calculus is a sound and complete procedure for nonparametric identification (Pearl, 1995; Huang & Valtorta, 2012). However, graph structure alone often tells an incomplete story of causal knowledge. Instrumental variable designs require monotonicity or linearity (Cragg & Donald, 1993), within-subjects designs require that latent confounders are shared across units (Loftus & Masson, 1994; Gelman, 2006), and regression discontinuity designs violate positivity assumptions that are implicit in the do-calculus (Lee & Lemieux, 2010). The do-calculus is inconclusive for these designs because it ignores the necessary restrictions on structural functions.

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, United States [2]Massachusetts Institute of Technology, Cambridge, United States. Correspondence to: Sam Witty <switty@cs.umass.edu>.

Historically, practitioners have circumvented the limitations of the do-calculus by either avoiding novel causal designs altogether or constructing cumbersome proofs from first principles. This paper proposes an alternative: (i) express causal assumptions as priors over structural causal models, implicitly inducing a joint distribution over observed and counterfactual random variables; and (ii) search for models that are likelihood equivalent, but that produce different effect estimates.

We introduce simulation-based identifiability (SBI), a flexible and automated identification technique that is compatible with any prior over structural causal models that: (i) can be used to sample data, and (ii) induces a differentiable likelihood function. SBI translates the problem of causal identification to an optimization procedure that seeks to maximize the data likelihood of two candidate structural causal models while maximizing the distance between their causal effect estimates. If the optimal solution to this procedure is two models that agree on effect estimates, then the causal effect is identifiable. Figure 1 provides intuition for how SBI uses gradient-based optimization to determine causal identifiability.

In Section 4 we prove that SBI is asymptotically sound and complete, assuming certain (strong) regularity conditions are met by the structural causal model, provide PAC-like bounds in the case of finite samples, and present an algorithm that can be implemented in any differentiable generative probabilistic programming language. In Section 5, we show that SBI is broadly applicable by presenting Bayesian variants of a suite of well-studied quasi-experimental designs such as regression discontinuity, instrumental variable, and within-subjects designs. We also present their semiparametric extensions using Gaussian processes, further demonstrating SBI's flexibility. We go on to show empirically that SBI correctly determines the identifiability of well-studied parameteric causal graphs and quasi-experimental designs, including those not covered by the do-calculus. Finally, we use SBI to extract quantitive insight about several Gaussian process quasi-experimental designs, which was previously unknown.



(a) Non-identifiable causal model and likelihood

(b) Identifiable causal model and likelihood

Figure 1: **Simulation-based identifiability discovers likelihood equivalent causal models.** Simulation-based identifiability maximizes the data likelihood *and* the distance between the causal estimates induced by two structural causal models, $\Delta\hat{Q}$. When causal effects are not identifiable (a) simulation-based identifiability discovers maximum likelihood models (blue and red) that estimate different causal effects. When causal effects are identifiable (b) the likelihood dominates and the two models converge to the same causal effect. Simulation-based identifiability is fully automated for any prior over structural causal models with a differentiable likelihood and causal effect, covering quasi-experimental designs that previously required custom proofs, including semiparametric versions using Gaussian process priors.

1.1. Related Work

Our work is not the first to automate identification for causal inference. The do-calculus (Pearl, 1995; Huang & Valtorta, 2012) determines nonparametric identifiability by manipulating expressions containing the *do* intervention operator symbolically using three rules, each of which depend only on the structure of a causal directed acyclic graph. On the other end of the spectrum of parametric assumptions, similar methods have been developed for linear causal graphs (Bollen, 2005; Kumor et al., 2019). However, these automated approaches provide no insight into semiparametric causal models, such as those using Gaussian processes, or models with richly structured, non-graphical assumptions. SBI expands the breadth of assumptions that can be reasoned about automatically, covering a suite of parametric and semiparametric models that previously required custom identification proofs.

Similar approaches for determining identifiability have been developed in other fields, such as neuroscience (Valdes-Sosa et al., 2011) and dynamical systems (Raue et al., 2009), by searching for likelihood equivalent parameters using gradient-based search. SBI differs from these approaches in two important ways. Methodologically, SBI uses a particle-based objective function to search for likelihood equivalent models globally, rather than locally near a single maximum likelihood solution. Second, SBI's objective function searches for models that estimate different causal effects, not only models that have different parameters. This distinction means that SBI can correctly determine identifiability even when models are semi-parametric (e.g see Section 4.1), or when effects are composed of many parameters. It is well known that causal effects can be identified even in settings where individual parameters can not (Pearl, 2009).

SBI requires that users express their assumptions as priors over structural causal models, which is an emerging practice in machine learning approaches to causal inference (Tran & Blei, 2018; Witty et al., 2020). This representation is well suited for probabilistic programming languages (Goodman et al., 2008; Mansinghka et al., 2014), which provide users with a syntax for expressing probabilistic models as code. Many of these languages provide support for automatic differentiation and gradient-based optimization (Bingham et al., 2018; Carpenter et al., 2017; Cusumano-Towner et al., 2019; Dillon et al., 2017), which can be used seamlessly with our optimization-based approach. While some probabilistic programming languages contain an explicit representation of interventions (Bingham et al., 2018; Perov et al., 2020; Tavares et al., 2019), none currently address the problem of causal identification.

2. Preliminaries

2.1. Structural Causal Models

SBI requires that causal assumptions be expressed as priors over structural causal models (SCMs). Given a fixed set of deterministic causal functions $\mathbf{F}$, SCMs induce a joint distribution $P(\mathbf{V}, \mathbf{V}_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}_{\mathbf{V}'=\mathbf{v}''})$ over observed, $\mathbf{V}$, and counterfactual, $\mathbf{V}_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}_{\mathbf{V}'=\mathbf{v}''}$, random variables (Pearl, 2009) by marginalizing over latent confounders, $\mathbf{U}$, and

Design	Description	Identifiable	Source
Unconfounded	Latent variables influence treatment, T, or outcome, Y, but not both.	Yes	(Pearl, 1995)
Confounded	A latent confounder, U, influences both T and Y.	No	(Pearl, 1995)
Backdoor	An observed confounder, X, influences T and Y.	Yes	(Pearl, 1995)
Frontdoor	U influences T and Y, but does not influence an observed mediator, X.	Yes	(Pearl, 1995)
Instrumental variable	U influences T and Y. An observed instrument, I, influences T, does not influence Y except through T, and is not influenced by U.	See § 5.2	(Angrist et al., 1996)
Within subjects	Each instance of U influences multiple instances of T and Y.	See § 5.2	(Draper, 1995)
Regression discontinuity	An observed confounder, X, influences T and Y. T depends on X being above or below a known threshold.	See § 5.2	(Rubin, 1977)

Table 1: **Causal designs covered by simulation-based identifiability.** A diverse set of quasi-experimental designs have been proposed in the literature, many of which can not be expressed in terms of graph structure alone. SBI translates identifiability, which previously required design-specific proofs, to gradient-based optimization, which can be automated.

exogenous noise, $\boldsymbol{\epsilon}$. The notation $\mathbf{V}_{\mathbf{V}'=\mathbf{v}'}$ refers to the set of random variables $\mathbf{V}$ induced by an intervention $do(\mathbf{V}' = \mathbf{v}')$. We distinguish between latent confounders, $\mathbf{U}$, which influence multiple observed variables, and exogenous noise, $\boldsymbol{\epsilon}$, which influence a single observed variable. Unless otherwise noted, we assume that $\mathbf{V}$ and $\mathbf{U}$ are composed of n independent and identically distributed random variables, denoted $\mathbf{V}_i$ and $\mathbf{U}_i$ respectively.

2.2. Priors over Functions

We consider two templates for priors over functions. The first, parametric priors, expresses a prior $P(f)$ over each $f \in \mathbf{F}$ in terms of a finite set of parameters θ. For example, $f_y(x_i, \epsilon_i) = \beta \cdot x_i + \epsilon_i$, $\beta \sim \mathcal{N}(0, 1)$ is a parametric prior.

The other template for priors are semiparametric priors, i.e. priors which can only be marginalized over to induce a joint distribution over observable and counterfactual random variables. These priors are often refered to as Bayesian nonparametric priors, although we refer to them as semiparametric to avoid confusion with nonparametric structural causal models. In Section 4.1 we show how SBI is compatible with flexible Gaussian process priors (Rasmussen, 2003). A Gaussian process is a distribution over deterministic functions, $y_i = f(x_i)$, $f \sim GP(m, k)$, given by a mean function, $m(x)$ and kernel covariance function, $k(x, x)$. By definition, any finite collection of draws from a Gaussian process prior are jointly Gaussian distributed, $Y \sim \mathcal{N}(\mu, \Sigma)$, where $\mu_i = m(x_i)$ and $\Sigma_{i,j} = k(x_i, x_j)$.

3. Identifiability in Bayesian Causal Inference

Given a prior over functions, latent confounders, and exogenous noise in a structural causal model, $P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$, we are interested in the posterior distribution of causal effects, $P(Q(\mathbf{V}_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}_{\mathbf{V}'=\mathbf{v}''})|\mathbf{V})$. Here, Q is a function that maps two sets of counterfactual outcomes to $\mathbb{R}$, also known as a causal estimand. For example, the sample average treatment effect, $\sum_{i=1}^{n}[Y_{i,T=t'} - Y_{i,T=t''}]$ is a causal estimand where Y is some outcome and T is some treatment in $\mathbf{V}$.

Let $\mathbf{F}^*, \mathbf{U}^*, \boldsymbol{\epsilon}^* \sim P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$ be a sample from the prior, and $\mathbf{V}^*, \mathbf{V}^*_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^*_{\mathbf{V}'=\mathbf{v}''}$ be the set of observed and counterfactual random variables generated by applying $\mathbf{F}^*$ to $\mathbf{U}^*$ and $\boldsymbol{\epsilon}^*$. Then we define identifiability as follows:

Definition 3.1. A causal estimand Q is identifiable given a sample $\mathbf{F}^*, \mathbf{U}^*, \boldsymbol{\epsilon}^* \sim P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$ if the posterior $Q(\mathbf{V}_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}_{\mathbf{V}'=\mathbf{v}''})|\mathbf{V}^*$ converges to $Q(\mathbf{V}^*_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^*_{\mathbf{V}'=\mathbf{v}''})$ as $n \to \infty$.

Even though Definition 3.1 is given in terms of posterior convergence, determining whether a causal estimand is identifiable does not require direct computation or approximation of the posterior. Instead, we show that a causal estimand is identifiable if and only if it is unique over all assignments of functions and latent confounders that maximize the conditional density of the data, i.e. the likelihood. All proofs are provided in the supplementary materials.

Theorem 3.1. In settings where $\mathbf{V}$ is comprised of n i.i.d instances, $\frac{P(\mathbf{V}^*|\mathbf{F}^\dagger,\mathbf{U}^\dagger)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} \to 0$ or 1 for all $(\mathbf{F}^\dagger, \mathbf{U}^\dagger) \in \text{supp}(P(\mathbf{F}, \mathbf{U}))$ as $n \to \infty$. [1]

Theorem 3.2. A causal estimand Q is identifiable given a sample $\mathbf{F}^*, \mathbf{U}^*, \boldsymbol{\epsilon}^* \sim P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$ if and only if there does not exist an $(\mathbf{F}^\dagger, \mathbf{U}^\dagger)$ such that $P(\mathbf{V}^*|\mathbf{F}^\dagger, \mathbf{U}^\dagger)$ converges to $P(\mathbf{V}^*|\mathbf{F}^*, \mathbf{U}^*)$ as $n \to \infty$, $Q(\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''}) \neq Q(\mathbf{V}^*_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^*_{\mathbf{V}'=\mathbf{v}''})$, and $P(\mathbf{F}^\dagger, \mathbf{U}^\dagger) > 0$. Here, $\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}$ and $\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''}$ are counterfactual outcomes generated by applying $\mathbf{F}^\dagger$ to $\mathbf{U}^*$ and $\boldsymbol{\epsilon}^\star$.

[1] All of our theoretical results hold under the condition that the likelihood ratio converge to 0 or 1 asymptotically, even if the observations are not sampled i.i.d.

Theorem 3.2 provides the foundation for the SBI procedure, relating identifiability to whether multiple maximum likelihood SCMs produce distinct causal effects. Importantly, Theorem 3.2 does not state that Q is not identifiable if there exists two maximum likelihood structural causal models, as they may produce the same effect estimates. Informally, functions or confounders that are *far away* from the treatment and outcome variables can be free to vary without consequence. SBI makes this distinction explicit, searching only for models that estimate different effects.

4. Simulation-Based Identifiability

Consider the following objective function:

$$\mathcal{L}(\mathbf{F}^1, \mathbf{U}^1, \mathbf{F}^2, \mathbf{U}^2, \mathbf{V}^*) = \log P(\mathbf{V}^*|\mathbf{F}^1, \mathbf{U}^1) + \log P(\mathbf{V}^*|\mathbf{F}^2, \mathbf{U}^2) + \lambda \Delta Q$$

Here, $\Delta Q = |Q^1 - Q^2|$, where Q^1 and Q^2 are the causal estimates resulting from each of the two SCMs, $(\mathbf{F}^1, \mathbf{U}^1, \boldsymbol{\epsilon}^*)$ and $(\mathbf{F}^2, \mathbf{U}^2, \boldsymbol{\epsilon}^*)$, respectively, and λ is a real-valued hyperparameter. In other words, ΔQ is the difference between estimated causal effects. Let $\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1, \hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2$ denote a solution that maximizes $\mathcal{L}$, and let $\Delta\hat{Q}$ be the corresponding ΔQ. All of the following theorems hold for any fixed λ.

First, we show that as n approaches infinity, $(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1)$ and $(\hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2)$ converge to maximum-likelihood solutions:

Theorem 4.1. For any $\mathbf{V}^* \sim P(\mathbf{V}|\mathbf{F}^*, \mathbf{U}^*)$, $P(\mathbf{V}^*|\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1)$ and $P(\mathbf{V}^*|\hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2)$ converge to $P(\mathbf{V}^*|\mathbf{F}^*, \mathbf{U}^*)$ as $n \to \infty$,

Next, we show that as n approaches infinity, the difference between maximum likelihood causal estimates is maximized by $\Delta\hat{Q}$. Here, $\mathbb{L}$ is the set of maximum likelihood functions and latent confounders, $(\mathbf{F}, \mathbf{U})$:

Theorem 4.2. $\Delta\hat{Q} \to \max_{(\mathbf{F}^1, \mathbf{U}^1, \mathbf{F}^2, \mathbf{U}^2) \in \mathbb{L}} \Delta Q$ as $n \to \infty$.

This leads us to our main asymptotic result. Namely, that:

Theorem 4.3. A causal estimand Q is identifiable given a prior $P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$ and a sample $\mathbf{F}^*, \mathbf{U}^*, \boldsymbol{\epsilon}^*$ from that prior if and only if $\Delta\hat{Q} \to 0$ as $n \to \infty$.

We now have a sufficient condition for determining identifiability asymptotically, which is justified in the i.i.d. setting by Theorem 3.1. However, given finite samples the conditions for Theorem 3.1 will not be satisfied, and $\Delta\hat{Q}$ may be greater than 0 even if the causal estimand is not identifiable. Therefore, we must rely on additional assumptions. Let $\mathbb{E}_{\mathbf{V}^*,n}\Delta\hat{Q}$ denote the expected difference between causal estimates for n instances, which we assume is less than the difference asymptotically, $\mathbb{E}_{\mathbf{V}^*,\infty}\Delta\hat{Q}$:

Assumption 4.1. $\mathbb{E}_{\mathbf{V}^*,n}\Delta\hat{Q} \geq \mathbb{E}_{\mathbf{V}^*,\infty}\Delta\hat{Q}$ for all n in $\mathbb{N}$.

Algorithm 1 Simulation-Based Identifiability

Input:
  Prior: $P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$
  Causal estimand and interventions: Q, $\mathbf{V}'$, $\mathbf{v}'$, $\mathbf{v}''$
  SBI hyperparameters: n, k, λ
  Threshold and acceptance probability: t, p
Procedure:
  $\mathbf{F}^*, \mathbf{U}^* \sim P(\mathbf{F}, \mathbf{U})$ ▷ prior sample
  for $i = 1$ **to** k **do**
    $\mathbf{V}_i^* \sim P(\mathbf{V}|\mathbf{F}^*, \mathbf{U}^*)$ ▷ n data instances
    $\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1, \hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2 \leftarrow \text{argmax } \mathcal{L}(\mathbf{F}^1, \mathbf{U}^1, \mathbf{F}^2, \mathbf{U}^2, \mathbf{V}_i^*)$
    ▷ via gradient descent
    $\hat{Q}_i^1 \leftarrow Q(\hat{\mathbf{V}}^1_{\mathbf{V}=\mathbf{v}'}, \hat{\mathbf{V}}^1_{\mathbf{V}=\mathbf{v}''})$ ▷ SCM 1 effect
    $\hat{Q}_i^2 \leftarrow Q(\mathbf{V}^2_{\hat{\mathbf{V}}=\mathbf{v}'}, \mathbf{V}^2_{\hat{\mathbf{V}}=\mathbf{v}''})$ ▷ SCM 2 effect
  $\hat{\mu}_{\Delta Q} \leftarrow \frac{1}{k}\sum_{i=1}^{k} |\hat{Q}_i^1 - \hat{Q}_i^2|$
  $\hat{\sigma}^2_{\Delta Q} \leftarrow \frac{1}{k-1}\sum_{i=1}^{k} (\hat{\mu}_{\Delta Q} - |\hat{Q}_i^1 - \hat{Q}_i^2|)^2$
  if $\Phi\left(\frac{(\hat{\mu}_{\Delta Q} - t)\sqrt{k}}{\hat{\sigma}^2_{\Delta Q}}\right) < p$ **then**
    return $TRUE$ ▷ Identifiable
  else
    return $FALSE$ ▷ Not identifiable

We approximate $\mathbb{E}_{\mathbf{V}^*,n}\Delta\hat{Q}$ by repeatedly simulating $\mathbf{V}^* \sim P(\mathbf{V}|\mathbf{F}^*, \mathbf{U}^*)$ and then optimizing $\mathcal{L}$ to determine $\Delta\hat{Q}$ for each sample. Given that the number of simulations, k, is large enough such that the sample mean is approximately normally distributed, Assumption 4.1 implies bounds on the probability that $\mathbb{E}_{\mathbf{V}^*,\infty}\Delta\hat{Q}$ is greater than some user-defined threshold, t.

This leads us to Algorithm 1, which works as follows. First, sample a set of functions, $\mathbf{F}^*$, and latent confounders, $\mathbf{U}^*$, from the prior. Then, repeatedly sample a set of observations, $\mathbf{V}^*$. For each set of observations optimize $\mathcal{L}$ jointly for two SCMs. Finally, return TRUE if the expected distance between causal estimates is greater than t with probability less p and return FALSE otherwise.

Theorem 4.4. If Algorithm 1 returns TRUE, $P(\mathbb{E}_{\mathbf{V}^*,\infty}\Delta\hat{Q} > t) < p$.[2]

Selecting the repulsion strength, λ. While the choice of repulsion strength, λ, does not influence our asymptotic results, this is not generally the case for any finite n. In our experiments in Section 5, we find that even small values of λ produce large $\Delta\hat{Q}$ for non-identifiable models. We suggest using the likelihood of $(\mathbf{F}^*, \mathbf{U}^*)$ to calibrate λ. In other words, if $P(\mathbf{V}^*|\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1)$ or $P(\mathbf{V}^*|\hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2)$ are significantly different from $P(\mathbf{V}^*|\hat{\mathbf{F}}^*, \hat{\mathbf{U}}^*)$, λ should be reduced.

[2]Here, we assume that the central limit theorem provides a sufficiently tight approximation to the distribution of the sample mean given a finite number of simulations, k.

Our results thusfar assume that we can optimize over the space of functions $\mathbf{F}$ directly. In the parametric setting optimization over $\mathbf{F}$ simply reduces to optimizing over parameters, θ. However, in the semiparametric case we do not have access to a closed-form expression for $P(\mathbf{V}^*|\mathbf{F}, \mathbf{U})$ by definition, and instead only have the conditional distribution of $\mathbf{V}^*$ given $\mathbf{F}$ evaluated at a finite collection of inducing points. Therefore, instead of searching over the space of θ, as in the parametric case, for any function $f \in \mathbf{F}$ with a semiparametric prior we instead search over the space of $f(x)$ evaluated at $x_1, ..., x_m \sim P(x)$. As $m \to \infty$ this approximation becomes exact.

4.1. Example: Confounded GP Regression

Consider a structural causal models over observed $\mathbf{V} = \{T_1, Y_1, ..., T_n, Y_n\}$ and latent $\mathbf{U} = \{U_1, ..., U_n\}$ random variables corresponding to the graph structure in Figure 1a. We assume that the function $Y_i = f(T_i, U_i, \epsilon_{T_i})$ is drawn from a zero-mean Gaussian process prior with a radial basis function (RBF) kernel, $k([T_i, U_i], [T_j, U_j]) = k(T_i, T_j; l_T, s_T) \cdot k(U_i, U_j; l_U, s_U)$, with additive exogenous noise and that treatment is a linear function of the latent confounder, $T_i = \gamma \cdot U_i + \epsilon_{T_i}$. Given the length-scale, l, and scale, s, hyperparameters the RBF kernel is given as follows: $k(x, x'; l, s) = s \cdot \exp[-(x - x')^2/l]$. Let K_U, K_T, $K_{T,t'}$, and $K_{T,t''}$ be $n \times n$ kernel matrices such that their i, j'th elements are given by $k(U_i, U_j; l_U, s_U)$, $k(T_i, T_j; l_T, s_T)$, $k(T_i, t'; l_T, s_T)$, and $k(T_i, t''; l_T, s_T)$ respectively. Let $k_{t',t''}$ be shorthand for $k(t', t'', l_T, s_T)$ and let $\mathbf{1}$ be a $n \times n$ block matrix of ones. Finally, let $\mathbf{t}$, $\mathbf{y}$, and $\mathbf{u}$ be the length n vectors of all treatment, outcome, and confounder instances respectively.

As this model includes a combination of parametric and semiparameteric priors over structural functions, optimizing over $\mathbf{F}$ involves optimizing over the set of parameters $\theta = \{l_T, s_T, l_U, s_U, \gamma, \sigma_T^2, \sigma_Y^2\}$ as well as the inducing counterfactual outcomes, $\mathbf{y}_{\text{cf}} = [\mathbf{y}', \mathbf{y}'']$, where $\mathbf{y}'$ and $\mathbf{y}''$ are the length n vectors of counterfactual outcomes given the interventions $do(T = t')$ and $do(T = t'')$ respectively. Therefore, the log likelihood is given by the following:

$$
\begin{aligned}
\log P(\mathbf{V}|\mathbf{F}, \mathbf{U}) &= \log \mathcal{N}(\mathbf{t}; \gamma\mathbf{u}, \sigma_T^2\mathbf{I}) + \log \mathcal{N}(\mathbf{y}; \mu_Y, \Sigma_Y) \\
\mu_Y &= K_* K^{-1} \mathbf{y}_{\text{cf}} \\
\Sigma_Y &= K - K_* K_{**}^{-1} K_*^t + \sigma_Y^2 \mathbf{I} \\
K &= K_T \odot K_U \\
K_* &= \begin{bmatrix} K_{T,t'} \\ K_{T,t''} \end{bmatrix} \odot \begin{bmatrix} K_U \\ K_U \end{bmatrix} \\
K_{**} &= \begin{bmatrix} s_T \mathbf{1} & k_{t',t''} \mathbf{1} \\ k_{t',t''} \mathbf{1} & s_T \mathbf{1} \end{bmatrix} \odot \begin{bmatrix} K_U & K_U \\ K_U & K_U \end{bmatrix}
\end{aligned}
$$

The above expression follows directly from the fact that factual and counterfactual outcomes are jointly Gaussian by definition, and that Gaussian distributions are closed under conditioning. Note also that factual and counterfactual outcomes are functions of the same set of latent confounders. This results in the confounders' contribution to the kernel covariance between factual and counterfactual outcomes, K_*, and the kernel covariance between counterfactual outcomes, K_{**}, being given by block-structured matrices with repeated blocks K_U.

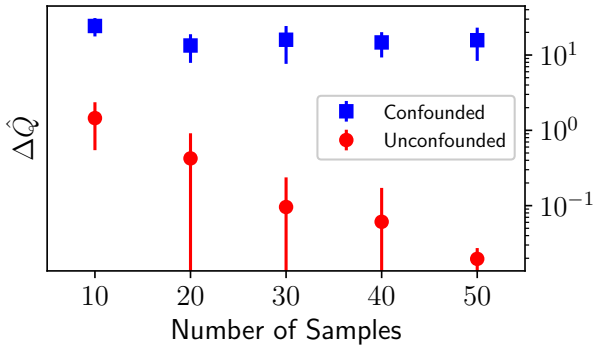


Figure 2: **Simulation-based identifiability is correct for flexible Gaussian process causal models.** High values of $\Delta\hat{Q}$ for the confounded Gaussian process model indicate that the treatment effect is not identifiable. Without confounding however, the two optimized particles converge.

Let us assume we are interested in the average treatment effect given the intervention $do(T = t')$ and $do(T = t'')$. Then, by definition Q is given by the mean difference in counterfactual outcomes, $\frac{1}{n} \sum_{i \in n} \mathbf{y}_i' - \mathbf{y}_i''$,

Given the above expressions for the log density of the observations and the causal estimand in terms of parameters, θ, counterfactual outcomes, $\mathbf{y}_{\text{cf}}$, and latent confounders, $\mathbf{U}$, we can now compute the partial derivative of the particle-based objective function, $\frac{\partial}{\partial s}\mathcal{L} = \frac{\partial}{\partial s} \log P(\mathbf{V}^*|\mathbf{F}^1, \mathbf{U}^1) + \frac{\partial}{\partial s} \log P(\mathbf{V}^*|\mathbf{F}^2, \mathbf{U}^2) + \lambda \frac{\partial}{\partial s} \Delta Q$ with respect to all $s \in \theta \cup \mathbf{U} \cup \mathbf{y}_{\text{cf}}$. Given an expression for each partial derivative, we can then apply standard gradient-descent algorithms to determine identifiability. Without loss of generality, the derivative of the repulsion term with respect to s for $\mathbf{F}^1$, $\mathbf{U}^1$ is given by the following:

$$
\frac{\partial}{\partial s} \Delta Q = \begin{cases} \frac{\partial Q}{\partial s} & Q^1 > Q^2 \\ -\frac{\partial Q}{\partial s} & \text{otherwise} \end{cases}
$$

$$
\frac{\partial Q}{\partial s} = \begin{cases} \frac{1}{n} & s \in \mathbf{y}_i' \ \ \forall i \in 1...n \\ -\frac{1}{n} & s \in \mathbf{y}_i'' \ \ \forall i \in 1...n \\ 0 & \text{otherwise} \end{cases}
$$

For the derivative of the log density we expand on standard identities of Gaussians as follows, where $L_{\mathbf{V}}$ is shorthand

for $\log P(\mathbf{V}|\mathbf{F}^1, \mathbf{U}^1)$, $L_{\mathbf{t}}$ is shorthand for $\log P(\mathbf{t}|\gamma\mathbf{U}, \sigma_T^2\mathbf{I})$ and $L_{\mathbf{y}}$ is shorthand for $\log P(\mathbf{y}|\mu_Y, \Sigma_Y)$:

$$
\begin{aligned}
\frac{\partial L_{\mathbf{V}}}{\partial s} &= \frac{\partial L_{\mathbf{t}}}{\partial s} + \frac{\partial L_{\mathbf{y}}}{\partial s} \\
\frac{\partial L_{\mathbf{t}}}{\partial s} &= \frac{1}{\sigma_T^2} \sum_{i=1}^{n} (T_i - \gamma U_i) \frac{\partial \gamma U_i}{\partial s} \\
&\quad - \frac{\partial \sigma_T^2}{\partial s} \frac{1}{2\sigma_T^2} \left(n - \frac{1}{\sigma_T^2} \right) \sum_{i=1}^{n} (T_i - \gamma U_i)^2 \\
\frac{\partial L_{\mathbf{y}}}{\partial s} &= \nabla_{\mu_Y} L_{\mathbf{y}} \cdot \frac{\partial \mu_Y}{\partial s} + \nabla_{\Sigma_Y} L_{\mathbf{y}} \cdot \frac{\partial \Sigma_Y}{\partial s} \\
\nabla_{\mu_Y} L_{\mathbf{y}} &= \Sigma_Y^{-1} (\mathbf{y} - \mu_Y) \\
\frac{\partial \mu_Y}{\partial s} &= \frac{\partial K_*}{\partial s} K^{-1} \mathbf{y}_{\text{cf}} + K_* K^{-1} \frac{\partial K}{\partial s} K^{-1} \mathbf{y}_{\text{cf}} \\
&\quad + K_* K^{-1} \frac{\partial \mathbf{y}_{\text{cf}}}{\partial s} \\
\nabla_{\Sigma_Y} L_{\mathbf{y}} &= -\frac{1}{2} (\Sigma_Y^{-1} - \Sigma_Y^{-1} (\mathbf{y} - \mu_Y)(\mathbf{y} - \mu_Y)^t \Sigma_Y^{-1}) \\
\frac{\partial \Sigma_Y}{\partial s} &= \frac{\partial K}{\partial s} - \frac{\partial K_*}{\partial s} K_{**}^{-1} K_*^t + K_* K_{**}^{-1} \frac{\partial K_{**}}{\partial s} K_{**}^{-1} K_*^t \\
&\quad + K_* K_{**}^{-1} \frac{\partial K_*}{\partial s}^t + \frac{\partial \sigma_Y^2}{\partial s} \mathbf{I}
\end{aligned}
$$

See the supplementary materials for additional details on the partial derivatives of the kernel covariance matrices. Note that although deriving this gradient is cumbersome and error-prone in general, it can be easily automated using standard automatic differentiation procedures, even for models using Gaussian process priors.

Figure 2 shows the results of Algorithm 1 with this prior over structural causal models using the Adam gradient descent algorithm (Kingma & Ba, 2014) to optimize $\mathcal{L}$. Unlike the unconfounded regression model, which is identical except that $\mathbf{U}$ has been omitted, we conclude that the confounded model is not identifiable. We expand on these examples in Section 5 and in the supplementary materials.

5. Experiments

We evaluated SBI on priors reflecting seven standard causal designs which are summarized in Table 1; unconfounded regression (UR), confounded regression (CR), backdoor adjusted (BA), frontdoor adjusted (FA), instrumental variable (IV), within-subjects (WS), and regression discontinuity designs (RD). For each of these seven designs we tested SBI using two sets of parametric assumptions; where each structural function is drawn from a prior over linear or quadratic functions respectively. In addition to these parametric priors, we evaluated SBI on Gaussian process versions of five of the seven designs, all but backdoor and frontdoor adjusted designs. The quadratic polynomial results, and additional experimental and baseline details are provided in the supplementary materials.

For each of the parametric causal models we ran Algorithm 1 for the sample average treatment effect, $Q = \frac{1}{n}\sum_{i=1}^{n}[Y_{i,T=t'} - Y_{i,T=t''}]$, with $n = 30{,}000$ samples, $k = 30$ trials, and using the Adam (Kingma & Ba, 2014) algorithm to optimize $\mathcal{L}$. We ran Adam with $\alpha = 0.01$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ for fifty epochs with a minibatch size of thirty instances.

For all of the linear parametric experiments, we assume that each function $V_i = f_V(Pa(V)_i, \epsilon_{V_i}) = \beta \cdot Pa(V)_i + \epsilon_{V_i}$, where each element of β is drawn from a normal prior. Here, $Pa(V)_i$ refers to the vector of all latent and observed arguments in the structural function, f_V. All of the experiments, including the Gaussian process models, assume that exogenous noise is normally distributed and additive.

For all of the Gaussian process experiments, we assume that each outcome function $Y_i = f_Y(Pa(Y), \epsilon_{Y_i})$ is drawn from the same Gaussian process prior described in Section 4.1. For each of the Gaussian process causal models we ran Algorithm 1 for the sample average treatment effect for $n = 50$ samples, $k = 30$ trials, and again using Adam to optimize $\mathcal{L}$ with $\alpha = 0.01$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ for 2,000 epochs with a minibatch size of ten instances and one gradient step per minibatch.

We compare SBI against a baseline using profile likelihood identification (Raue et al., 2009), which alternates between parameter perturbations and maximum likelihood optimization. We implemented Algorithm 1, all designs, and the baseline using the Gen probabilistic programming

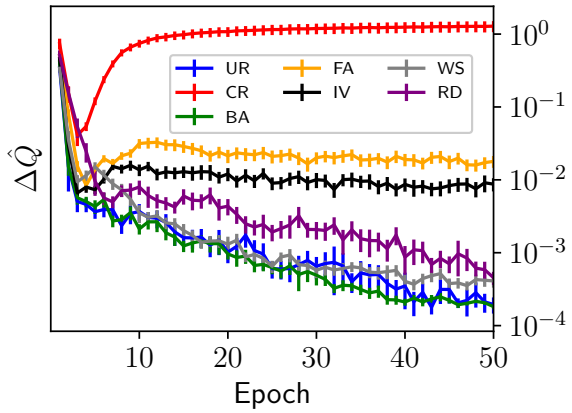


Figure 3: **Simulation-based identifiability agrees with known parametric results.** SBI correctly determines that of the causal designs we consider, the average treatment effect is identifiable for all except the confounded regression model. Results are normalized by the true causal effect.

Design	Linear			Gaussian Process			Nonparametric
	$\Delta\hat{Q}_{\text{SBI}}$	$\Delta\hat{Q}_{\text{Baseline}}$	Ground Truth	$\Delta\hat{Q}_{\text{SBI}}$	$\Delta\hat{Q}_{\text{Baseline}}$	Ground Truth	Ground Truth
Unconfounded	$\mathbf{.00 \pm .00}$	$\mathbf{.02 \pm .00}$	ID	$\mathbf{.02 \pm .24}$	$.28 \pm .01$	ID	ID
Confounded	$\mathbf{.66 \pm .12}$	$\mathbf{.22 \pm .06}$	Not ID	$\mathbf{1.2 \pm .33}$	$\mathbf{1.1 \pm .38}$	Not ID	Not ID
Backdoor	$\mathbf{.00 \pm .00}$	$.06 \pm .00$	ID	n/a	n/a	ID	ID
Frontdoor	$\mathbf{.02 \pm .00}$	$.12 \pm .01$	ID	n/a	n/a	ID	ID
Inst. Var.	$\mathbf{.01 \pm .00}$	$\mathbf{.04 \pm .00}$	ID	$\mathbf{.35 \pm .16}$	$\mathbf{.25 \pm .05}$	Unknown	Not ID
Within Subj.	$\mathbf{.01 \pm .00}$	$\mathbf{.03 \pm .00}$	ID	$.10 \pm .07$	$.99 \pm .06$	Unknown	ID
Regr. Disc.	$\mathbf{.00 \pm .00}$	$.05 \pm .00$	ID	$\mathbf{.37 \pm .08}$	$\mathbf{1.6 \pm .20}$	Unknown	Not ID

Table 2: **Simulation-based identifiability accurately determines identifiability for diverse causal designs.** SBI outperforms the baseline, resulting in near-perfect identification results for the average treatment effect. Values are shown in bold if SBI or the baseline correctly determine identifiability using a threshold of 5 percent of the true effect. SBI provides novel insight that the Gaussian process quasi-experimental designs agree with nonparametric identification results.

language (Cusumano-Towner et al., 2019), which provides the necessary support for sampling and automatic differentiation. Using a threshold of 5 percent of the true causal effect, SBI correctly determines the identifiability of all designs except for the Gaussian process within-subjects design, performing significantly better than the baseline. Our experiments demonstrate that SBI agrees with the do-calculus in settings where graph structure alone is sufficient, and produces correct identification results in settings that require non-graphical assumptions and previously required custom identification proofs. Finally, we present the first known identification results for Gaussian process quasi-experimental designs, demonstrating agreement with widely held intuition. See Table 2 for a summary of identification results for the average treatment effect.

5.1. Causal graphical models

In addition to the unconfounded and confounded regression designs presented in Section 4.1, we evaluated SBI on two models that are covered by the do-calculus, backdoor adjusted and frontdoor adjusted designs.

Backdoor adjusted designs represent settings where all of the random variables that confound the relationship between treatment and outcome are observed, blocking all *backdoor* paths. Unlike backdoor adjusted designs, frontdoor adjusted designs can include latent confounding between treatment and outcome, as long as there exists an observed mediator that is not confounded, as in Figure 1b. Despite this confounding, average treatment effects are nonparametrically identifiable (Pearl, 2009). For these and the two models discussed in Section 4.1, SBI agrees with ground truth, even though the baseline fails for the unconfounded Gaussian process model.

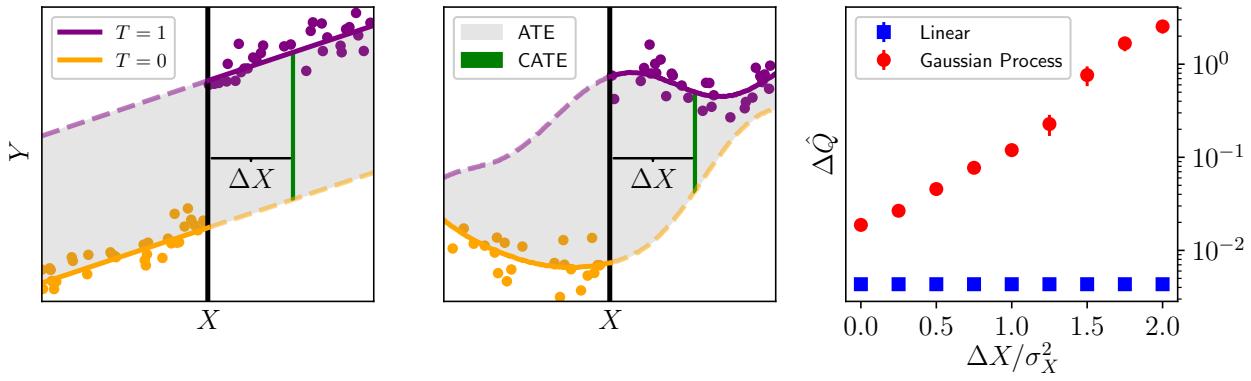
5.2. Linear Quasi-Experiment Designs

We evaluated SBI and the baseline on three well-studied, but previously disparate, linear quasi-experimental designs.

Instrumental variable designs are quasi-experimental designs in which an observed variable, known as the instrument, influences a treatment. Two conditions must be satisfied for the instrument to enable identification: (i) the instrument and the treatment must not be confounded; and (ii) all influence from the instrument to the outcome is mediated through the treatment. While these first two assumptions can be expressed graphically, additional parametric assumptions are needed for effects to be identifiable (Pearl, 2009). For example, if all functions are assumed to be linear, then the average treatment effect can be identified.

Within-subjects designs involve hierarchically structured data in which individual instances (e.g., students) are affiliated with one of several objects (e.g., schools). Treatment effects for linear models in these kinds of hierarchical settings can be identified even if treatment and outcome are confounded, as long as the confounders are shared across all instances belonging to the same object (Witty et al., 2020). Within-subjects designs can be described as the family of structural causal models; $T_i = f_T(U_{o(i)}, \epsilon_{T_i})$ $Y_i = f_Y(T_i, U_{o(i)}, \epsilon_{Y_i})$, where $U_{o(i)}$ refers to the shared value of the latent confounder corresponding to instance i. Hierarchically structured confounding is applicable to a wide variety of common causal designs (Jensen et al., 2020): including twin studies (Boomsma et al., 2002), difference-in-differences designs (Shadish et al., 2008), and multi-level-modeling (Gelman, 2006).

Regression discontinuity designs are quasi-experimental designs in which the treatment assignment depends on a particular observed covariate being above or below a known threshold. In this example we consider a sharp deterministic discontinuity in that the binary treatment assignment is deterministically 1 if $X_i > 0$, and 0 otherwise. Regression discontinuity designs can be described as the family of structural causal models $X_i = f_X(\epsilon_{X_i})$ $T_i = \mathbf{1}[X_i > 0]$, and $Y_i = f_Y(T_i, X_i, \epsilon_{T_i})$, where $\mathbf{1}$ is the indicator function. Even though all confounders are observed, estimating treatment effects requires estimating



(a) Linear RDD (b) GP RDD (c) CATE Results

Figure 4: **Simulation based identifiability agrees with widely-held intuition for regression discontinuity designs.** Conditional average treatment effects (CATE) are only nonparametrically identifiable infinitessimally close to the discontinuity (vertical black bar). While conditional average treatment effects for linear models (a) are well known to be identifiable, the identifiability of Gaussian process versions (b) was previously unknown. SBI provides novel and intuitive identification results (c), that effects becomes *less identifiable* as we condition on covariates further from the discontinuity.

conditional distributions for configurations of T and X which are never observed, as shown in Figures 4a and 4b. This violates the positivity assumption, which is a necessary assumption for soundness of the do-calculus.

Although these parametric quasi-experimental designs are not covered by the do-calculus, their identification results are well-known (Angrist et al., 1996; Draper, 1995; Rubin, 1977). Similar to the causal graphical models, SBI correctly recovers that average treatment effects are identifiable for all three of these designs, despite their diversity.

5.3. Gaussian Process Quasi-Experimental Designs

In addition to the causal graphical models and linear quasi-experimental designs, which have known identification results, we used SBI to determine the previously unknown identifiability of Gaussian process versions of quasi-experimental designs. By assuming a particular kernel we place an inductive bias on the class of structural functions, which could in principle enable identification. In this setting SBI instead confirms that the identifibility of Gaussian process models agrees with their nonparametric counterparts for the average treatment effect.

We also evaluated SBI on the conditional average treatment effect for linear and Gaussian process version of the regression discontinuity design. In the linear case, shown in Figure 4a, conditional average treatment effects can be unambiguously estimated by extrapolating the linear relationship between X and Y for each treatment. However, in the Gaussian process version, observations in one region of X provide only partial information about counterfactuals in another. SBI's results in Figure 4c agree with this intuition, demonstrating that the distance between likelihood equivalent conditional average treatment effects increases as we condition on covariates further from the discontinuity.

6. Discussion

In addition to being more flexible than prior automated approaches, SBI can used as a kind of sensitivity analysis (Franks et al., 2019; Kallus et al., 2019; Robins et al., 2000), bounding the range of causal effects that are likelihood equivalent. Our regression discontinuity design experiments in Figure 4c emphasize this capability, showing that irreducible uncertainty in effect estimates increases as we condition on covariates further from the discontinuity.

SBI builds on a long history of optimization-focused machine learning research. Reducing identifiability to optimization in this way provides a path towards reasoning about causal designs at previously unattainable scales, which is an exciting area of future work.

7. Conclusion

In this paper we presented simulation-based identifiability, an automated approach for causal identification. In Section 4 we proved that SBI is sound and complete in the limit of infinite samples and compute, and provided practical finite-sample bounds. In Section 5 we demonstrated that SBI correctly determines identifiability empirically with seven well-studied causal designs, three of which are out-of-scope for the do-calculus. We hope that SBI can embolden practitioners and researchers to rapidly iterate and explore novel designs for causal inference.

Acknowledgments

Thanks to Kenta Takatsu, Alex Lew, Cameron Freer, Marco Cusumano-Towner, Tan Zhi-Xuan, Jameson Quinn, Veronica Weiner, Sharan Yalburgi, Przemyslaw Grabowicz, Purva Pruthi, Sankaran Vaidyanathan, Erica Cai, and Jack Kenney for their helpful feedback and suggestions. Sam Witty and David Jensen were supported by DARPA and the United States Air Force under the XAI (Contract No. HR001120C0031), CAML (Contract No. FA8750-17-C-0120), and SAIL-ON (Contract No. w911NF-20-2-0005) programs. Vikash Mansinghka was supported by DARPA under the SD2 program (Contract No. FA8750-17-C-0239), a philanthropic gift from the Aphorism Foundation, a project under the MIT-Takeda program (Proposal No. 51135) and Intel (Agreement No. 6939564). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the United States Air Force.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434): 444–455, 1996.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., and Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

Bollen, K. A. Structural equation models. *Encyclopedia of biostatistics*, 7, 2005.

Boomsma, D., Busjahn, A., and Peltonen, L. Classical twin studies and beyond. *Nature Reviews Genetics*, 3 (11):872–882, 2002.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

Cragg, J. G. and Donald, S. G. Testing identifiability and specification in instrumental variable models. *Econometric Theory*, pp. 222–240, 1993.

Cusumano-Towner, M. F., Saad, F. A., Lew, A. K., and Mansinghka, V. K. Gen: A general-purpose probabilistic programming system with programmable inference. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI 2019, pp. 221–236, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367127. doi: 10.1145/3314221.3314642. URL https://doi.org/10.1145/3314221.3314642.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M., and Saurous, R. A. Tensorflow distributions. *arXiv preprint arXiv:1711.10604*, 2017.

Doob, J. L. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pp. 23–27, 1949.

Draper, D. Inference and hierarchical modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20(2):115–147, 1995.

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

Gelman, A. Multilevel (hierarchical) modeling: What it can and cannot do. *Technometrics*, 48(3):432–435, 2006.

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., and Tenenbaum, J. B. Church: a language for generative models. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, pp. 220–229, 2008.

Huang, Y. and Valtorta, M. Pearl's calculus of intervention is complete. *arXiv preprint arXiv:1206.6831*, 2012.

Jensen, D., Burroni, J., and Rattigan, M. Object conditioning for causal inference. In *Uncertainty in Artificial Intelligence*, pp. 1072–1082. PMLR, 2020.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kumor, D., Chen, B., and Bareinboim, E. Efficient identification in linear structural causal models with instrumental cutsets. In *Advances in Neural Information Processing Systems*, pp. 12477–12486, 2019.

Lee, D. S. and Lemieux, T. Regression discontinuity designs in economics. *Journal of economic literature*, 48 (2):281–355, 2010.

Loftus, G. and Masson, M. Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, 1(4):476–490, 1994.

Mansinghka, V., Selsam, D., and Perov, Y. Venture: a higher-order probabilistic programming platform with programmable inference. *arXiv preprint arXiv:1404.0099*, 2014.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. *Causality*. Cambridge university press, 2009.

Perov, Y., Graham, L., Gourgoulias, K., Richens, J., Lee, C., Baker, A., and Johri, S. Multiverse: causal reasoning using importance sampling in probabilistic programming. In *Symposium on advances in approximate bayesian inference*, pp. 1–36. PMLR, 2020.

Rasmussen, C. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pp. 63–71. Springer, 2003.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

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

Rubin, D. B. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

Shadish, W., Clark, M., and Steiner, P. Can nonrandomized experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103(484):1334–1344, 2008.

Tavares, Z., Zhang, X., Koppel, J., and Lezama, A. S. A language for counterfactual generative models. 2019.

Tran, D. and Blei, D. M. Implicit causal models for genome-wide association studies. In *International Conference on Learning Representations*, 2018.

Valdes-Sosa, P. A., Roebroeck, A., Daunizeau, J., and Friston, K. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361, 2011.

Witty, S., Takatsu, K., Jensen, D., and Mansinghka, V. Causal inference using gaussian processes with structured latent confounders. In *International Conference on Machine Learning*, pp. 10313–10323. PMLR, 2020.

8. Supplementary Materials

8.1. Asymptotic Soundness and Completeness

In this section, we prove Theorems 3.1, 3.2, 4.1, 4.2, 4.3, and 4.4.

Theorem 3.1. In settings where $\mathbf{V}$ is comprised of n i.i.d instances, $\frac{P(\mathbf{V}^*|\mathbf{F}^\dagger,\mathbf{U}^\dagger)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} \to 0$ or 1 for all $(\mathbf{F}^\dagger,\mathbf{U}^\dagger) \in \text{supp}(P(\mathbf{F},\mathbf{U}))$ as $n \to \infty$.

Proof. Let $p = \mathbb{E}[\frac{P(\mathbf{V}_i^*|\mathbf{F}^\dagger,\mathbf{U}^\dagger)}{P(\mathbf{V}_i^*|\mathbf{F}^*,\mathbf{U}^*)}]$ for a single data instance. Then $\mathbb{E}[\frac{P(\mathbf{V}^*|\mathbf{F}^\dagger,\mathbf{U}^\dagger)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)}] = p^n$ for n i.i.d data instances. As $0 \leq p \leq 1$, $lim_{n\to\infty} p^n = 0$ or 1. By the weak law of large numbers, we have that $\frac{P(\mathbf{V}^*|\mathbf{F}^\dagger,\mathbf{U}^\dagger)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} \to 0$ or 1 for all $(\mathbf{F}^\dagger,\mathbf{U}^\dagger) \in \text{supp}(P(\mathbf{F},\mathbf{U}))$ as $n \to \infty$. □

Theorem 3.2 A causal estimand Q is identifiable given a sample $\mathbf{F}^*, \mathbf{U}^*, \boldsymbol{\epsilon}^* \sim P(\mathbf{F},\mathbf{U},\boldsymbol{\epsilon})$ if and only if there does not exist an $(\mathbf{F}^\dagger,\mathbf{U}^\dagger)$ such that $P(\mathbf{V}^*|\mathbf{F}^\dagger,\mathbf{U}^\dagger)$ converges to $P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)$ as $n \to \infty$, $Q(\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''}) \neq Q(\mathbf{V}^*_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^*_{\mathbf{V}'=\mathbf{v}''})$, and $P(\mathbf{F}^\dagger,\mathbf{U}^\dagger) > 0$. Here, $\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}$ and $\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''}$ are counterfactual outcomes generated by applying $\mathbf{F}^\dagger$ to $\mathbf{U}^*$ and $\boldsymbol{\epsilon}^\star$.

Proof. The fact that Q is identifiable if there does not exist such an $(\mathbf{F}^\dagger,\mathbf{U}^\dagger)$ follows directly from the Bernstein-von Mises Theorem (Doob, 1949), which states that under certain regularity conditions (e.g. uniqueness of the maximum likelihood) the posterior distribution of any random variable is asymptotically consistent.

To show that Q is identifiable only if there does not exist such an $(\mathbf{F}^\dagger,\mathbf{U}^\dagger)$, we have that for all $(\mathbf{F}^\dagger,\mathbf{U}^\dagger) \in \text{supp}(P(\mathbf{F},\mathbf{U}))$:

$$P(Q(\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''})|\mathbf{V}^*) = \int_{(\mathbf{F},\mathbf{U})\in\mathbb{A}^\dagger} P(\mathbf{F},\mathbf{U}|\mathbf{V}^*) d\mathbf{F} d\mathbf{U} \tag{1}$$

$$= \int_{(\mathbf{F},\mathbf{U})\in\mathbb{A}^\dagger} \frac{P(\mathbf{V}^*|\mathbf{F},\mathbf{U})P(\mathbf{F},\mathbf{U})}{P(\mathbf{V}^*)} d\mathbf{F} d\mathbf{U} \tag{2}$$

$$= \frac{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)}{P(\mathbf{V}^*)} \int_{(\mathbf{F},\mathbf{U})\in\mathbb{A}^\dagger} \frac{P(\mathbf{V}^*|\mathbf{F},\mathbf{U})}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} P(\mathbf{F},\mathbf{U}) d\mathbf{F} d\mathbf{U} \tag{3}$$

$$\geq \frac{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)}{P(\mathbf{V}^*)} \int_{(\mathbf{F},\mathbf{U})\in\mathbb{A}^\dagger\cap\mathbb{L}^*} P(\mathbf{F},\mathbf{U}) d\mathbf{F} d\mathbf{U} \tag{4}$$

$$\frac{P(Q(\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''})|\mathbf{V}^*)}{P(Q(\mathbf{V}^*_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^*_{\mathbf{V}'=\mathbf{v}''})|\mathbf{V}^*)} > 0 \text{ if } \mathbb{A}^\dagger \cap \mathbb{L}^* \neq \emptyset \tag{5}$$

where $\mathbb{A}^\dagger$ is the set of tuples that induce the same effect as $(\mathbf{F}^\dagger,\mathbf{U}^\dagger)$, i.e. $\{(\mathbf{F},\mathbf{U}) \in \text{supp}(P(\mathbf{F},\mathbf{U})) : Q(\mathbf{V}_{\mathbf{V}'=\mathbf{v}'_1}, \mathbf{V}_{\mathbf{V}'=\mathbf{v}'_2}) = Q(\mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}'}, \mathbf{V}^\dagger_{\mathbf{V}'=\mathbf{v}''})\}$ and $\mathbb{L}^*$ is the set of tuples that maximize the likelihood of the data asymptotically, i.e. $\{(\mathbf{F},\mathbf{U}) \in \text{supp}(P(\mathbf{F},\mathbf{U})) : \frac{P(\mathbf{V}^*|\mathbf{F},\mathbf{U})}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} \to 1 \text{ as } n \to \infty\}$. □

Theorem 4.1. For any $\mathbf{V}^* \sim P(\mathbf{V}|\mathbf{F}^*,\mathbf{U}^*)$, $P(\mathbf{V}^*|\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1)$ and $P(\mathbf{V}^*|\hat{\mathbf{F}}^2,\hat{\mathbf{U}}^2)$ converge to $P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)$ as $n \to \infty$,

Proof. Without loss of generality, toward a contradiction assume that $P(\mathbf{V}^*|\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1) \not\to P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)$ as $n \to \infty$. Therefore, by Theorem 3.1 we have that $\frac{P(\mathbf{V}^*|\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} \to 0$ as $n \to \infty$. $\mathcal{L}(\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1,\hat{\mathbf{F}}^2,\hat{\mathbf{U}}^2,\mathbf{V}^*) \geq \mathcal{L}(\mathbf{F}^*,\mathbf{U}^*,\mathbf{F}^*,\mathbf{U}^*,\mathbf{V}^*)$ implies that $\log P(\mathbf{V}^*|\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1) + \log P(\mathbf{V}^*|\hat{\mathbf{F}}^2,\hat{\mathbf{U}}^2) + \lambda|\hat{Q}^1 - \hat{Q}^2| \geq 2\log P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*) + \lambda|Q^* - Q^*| = 2\log P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)$. Or equivalently, $0 \leq \log P(\mathbf{V}^*|\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1) + \log P(\mathbf{V}^*|\hat{\mathbf{F}}^2,\hat{\mathbf{U}}^2) + \lambda|\hat{Q}^1 - \hat{Q}^2| - 2\log P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*) = \log \frac{P(\mathbf{V}^*|\hat{\mathbf{F}}^1,\hat{\mathbf{U}}^1)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} + \log \frac{P(\mathbf{V}^*|\hat{\mathbf{F}}^2,\hat{\mathbf{U}}^2)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} + \lambda|\hat{Q}^1 - \hat{Q}^2| \to \log(0) + \log \frac{P(\mathbf{V}^*|\hat{\mathbf{F}}^2,\hat{\mathbf{U}}^2)}{P(\mathbf{V}^*|\mathbf{F}^*,\mathbf{U}^*)} + \lambda|\hat{Q}^1 - \hat{Q}^2| = -\infty$ as n goes to ∞, which is a contradiction. □

Theorem 4.2. $\Delta\hat{Q} \to \max_{(\mathbf{F}^1,\mathbf{U}^1,\mathbf{F}^2,\mathbf{U}^2)\in\mathbb{L}} \Delta Q$ as $n \to \infty$.

Proof. Toward a contradiction assume that there exists some $(\mathbf{F}^1, \mathbf{U}^1, \mathbf{F}^2, \mathbf{U}^2)$ such that $\mathcal{L}(\mathbf{F}^1, \mathbf{U}^1, \mathbf{F}^2, \mathbf{U}^2, \mathbf{V}^*) \leq \mathcal{L}(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1, \hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2, \mathbf{V}^*)$ and $|Q^1 - Q^2| > |\hat{Q}^1 - \hat{Q}^2|$. By Theorems 3.1 and 4.1, we have that $P(\mathbf{V}^*|\mathbf{F}^1, \mathbf{U}^1), P(\mathbf{V}^*|\mathbf{F}^2, \mathbf{U}^2), P(\mathbf{V}^*|\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1), P(\mathbf{V}^*|\hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2)$ all converge to $P(\mathbf{V}^*|\mathbf{F}^*, \mathbf{U}^*)$ as $n \to \infty$. Therefore, by definition of $\mathcal{L}$, we have that as $n \to \infty$, $2 \log P(\mathbf{V}^*|\mathbf{F}^*, \mathbf{U}^*) + |Q^1 - Q^2| \leq 2 \log P(\mathbf{V}^*|\mathbf{F}^*, \mathbf{U}^*) + |\hat{Q}^1 - \hat{Q}^2|$, which implies that $|Q^1 - Q^2| \leq |\hat{Q}^1 - \hat{Q}^2|$, which is a contradiction. □

Theorem 4.3 A causal estimand Q is identifiable given a prior $P(\mathbf{F}, \mathbf{U}, \boldsymbol{\epsilon})$ and a sample $\mathbf{F}^*, \mathbf{U}^*, \boldsymbol{\epsilon}^*$ from that prior if and only if $\Delta\hat{Q} \to 0$ as $n \to \infty$.

Proof. By Theorem 4.1 we have that $(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1)$ and $(\hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2)$ are in $\mathbb{L}^*$, i.e. the set of functions that maximize the log likelihood of the data asymptotically. Therefore, if $|\hat{Q}^1 - \hat{Q}^2| > 0$, then at least one of $(\hat{\mathbf{F}}^1, \hat{\mathbf{U}}^1)$ or $(\hat{\mathbf{F}}^2, \hat{\mathbf{U}}^2)$ are a $(\mathbf{F}^\dagger, \mathbf{U}^\dagger)$ that satisfy Theorem 3.2. By Theorem 4.2 we have that $|\hat{Q}^1 - \hat{Q}^2|$ maximizes the distance between induced causal effects. Therefore, if $|\hat{Q}^1 - \hat{Q}^2| \to 0$ as $n \to \infty$, no such $(\mathbf{F}^\dagger, \mathbf{U}^\dagger)$ exists. □

Theorem 4.4 If Algorithm 1 returns TRUE, $P(\mathbb{E}_{\mathbf{V}^*,\infty}\Delta\hat{Q} > t) < p$.

Proof. By the central limit theorem, we have that $P(\mathbb{E}_{\mathbf{V}^*,n}\Delta\hat{Q} > t) = \Phi\left(\frac{(\hat{\mu}_{\Delta Q} - t)\sqrt{n}}{\hat{\sigma}_{\Delta Q}}\right)$. Therefore, by Assumption 4.1 we have that $P(\mathbb{E}_{\mathbf{V}^*,\infty}\Delta\hat{Q} > t) < \Phi\left(\frac{(\hat{\mu}_D - t)\sqrt{n}}{\hat{\sigma}_D}\right)$. □

8.2. Confounded Gaussian Process Kernel Partial Derivatives

Here, we present the partial derivatives of the kernel matrices K, K_*, and K_{**} for the confounded Gaussian process example with respect to all parameters, latent confounders, and inducing counterfactuals, $s \in \theta \cup \mathbf{U} \cup \mathbf{y}$.

$$\frac{\partial K}{\partial s} = \frac{\partial K_T}{\partial s} \odot K_U + K_T \odot \frac{\partial K_U}{\partial s}$$

$$\frac{\partial K_*}{\partial s} = \begin{bmatrix} \frac{\partial K_{T,t'}}{\partial s} \\ \frac{\partial K_{T,t''}}{\partial s} \end{bmatrix} \odot \begin{bmatrix} K_U \\ K_U \end{bmatrix} + \begin{bmatrix} K_{T,t'} \\ K_{T,t''} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial K_U}{\partial s} \\ \frac{\partial K_U}{\partial s} \end{bmatrix}$$

$$\frac{\partial K_{**}}{\partial s} = \begin{bmatrix} \frac{\partial s_T}{\partial s}\mathbf{1} & \frac{\partial k_{t',t''}}{\partial s}\mathbf{1} \\ \frac{\partial k_{t',t''}}{\partial s}\mathbf{1} & \frac{\partial s_T}{\partial s}\mathbf{1} \end{bmatrix} \odot \begin{bmatrix} K_U & K_U \\ K_U & K_U \end{bmatrix} + \begin{bmatrix} s_T\mathbf{1} & k_{t',t''}\mathbf{1} \\ k_{t',t''}\mathbf{1} & s_T\mathbf{1} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial K_U}{\partial s} & \frac{\partial K_U}{\partial s} \\ \frac{\partial K_U}{\partial s} & \frac{\partial K_U}{\partial s} \end{bmatrix}$$

$$\frac{\partial K_{T,i,j}}{\partial s} = \begin{cases} \frac{(T_i - T_j)^2}{l_T} K_{T,i,j} & s = l_T \\ \frac{1}{s} K_{T,i,j} & s = s_T \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial K_{T,t',i,j}}{\partial s} = \begin{cases} \frac{(T_i - t')^2}{l_T} K_{T,t',i,j} & s = l_T \\ \frac{1}{s} K_{T,t',i,j} & s = s_T \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial k_{t',t''}}{\partial s} = \begin{cases} \frac{(t' - t'')^2}{l_T} k_{t',t''} & s = l_T \\ \frac{1}{s} k_{t',t''} & s = s_T \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{\partial K_{U,i,j}}{\partial s} = \begin{cases} \frac{(U_i - U_j)^2}{l_U} K_{U,i,j} & s = l_U \\ \frac{1}{s} K_{U,i,j} & s = s_U \\ \text{-2}(U_i - U_j) K_{U,i,j} & s = U_i \\ 2(U_i - U_j) K_{U,i,j} & s = U_j \\ 0 & \text{otherwise} \end{cases}$$

8.3. Structural Causal Models

Here, we present a mathematical description for the structural causal models underlying the unconfounded regression and the backdoor adjusted designs which is agnostic to any particular choice of functions, and that we expand on for particular choices of functions in Section 8.4 of this supplementary materials. The remaining five designs in Table 1 are presented throughout the main body of the paper.

Unconfounded Regression. The unconfounded regression design is identical to the confounded regression design, except that the latent confounded has been omitted. Specifically, we have that $T_i = f_T(\epsilon_{T_i})$ and $Y_i = f_Y(T_i, \epsilon_{Y_i})$.

Backdoor Adjusted Design. The backdoor adjusted design includes an observed confounder, X, that influence both treatment, T, and outcome Y, but no latent confounders. Specifically, we have that $X_i = f_X(\epsilon_{X_i})$, $T_i = f_T(X_i, \epsilon_{T_i})$, and $Y_i = f_Y(T_i, X_i, \epsilon_{Y_i})$.

8.4. Experiments

In this section we provide additional detail for the linear, quadratic polynomial, and Gaussian process experiments, and results for the quadratic polynomial experiments. For each of the seven designs we assume that treatment, T, outcome, Y, and where applicable covariates, X, and instuments, I, are observed. All other random variables are latent.

8.4.1. Linear Structural Causal Models

For all of the linear parametric experiments, we assume that each function $V_i = f_V(Pa(V)_i, \epsilon_{V_i}) = \beta_V \cdot Pa(V)_i + \epsilon_{V_i}$, where each element of β_V is drawn from an independent Gaussian prior. For example, for the linear confounded model we have that $T_i = \beta_T \cdot U_i + \epsilon_{T_i}$ and $Y_i = \beta_Y \cdot [T_i, U_i] + \epsilon_{Y_i}$, where β_T and β_Y are one dimensional and two dimensional vectors respectively. All exogenous noise variables are sampled from mean-zero independent Gaussian distributions.

Here, we provide the full prior over linear structural causal models for each of the seven designs in Table 1.

Unconfounded Regression.

$$\beta_Y \sim \mathcal{N}(1, 0.3) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$T_i = \epsilon_{T_i} \qquad Y_i = \beta_Y \cdot T_i + \epsilon_{Y_i}$$

Confounded Regression.

$$\beta_T \sim \mathcal{N}(.5, 0.3) \quad \beta_Y \sim \mathcal{N}([1, .5]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad U_i \overset{iid}{\sim} \mathcal{N}(0, 0.3)$$
$$\epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \qquad T_i = \beta_T \cdot U_i + \epsilon_{T_i} \qquad Y_i = \beta_Y \cdot [T_i, U_i] + \epsilon_{Y_i}$$

Backdoor Adjusted.

$$\beta_T \sim \mathcal{N}(.5, 0.3) \quad \beta_Y \sim \mathcal{N}([1, .5]^t, 0.3\mathbf{I}) \quad \log(\sigma_X^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3)$$
$$\epsilon_{X_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2) \qquad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \qquad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$X_i = \epsilon_{X_i} \qquad T_i = \beta_T \cdot X_i + \epsilon_{T_i} \qquad Y_i = \beta_Y \cdot [T_i, X_i] + \epsilon_{Y_i}$$

Frontdoor Adjusted.

$$\beta_T \sim \mathcal{N}(.5, 0.3) \quad \beta_X \sim \mathcal{N}(1, 0.3) \quad \beta_Y \sim \mathcal{N}([1, .5]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}2, 0.3) \quad \log(\sigma_X^2) \sim \mathcal{N}(\text{-}0, 0.3)$$
$$\log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{X_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \quad U_i \overset{iid}{\sim} \mathcal{N}(0, 0.3)$$
$$T_i = \beta_T \cdot U_i + \epsilon_{T_i} \qquad X_i = \beta_X \cdot T_i + \epsilon_{X_i} \qquad Y_i = \beta_Y \cdot [X_i, U_i] + \epsilon_{Y_i}$$

Instrumental Variable. Note that here I_i refers to the i'th instance of the instrumental random variable I, and $\mathbf{I}$ refers to the identity matrix.

$$\beta_T \sim \mathcal{N}([2,.5]^t, 0.3\mathbf{I}) \quad \beta_Y \sim \mathcal{N}([1,.5]^t, 0.3\mathbf{I}) \quad \log(\sigma_I^2) \sim \mathcal{N}(0, 0.3) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3)$$
$$\log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{I_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_I^2) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$U_i \overset{iid}{\sim} \mathcal{N}(0, 0.3) \quad I_i = \epsilon_{I_i} \quad T_i = \beta_T \cdot [I_i, U_i] + \epsilon_{T_i} \quad Y_i = \beta_Y \cdot [T_i, U_i] + \epsilon_{Y_i}$$

Within Subjects. Here, $U_{o(i)}$ refers to the shared value of the latent confounder, U_o, associated with instance i. For these and all other experiments, we assume that each object instance, o, is shared between 25 instances of treatment and outcome.

$$\beta_T \sim \mathcal{N}(.5, 0.3) \quad \beta_Y \sim \mathcal{N}([1,.5]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad U_o \overset{iid}{\sim} \mathcal{N}(0, 0.3)$$
$$\epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \quad T_i = \beta_T \cdot U_{o(i)} + \epsilon_{T_i} \quad Y_i = \beta_Y \cdot [T_i, U_{o(i)}] + \epsilon_{Y_i}$$

Regression Discontinuity Design. Here, $\mathbf{1}[X_i > 0]$ refers to the indicator function that returns 1 if $X_i > 0$ and 0 otherwise.

$$\beta_Y \sim \mathcal{N}([1,.5]^t, 0.3\mathbf{I}) \quad \log(\sigma_X^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{X_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$X_i = \epsilon_{X_i} \quad T_i = \mathbf{1}[X_i > 0] \quad Y_i = \beta_Y \cdot [T_i, X_i] + \epsilon_{Y_i}$$

8.4.2. Quadratic Polynomial Structural Causal Models

For all of the quadratic polynomial experiments, we assume that each function $V_i = f_V(Pa(V)_i, \epsilon_{V_i}) = \beta \cdot \phi(Pa(V)_i) + \epsilon_{V_i}$, where $\phi(Pa(V)_i)$ is a quadratic basis expansion of the vector of all latent and observed arguments to f_V. For example, for the function $f_Y(T_i, U_i, \epsilon_{Y_i})$, $\phi(Pa(V)_i) = [T_i, T_i^2, U_i, U_i^2, T_i \cdot U_i]$. Again, each element of β is drawn from an independent Gaussian prior.

Here, we provide the full prior over quadratic polynomial structural causal models for each of the seven designs in Table 1.

Unconfounded Regression.

$$\beta_Y \sim \mathcal{N}([1,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$T_i = \epsilon_{T_i} \quad Y_i = \beta_Y \cdot [T_i, T_i^2] + \epsilon_{Y_i}$$

Confounded Regression.

$$\beta_T \sim \mathcal{N}([1,0]^t, 0.3\mathbf{I}) \quad \beta_Y \sim \mathcal{N}([1,0,.5,0,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3)$$
$$U_i \overset{iid}{\sim} \mathcal{N}(0, 0.3) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \quad T_i = \beta_T \cdot [U_i, U_i^2] + \epsilon_{T_i}$$
$$Y_i = \beta_Y \cdot [T_i, T_i^2, U_i, U_i^2, T_i \cdot U_i] + \epsilon_{Y_i}$$

Backdoor Adjusted.

$$\beta_T \sim \mathcal{N}([1,0]^t, 0.3\mathbf{I}) \quad \beta_Y \sim \mathcal{N}([1,0,.5,0,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_X^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3)$$
$$\log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{X_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$X_i = \epsilon_{X_i} \quad T_i = \beta_T \cdot [X_i, X_i^2] + \epsilon_{T_i} \quad Y_i = \beta_Y \cdot [T_i, T_i^2, X_i, X_i^2, T_i \cdot X_i] + \epsilon_{Y_i}$$

Frontdoor Adjusted.

$$\beta_T \sim \mathcal{N}([1,0]^t, 0.3\mathbf{I}) \quad \beta_X \sim \mathcal{N}([1,0]^t, 0.3\mathbf{I}) \quad \beta_Y \sim \mathcal{N}([1,0,.5,0,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}2, 0.3)$$
$$\log(\sigma_X^2) \sim \mathcal{N}(\text{-}0, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{X_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2)$$
$$\epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \quad U_i \overset{iid}{\sim} \mathcal{N}(0, 0.3) \quad T_i = \beta_T \cdot [U_i, U_i^2] + \epsilon_{T_i} \quad X_i = \beta_X \cdot [T_i, T_i^2] + \epsilon_{X_i}$$
$$Y_i = \beta_Y \cdot [X_i, X_i^2, U_i, U_i^2, X_i \cdot U_i] + \epsilon_{Y_i}$$

Instrumental Variable. Again, I_i refers to the i'th instance of the instrumental random variable I, and $\mathbf{I}$ refers to the identity matrix.

$$\beta_T \sim \mathcal{N}([1,0,.5,0,0]^t, 0.3\mathbf{I}) \quad \beta_Y \sim \mathcal{N}([1,0,.5,0,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_I^2) \sim \mathcal{N}(0, 0.3) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3)$$
$$\log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{I_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_I^2) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2) \quad \epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2)$$
$$U_i \overset{iid}{\sim} \mathcal{N}(0, 0.3) \quad I_i = \epsilon_{I_i}$$
$$T_i = \beta_T \cdot [I_i, I_i^2, U_i, U_i^2, I_i \cdot U_i] + \epsilon_{T_i} \quad Y_i = \beta_Y \cdot [T_i, T_i^2, U_i, U_i^2, T_i \cdot U_i] + \epsilon_{Y_i}$$

Within Subjects. Here, $U_{o(i)}$ refers to the shared value of the latent confounder, U_o, associated with instance i.

$$\beta_T \sim \mathcal{N}([1,0]^t, 0.3\mathbf{I}) \quad \beta_Y \sim \mathcal{N}([1,0,.5,0,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_T^2) \sim \mathcal{N}(\text{-}1, 0.3)$$
$$\log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad U_o \overset{iid}{\sim} \mathcal{N}(0, 0.3) \quad \epsilon_{T_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_T^2)$$
$$\epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \quad T_i = \beta_T \cdot [U_{o(i)}, U_{o(i)}^2] + \epsilon_{T_i} \quad Y_i = \beta_Y \cdot [T_i, T_i^2, U_{o(i)}, U_{o(i)}^2, T_i \cdot U_{o(i)}] + \epsilon_{Y_i}$$

Regression Discontinuity Design. Again, $\mathbf{1}[X_i > 0]$ refers to the indicator function that returns 1 if $X_i > 0$ and 0 otherwise.

$$\beta_Y \sim \mathcal{N}([1,.5,0,0]^t, 0.3\mathbf{I}) \quad \log(\sigma_X^2) \sim \mathcal{N}(\text{-}1, 0.3) \quad \log(\sigma_Y^2) \sim \mathcal{N}(\text{-}3, 0.3) \quad \epsilon_{X_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_X^2)$$
$$\epsilon_{Y_i} \overset{iid}{\sim} \mathcal{N}(0, \sigma_Y^2) \quad X_i = \epsilon_{X_i} \quad T_i = \mathbf{1}[X_i > 0] \quad Y_i = \beta_Y \cdot [T_i, X_i, X_i^2, T_i \cdot X_i] + \epsilon_{Y_i}$$

Quadratic Polynomial Results We present the results of the quadratic polynomial experiments in Table 3, which use the exact same configurations as the linear experiments described in Section 5 of the main body. Again, SBI significantly outperforms the baseline.

Design	Quadratic Polynomial		
	$\Delta\hat{Q}_{\text{SBI}}$	$\Delta\hat{Q}_{\text{Baseline}}$	Ground Truth
Unconfounded	$\mathbf{.00 \pm .00}$	$.07 \pm .00$	ID
Confounded	$\mathbf{.95 \pm .15}$	$\mathbf{.2 \pm .026}$	Not ID
Backdoor	$\mathbf{.00 \pm .00}$	$.07 \pm .00$	ID
Frontdoor	$.10 \pm .01$	$.13 \pm .01$	ID
Inst. Var.	$\mathbf{.15 \pm .02}$	$\mathbf{.05 \pm .00}$	Not ID
Within Subj.	$\mathbf{.02 \pm .01}$	$.07 \pm .00$	ID
Regr. Disc.	$\mathbf{.00 \pm .00}$	$\mathbf{.04 \pm .00}$	ID

Table 3: **Simulation-based identifiability accurately determines identifiability for diverse causal designs.** SBI outperforms the baseline, resulting in near-perfect identification results for the average treatment effect. Values are shown in bold if SBI or the baseline correctly determine identifiability using a threshold of 5 percent of the true effect.

8.4.3. Gaussian Process Structural Causal Models

For each of the experiments using Gaussian process priors over structural causal models we use the same prior over linear structural causal models for all functions except the outcome function f_Y, which is drawn from a Gaussian process prior with a radial basis function kernel. In this section we describe the priors over kernel hyperparameters for each of the seven designs.

In our experiments we approximated the gradient of the average treatment effect using a single instance for both counterfactual outcomes. We found that this approximation produced more stable gradient estimates, resulting in faster and more consistent convergence of the two model particles. Note that this downsampling does not introduce bias, as the instance is selected at random.

Unconfounded Regression.

$$\log(l_T) \sim \mathcal{N}(0, 0.3) \qquad \log(s_T) \sim \mathcal{N}(0, 0.3)$$

Confounded Regression.

$$\log(l_T) \sim \mathcal{N}(0, 0.3) \qquad \log(s_T) \sim \mathcal{N}(0, 0.3) \qquad \log(l_U) \sim \mathcal{N}(0, 0.3) \qquad \log(s_U) \sim \mathcal{N}(0, 0.3)$$

Instrumental Variable.

$$\log(l_T) \sim \mathcal{N}(0, 0.3) \qquad \log(s_T) \sim \mathcal{N}(0, 0.3) \qquad \log(l_U) \sim \mathcal{N}(0, 0.3) \qquad \log(s_U) \sim \mathcal{N}(0, 0.3)$$

Within Subjects. Note that the kernel function $k(U_o(i), U_o(j), l_U, s_U) = 1$ by definition when $o(j) = o(i)$. This represents the fact that confounders are shared across instances belonging to the same object. This Gaussian process model is a simplified version of recent work using Gaussian process priors for hierarchical causal inference (Witty et al., 2020).

$$\log(l_T) \sim \mathcal{N}(0, 0.3) \qquad \log(s_T) \sim \mathcal{N}(0, 0.3) \qquad \log(l_U) \sim \mathcal{N}(0, 0.3) \qquad \log(s_U) \sim \mathcal{N}(0, 0.3)$$

Regression Discontinuity Design.

$$\log(l_T) \sim \mathcal{N}(0, 0.3) \qquad \log(s_T) \sim \mathcal{N}(0, 0.3) \qquad \log(l_X) \sim \mathcal{N}(0, 0.3) \qquad \log(s_X) \sim \mathcal{N}(0, 0.3)$$

8.5. Baseline

For our baseline identification method, we used an approach based on profile likelihood identification (Raue et al., 2009). For each model the baseline is identical to the SBI in all respects, except that it uses only a single particle with no repulsion term. Instead, to traverse the likelihood surface the baseline first performs 500 steps of the Adam optimization method using the gradient of the log-likelihood to find a single maximum likelihood solution. Then, for each parameter $s \in \theta$, we increment the parameter by a small amount $s \leftarrow s + \Delta s$ and then again run the Adam optimization method using the gradient of the log-likelihood with respect to all parameters except for s for 100 steps. In our experiments we use $\Delta s = 0.01$. We report the range over estimated causal effects after repeating this procedure 100 times for all $s \in \theta$. Intuitively, if the likelihood surface is on a *ridge* of equivalent maximum likelihood models then alternating between perturbations and optimization will find other locations on that maximum likelihood surface. We discuss limitations of this kind of approach in Section 1, and show empirically that SBI outperforms it in Section 5.