The Hardness of Validating Observational Studies with Experimental Data

Jake Fawkes[*]
University of Oxford

Michael O'Riordan
Spotify

Athanasios Vlontzos
Spotify &
Imperial College London

Oriol Corcoll
Spotify

Ciarán M. Gilligan-Lee
Spotify &
University College London

Abstract

Observational data is often readily available in large quantities, but can lead to biased causal effect estimates due to the presence of unobserved confounding. Recent works attempt to remove this bias by supplementing observational data with experimental data, which, when available, is typically on a smaller scale due to the time and cost involved in running a randomised controlled trial. In this work, we prove a theorem that places fundamental limits on this "best of both worlds" approach. Using the framework of impossible inference, we show that although it is possible to use experimental data to *falsify* causal effect estimates from observational data, in general it is not possible to *validate* such estimates. Our theorem proves that while experimental data can be used to detect bias in observational studies, without additional assumptions on the smoothness of the correction function, it can not be used to remove it. We provide a practical example of such an assumption, developing a novel Gaussian Process approach to construct intervals which contain the true treatment effect with high probability, both inside and outside of the support of the experimental data. We demonstrate our methodology on both simulated and semi-synthetic datasets and make the code available.

1 INTRODUCTION

It is often said that randomised controlled trials (RCTs) are the gold standard for establishing causal relationships (Hariton and Locascio, 2018; Gilligan-Lee et al., 2022), and estimating treatment effects (Aronow et al., 2021; Gilligan-Lee, 2020). However, in many cases, it is prohibitively costly, slow, or even unethical to run experiments that are large enough to accurately estimate such effects. Meanwhile, observational data is abundant, being significantly easier and cheaper to obtain. Unfortunately, such data is often subject to unmeasured confounding. This leaves treatment effects unidentifiable, and naively attempting to use the observational data despite this will lead to biased estimates of causal effects.

As a response to these problems, there has been substantial recent research effort to develop methodology for combining experimental and observational data sources, aiming to get the strengths of each (Colnet et al., 2024; Lin and Evans, 2023; Van Goffrier et al., 2023; Jeunen et al., 2022). These approaches aim to use the experimental data to de-bias the observational data (Yang et al., 2020; Kallus et al., 2018), falsify observational studies Hussain et al. (2022, 2023), or benchmark the level of unmeasured confounding (De Bartolomeis et al., 2024a,b).

In this work, we prove a selection of fundamental limitations of this approach. We use the framework of impossible inference (Bahadur and Savage, 1956; Bertanha and Moreira, 2020), a popular tool in econometrics (Canay et al., 2013). This field studies when it is possible for a non-trivial hypothesis tests to exist for a problem, where trivial hypothesis tests are those which are unable to distinguish *any* alternative from the null. We apply this to show that whilst non-trivial tests exist to *falsify* estimates from observational studies, we cannot *validate* heterogeneous treatment effects

[*] Work done during internship at Spotify. Correspondence to: jake.fawkes@st-hughs.ox.ac.uk.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

estimates using experimental data without additional assumptions. In terms of benchmarking confounding with a causal sensitivity model, our result corresponds to the statement that it is possible to form valid *lower bounds* of the sensitivity parameter but that it is impossible to form non-trivial upper bounds—again, absent additional assumptions.

Our hardness proof relies on the lack of smoothness in the unknown correction function. Therefore, as an example of an assumption that does permit this type of inference, we take the corrective function to be a sample from a Gaussian Process, a probabilistic function family with inherent smoothness guarantees. Developing this into a workable and practical methodology leads us to create a novel Gaussian Process based approach to learning from pseudo-outcomes (Kennedy, 2023), which correctly accounts for the unwieldy error distribution of pseudo-outcomes. We experimentally evaluate our approach against other Gaussian Process effect estimation approaches, showing strong improvement in predictive performance and uncertainty calibration. In short, to summarise our contributions:

- A proof of the limits of current approaches that use a learned corrective term to reconcile experimental and observational data.
- A demonstration that the smoothness properties of Gaussian Processes circumvent the assumptions required for the above proof and provide guarantees that Gaussian Processes give intervals which contain the treatment effect over the whole observational support with high probability.
- A novel Gaussian Process method to learn from inverse propensity weighted pseudo-outcomes—which may be of independent interest.
- Extensive experimentation validating the aforementioned novel method on synthetic and semi-synthetic data.

2 BACKGROUND AND NOTATION

2.1 Notation

We let the random variables X, T, and Y represent the covariates, treatment, and outcomes, with domains $\mathcal{X}$, $\{0, 1\}$, and $\mathbb{R}$ respectively, and use $\mathbf{x}$, t, and y to denote realisations of the variables. We suppose we have two datasets, the observational $\mathcal{D}_e = \{\mathbf{x}_i^o, t_i^o, y_i^o\}_{i=1}^{n_o}$ and the experimental $\mathcal{D}_o = \{\mathbf{x}_i^e, t_i^e, y_i^e\}_{i=1}^{n_e}$ which are drawn from distributions $P_e(\mathbf{x}, t, y)$ and $P_o(\mathbf{x}, t, y)$ respectively, with $\mathcal{D} = \mathcal{D}_o \cup \mathcal{D}_e$ denoting the full dataset of size $n = n_o + n_e$. We will use a variable E to denote if we are in the observational or experimental regime, so that, for example, $P_e(\mathbf{x}, t, y) = P(\mathbf{x}, t, y \mid E = e)$. Letting $\star \in \{o, e\}$ we use $\mathcal{X}^\star \subset \mathcal{X}$ for the support of $P_\star(\mathbf{x})$ and assume that $\mathcal{X}^e \subset \mathcal{X}^o$. Vectors of observations and treatment in a dataset, $\mathcal{D}_\star$, are denoted by $\mathbf{y}_\star$ and $\mathbf{t}_\star$ respectively, while $\mathbf{X}_\star$ refers to the data matrix, so that $\mathbf{y}_\star = (y_i)_{i=1}^{n_e}$, $\mathbf{t}_\star = (t_i)_{i=1}^{n_e}$, and $\mathbf{X}_\star = (\mathbf{x}_i)_{i=1}^{n_e}$. We let $\omega_\star(\mathbf{x})$ be defined as:

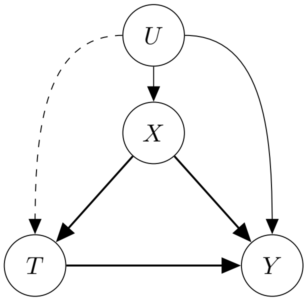


Figure 1: Causal Structure for generating the experimental and observational datasets. Dashed edges are only present in the observational dataset, whilst all others are present and fixed across both datasets.

$$\omega_\star(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}, T = 1, E = \star] \tag{1}$$
$$- \mathbb{E}[Y|X = \mathbf{x}, T = 0, E = \star], \tag{2}$$

be the difference between expected conditional outcomes for each treatment conditional on $E = \star$ and $X = \mathbf{x}$. We use potential outcomes (Rubin, 1974), so that $Y(t)$ represents the outcome from setting $T = t$[*].

2.2 Objectives and Assumptions

Throughout we focus on estimating the *Conditional Average Treatment Effect* (CATE) for the observational datasets, which is given by:

$$\tau(\mathbf{x}) := \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}, E = o\right]. \tag{3}$$

We assume the datasets are generated according to the causal structure in Figure 1, where dashed edges are present only in the observational dataset. Importantly, this implies $Y(t) \perp E \mid X = \mathbf{x}$, which ensures that the CATE is fixed across environments as:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}, E = o\right] \tag{4}$$
$$= \mathbb{E}\left[Y(1) - Y(0) | X = \mathbf{x}, E = e\right]. \tag{5}$$

We demonstrate this in Appendix A.1.

For the experimental study we assume that treatment is randomised according to a known propensity score

[*] We make use of the SWIG framework to combine causal graphical models with potential outcomes. More details can be found in Richardson and Robins (2013).

$\pi(\mathbf{x}) = P_e(T = 1 \mid X = x)$ which we assume to satisfy *strict overlap* (D'Amour et al., 2021). That is we assume that there exists a $\delta > 0$ such that:

$$\delta < \pi(\mathbf{x}) < 1 - \delta \text{ for all } \mathbf{x} \in \mathcal{X}^e. \tag{6}$$

Under the additional assumption of consistency ($Y(t) = Y$ when $T = t$) we have that the CATE is *identified* (Pearl, 2009; Richardson and Robins, 2013) within the support of the experimental dataset, and given by:

$$\tau(\mathbf{x}) = \omega_e(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{X}^e. \tag{7}$$

This is not the case in the observational dataset, where the hidden confounding induced by U means that $\tau(\mathbf{x}) \neq \omega_o(\mathbf{x})$ in general. We use $\Delta(\mathbf{x})$ to denote this gap as:

$$\Delta(\mathbf{x}) = \tau(\mathbf{x}) - \omega_o(\mathbf{x}). \tag{8}$$

So that if $\Delta(\mathbf{x}) = 0$, an unbiased estimate of $\omega_o(\mathbf{x})$ from the observational study is an unbiased estimate of the true CATE. Throughout, we will assume a model $\hat{\omega}_o(\mathbf{x})$ has been fit for $\omega_o(\mathbf{x})$ from the observational sample and let $\hat{\Delta}(\mathbf{x}) \coloneqq \tau(\mathbf{x}) - \hat{\omega_o}(\mathbf{x})$.

The idea is that $\Delta(\mathbf{x})$ should be simpler than the true CATE function, $\tau(\mathbf{x})$. Under such circumstances, it should be more efficient to use to the experimental dataset to estimate or bound $\hat{\Delta}(\mathbf{x})$ and combine it with $\hat{\omega_o}(\mathbf{x})$ than use the small experimental dataset to learn the CATE directly(Yang et al., 2020). Ideally, we would want these to be extendable from the support of the experimental distribution to the support of the observational distribution (Kallus et al., 2018), potentially by incorporating bounds on the correction function as opposed to point estimates. This would give us expression for CATE over all of $\mathcal{X}^o$.

Finally, we introduce the IPW pseudo-outcome (Kennedy, 2023; Curth and Van der Schaar, 2021), which throughout we will only refer to relative to the experimental distribution, as follows:

Definition 2.1 (IPW Pseudo-Outcome)**.** The IPW Pseudo-Outcome is given by:

$$\tilde{Y} \coloneqq \left(\frac{T - \pi(X)}{\pi(X)\,(1 - \pi(X))} \right) Y \tag{9}$$

Where $\pi(X) = P(T = 1 \mid X, E = e)$. $\tilde{Y}$ has the property that $\mathbb{E}[\tilde{Y} | X = \mathbf{x}, E = e] = \tau(\mathbf{x})$.

We assume that $\pi(X)$ is known, which is common for a well conducted randomised controlled trial.

2.3 Related work bounding $\Delta(\mathbf{x})$

Our setting is first considered in Kallus et al. (2018), where they assume that $\Delta(\mathbf{x})$ is linear in order to allow for extrapolation from the experimental sample. By taking $\Delta(\mathbf{x}) = \beta_0^\top \mathbf{x}$ and assuming β_0 is identifiable from experimental data, they can obtain an estimate $\Delta(\mathbf{x})$ that generalises beyond the experimental sample by performing a linear regression of $\{\mathbf{x}_i^e\}_{i=1}^{n_e}$ onto $\{\tilde{y}_i^e - \hat{\omega}_o(\mathbf{x})\}_{i=1}^{n_e}$. Kallus et al. (2018) prove that as the number of experimental and observational samples tend to infinity this will converge to the true CATE at a faster rate than using the experimental sample alone. This has been extended to the semiparametric (Yang et al., 2020) and nonparametric case (Wu and Yang, 2022), however extrapolation still requires the function to be uniquely identified from the experimental study.

A strongly related area of work aims to use data from the experimental study to test causal effect estimates from observational studies. One approach aims to falsify causal estimates from observational studies using experimental data (Hussain et al., 2023, 2022). This work converts a variety of assumptions regarding the validity of the observational study, consistency of the CATE across studies and external validity of the RCT into testable statistical hypothesis, which can then be falsified. Another approach aims to estimate how much unmeasured confounding must be present in an observational study for it to be consistent with the RCT (De Bartolomeis et al., 2024a,b). This is achieved using causal sensitivity models (Rosenbaum and Rubin, 1983), which uses a single parameter to control the strength of unmeasured confounding. The goal is then to use the RCT to lower bound this parameter. A significant portion of work in both these areas utilises non-parametric tests of conditional moment restrictions (Muandet et al., 2020).

3 THE HARDNESS OF VALIDATING OBSERVATIONAL STUDIES

We now provide some theoretical limits on using experimental data to measure the level of unmeasured confounding in an observational study. We do this using the framework of impossible inference (Bertanha and Moreira, 2020), a popular tool in econometrics (Canay et al., 2013). This field studies when it is possible for a non-trivial hypothesis tests to exist for a problem, where trivial hypothesis tests are those which are unable to distinguish *any* alternative from the null.

Impossible inference has been applied within causal inference to show the hardness of conditional independence testing (Shah and Peters, 2020). In our case, the conditional independence $Y \perp E \mid X, T$ would imply a total lack of unmeasured confounding as:

$$\begin{aligned} P_o(Y \mid X = \mathbf{x}, T = t) &= P_e(Y \mid X = \mathbf{x}, T = t, E = e) \\ &= P(Y(t) \mid X = \mathbf{x}, E = e). \end{aligned}$$

Therefore, the hardness of this testing problem already demonstrates that there are no non-trivial tests for full unconfoundedness in the observational distribution.

However, this still doesn't preclude us from being able to estimate CATE in an unbiased manner from the observational dataset. Or, failing full estimation, bounding the correction function, $\hat{\Delta}(\cdot)$ to return intervals for CATE. Therefore, in this section, we apply the same techniques to focus on what experimental data allows us to test for in term unbiasedness of CATE estimates. This translates to estimating or bounding $\hat{\Delta}(\mathbf{x})$ with confidence guarantees, which we do via the following definition:

Definition 3.1. We say that that $\hat{\Delta}(\cdot)$ is **controlled** by functions $\bar{f}, \underline{f} : \mathcal{X} \to \mathbb{R}$ if we have:

$$\hat{\Delta}(\mathbf{x}) \in \left[\underline{f}(\mathbf{x}), \bar{f}(\mathbf{x})\right] \text{ a.s for } x \sim P_o(\mathbf{x}) \tag{10}$$

Where $\underline{f}(\mathbf{x}) \leq \bar{f}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

If the unmeasured confounding were controlled by $\underline{f}(\mathbf{x}) = \bar{f}(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$ then we would have that the CATE estimate from the observational study is unbiased to the true CATE. Going further than this, if we knew functions that controlled the confounding, then we could use them alongside the CATE estimate from the observational study to give intervals which contain the true CATE. Therefore, the goal is to understand how we can use the observational study to learn the functions, $\bar{f}, \underline{f}$, that control the confounding.

3.1 Testing Notation and Background

First, we will assume a fixed propensity score function $\pi : \mathcal{X} \to [0, 1]$. Now we define $\mathcal{E}_{M,\pi}$ to be the set of distributions over (X, T, Y) which are absolutely continuous in X with respect to Lebesgue measure, bounded above in ℓ_∞ norm by M, and whose propensity score is given by π. As before, the domains of T, Y, E are $\{0, 1\}, \mathbb{R}$ and $\{o, e\}$ respectively. For this section, we will use the notation $\mathbb{P}_P$ to denote a probability taken with respect to $P \in \mathcal{E}_{M,\pi}$.

A potentially randomised test ψ_n is a measurable function which takes in a dataset, $\mathcal{D}$, a random variable $U \sim U[0, 1]$ that represents the randomness of the test—a choice of permutations in permutation testing, for example—and outputs a result in $\{0, 1\}$ where 1 corresponds to a rejection. So we write $\psi_n(\mathcal{D}, U)$ for the result of the test ψ_n with dataset $\mathcal{D}$ and U as input.

Suppose we observe an experimental dataset $\mathcal{D}$ sampled i.i.d from a distribution $P_0 \in \mathcal{E}_{M,\pi}$ and wish to test the null hypothesis $H_0 : P_0 \in \mathcal{N} \subset \mathcal{E}_{M,\pi}$ against the alternative hypothesis $H_1 : P_0 \in \mathcal{A} \subset \mathcal{E}_{M,\pi}$. Then we have the following important definitions:

Definition 3.2. Let ψ_n be a randomised test which takes in a dataset $\mathcal{D}$ of size n. We say ψ_n has **level** α at size n if we have $\sup_{P \in \mathcal{N}} \mathbb{P}_{\mathcal{D} \sim P^n} (\psi_n(\mathcal{D}, U) = 1) \leq \alpha$. For an alternative distribution $P \in \mathcal{A}$, we define $\mathbb{P}_{\mathcal{D} \sim P^n} (\psi_n(\mathcal{D}, U) = 1)$ as the **point-wise power** against P at size n.

Ideally, we would want a test to have power against as many alternatives as possible, preferably uniformly so that $\inf_{P \in \mathcal{A}} \mathbb{P}_{\mathcal{D} \sim P^n} (\psi_n(\mathcal{D}, U) = 1) \to 1$. For non-parametric hypothesis, restrictions on the alternative such as smoothness conditions are often required to achieve this (Balakrishnan and Wasserman, 2019).

A particular problem given by sets $(\mathcal{N}, \mathcal{A})$ is known as **untestable** if for all tests the point-wise power is bounded by the level for any alternative. In this case, there exists no test that can distinguish the null from *any* alternative, therefore the null has to be restricted for there to be an informative test.

3.2 Setting the Testing Problem for Unmeasured Confounding

We now apply the above to testing for bias in treatment effect estimation due to unmeasured confounding. Fixing the observational estimate $\hat{\omega}(\cdot)$ and so the correction function $\hat{\Delta}(\cdot)$, we define the following sets of distributions:

Definition 3.3. Let $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$. We define:

$$\mathcal{P}_{M,\pi}(\underline{f}, \bar{f}) = \left\{ P \in \mathcal{E}_{M,\pi} : \ \hat{\Delta}(\cdot) \text{ controlled by } \underline{f}, \bar{f} \ \right\}$$
$$\mathcal{Q}_{M,\pi}(\underline{f}, \bar{f}) = \mathcal{E}_{M,\pi} \setminus \mathcal{P}_{M,\pi}$$

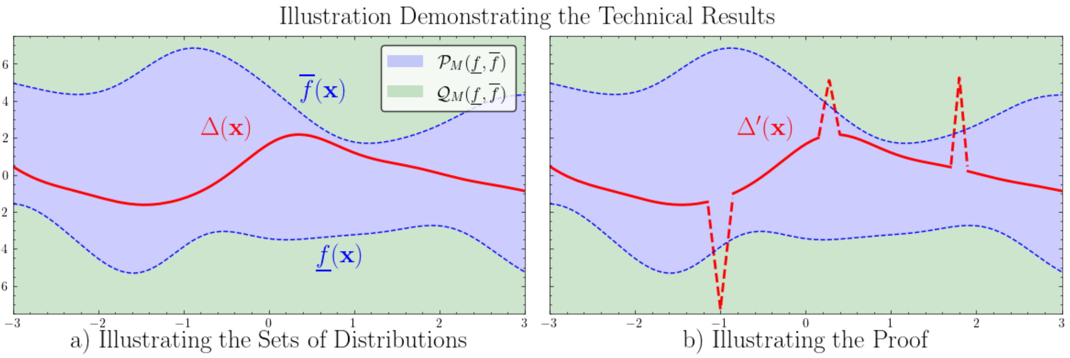
Which way round to test? Now with both sets of distributions , the question is what to take as the null and alternative? We have the following choices:

$$\text{Test 1:} \Big\{ H_0 : P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f}), \quad H_1 : P \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f}),$$

$$\text{Test 2:} \Big\{ H_0 : P \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f}), \quad H_1 : P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f}),$$

Rejecting under test 1 is more standard, and would correspond to *falsifying the hypothesis* that the observational study has a level of confounding controlled by $\underline{f}, \bar{f}$. This would mean finding statistical evidence that bias in the observational CATE estimate is not contained in the intervals given by $(\underline{f}, \bar{f})$.

However, failing to reject under test 1 *does not provide evidence* that the confounding is controlled by $(\underline{f}, \bar{f})$. We may fail to reject because of other reasons, such as a lack of data or the test having limited power against the true distribution. In an ideal world, we would like to reject the hypothesis that confounding is above a certain level. Fixing this level using some $\underline{f}, \bar{f}$, this



a) Illustrating the Sets of Distributions

b) Illustrating the Proof

Figure 2: Illustration of both the sets of distributions and proof of the technical result in Section 3. The first figure demonstrates the sets of distributions $\mathcal{P}_{M,\pi}(\underline{f},\bar{f}), \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$. $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ is the set of distributions where $\Delta(\cdot)$ is always contained in the blue region, and $\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ is the set of all other distributions, so those where $\Delta(\cdot)$ leaves the blue region. To prove the hardness of validating observational study estimates, we show that for any $P \in \mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ we can find distributions $Q \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ that are arbitrarily close by adding spikes as in Figure b.

would correspond to test 2, where the data is used to reject the hypothesis that $\hat{\Delta}$ is not controlled by $\underline{f}, \bar{f}$.

The formulation of test 2 is strongly related to bioequivalence testing[†] (Wellek, 2002; Chow and Liu, 2008). In this field, the goal is to find statistical evidence that one medical treatment works almost equivalently to another, where there is some tolerance specified due to working with finite samples. This is used to approve generic drugs, which have the same active ingredient as a branded drug but can only be sold once the branded drug's patent expires. Here the aim is to ensure the generic drug works as well as the branded one, and so consumers can use them interchangeably.

In our context, we have a similar goal, in that we would like to show that the observational CATE estimate with the adjustment from the RCT is "good enough" up to some tolerance. We specify this tolerance by functions, $(\underline{f}, \bar{f})$, that control Δ as in Definition 3.1.

Following this discussion we provide the following definitions, we refer to tests of type 1 as **falsification tests** and tests of type 2 as **equivalence test**.

3.3 Limits on Testing

Having laid out the two testing problems in Section 3.2, we now apply the impossible inference framework detailed in Section 3.1 to these problems. Firstly, we demonstrate that whilst equivalence testing is more aligned with the aim of the field, the problem as set out is untestable:

Theorem 3.4. *Fix any $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$ and let ψ_n be an equivalence test with null $Q_M(\underline{f},\bar{f})$ and alternative $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$. If the level of this test is, α we have that:*

$$\mathbb{P}_P(\psi_n = 1) \leq \alpha, \tag{11}$$

for any $P \in \mathcal{P}_{M,\pi}(\underline{f},\bar{f})$. That is ψ_n does not have power against any alternative.

This shows that *any* equivalence test which has level α against the null will fail to distinguish *any* alternative. This means that there is no hypothesis test that can confirm from data that the difference function Δ is controlled by any pair of functions $(\underline{f}, \bar{f})$. We visualise the proof of Theorem 3.4 in Figure 2. The idea is that for any distribution $P \in \mathcal{P}_{M,\pi}(\underline{f},\bar{f})$, we can construct a series of distributions $\{Q_i\}_{i=1}^{\infty} : Q_i \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ that tends to P in total variation distance. Therefore, not statistical procedure can distinguish the two.

This result has implications for falsification tests:

Corollary 3.5. *For fixed $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$, any falsification test ψ_n with null $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ and alternative $\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ has:*

$$\inf_{Q \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})} \mathbb{P}_Q(\psi_n = 1) \leq \alpha, \tag{12}$$

where α is the level of ψ_n.

This means that for any n, any falsification test will always fail to distinguish some set of alternatives from the null with power distinctly above the level. However, the next result shows there are alternatives which can be distinguished from the null in falsification tests:

Proposition 3.6. *There exists a distribution $Q \in \mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ such that $\mathrm{TV}(Q, \mathrm{co}(\mathcal{P}_{M,\pi}(\underline{f},\bar{f}))) \geq \beta$ for some $\beta > 0$ where $\mathrm{co}(\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ is the convex hull of $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$. Following Bertanha and Moreira (2020),*

[†] Also referred to as simply equivalence testing.

this guarantees that there is a test with $\beta + \alpha$ against Q where α is the level of the test.

This demonstrates that, unlike validation, falsification is possible relative to certain alternatives.

3.4 Implications for Sensitivity Models

Finally, we apply the results in Section 3.3 to approaches that make use of causal sensitivity models to measure the degree of unmeasured confounding. First, defining a generalised sensitivity model as follows:

Definition 3.7. A **sensitivity model** is a parameterised set of pairs of functions:

$$\left\{(\underline{f}_\gamma, \bar{f}_\gamma) : \underline{f}_\gamma, \bar{f}_\gamma : \mathcal{X} \to \mathbb{R}, \gamma \in [\Gamma_0, \Gamma_1]\right\}, \tag{13}$$

where for fixed $\mathbf{x}$, $\underline{f}_\gamma(\mathbf{x}), \bar{f}_\gamma(\mathbf{x})$ are continuously decreasing/increasing respectively in γ. Moreover, that $\underline{f}_{\Gamma_0}(\mathbf{x}), \bar{f}_{\Gamma_0}(\mathbf{x}) = 0$ and $\underline{f}_{\Gamma_1}(\mathbf{x}), \bar{f}_{\Gamma_1}(\mathbf{x})$ are $-M$ and M respectively. For a distribution in $P \in \mathcal{E}_M$ we define:

$$\Gamma(P) = \inf\left\{\gamma : (\underline{f}_\gamma, \bar{f}_\gamma) \text{ controls } \Delta(\cdot)\right\} \tag{14}$$

Viewed this way, a sensitivity model is a way of constructing intervals around the confounded CATE, $\omega_o(\mathbf{x})$, that contain the true CATE. Moreover, previous work in this area can be seen as aiming to use the experimental dataset to perform inference on $\Gamma(P)$. Specifically, De Bartolomeis et al. (2024a,b) both look to use the experimental data to construct probabilistic lower bounds on $\Gamma(P)$. We now apply results Section 3.3 to constructing confidence intervals for $\Gamma(P)$, showing non-trivial upper bounds are not possible:

Theorem 3.8. *Fix a sensitivity model and let $\mathcal{D}$ be a dataset sampled from $P^{(n)}$ where $P \in \mathcal{E}_{M,\pi}$. Let $\left[\underline{C}(\mathcal{D}), \bar{C}(\mathcal{D})\right]$ be a confidence interval for $\Gamma(P)$ in that it satisfies the following coverage requirement:*

$$\inf_{P \in \mathcal{E}_{0,M}} \mathbb{P}_{\mathcal{D}\sim P^{(n)}}\left(\Gamma(P) \in C\left(\mathcal{D}_n\right)\right) \geq 1 - \alpha \tag{15}$$

Then $\bar{C}(\mathcal{D}) = \Gamma_1$ with probability $1-\alpha$. That is, there are no non-trivial upper bounds on $\Gamma(P)$.

This demonstrates that in contrast to the lower bound case, non-trivial probabilistic upper bounds on $\Gamma(P)$ are not possible without further assumptions on the set of distributions, $\mathcal{E}_M$. This creates difficulties in using sensitivity models to benchmark unmeasured confounding, as a lower bound represents the smallest amount of confounding that can explain the data. Therefore, it *does not* allow us to say with confidence that a treatment effect is contained in some interval.

4 PSEUDO-OUTCOME GAUSSIAN PROCESSES AND UNIFORM ERROR BOUNDS

The results presented in Section 3 show that without further assumptions, we cannot produce intervals which contain the true CATE with high probability. The proof relied on constructing arbitrary peaks in the correction function, which was possible due to a lack of smoothness. As an example of an assumption that can permit this type of inference, we leverage Gaussian process (GPs) which come with inherent smoothness constraints. We then adapt uniform error bounds for Gaussian processes (Fiedler et al., 2021; Lederer et al., 2019) in order to get functions which control $\Delta(\cdot)$.

Before doing this, we develop a Gaussian process approach to learning CATE from pseudo-outcomes. To this best of our knowledge, this is first example of such a method. This is important as it allows us to learn the difference function directly from an estimate of $\omega_o(\mathbf{x})$ in the observational study. Alternative causal Gaussian process approaches, such as Alaa and Van Der Schaar (2017), would require access to an estimate of $\mathbb{E}\left[Y \mid X = \mathbf{x}, T = t, E = o\right]$, to learn a separate correction for each t. By using the pseudo-outcome approach, we sidestep this issue and so allow users full flexibility on how to model $\omega_o(\mathbf{x})$.

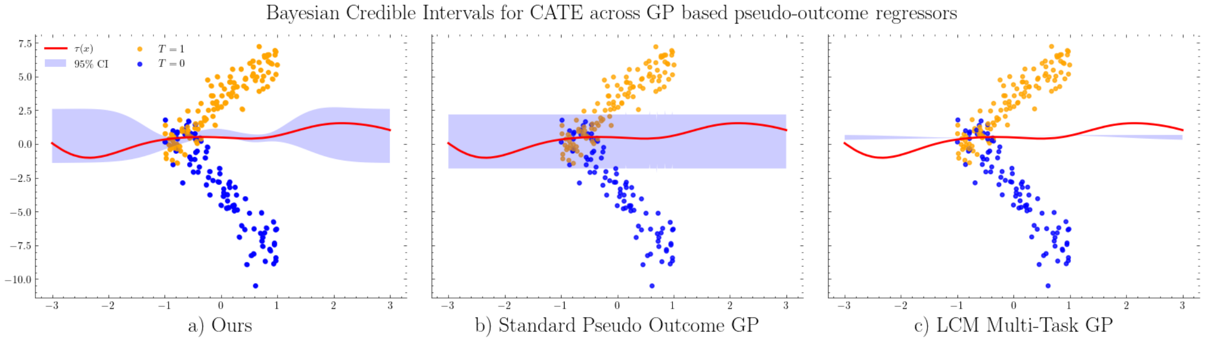
4.1 Pseudo Outcome Regression with Gaussian Processes

We now turn to designing a GP based pseudo-outcome approach. pseudo-outcomes are designed so that the minimiser of the pseudo-outcome mean squared error is the same as the minimiser of the mean squared error under the true unobserved CATE. If the propensity score is correctly specified, the pseudo mean square error will converge optimally to the true CATE mean square error (Kennedy, 2023). However, GP based methods are fit via maximum likelihood, with closed form solutions to the posterior requiring Gaussian errors. This creates a problem for applying them directly to pseudo-outcomes, which have distinctly non Gaussian errors. Specifically, they may be written as:

$$\tilde{Y} = \tau(X) + \epsilon \tag{16}$$

Where $\mathbb{E}[\epsilon \mid X] = 0$ but $\mathbb{E}[\epsilon \mid X, T] \neq 0$. As we show in Appendix C, this breaks Gaussianity assumptions even in cases where the errors on Y are Gaussian. However, the next proposition suggests a simple correction term which allows us to recover well-behaved errors:

Proposition 4.1. *There exists a function $\phi : \mathcal{X} \to \mathbb{R}$*



a) Ours b) Standard Pseudo Outcome GP c) LCM Multi-Task GP

Figure 3: A particularly pathological example of the behaviour we observe for each method in our simulated experiment of Section 5.1. For the standard GP, hyperparameter optimisation leads to uninformative predictions as it cannot account for close $\mathbf{x}$ values with seemingly no correlation. For the trained LCM, we get strong predictive performance but poor uncertainty quantification, especially out of distribution. Our approach gets the best of both scenarios, with strong predictive performance and calibrated uncertainty out of distribution.

such that the IPW pseudo-outcome can be written as:

$$\tilde{Y} = \tau(X) + \left(\frac{T - \pi(X)}{\pi(X)\,(1 - \pi(X))}\right)\phi(X) + \tilde{\epsilon}$$

Where $\mathbb{E}[\tilde{\epsilon} \mid X, T] = 0$ and $\tilde{\epsilon} \mid T, X$ is Gaussian if the original errors on Y are.

Using the decomposition provided by this proposition we may model this using independent Gaussian Processes for $\hat{\Delta}$ and ϕ, so $\hat{\Delta} \sim \mathrm{GP}(0, k_\theta), \phi \sim \mathrm{GP}(0, l_\eta)$. This is equivalent to using a vector valued GP (Alvarez et al., 2012) for multitask regression, where we take the three tasks to be predicting the pseudo-outcome when $T = 0$, predicting the pseudo-outcome when $T = 1$, and the finally predicting the CATE. This corresponds to using the following LCM multitask kernel (Alvarez et al., 2012):

$$\mathbf{K} = \mathbf{a}\mathbf{a}^\top k_\theta + \mathbf{b}\mathbf{b}^\top l_\eta \tag{17}$$

$$\mathbf{a} = \begin{bmatrix}1 & 1 & 1\end{bmatrix} \tag{18}$$

$$\mathbf{b} = \begin{bmatrix}\frac{-1}{1-\pi(X)} & \frac{1}{\pi(X)} & 0\end{bmatrix} \tag{19}$$

From here, we can compute the posterior for $\hat{\Delta}$ given the pseudo-outcomes whilst marginalising out the effect of ϕ. That can all be done in closed form, and gives a posterior of the form $\hat{\Delta}(\mathbf{x}) \sim GP(\tilde{\Delta}_{\mathcal{D}_e}(\cdot), k_{\mathcal{D}_e}(\cdot, \cdot))$ with expressions for $\tilde{\Delta}_{\mathcal{D}_e}(\cdot), k_{\mathcal{D}_e}(\cdot, \cdot)$ in Appendix C.

4.2 Uniform Error Bounds

We now turn to the main purpose of using Gaussian Processes in that we provide a set of assumptions under which our method we can learn functions that control $\Delta(\cdot)$ from experimental data. We do this by adapting the uniform error bounds for GPs from Lederer et al. (2019) to our specific model. First, providing the assumption which makes inference possible:

Assumption 4.2. The unknown $\hat{\Delta}$ and ϕ are samples from Gaussian processes with kernel k and l respectively, i.e $\hat{\Delta}(\cdot) \sim \mathrm{GP}(0, k)$ and $\phi \sim \mathrm{GP}(0, l)$. Further we assume $\mathcal{X}_o$ is compact, the errors have distribution $\mathcal{N}(0, \sigma_t^2)$ given $T = t$, k has Lipschitz constant L_k, and $\hat{\Delta}$ has a Lipschitz constant $L_{\hat{\Delta}}$.

This implies the following bounds on the $\hat{\Delta}(\cdot)$:

Theorem 4.3. *Let the posterior for $\hat{\Delta}(\cdot)$ from the GP model defined in Section 4.1 be given pointwise by $\mathcal{N}(\tilde{\Delta}(\mathbf{x}), \sigma^2(\mathbf{x}))$ where $\tilde{\Delta}, \sigma : \mathcal{X}_o \to \mathbb{R}$ and let $\hat{\tau}(\mathbf{x}) = \hat{\omega}(\mathbf{x}) + \tilde{\Delta}(\mathbf{x})$. Then, under assumption 4.2, for fixed $\delta \in (0, 1), \tau \in \mathbb{R}^+$ we have:*

$$P(|\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| \leq B(\mathbf{x}) \;\; \forall \mathbf{x} \in \mathcal{X}_o) > 1 - \delta$$

$$B(\mathbf{x}) = \sqrt{2 \log\left(\frac{M(\tau, \mathcal{X}_o)}{\delta}\right)}\sigma(\mathbf{x}) + \gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$$

Where $M(\tau, \mathcal{X}_o)$ is the τ covering number of $\mathcal{X}_o$, defined as the minimum number of spherical balls of radius τ needed to cover $\mathcal{X}_o$, and $\gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$ is defined in Appendix C.

There is not a uniformly optimal value of τ due to dependence on other constants. However, if our covariate space is a hypercube of length r with dimension d, we have that $M(\tau, \mathcal{X}_o) \leq (1 + \frac{r}{\tau})^d$ and as we show in Appendix C, we have that $\gamma(\tau, \mathbf{X}_e, L_k, L_\Delta) = o(\tau^{\frac{1}{2}})$. This ensures we can always recover log dependency in δ and τ. Further, the kernel choice gives us knowledge of L_k and allows us to probabilistically bound L_f as in Lederer et al. (2019) and shown in Appendix C.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.77 \pm 0.01}$	0.785 ± 0.04	$\mathbf{3.31 \pm 0.02}$
Naive GP	2.05 ± 0.02	$\mathbf{0.796 \pm 0.04}$	3.65 ± 0.03
LCM	1.91 ± 0.01	0.303 ± 0.11	1.09 ± 0.06

Table 1: Results for the simulated experiment in Section 5.1 with $d = 10$ and $n_e = 1000$, averaged over 200 runs. Our approach leads to both the best predictive performance and well calibrated uncertainty, achieving a similar coverage to the standard GP with smaller predictive intervals.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.10 \pm 0.04}$	$\mathbf{0.831 \pm 0.008}$	$\mathbf{2.66 \pm 0.02}$
Naive GP	2.19 ± 0.13	0.752 ± 0.010	3.38 ± 0.03
LCM	1.39 ± 0.05	0.828 ± 0.010	3.00 ± 0.05

Table 2: Results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs. We again find that our method has the best predictive performance and most informative coverage intervals, in the sense that they contain the true CATE with high probability whilst also being significantly smaller.

5 Experiments

For the experiments, we demonstrate the improvements in predictive performance and calibrated uncertainty provided by our pseudo-outcome GP approach when compared against other causal GP approaches. We compare against fitting a naive standard GP, and a GP with a multitask LCM kernel to pseudo-outcomes. This second approach which can be viewed as a scaled version of the causal multitask Gaussian process of (Alaa and Van Der Schaar, 2017),which represents the state of the art in GP's for CATE estimation[‡]. For all models we tune the free hyperparameters using gradient descent on the marginal log likelihood, details in Appendix D.1. For results on the coverage of uniform error bounds for our model specifically, see F.

5.1 Simulated Experiment

Firstly, we use an adaptation of the simulated provided in Kallus et al. (2018). We let the experimental and observational covariate distribution be $\mathcal{U}([-1,1]^d)$ and $\mathcal{U}([-3,3]^d)$ respectively, where d is the covariate dimension, and $T \sim \text{Ber}(\frac{1}{2})$. The observational outcomes are simulated using a quadratic in the first component of X and T with normal noise. For the experimental outcomes, we simulate use the same polynomial but add a sample from a GP. We do this in order to test our methodology in setting where assumptions are satisfied. Full details are given in Appendix D.

Finally, as we focus on assessing the GP portion of the model, we use the true $\omega_o(\mathbf{x})$ for this experiment. This represents a case where n_o is so large, we approach fitting a perfect model. We later relax this.

5.1.1 Results

We present results for this experiment with $d = 10$ and $n_e = 1000$ in Table 1, alongside an illustrative figure for $d = 1$ and $n_e = 200$ in Figure 3. In Table 1 we compare the mean squared error to the true CATE, the coverage of 95% Bayesian credible intervals, and the width of these intervals. We use Bayesian credible intervals to form a fair comparison for uncertainty quantification, as uniform error bounds are only available for our model. Additional results including varying dimension, varying sample size, and extrapolation beyond the experimental sample in Appendix E.1.

Across all settings, our results show the following trends: i) The naive GP shows poor predictive performance and is totally uninformative in some settings. This is because it is unable to correctly optimise hyperparameters to capture the highly variable noise distribution in the pseudo-outcomes, and so it reverts to the prior ii) The trainable LCM multitask kernel has good predictive performance but poorly calibrated uncertainty, This is because the hyperparameter optimisation either leads to overly confident or overly wide credible intervals, depending on the dimension and the number of samples. Further, this optimisation leads it to overfit training data and extrapolate poorly. iii) Our approach is able to incorporate both strong predictive performance and calibrated uncertainty, with intervals that have the coverage guarantees of other methods that produce much wider intervals.

5.2 Semi-Synthetic Experiments

To assess our method in a more realistic setting, we use the Infant Health and Development Program (IHDP) dataset (Louizos et al., 2017), similarly to Hussain et al. (2022, 2023). The IHDP dataset comes from a randomised controlled trial, and it contains $n = 985$ samples and a 28 dimensional covariate distribution with 7 continuous covariates. The dataset comes with a treatment allocation, but outcomes need to be simulated.

We form our data from the IHDP dataset as follows: for the observational sample, we uniformly sample the covariates and treatment with replacement until reaching the desired sample size. For the experimental study, we do a weighted sampling to ensure a covariate shift and then randomly sample the treatment from $T \sim \text{Ber}(p)$. For the outcomes in the observational dataset, we simulate from a sparse linear model in X and T. We do this as we want to emulate a scenario

[‡] For more on the comparison between the LCM GP and Causal Multitask GP's see Appendix C.3

where we have relatively low error in estimating $\omega_o(\mathbf{x})$ due to a large n_o, but we do not want to repeat the small dataset multiple times. Finally, we again simulate the difference between observational and experimental distributions with a GP as in Section 5.1. Full details available in Appendix D.3.

5.2.1 Results

We present the results for the experiment with $n_e = 400$ in Table 2, with other results for varying sample size, treatment proportion and out of distribution generalisation in Appendix E.2. We can see that our proposed methodology outperforms both alternatives in terms of predictive performance and calibrated uncertainty, having the joint best coverage with the smallest intervals. The advantage becomes even more clear when extrapolating beyond the experimental study to the observational study, as shown in Table E.2 in Appendix E.2. In this case, the trained GP predicts overly broad intervals due to over-fitting the hyperparameters to the experimental data.

5.3 Additional Results

Finally, we highlight some additional results available in the Appendices. In Appendix E.1, we present our simulated experiment over different dimensionality and sample sizes, in Appendix E.2 we present the IHDP study for different treatment proportions, and for a varying quality of fit of the observational model and in Appendix E.3 we present results for both scenarios, varying the difference function so that it is not a GP. Finally, in Appendix E.3, we provide results on our uniform error bounds for our proposed GP model.

6 CONCLUSION

It is well known in causal inference that there are no observational tests for unmeasured confounding. In this work, we showed that even with experimental data, there are fundamental limits to testing for unmeasured confounding. We showed that in order to validate observational studies, one needs to make assumptions on the smoothness of the correction function. Following this, we developed a Gaussian Process approach to learning from pseudo-outcomes, and assumptions arising from this model which produce intervals that contain the CATE with high probability.

Acknowledgements

JF gratefully acknowledges funding from the EPSRC.

References

A. M. Alaa and M. Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in neural information processing systems*, 30, 2017.

M. A. Alvarez, L. Rosasco, N. D. Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

P. Aronow, J. M. Robins, T. Saarinen, F. Sävje, and J. Sekhon. Nonparametric identification is not enough, but randomized controlled trials are. *arXiv preprint arXiv:2108.11342*, 2021.

R. R. Bahadur and L. J. Savage. The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122, 1956.

S. Balakrishnan and L. Wasserman. Hypothesis testing for densities and high-dimensional multinomials. *The Annals of Statistics*, 47(4):1893–1927, 2019.

M. Bertanha and M. J. Moreira. Impossible inference in econometrics: Theory and applications. *Journal of Econometrics*, 218(2):247–270, 2020.

E. V. Bonilla, K. Chai, and C. Williams. Multi-task gaussian process prediction. *Advances in neural information processing systems*, 20, 2007.

I. A. Canay, A. Santos, and A. M. Shaikh. On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6): 2535–2559, 2013.

S.-C. Chow and J.-p. Liu. Design and analysis of bioavailability and bioequivalence studies. 2008.

B. Colnet, I. Mayer, G. Chen, A. Dieng, R. Li, G. Varoquaux, J.-P. Vert, J. Josse, and S. Yang. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical science*, 39(1):165–191, 2024.

A. Curth and M. Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1810–1818. PMLR, 2021.

P. De Bartolomeis, J. Abad, K. Donhauser, and F. Yang. Detecting critical treatment effect bias in small subgroups. *arXiv preprint arXiv:2404.18905*, 2024a.

P. De Bartolomeis, J. A. Martinez, K. Donhauser, and F. Yang. Hidden yet quantifiable: A lower bound for confounding strength using randomized trials. In *International Conference on Artificial Intelligence and Statistics*, pages 1045–1053. PMLR, 2024b.

A. D'Amour, P. Ding, A. Feller, L. Lei, and J. Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

C. Fiedler, C. W. Scherer, and S. Trimpe. Practical and rigorous uncertainty bounds for gaussian process regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7439–7447, 2021.

C. Gilligan-Lee. Causing trouble. *New Scientist*, 246 (3279):32–35, 2020.

C. M. Gilligan-Lee, C. Hart, J. Richens, and S. Johri. Leveraging directed causal discovery to detect latent common causes in cause-effect pairs. *IEEE Transactions on Neural Networks and Learning Systems*, 35(4):4938–4947, 2022.

E. Hariton and J. J. Locascio. Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13):1716, 2018.

Z. Hussain, M.-C. Shih, M. Oberst, I. Demirel, and D. Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. In *International Conference on Artificial Intelligence and Statistics*, pages 5869–5898. PMLR, 2023.

Z. M. Hussain, M. Oberst, M.-C. Shih, and D. Sontag. Falsification before extrapolation in causal effect estimation. *Advances in Neural Information Processing Systems*, 35:6161–6174, 2022.

O. Jeunen, C. Gilligan-Lee, R. Mehrotra, and M. Lalmas. Disentangling causal effects from sets of interventions in the presence of unobserved confounders. *Advances in Neural Information Processing Systems*, 35:27850–27861, 2022.

N. Kallus, A. M. Puli, and U. Shalit. Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31, 2018.

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

A. Lederer, J. Umlauft, and S. Hirche. Uniform error bounds for gaussian process regression with application to safe control. *Advances in Neural Information Processing Systems*, 32, 2019.

X. Lin and R. J. Evans. Many data: Combine experimental and observational data through a power likelihood. *arXiv preprint arXiv:2304.02339*, 2023.

C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

K. Muandet, W. Jitkrittum, and J. Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020.

J. Pearl. Causal inference in statistics: An overview. 2009.

T. S. Richardson and J. M. Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128 (30):2013, 2013.

J. P. Romano. On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584, 2004.

P. R. Rosenbaum and D. B. Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. 2020.

G. Van Goffrier, L. Maystre, and C. M. Gilligan-Lee. Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding. In *Conference on Causal Learning and Reasoning*, pages 791–813. PMLR, 2023.

S. Wellek. *Testing statistical hypotheses of equivalence.* Chapman and Hall/CRC, 2002.

L. Wu and S. Yang. Integrative r-learner of heterogeneous treatment effects combining experimental and observational studies. In *Conference on Causal Learning and Reasoning*, pages 904–926. PMLR, 2022.

S. Yang, D. Zeng, and X. Wang. Improved inference for heterogeneous treatment effects using real-world data subject to hidden confounding. *arXiv preprint arXiv:2007.12922*, 2020.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

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

(a) Citations of the creator If your work uses existing assets. [Not Applicable]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

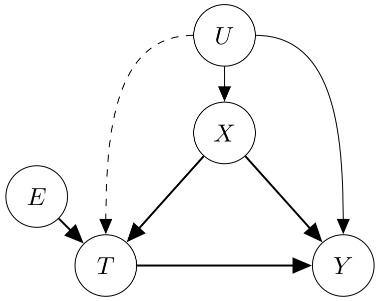


Figure 4: Causal Structure for generating the experimental and observational datasets with environment node drawn in.

A Causal Assumptions

A.1 Constant CATE

Drawing the graph with an environment node in Figure 4, we can read off the graph using the SWIG framework (Richardson and Robins, 2013) for conditional indecencies of potential outcomes from graphical models that $Y(t) \perp E \mid X$. This implies:

$$\mathbb{E}\left[Y(t) \mid X, E = e\right] = \mathbb{E}\left[Y(t) \mid X, E = o\right] \tag{20}$$

And so constant CATE.

B Hardness of Testing

Theorem 3.4. *Fix any $\underline{f}, \bar{f} : \mathcal{X} \to \mathbb{R}$ and let ψ_n be an equivalence test with null $\mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ and alternative $\mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. If the level of this test is α we have that:*

$$\mathbb{P}_P(\psi_n = 1) \leq \alpha, \tag{21}$$

for any $P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$. That is ψ_n does not have power against any alternative.

Proof. Following Romano (2004), we need to show that the null is dense in the alternative in total variation distance. This corresponds to showing that for $P \in \mathcal{P}_{M,\pi}(\underline{f}, \bar{f})$ we can find a sequence of distributions $\{Q_n\}_{n=1}^{\infty} \in \mathcal{Q}_{M,\pi}(\underline{f}, \bar{f})$ such that $d_{\mathrm{TV}}(P, Q_n) \to 0$.

Firstly, note that as Y is bounded by M we must have that $\left|\hat{\Delta}(\mathbf{x}) + \hat{\omega}_o(\mathbf{x})\right| \leq 2M$ which implies $-\left(\hat{\omega}_o(\mathbf{x}) + 2M\right) \leq \hat{\Delta}(\mathbf{x}) \leq 2M - \hat{\omega}_o(\mathbf{x})$. So in order for the bounds to be non vacuous we need $\bar{f}(\mathbf{x}) < 2M - \hat{\Delta}(\mathbf{x}) \leq \hat{\omega}_o(\mathbf{x})$ or $\underline{f}(\mathbf{x}) > -\left(\hat{\omega}_o(\mathbf{x}) + 2M\right)$ on some set of positive measure. Assume wlog that we have $\bar{f}(\mathbf{x}) < 2M - \hat{\omega}_o(\mathbf{x})$ and then let $\mathcal{A}_n =\subset \mathcal{X}$ be a measurable set such that $P(\mathcal{A}_n) = \epsilon_n$ where $0 < \epsilon_n \leq \frac{1}{n^2}$ such that this holds. We can find such a set as P is absolutely continuous in $\mathbf{x}$ with respect to the Lebesgue measure.

Now define R_n as a distribution over (X, T, Y) where X is uniform on $\mathcal{A}_n$, conditional distribution of T given X coming from π and the distribution over Y is such that the true CATE is $2M$. Now we let:

$$Q_n := \frac{n-1}{n} P + \frac{1}{n} R_n \tag{22}$$

Now consider the expectation of $\tilde{Y}$ under this distribution, given $X \in \mathcal{A}_n$:

$$\mathbb{E}_{Q_n}[\tilde{Y} \mid X \in \mathcal{A}_n] = \mathbb{E}_P[\tilde{Y} \mid X \in \mathcal{A}_n]\mathbb{P}\left((\tilde{Y}, X) \sim P \mid X \in \mathcal{A}_n\right) \tag{23}$$

$$+ \mathbb{E}_{R_n}[\tilde{Y} \mid X \in \mathcal{A}_n]\mathbb{P}\left((\tilde{Y}, X) \sim R_n \mid X \in \mathcal{A}_n\right) \tag{24}$$

We have the following:

$$\mathbb{P}\left((\tilde{Y},X)\sim P \mid X\in\mathcal{A}_n\right) = \frac{\mathbb{P}\left(X\in\mathcal{A}_n \mid (\tilde{Y},X)\sim P\right)\mathbb{P}\left((\tilde{Y},X)\sim P\right)}{\mathbb{P}\left(X\in\mathcal{A}_n\right)} \tag{25}$$

$$= \frac{\epsilon_n\frac{n-1}{n}}{\epsilon_n\frac{n-1}{n}+\frac{1}{n}} \tag{26}$$

$$= \frac{\epsilon_n(n-1)}{\epsilon_n(n-1)+1} \tag{27}$$

And:

$$\mathbb{P}\left((\tilde{Y},X)\sim Q_n \mid X\in\mathcal{A}_n\right) = 1-\mathbb{P}\left((\tilde{Y},X)\sim P \mid X\in\mathcal{A}_n\right) \tag{28}$$

$$= \frac{1}{\epsilon_n(n-1)+1} \tag{29}$$

So that:

$$\mathbb{E}_{Q_n}[\tilde{Y}\mid X\in\mathcal{A}_n] = \frac{\epsilon_n(n-1)\mathbb{E}_P[Z\mid X\in\mathcal{A}_n]}{\epsilon_n(n-1)+1}+\frac{\mathbb{E}_{Q_n}[Z\mid X\in\mathcal{A}_n]}{\epsilon_n(n-1)+1} \tag{30}$$

$$= \frac{\epsilon_n(n-1)\mathbb{E}_P[\tilde{Y}\mid X\in\mathcal{A}_n]}{\epsilon_n(n-1)+1}+\frac{2M}{\epsilon_n(n-1)+1} \tag{31}$$

Now as $\epsilon_n(n-1)\to 0$ as $n\to\infty$ we have $\mathbb{E}_{P_n}[\tilde{Y}\mid X\in\mathcal{A}_n]\to 2M$. As we have $\hat{\Delta}(\mathbf{x})=\mathbb{E}_P[\tilde{Y}\mid\mathbf{x}]-\hat{\omega}_o(\mathbf{x})$, we have that $\hat{\Delta}(\mathbf{x})\to 2M-\hat{\omega}_o(\mathbf{x})$ for $\mathbf{x}\in\mathcal{A}_n$. Therefore, by taking the cutoff sequence $\{Q_n\}_{n\geq k}^{\infty}$ for some k we have a sequence of distributions such that $\hat{\Delta}(\mathbf{x})$ expectation is greater than $\underline{f}(\mathbf{x})$ on a set of positive measure and that $d_{\mathrm{TV}}(Q,P_n)\to 0$. □

Corollary 3.5. *For fixed $\underline{f},\bar{f}:\mathcal{X}\to\mathbb{R}$, any falsification test ψ_n with null $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ and alternative $\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ has:*

$$\inf_{Q\in\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})}\mathbb{P}_Q(\psi_n=1)\leq\alpha, \tag{32}$$

where α is the level of ψ_n.

Proof. This follows as a direct result of the fact that the alternative is dense in the null, as shown in the previous proposition □

Proposition 3.6. *There exists a distribution $Q\in\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ such that $\mathrm{TV}(Q,\mathrm{co}(\mathcal{P}_{M,\pi}(\underline{f},\bar{f})))\geq\beta$ for some $\beta>0$ where $\mathrm{co}(\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$ is the convex hull of $\mathcal{P}_{M,\pi}(\underline{f},\bar{f})$. Following Bertanha and Moreira (2020), this guarantees that there is a test with $\beta+\alpha$ against Q where α is the level of the test.*

Proof. As in the proof of Theorem 3.4 assume wlog that we have $\bar{f}(\mathbf{x})<2M-\hat{\omega}_o(\mathbf{x})$ for some measurable set $\mathcal{A}$. Now define a distribution $Q\in\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ as follows:

- X is uniform on $\mathcal{A}$.
- $T|X\sim\pi(X)$.
- $Y|T=M(2T-1)$.

Now let $\tilde{A}=\mathrm{supp}(Q)$ be the support of Q. We have that $\mathbb{E}_P[\tilde{Y}|(X,T,Y)\in\tilde{\mathcal{A}}]=2M$ for any distribution P. Now for $P\in\mathcal{Q}_{M,\pi}(\underline{f},\bar{f})$ we have that $\mathbb{E}_P[\tilde{Y}|(X)\in\mathcal{A}]\leq\bar{f}(\mathbf{x})<2M-\hat{\omega}_o(\mathbf{x})$. Now we have:

$$\mathbb{E}_P[\tilde{Y}|X\in\mathcal{A}]=\mathbb{E}_P[\tilde{Y}|(X,T,Y)\in\tilde{\mathcal{A}}]\mathbb{P}_P((X,T,Y)\in\tilde{\mathcal{A}}\mid X\in\mathcal{A}) \tag{33}$$

$$+\mathbb{E}_P[\tilde{Y}|(X,T,Y)\in\mathcal{A}\setminus\tilde{\mathcal{A}}]\left(1-\mathbb{P}_P((X,T,Y)\in\tilde{\mathcal{A}}\mid X\in\mathcal{A})\right) \tag{34}$$

$$\geq 2M\left(2\mathbb{P}_P((X,T,Y)\in\tilde{\mathcal{A}}\mid X\in\mathcal{A})-1\right) \tag{35}$$

As $\mathbb{E}_P[\tilde{Y}|(X,T,Y) \in \mathcal{A} \setminus \tilde{\mathcal{A}}] \geq -2M$. Putting this together implies:

$$2M - \hat{\omega}_o(\mathbf{x}) \geq 2M \left(2\mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A}) - 1 \right) \tag{36}$$

$$\implies \mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}} \mid X \in \mathcal{A}) \leq 1 - \frac{\hat{w}_o(\mathbf{x})}{4M} \tag{37}$$

$$\implies \mathbb{P}_P((X,T,Y) \in \tilde{\mathcal{A}}) \leq 1 - \frac{\hat{w}_o(\mathbf{x})}{4M} \tag{38}$$

Which completes the proof. □

Theorem 3.8. *Fix a sensitivity model and let $\mathcal{D}$ be a dataset sampled from $P^{(n)}$ where $P \in \mathcal{E}_{M,\pi}$. Let $\left[\underline{C}\left(\mathcal{D}\right), \bar{C}\left(\mathcal{D}\right)\right]$ be a confidence interval for $\Gamma(P)$ in that it satisfies the following coverage requirement:*

$$\inf_{P \in \mathcal{E}_{0,M}} \mathbb{P}_{\mathcal{D} \sim P^{(n)}} \left(\Gamma(P) \in C\left(\mathcal{D}_n\right) \right) \geq 1 - \alpha \tag{39}$$

Then $\bar{C}\left(\mathcal{D}\right) = \Gamma_1$ with probability $1-\alpha$. That is, there are no non trivial upper bounds on $\Gamma(P)$.

Proof. This follows from the fact that for any $\gamma \in [\Gamma_0, \Gamma_1]$ we have that $\mathcal{Q}_{M,\pi}(\underline{f}_\gamma, \bar{f}_\gamma)$ is dense in $\mathcal{P}_{M,\pi}(\underline{f}_\gamma, \bar{f}_\gamma)$. Therefore, for any distribution $P \in \mathcal{E}_{0,M}$ is arbitrarily close in total variation to a distribution whose true sensativity parameter is arbitrarily high. Therefore, if we are to satisfy the coverage requirment uniformly over all distributions we must have $\bar{C}\left(\mathcal{D}\right) = \Gamma_1$ with probability $1-\alpha$. □

C Gaussian Process

Proposition 4.1. *There exists a function $\phi : \mathcal{X} \to \mathbb{R}$ such that the IPW pseudo-outcome can be written as:*

$$\tilde{Y} = \tau(X) + \left(\frac{T - \pi(X)}{\pi(X)\left(1 - \pi(X)\right)} \right) \phi(X) + \tilde{\epsilon}$$

Where $\mathbb{E}[\tilde{\epsilon} \mid X, T] = 0$ and $\tilde{\epsilon} \mid T, X$ is Gaussian if the original errors on Y are.

Proof. Suppose we have:

$$Y = \mu(X,T) + \epsilon \tag{40}$$

Where $\mu(X,T) = \mathbb{E}[Y \mid X, T]$ so that $\mathbb{E}[\epsilon \mid X, T] = 0$. Now the error of the pseudo-outcome from $\tau(X)$ is:

$$\frac{T - e(X)}{e(X)\left(1 - e(X)\right)} Y - \left(\mu(X,1) - \mu(X,0)\right) = \frac{\left(T - e(X)\right)\left(\mu(X,T) + \epsilon\right)}{e(X)\left(1 - e(X)\right)} - \left(\mu(X,1) - \mu(X,0)\right) \tag{41}$$

$$= \begin{cases} \frac{(\mu(X,1)+\epsilon)}{e(X)} - \left(\mu(X,1) - \mu(X,0)\right) & \text{if } T = 1 \\ -\frac{(\mu(X,0)+\epsilon)}{1-e(X)} - \left(\mu(X,1) - \mu(X,0)\right) & \text{if } T = 0 \end{cases} \tag{42}$$

$$= \begin{cases} \frac{1}{e(X)} \left(\left(1 - e(X)\right)\mu(X,1) + e(X)\mu(X,0) + \epsilon \right) & \text{if } T = 1 \\ \frac{-1}{1-e(X)} \left(e(X)\mu(X,0) + \left(1 - e(X)\right)\mu(X,1) - \epsilon \right) & \text{if } T = 0 \end{cases} \tag{43}$$

If we let $\phi(X) = \left(1 - e(X)\right)\mu(X,1) + e(X)\mu(X,0)$ then we have that (43) is equal to:

$$= \frac{T - e(X)}{e(X)\left(1 - e(X)\right)} \left(\phi(X) + (-1)^{T+1} \epsilon \right) \tag{44}$$

Now if we let $\tilde{\epsilon} = (-1)^{T+1} \epsilon \frac{T - e(X)}{e(X)(1 - e(X))}$ we can see that we now have $\mathbb{E}\left[\tilde{\epsilon} \mid X, T\right] = 0$. Moreover, since the distribution of $\tilde{\epsilon}$ given X, T is just a constant scaled version of ϵ we have that $\tilde{\epsilon} \mid X, T$ is Gaussian if and only if $\epsilon \mid X, T$ is. □

C.1 Closed Form Posterior Expressions

Now, for the closed form expressions have as follows, were the training dataset is $\{\mathbf{x}_i, t_i, \tilde{y}_i - \hat{\omega}_o(\mathbf{x}_i)\}_{i=1}^{n_e}$:

$$\mathbf{M}_N = \left((\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j))_{t_i, t_j} \right)_{i,j} \tag{45}$$

$$\mathbf{\Sigma}_N = \operatorname{diag}\left(\left(\sigma_{t_i}^2\right)_i \right) \tag{46}$$

$$\mathbf{y}_N = (\tilde{y}_i - \hat{\omega}_o(\mathbf{x}_i))_i \tag{47}$$

$$\boldsymbol{k}_N(\mathbf{x}) = (k(\mathbf{x}_i, \mathbf{x}))_i \tag{48}$$

$$\tilde{\Delta}_{\mathcal{D}_e}(\mathbf{x}) = \boldsymbol{k}_N(\mathbf{x})^\top (\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1} \mathbf{y}_N \tag{49}$$

$$k_{\mathcal{D}_e}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \boldsymbol{k}_N(\mathbf{x})^\top (\mathbf{M}_N + \mathbf{\Sigma}_N)^{-1} \boldsymbol{k}_N(\mathbf{x}') \tag{50}$$

C.2 Closed Form Bounds

Theorem 4.3. *Let the posterior for $\Delta(\cdot)$ from the GP model defined in Section 4.1 be given pointwise by $\mathcal{N}(\tilde{\Delta}(\mathbf{x}), \sigma^2(\mathbf{x}))$ where $\tilde{\Delta}, \sigma : \mathcal{X}_o \to \mathbb{R}$. Then, under assumption 4.2, for fixed $\delta \in (0,1), \tau \in \mathbb{R}^+$ we have:*

$$P(|\Delta(\mathbf{x}) - \tilde{\Delta}(\mathbf{x})| \leq B(\mathbf{x}) \;\; \forall \mathbf{x} \in \mathcal{X}_o) > 1 - \delta$$
$$B(\mathbf{x}) = \sqrt{2 \log\left(\frac{M(\tau, \mathcal{X}_o)}{\delta} \right)} \sigma(\mathbf{x})$$
$$+ \gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$$

Where $M(\tau, \mathcal{X}_o)$ is the τ covering number of $\mathcal{X}_o$, defined as the minimum number of spherical balls of radius τ needed to cover $\mathcal{X}_o$, and $\gamma(\tau, \mathbf{X}_e, L_k, L_\Delta)$ is defined in Appendix C.

Proof. This follows from theorem 3.1 in Lederer et al. (2019) which we reproduce here:

Theorem (Lederer et al. (2019)). *Consider a zero mean Gaussian process defined through the continuous covariance kernel $k(\cdot, \cdot)$ with Lipschitz constant L_k on the compact set $\mathbb{X}$. Furthermore, consider a continuous unknown function $f : \mathbb{X} \to \mathbb{R}$ with Lipschitz constant L_f and $N \in \mathbb{N}$ observations y_i satisfying:*

Assumption C.1. The unknown function $f(\cdot)$ is a sample from a Gaussian process $\mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}'))$ and observations $y = f(\boldsymbol{x}) + \epsilon$ are perturbed by zero mean i.i.d. Gaussian noise ϵ with variance σ_n^2.

Then, the posterior mean function $\nu_N(\cdot)$ and standard deviation $\sigma_N(\cdot)$ of a Gaussian process conditioned on the training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ are continuous with Lipschitz constant L_{ν_N} and modulus of continuity $\omega_{\sigma_N}(\cdot)$ on $\mathbb{X}$ such that:

$$L_{\nu_N} \leq L_k \sqrt{N} \left\| \left(k(\boldsymbol{X}_N, \boldsymbol{X}_N) + \sigma_n^2 \boldsymbol{I}_N \right)^{-1} \boldsymbol{y}_N \right\|$$
$$\omega_{\sigma_N}(\tau) \leq \sqrt{2\tau L_k \left(1 + N \left\| \left(k(\boldsymbol{X}_N, \boldsymbol{X}_N) + \sigma_n^2 \boldsymbol{I}_N \right)^{-1} \right\| \max_{\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X}} k(\boldsymbol{x}, \boldsymbol{x}') \right)}$$

Moreover, pick $\delta \in (0,1), \tau \in \mathbb{R}_+$ and set

$$\beta(\tau) = 2 \log\left(\frac{M(\tau, \mathbb{X})}{\delta} \right)$$
$$\gamma(\tau) = (L_{\nu_N} + L_f)\tau + \sqrt{\beta(\tau)} \omega_{\sigma_N}(\tau)$$

Then, it holds that

$$P\left(|f(\boldsymbol{x}) - \nu_N(\boldsymbol{x})| \leq \sqrt{\beta(\tau)} \sigma_N(\boldsymbol{x}) + \gamma(\tau), \forall \boldsymbol{x} \in \mathbb{X} \right) \geq 1 - \delta$$

In our case, the bounds on the Lipschitz constants and modulus of continuity are different. However by the same argument as Lederer et al. (2019) we have:

$$L_{\tilde{\Delta}_{\mathcal{D}_e}} \leq L_k\sqrt{N}\|(\mathbf{M}_N+\mathbf{\Sigma}_N)^{-1}\mathbf{y}_N\| \tag{51}$$

$$\omega_{\sigma_N}(\tau) \leq \sqrt{2\tau L_k\left(1+N\left\|(\mathbf{M}_N+\mathbf{\Sigma}_N)^{-1}\right\|\max_{\boldsymbol{x},\boldsymbol{x}'\in\mathbb{X}} k\left(\boldsymbol{x},\boldsymbol{x}'\right)\right)} \tag{52}$$

Putting this in to the above Theorem of Lederer et al. (2019) that, under assumption 4.2, for fixed $\delta \in (0,1), \tau \in \mathbb{R}^+$ we have

$$\begin{aligned}
&P(|\Delta(\mathbf{x})-\tilde{\Delta}(\mathbf{x})| \leq B(\mathbf{x}) \;\; \forall \mathbf{x}\in\mathcal{X}_o) > 1-\delta \\
&B(\mathbf{x}) = \sqrt{2\log\left(\frac{M\left(\tau,\mathcal{X}_o\right)}{\delta}\right)}\sigma(\mathbf{x}) \\
&\quad + \left(L_k\sqrt{N}\|(\mathbf{M}_N+\mathbf{\Sigma}_N)^{-1}\mathbf{y}_N\|+L_f\right)\tau \\
&\quad + \sqrt{\beta(\tau)}\sqrt{2\tau L_k\left(1+N\left\|(\mathbf{M}_N+\mathbf{\Sigma}_N)^{-1}\right\|\max_{\boldsymbol{x},\boldsymbol{x}'\in\mathbb{X}} k\left(\boldsymbol{x},\boldsymbol{x}'\right)\right)}
\end{aligned}$$

Where we can see the claimed convergence property in τ. □

C.3 LCM Kernel and Causal Multitask Kernel of Alaa and Van Der Schaar (2017)

In this the work of Alaa and Van Der Schaar (2017), CATE is modelled using a multitask Gaussian process (Bonilla et al., 2007). Multitask Gaussian Processes use a GP in vector-valued Reproducing Kernel Hilbert Space (vv-RKHS) to share information between tasks (Alvarez et al., 2012). In Alaa and Van Der Schaar (2017), learning the conditional outcome function for each treatment is seen as a separate task, so we jointly model:

$$Y|\mathbf{x},t \sim \mathcal{N}(0, f_t(\mathbf{x}), \sigma_t^2) \tag{53}$$

Where each f_t is a Gaussian Process. The kernel $\tilde{\mathbf{K}}_\eta : \mathcal{X}\times\mathcal{X}\to\mathbb{R}^{2\times 2}$ is now a symmetric positive semi-definite matrix-valued function, with hyper-parameters η. In the case of Alaa and Van Der Schaar (2017) they use a *linear model of coregionalization*[§], giving the kernel as:

$$\tilde{\mathbf{K}}(\mathbf{x},\mathbf{x}') = \mathbf{A}_0 k(\mathbf{x},\mathbf{x}') + \mathbf{A}_1 l(\mathbf{x},\mathbf{x}') \tag{54}$$

Where $\mathbf{A}_t$ is given by:

$$\tilde{\mathbf{A}}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \tilde{\mathbf{A}}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{55}$$

And $\mathbf{A}_t$ are now free hyperparameters to learn.

This is equivalent to the LCM kernel that we make use of for our experiments, however we regress onto pseudo-outcomes as opposed to observed outcomes. This is equivalent to scaling the task t is scaled by $\frac{t-\pi(\mathbf{x}))}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}$ and leaving all hyper-parameters free to learn. As scaled Gaussian processes are still Gaussian processes this is same as using the kernel:

$$\mathbf{K}(\mathbf{x},\mathbf{x}') = \mathbf{M}\mathbf{A}_0 k(\mathbf{x},\mathbf{x}') + \mathbf{M}\mathbf{A}_1 l(\mathbf{x},\mathbf{x}') \tag{56}$$

Where:

$$\mathbf{M} = \left(\left(\frac{t-\pi(\mathbf{x}))}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}\right)\left(\frac{t'-\pi(\mathbf{x}))}{\pi(\mathbf{x})(1-\pi(\mathbf{x}))}\right)\right)_{t,t'=0,1} \tag{57}$$

This demonstrates an equivalence between the Pseudo-outcome based LCM method we use for the experiments and the methods of (Alaa and Van Der Schaar, 2017).

[§] See Alvarez et al. (2012) for more details.

D Experiment Details

D.1 Model Tuning Details

For each of the models we regress from $\mathbf{x}, \mathbf{t}$ onto $\tilde{y} - \hat{w}_o(\mathbf{x})$ where $\hat{w}_o(\mathbf{x})$ is our estimate of $w_o(\mathbf{x})$. We do so as:

$$\tilde{Y} - \hat{w}_o(X) \sim \mathcal{N}(f_t(X), \sigma_t^2) \tag{58}$$

With specific model details as follows:

1. **Standard or Naive GP** Taking $f_0 = f_1 = f$ and $\sigma_0^2 = \sigma_1^2 = \sigma^2 2$. We then model f directly using a $\text{GP}(0, k_\theta)$ where k_θ is a kernel with hyper parameters θ. Described throughout as standard or naive GP. The hyperparmeters are given by θ, σ.

2. **LCM GP** Modelling f_t using a multitask GP. This corresponds to using the vector valued kernel:

   $$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \mathbf{A}_0 k_\theta(\mathbf{x}, \mathbf{x}') + \mathbf{A}_1 l_\eta(\mathbf{x}, \mathbf{x}') \tag{59}$$

   Where $\mathbf{A}_t$ is given by:

   $$\mathbf{A}_0 = \begin{bmatrix} \theta_{00}^2 & \rho_0 \\ \rho_0 & \theta_{01}^2 \end{bmatrix}, \mathbf{A}_1 = \begin{bmatrix} \theta_{10}^2 & \rho_1 \\ \rho_1 & \theta_{11}^2 \end{bmatrix}. \tag{60}$$

   CATE differences are then formed as the weighted average between both treatments. So modelled as:

   $$\Delta(\mathbf{x}) = \sum_{t=0}^{1} f_t(\mathbf{x}) \tag{61}$$

   Where f_t is the prediction for task t. For this method the hyper-parameters to learn are $\theta, \eta, \mathbf{A}_0, \mathbf{A}_1, \sigma_0, \sigma_1$.

3. **Our Approach**. We use a multitask Gaussian process given by:

   $$\mathbf{K} = \mathbf{a}\mathbf{a}^\top k_\theta + \mathbf{b}\mathbf{b}^\top l_\eta \tag{62}$$

   $$\mathbf{a} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \tag{63}$$

   $$\mathbf{b} = \begin{bmatrix} \frac{-1}{1-\pi(X)} & \frac{1}{\pi(X)} & 0 \end{bmatrix} \tag{64}$$

   Where the first task models f_0, the second f_1, and the final is the CATE gap, $\Delta(\mathbf{x})$. Using the decomposition:

   $$\tilde{Y} = \tau(X) + \left(\frac{T - \pi(X)}{\pi(X)\,(1 - \pi(X))} \right) \phi(X) + \tilde{\epsilon}$$

   This is equivalent to modelling $\tau(X) \sim \text{GP}(0, k_\theta)$ and $\phi(X) \sim \text{GP}(0, l_\eta)$. The hyper-parameters for this method are $\eta, \theta, \sigma_1, \sigma_0$.

D.2 Simulation Details

For the first experiment, we simulate data as follows

$$X|E = o \sim \mathcal{U}([-3, 3]^d), \;\; X|E = e \sim \mathcal{U}([-1, 1]^d) \tag{65}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{2} \sum_{j=1}^{1} \beta_{i,j}^\top (t^j \odot \mathbf{x}^j) + \epsilon \tag{66}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{2} \sum_{j=1}^{1} \beta_{i,j}^\top (t^j \odot \mathbf{x}^j) + \mathbf{f}_t(\mathbf{x}) + \epsilon \tag{67}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_0), \mathbf{f}_t \sim GP(0, k_{\theta_0}) \tag{68}$$

Where $\beta_{i,j} = 1$ for each i, j and $\sigma_0 = 0.5$

D.3 IHDP details

We simulate covariates following a similar approach to Hussain et al. (2022). For the observational dataset we uniformly sample from the IHDP covariate distribution. For the experimental covariate distribution we sample with weights:

$$w_i = 0.8^{\mathbb{1}\{\text{mother is smoker}\}+\mathbb{1}\{\text{is male}\}} \tag{69}$$

So that the experiment dataset is significantly more likely to include male babies whose mothers are smokers. For the experimental dataset treatment is simulated as $\text{Ber}(p)$. The outcome is then simulated as:

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^j) + \epsilon \tag{70}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^j) + \mathbf{f}_t(\mathbf{x}) + \epsilon \tag{71}$$

$$\epsilon \sim \mathcal{N}(0,\sigma_0), \mathbf{f}_t \sim GP(0, k_{\theta_0}) \tag{72}$$

Where, $\beta_{i,j} = Z_i N_j$ where $Z_i \sim \text{Ber}(0.3)$ and $N_j \sim \mathcal{N}(0,1)$ and $\sigma_0 = 0.5$.

E Additional Results

E.1 Simulated Experiment Additional Results

Dim X	MSE			Coverage			Interval Width		
	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	0.301 ± 0.006	1.21 ± 0.06	0.352 ± 0.007	0.935 ± 0.003	0.858 ± 0.008	0.932 ± 0.003	1.97 ± 0.02	3.22 ± 0.02	2.15 ± 0.02
10	$1.77 + \pm 0.0157$	2.05 ± 0.02	1.91 ± 0.02	$0.785 + \pm 0.005$	0.796 ± 0.007	0.303 ± 0.021	3.31 ± 0.03	3.65 ± 0.04	1.09 ± 0.08
25	2.15 ± 0.02	2.27 ± 0.03	2.02 ± 0.01	0.779 ± 0.004	0.754 ± 0.008	0.128 ± 0.012	3.60 ± 0.02	3.50 ± 0.04	0.463 ± 0.045

Table 3: In distribution results for $n_{\text{exp}} = 1000$ across dimension, average across 200 runs with 95% confidence interval.

Dim X	MSE			Coverage			Interval Width		
	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	2.10 ± 0.05	2.17 ± 0.06	2.31 ± 0.07	0.825 ± 0.010	0.817 ± 0.010	0.997 ± 0.01	3.91 ± 0.00	3.92 ± 0.00	9.65 ± 0.20
10	2.01 ± 0.02	2.05 ± 0.02	2.38 ± 0.02	0.832 ± 0.002	0.829 ± 0.003	0.720 ± 0.041	3.91 ± 0.02	3.91 ± 0.02	3.85 ± 0.15
25	2.01 ± 0.01	2.03 ± 0.01	2.04 ± 0.00	0.832 ± 0.001	0.831 ± 0.002	0.239 ± 0.030	3.92 ± 0.00	3.92 ± 0.00	0.910 ± 0.14

Table 4: Out of distribution results for $n_{\text{exp}} = 1000$ across dimension, average across 200 runs with 95% confidence interval.

Dim X	MSE			Coverage			Interval Width		
	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	0.155 ± 0.002	0.808 ± 0.06	0.290 ± 0.010	0.950 ± 0.002	0.897 ± 0.006	0.892 ± 0.005	1.48 ± 0.01	2.86 ± 0.08	1.76 ± 0.03
10	$1.49 + \pm 0.01$	1.94 ± 0.02	1.69 ± 0.02	$0.807 + \pm 0.003$	0.812 ± 0.006	0.481 ± 0.019	3.17 ± 0.02	3.67 ± 0.04	1.70 ± 0.07
25	2.08 ± 0.01	2.19 ± 0.02	2.01 ± 0.02	0.804 ± 0.002	0.775 ± 0.005	0.104 ± 0.008	3.742 ± 0.01	3.587 ± 0.03	0.373 ± 0.03

Table 5: In distribution results for $n_{\text{exp}} = 2500$ across dimension, average across 200 runs with 95% confidence interval.

Dim X	MSE			Coverage			Interval Width		
	Ours	Standard	LCM	Ours	Standard	LCM	Ours	Standard	LCM
5	2.13 ± 0.11	2.21 ± 0.12	2.76 ± 0.27	0.819 ± 0.011	0.810 ± 0.013	0.999 ± 0.001	3.92 ± 0.00	3.92 ± 0.00	11.42 ± 0.32
10	2.01 ± 0.02	2.03 ± 0.02	2.33 ± 0.02	0.833 ± 0.002	0.831 ± 0.003	0.931 ± 0.011	3.92 ± 0.00	3.92 ± 0.00	5.80 ± 0.18
25	1.99 ± 0.01	2.01 ± 0.02	2.01 ± 0.00	0.834 ± 0.001	0.832 ± 0.002	0.216 ± 0.011	3.92 ± 0.00	3.92 ± 0.00	0.794 ± 0.18

Table 6: Out of distribution results for $n_{\text{exp}} = 2500$ across dimension, average across 200 runs with 95% confidence interval.

E.2 IHDP Experiment Additional Results

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.19 \pm 0.04}$	$\mathbf{0.795 \pm 0.008}$	$\mathbf{2.572 \pm 0.02}$
Naive GP	2.43 ± 0.13	0.752 ± 0.010	3.42 ± 0.03
LCM	1.57 ± 0.05	0.752 ± 0.010	3.21 ± 0.05

Table 7: In of distribution results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.93 \pm 0.04}$	$\mathbf{0.812 \pm 0.008}$	$\mathbf{3.65 \pm 0.02}$
Naive GP	2.27 ± 0.13	0.797 ± 0.010	3.81 ± 0.03
LCM	2.50 ± 0.05	0.961 ± 0.010	8.03 ± 0.05

Table 8: Out of distribution results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.63 \pm 0.04}$	$\mathbf{0.643 \pm 0.008}$	$\mathbf{2.27 \pm 0.02}$
Naive GP	2.59 ± 0.13	0.732 ± 0.010	3.48 ± 0.03
LCM	2.04 ± 0.05	0.931 ± 0.010	6.33 ± 0.05

Table 9: Out of distribution results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{2.07 \pm 0.04}$	$\mathbf{0.778 \pm 0.008}$	$\mathbf{3.58 \pm 0.02}$
Naive GP	2.37 ± 0.13	0.789 ± 0.010	3.82 ± 0.03
LCM	2.58 ± 0.05	0.931 ± 0.010	6.33 ± 0.05

Table 10: Out of distribution results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

E.3 Robustness Results

We now repeat the experiment for the IHDP dataset but we add squared terms to the simulation as follows:

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^j) + \gamma_{i,j}^{\top}(t^j \odot \mathbf{x}^j)^2 + \epsilon \tag{73}$$

$$Y_{|X=\mathbf{x},T=t,E=o} = \sum_{i=0}^{1}\sum_{j=1}^{1} \beta_{i,j}^{\top}(t^j \odot \mathbf{x}^j) + \gamma_{i,j}^{\top}(t^j \odot \mathbf{x}^j)^2 + \mathbf{f}_t(\mathbf{x}) + \epsilon \tag{74}$$

$$\epsilon \sim \mathcal{N}(0, \sigma_0), \mathbf{f}_t \sim GP(0, k_{\theta_0}) \tag{75}$$

We still fit a linear model for $\omega_o(\mathbf{x})$ which ensures that $\Delta(\mathbf{x})$ is not a GP.

Model	MSE	Coverage	Interval Width
Ours	$\mathbf{1.10 \pm 0.03}$	$\mathbf{0.824 \pm 0.03}$	$\mathbf{2.63 \pm 0.02}$
Naive GP	2.13 ± 0.10	0.761 ± 0.01	3.39 ± 0.04
LCM	1.34 ± 0.04	0.832 ± 0.01	2.96 ± 0.04

Table 11: In of distribution results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

Model	**MSE**	**Coverage**	**Interval Width**
Ours	$\mathbf{1.91 \pm 0.03}$	$\mathbf{0.821 \pm 0.003}$	$\mathbf{3.66 \pm 0.05}$
Naive GP	2.15 ± 0.04	0.805 ± 0.004	3.80 ± 0.014
LCM	2.22 ± 0.09	0.832 ± 0.01	7.76 ± 0.27

Table 12: In of distribution results for the IHDP setting described in Section 5.2 with $n_e = 400$ averaged over 100 runs where now the experimental treatment proportion is 0.7.

F Uniform Error Bounds

Finally, we repeat the experiment in Section E.3 but with the uniform error bounds.

n_e	500	1000	5000	10000
Whole Function Coverage	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
Interval width in Distribution	21.3 ± 0.12	14.0 ± 0.01	5.36 ± 0.01	3.76 ± 0.03
Interval width out of distribution	31.7 ± 0.04	30.4 ± 0.04	28.9 ± 0.05	28.6 ± 0.1

Table 13: Uniform error bounds averaged over 100 runs. We vary the sample size to show that the in distribution bounds decrease in width as n_e increases.