Causal-BALD: Deep Bayesian Active Learning of Outcomes to Infer Treatment-Effects from Observational Data

Andrew Jesson* OATML University of Oxford

Panagiotis Tigas* OATML University of Oxford

Joost van Amersfoort OATML University of Oxford

Andreas Kirsch OATML University of Oxford

Uri Shalit Machine Learning and Causal Inference in Healthcare Lab Technion

Yarin Gal OATML University of Oxford

Abstract

Estimating personalized treatment effects from high-dimensional observational data is essential in situations where experimental designs are infeasible, unethical, or expensive. Existing approaches rely on fitting deep models on outcomes observed for treated and control populations. However, when measuring individual outcomes is costly, as is the case of a tumor biopsy, a sample-efficient strategy for acquiring each result is required. Deep Bayesian active learning provides a framework for efficient data acquisition by selecting points with high uncertainty. However, existing methods bias training data acquisition towards regions of non-overlapping support between the treated and control populations. These are not sample-efficient because the treatment effect is not identifiable in such regions. We introduce causal, Bayesian acquisition functions grounded in information theory that bias data acquisition towards regions with overlapping support to maximize sample efficiency for learning personalized treatment effects. We demonstrate the performance of the proposed acquisition strategies on synthetic and semi-synthetic datasets IHDP and CMNIST and their extensions, which aim to simulate common dataset biases and pathologies.

1 Introduction

How will a patient's health be affected by taking a medication [37]? How will a user's question be answered by a search recommendation [34]? We can gain insight into these questions by learning about personalized treatment effects. Estimating personalized treatment effects from observational data is essential when experimental designs are infeasible, unethical, or expensive. Observational data represent a population of individuals described by a set of pre-treatment covariates (age, blood pressure, socioeconomic status), an assigned treatment (medication, no medication), and a post-treatment outcome (severity of migraines). An ideal personalized treatment effect is the difference between the post-treatment outcome if the individual receives treatment and the post-treatment outcome if they do not receive treatment. However, it is impossible to observe both outcomes for an individual; therefore, the difference is estimated between populations instead. In the setting of binary treatments, data belong to either the *treatment group* (individuals that received the treatment) or the *control group* (individuals who did not). The personalized treatment effect is the expected difference

*Equal contribution. Correspondence to {andrew.jesson, panagiotis.tigas}@cs.ox.ac.uk

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

in outcomes between treated and controlled individuals who share the same (or similar) measured covariates; as an illustration, see the difference between the solid lines in Fig. 1 (middle pane).

The use of pre-treatment covariates assembled from high-dimensional, heterogeneous measurements such as medical images and electronic health records is increasing [44]. Deep learning methods have been shown capable of learning personalized treatment effects from such data [42, 43, 21]. However, a problem in deep learning is data efficiency. While modern methods are capable of impressive performance, they need a significant amount of labeled data. Acquiring labeled data can be expensive, requiring specialist knowledge or an invasive procedure to determine the outcome. Therefore, it is desirable to minimize the amount of labeled data needed to obtain a well-performing model. Active learning provides a principled framework to address this concern [8]. In active learning for treatment effects [10, 45, 39] a model is trained on available labeled data consisting of covariates, assigned treatments, and acquired outcomes. The model predictions decide the most informative examples from data comprised of only covariates and treatment indicators. Outcomes are acquired, e.g., by performing a biopsy for the selected patients, and the model is retrained and evaluated. This process repeats until either a satisfactory performance level is achieved or the labeling budget is exhausted.

At first sight, this might seem simple; however, active learning induces biases that result in divergence between the distribution of the acquired training data and the distribution of the pool set data [13]. In the context of learning causal effects, such bias can have both positive and negative consequences. For example, while random acquisition active learning results in an unbiased sample of the training data, we demonstrate how it can lead to over-allocation of resources to the mode of the data at the expense of learning about underrepresented data. Conversely, while biasing acquisitions toward lower density regions of the pool data can be desirable, it can also lead to outcome acquisition for data with unidentifiable treatment effects, which leads to uninformed, potentially harmful, personalized recommendations.

To see how training data bias can benefit treatment effect estimation, consider one difference between experimental and observational data: the treatment assignment mechanism is unavailable for observational data. In observational data, variables that affect treatment assignment (an untestable condition) may be unobserved. Moreover, the relative proportion of treated to controlled individuals varies across different sub-populations of the data. Fig. 1 illustrates the latter point, where there are relatively equal proportions of treated and controlled examples for data in region 3. However, the proportions become less balanced as we move to either the left or the right. In extreme cases, say if a group described by some covariate values were systematically excluded from treatment, the treatment effect for that group *cannot be known* [38]. Fig. 1 illustrates this in region 1, where only controlled examples reside, and in region 5, where only treated cases occur. In the language of causal inference, the necessity of seeing both treated and untreated examples for each subpopulation corresponds to satisfying the overlap (or positivity) assumption (see 2.3). The data available in the pool set limits overlap when treatments cannot be assigned. In this setting, regions 2 and 4 of Fig. 1 are very interesting because while either the treated or control group are underrepresented, there may still be sufficient coverage to estimate treatment effects. D'Amour and Franks [9] have described such regions as having weak overlap. Training data bias towards such regions can benefit treatment effect estimation for underrepresented data by acquiring low-frequency data with sufficient overlap.

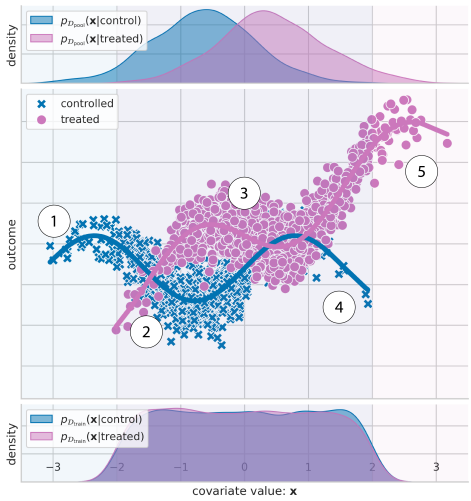


Figure 1: Observational data. Top: data density of treatment (right) and control (left) groups. Middle: observed outcome response for treatment (circles) and control (x's) groups. Bottom: data density for active learned training set after a number of acquisition steps.

We hypothesize that the efficient acquisition of unlabeled data for treatment effect estimation focuses on only exploring regions with sufficient overlap, and uncertainty should be high for areas with non-overlapping support. The bottom pane of Fig. 1 imagines what a resulting training set distribution

could look like at an intermediate active learning step. It is not trivial to design such acquisition functions: naively applying active learning acquisition functions results in suboptimal and sample inefficient acquisitions of training examples, as we show below. To this end, we develop epistemic uncertainty-aware methods for active learning of personalized treatment effects from high dimensional observational data. In contrast to previous work that uses only information gain as the acquisition objective, we propose ρBALD and $\mu\rho$BALD as "Causal BALD" objectives because they consider both the information gain and overlap between treated and control groups. We demonstrate the performance of the proposed acquisition strategies using synthetic and semi-synthetic datasets.

2 Background

2.1 Estimation of Personalized Treatment-Effects

Personalized treatment-effect estimation seeks to know the effect of a treatment $\mathrm{T} \in \mathcal{T}$ on the outcome $\mathrm{Y} \in \mathcal{Y}$ for individuals described by covariates $\mathbf{X} \in \mathcal{X}$. In this work, we consider the random variable (r.v.) T to be binary ($\mathcal{T} = \{0, 1\}$), the r.v. Y to be part of a bounded set $\mathcal{Y}$, and $\mathbf{X}$ to be a multi-variate r.v. of dimension d ($\mathcal{X} = \mathbb{R}^d$). Under the Neyman-Rubin causal model [33, 41], the individual treatment effect (ITE) for a person u is defined as the difference in potential outcomes $\mathrm{Y}^1(u) - \mathrm{Y}^0(u)$, where the r.v. Y^1 represents the potential outcome were they *treated*, and the r.v. Y^0 represents the potential outcome were they *controlled* (not treated). Realizations of the random variables $\mathbf{X}$, T, Y, Y^0, and Y^1 are denoted by $\mathbf{x}$, t, y, y^0, and y^1, respectively.

The ITE is a fundamentally unidentifiable quantity, so instead we look at the expected difference in potential outcomes for individuals described by $\mathbf{X}$, or the Conditional Average Treatment Effect (CATE): $\tau(\mathbf{x}) \equiv \mathbb{E}[\mathrm{Y}^1 - \mathrm{Y}^0 \mid \mathbf{X} = \mathbf{x}]$ [1]. The CATE is identifiable from an observational dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)\}_{i=1}^n$ of samples $(\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i)$ from the joint empirical distribution $P_{\mathcal{D}}(\mathbf{X}, \mathrm{T}, \mathrm{Y}^0, \mathrm{Y}^1)$, under the following three assumptions [41]:

Assumption 2.1. *(Consistency)* $\mathrm{y} = \mathrm{t}\mathrm{y}^{\mathrm{t}} + (1 - \mathrm{t})\mathrm{y}^{1-\mathrm{t}}$*, i.e. an individual's observed outcome* y *given assigned treatment* t *is identical to their potential outcome* y^{t}.

Assumption 2.2. *(Unconfoundedness)* $(\mathrm{Y}^0, \mathrm{Y}^1) \perp\!\!\!\perp \mathrm{T} \mid \mathbf{X}$.

Assumption 2.3. *(Overlap)* $0 < \pi_{\mathrm{t}}(\mathbf{x}) < 1 : \forall \mathrm{t} \in \mathcal{T}$,

where $\pi_{\mathrm{t}}(\mathbf{x}) \equiv \mathrm{P}(\mathrm{T} = \mathrm{t} \mid \mathbf{X} = \mathbf{x})$ is the **propensity for treatment** for individuals described by covariates $\mathbf{X} = \mathbf{x}$. When these assumptions are satisfied, $\widehat{\tau}(\mathbf{x}) \equiv \mathbb{E}[Y \mid \mathrm{T} = 1, \mathbf{X} = \mathbf{x}] - \mathbb{E}[Y \mid \mathrm{T} = 0, \mathbf{X} = \mathbf{x}]$ is an unbiased estimator of $\tau(\mathbf{x})$ and is identifiable from observational data.

A variety of parametric [40, 46, 42] and non-parametric estimators [17, 49, 3, 14] have been proposed for CATE. Here, we focus on parametric estimators for compactness. Parametric CATE estimators assume that outcomes y are generated according to a likelihood $p_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$, given measured covariates $\mathbf{x}$, observed treatment t, and model parameters $\boldsymbol{\omega}$. For continuous outcomes, a Gaussian likelihood can be used: $\mathcal{N}(\mathrm{y} \mid \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \widehat{\sigma}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}))$. For discrete outcomes, a Bernoulli likelihood can be used: $\mathrm{Bern}(\mathrm{y} \mid \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}))$. In both cases, $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$ is a parametric estimator of $\mathbb{E}[Y \mid \mathrm{T} = \mathrm{t}, \mathbf{X} = \mathbf{x}]$, which leads to: $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}) \equiv \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0)$, a parametric CATE estimator.

Jesson et al. [20] have shown that Bayesian inference over the model parameters $\boldsymbol{\omega}$, treated as stochastic instances of the random variable $\boldsymbol{\Omega} \in \mathcal{W}$, yields a model capable of quantifying when assumption 2.3 (overlap) does not hold. Moreover, they show that such models can quantify when there is insufficient knowledge about the treatment effect $\tau(\mathbf{x})$ because the observed value $\mathbf{x}$ lies far from the support of $P_{\mathcal{D}}(\mathbf{X}, \mathrm{T}, \mathrm{Y}^0, \mathrm{Y}^1)$. Such methods seek to enable sampling from the posterior distribution of the model parameters given the data, $p(\boldsymbol{\Omega} \mid \mathcal{D})$. Each sample, $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D})$ induces a unique CATE function $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x})$. Jesson et al. [20] propose $\mathrm{Var}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D})}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0))$ as a measure of epistemic uncertainty (i.e., how much the functions "disagree" with one another at a given value $\mathbf{x}$) [23] for the CATE estimator.

2.2 Active Learning

Formally, an active learning setup consists of an unlabeled dataset $\mathcal{D}_{\mathrm{pool}} = \{\mathbf{x}_i\}_{i=1}^{n_{\mathrm{pool}}}$, a labeled training set $\mathcal{D}_{\mathrm{train}} = \{\mathbf{x}_i, \mathrm{y}_i\}_{i=1}^{n_{\mathrm{train}}}$, and a predictive model with likelihood $p_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x})$ parameterized by $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\mathrm{train}})$. The setup also assumes that an oracle exists to provide outcomes y for any

data point in $\mathcal{D}_{\text{pool}}$. After model training, a batch of data $\{\mathbf{x}_i^*\}_{i=1}^b$ is selected from $\mathcal{D}_{\text{pool}}$ using an acquisition function a according to the informativeness of the batch.

By including the treatment, we depart from the standard active learning setting. For active learning of treatment effects, we define $\mathcal{D}_{\text{pool}} = \{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^{n_{\text{pool}}}$, a labeled training set $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^{n_{\text{train}}}$, and a predictive model with likelihood $p_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$ parameterized by $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})$. The acquisition function takes as input $\mathcal{D}_{\text{pool}}$ and returns a batch of data $\{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^b$ which are labelled using an oracle and added to $\mathcal{D}_{\text{train}}$. We are specifically examining the case when there is access to only the treatments observed in the pool data $\{\mathrm{t}_i\}_{i=1}^{n_{\text{pool}}}$: i.e., scenarios where treatment assignment is not possible.

An intuitive way to define informativeness is using the estimated uncertainty of our model. In general, we can distinguish two sources of uncertainty: epistemic and aleatoric uncertainty [11, 23]. Epistemic (or model) uncertainty, arises from ignorance about the model parameters. For example, this is caused by the model not seeing similar data points during training, so it is unclear what the correct label would be. We focus on using epistemic uncertainty to identify the most informative points for label acquisition.

Bayesian Active Learning by Disagreement (BALD) [18] defines an acquisition function based on epistemic uncertainty. Specifically, it uses the mutual information (MI) between the unknown output and model parameters as a measure of disagreement:

$$\mathrm{I}(\mathrm{Y}; \boldsymbol{\Omega} \mid \mathbf{x}, \mathcal{D}_{\text{train}}) = \mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \mathcal{D}_{\text{train}}) - \mathbb{E}_{\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})} \left[\mathrm{H}(\mathrm{Y} \mid \mathbf{x}, \boldsymbol{\omega}) \right], \tag{1}$$

where H is the entropy function; a straightforward estimand for discrete outcomes with Bernoulli or Categorical likelihoods.

The general acquisition function based on BALD for acquiring a batch of data points given the pool dataset and the model parameters is given by the joint mutual information between the set $\{\mathrm{Y}_i\}$ and the model parameters [26]:

$$a_{\text{BALD}}(\mathcal{D}_{\text{pool}}, p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})) = \underset{\{\mathbf{x}_i\}_{i=1}^b \subseteq \mathcal{D}_{\text{pool}}}{\arg\max} \; \mathrm{I}(\{\mathrm{Y}_i\}; \boldsymbol{\Omega} \mid \{\mathbf{x}_i\}, \mathcal{D}_{\text{train}}). \tag{2}$$

This batch acquisition function can be upper-bounded by scoring each point in $\mathcal{D}_{\text{pool}}$ independently and taking the top b; however, this bound ignores correlations between the samples. In fact, for datasets with significant repetition, this approach can perform worse than random acquisition, and computing the joint mutual information (introduced as *BatchBALD*) rectifies the issue [26].

Estimating the joint mutual information is computationally expensive, as evaluating the joint entropy over all possible outcomes (for classification) or a covariance matrix over all inputs (for regression) is required. An alternative approach is to use softmax-BALD, which involves importance weighted sampling across $\mathcal{D}_{\text{pool}}$ with the individual importance weights given by BALD [25]. We use softmax-BALD for batch acquisition because it is computationally more efficient and performs competitively with BatchBALD. We discuss how BALD maps onto epistemic uncertainty quantification in CATE and the arising complications stemming from the question of overlap in Section 3.

3 Methods

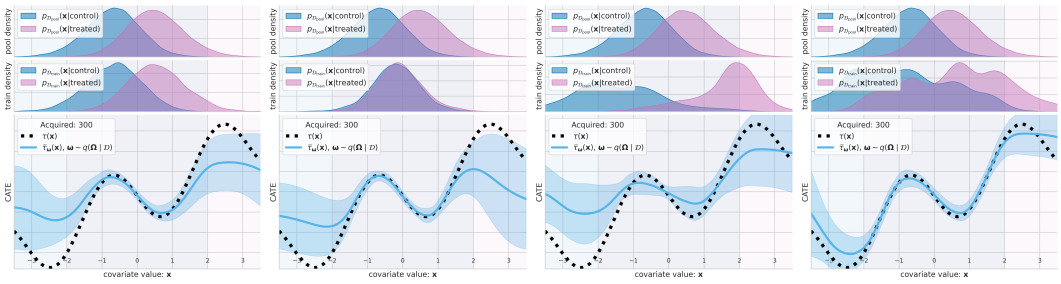
In this section: we introduce several acquisition functions, we then analyze how they bias the acquisition of training data, and we show the resulting CATE functions learned from such training data. We are interested in acquisition functions conditioned on realizations of both $\mathbf{x}$ and t:

$$a(\mathcal{D}_{\text{pool}}, p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})) = \underset{\{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^b \subseteq \mathcal{D}_{\text{pool}}}{\arg\max} \; \mathrm{I}(\bullet \mid \{\mathbf{x}_i, \mathrm{t}_i\}, \mathcal{D}_{\text{train}}), \tag{3}$$

where $\mathrm{I}(\bullet \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}})$ is a measure of disagreement between parametric function predictions given $\mathbf{x}$ and t over samples $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D})$. We make assumptions 2.1 and 2.2 (consistency, and unconfoundedness). We relax assumption 2.3 (overlap) by allowing for its violation over subsets of the support of $\mathcal{D}_{\text{pool}}$. We present all theorems, proofs, and detailed assumptions in Appendix A.

3.1 How do naive acquisition functions bias the training data?

To motivate Causal–BALD, we first look at a set of naive acquisition functions. A random acquisition function selects data points uniformly at random from $\mathcal{D}_{\text{pool}}$ and adds them to $\mathcal{D}_{\text{train}}$. In Fig. 2a we



(a) Random (b) Propensity (c) τBALD (d) μBALD

Figure 2: Naive acquisition functions: How the training set is biased and how this effects the CATE function with a fixed budget of 300 acquired points.

have acquired 300 such examples from a synthetic dataset and trained a deep-kernel Gaussian process [47] on those labeled examples. Comparing the top two panels, we see that $\mathcal{D}_{\text{train}}$ (middle) contains an unbiased sample of the data in $\mathcal{D}_{\text{pool}}$ (top). However, in the bottom panel, we see that while the CATE estimator is accurate and confident near the modes of $\mathcal{D}_{\text{pool}}$, it becomes less accurate as we move to lower-density regions. In this way, the random acquisition of data reflects the biases inherent in $\mathcal{D}_{\text{pool}}$ and over-allocates resources to the modes of the distribution. If the mode were to coincide with a region of non-overlap, the function would most frequently acquire uninformative examples.

Next, we look at using the propensity score to bias data acquisition toward regions where the overlap assumption is satisfied.

Definition 3.1. *Counterfactual Propensity Acquisition*

$$\mathrm{I}(\widehat{\pi}_{\mathrm{t}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \equiv 1 - \widehat{\pi}_{\mathrm{t}}(\mathbf{x}) \tag{4}$$

Intuitively, this function prefers points where the propensity for observing the counterfactual is high. We are considering the setup where $\mathcal{D}_{\text{pool}}$ contains observations of both $\mathbf{X}$ and T, so it is straightforward to train an estimator for the propensity, $\widehat{\pi}_{\mathrm{t}}(\mathbf{x})$. Figure 2b shows that while propensity score acquisition matches the treated and control densities in the train set, it still biases data selection towards the modes of $\mathcal{D}_{\text{pool}}$.

The goal of BALD is to acquire data $(\mathbf{x}, \mathrm{t})$ that maximally reduces uncertainty in the model parameters $\mathbf{\Omega}$ used to predict the treatment effect. The most direct way to apply BALD is to use our uncertainty over the predicted treatment effect, expressed using the following information theoretic quantity:

Definition 3.2. *τBALD*

$$\mathrm{I}(\mathrm{Y}^1 - \mathrm{Y}^0; \mathbf{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \approx \operatorname*{Var}_{\boldsymbol{\omega} \sim p(\mathbf{\Omega} \mid \mathcal{D}_{\text{train}})} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 1) - \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, 0)\right). \tag{5}$$

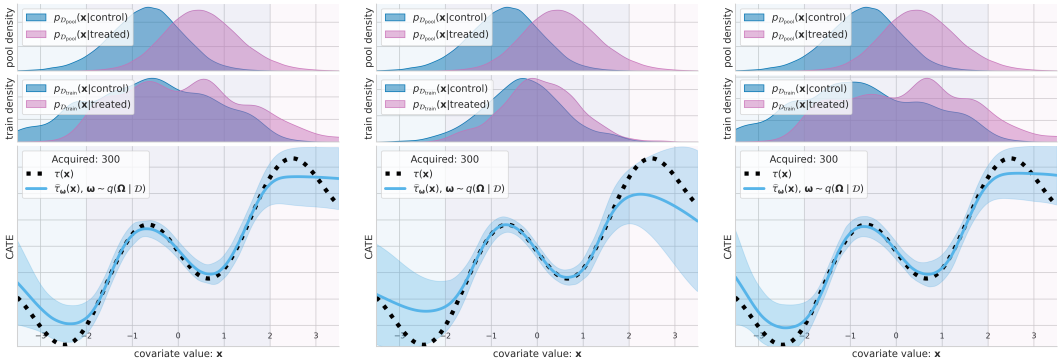
Building off the result in [20], we show how the LHS measure about the *unobservable potential outcomes* is estimated by the variance over $\mathbf{\Omega}$ of the *identifiable difference in expected outcomes* in Theorem 1 of the appendix. Alaa and van der Schaar [4] propose a similar result is for non-parametric models. Intuitively, this measure represents the information gain for $\mathbf{\Omega}$ if we observe the difference in potential outcomes $\mathrm{Y}^1 - \mathrm{Y}^0$ for a given measurement $\mathbf{x}$ and $\mathcal{D}_{\text{train}}$.

However, a fundamental flaw with this measure exists: observing labels for the random variable $\mathrm{Y}^1 - \mathrm{Y}^0$ is impossible. Thus, τBALD represents an irreducible measure of uncertainty. That is, τBALD will be high if it is uncertain about the label given the unobserved treatment t', regardless of its certainty about the outcome given the factual treatment t, which makes τBALD highest for low-density regions and regions with no overlap. Figure 2c illustrates these consequences. We see the acquisition biases the training data away from the modes of the $\mathcal{D}_{\text{pool}}$, where we cannot know the treatment effect (no overlap). In datasets where we have limited overlap, it leads to uninformative acquisitions.

One remedy to the issues of τBALD is to only focus on reducible uncertainty:

Definition 3.3. *μBALD*

$$\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \mathbf{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \approx \operatorname*{Var}_{\boldsymbol{\omega} \sim p(\mathbf{\Omega} \mid \mathcal{D}_{\text{train}})} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})\right). \tag{6}$$



(a) $\mu\pi$BALD (b) ρBALD (c) $\mu\rho$BALD

Figure 3: Causal–BALD acquisition functions: How the training set is biased and how this effects the CATE function with a fixed budget of 300 acquired points.

This measure represents the information gain for the model parameters $\mathbf{\Omega}$ if we obtain a label for the observed potential outcome Y^{t} given a data point $(\mathbf{x}, \mathrm{t})$ and $\mathcal{D}_{\text{train}}$. We give proof for these results in Theorem 2 of the appendix.

μBALD only contains observable quantities; however, it does not account for our belief about the counterfactual outcome. As illustrated in Fig. 2d, this approach can prefer acquiring $(\mathbf{x}, \mathrm{t})$ when we are also very uncertain about $(\mathbf{x}, \mathrm{t}')$, even if $(\mathbf{x}, \mathrm{t}')$ is not in $\mathcal{D}_{\text{pool}}$. Since we can neither reduce uncertainty over such $(\mathbf{x}, \mathrm{t}')$ nor know the treatment effect, the acquisition function would not be optimally data efficient.

3.2 Causal–BALD.

In the previous section, we looked at naive methods that either considered overlap or considered information gain. In this section, we present three measures that account for both factors when choosing a new point to acquire for model training.

A straightforward to combine knowledge about a data point's information gain and overlap is to simply multiply μBALD(6) by the propensity acquisition term (4):

Definition 3.4. *$\mu\pi$BALD*

$$\mathrm{I}(\mu\pi \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \equiv (1 - \widehat{\pi}_{\mathrm{t}}(\mathbf{x})) \underset{\boldsymbol{\omega} \sim p(\mathbf{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} (\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})) \tag{7}$$

We can see in Fig. 3a that the acquisition of training data results in matched sampling that we saw for propensity acquisition in Fig. 2b. However, the tails of the overlapping distributions extend further into the low-density regions of the pool set support where the overlap assumption is satisfied.

Alternatively, we can take an information-theoretic approach to combine knowledge about a data point's information gain and overlap. Let $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})$ be an instance of the random variable $\widehat{\mu}^{\mathrm{t}}_{\mathbf{\Omega}} \in \mathbb{R}$ corresponding to the expected outcome conditioned on t. Further, let $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x})$ be an instance of the random variable $\widehat{\tau}_{\mathbf{\Omega}} = \widehat{\mu}^{1}_{\mathbf{\Omega}} - \widehat{\mu}^{0}_{\mathbf{\Omega}}$ corresponding to the CATE. Then,

Definition 3.5. *ρBALD*

$$\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \widehat{\tau}_{\mathbf{\Omega}} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \gtrapprox \frac{1}{2} \log \left(\frac{\mathrm{Var}_{\boldsymbol{\omega}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}) \right) - 2 \, \mathrm{Cov}_{\boldsymbol{\omega}} (\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}{\mathrm{Var}_{\boldsymbol{\omega}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}') \right)} + 1 \right). \tag{8}$$

This measure represents the information gain for the CATE $\tau_{\mathbf{\Omega}}$ if we observe the outcome Y for a datapoint $(\mathbf{x}, \mathrm{t})$ and the data we have trained on $\mathcal{D}_{\text{train}}$. We give proof for this result in Theorem 3.

In contrast to μ-BALD, this measure accounts for overlap in two ways. First, ρ–BALD will be scaled by the inverse of the variance of the expected counterfactual outcome $\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}')$. This scaling biases acquisition towards examples for which we know about the counterfactual outcome, so we can assume that overlap is satisfied for observed $(\mathbf{x}, \mathrm{t})$. Second, ρ–BALD is discounted by $\mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))$. This discounting is a concept that we will leave for future discussion.

In Fig. 3b we see that ρ–BALD has matched the distributions of the treated and control groups similarly to propensity acquisition in Fig. 2b. Further, we see that the CATE estimator is more accurate over the support of the data.

There is, however, a shortcoming of ρ–BALD that may lead to suboptimal data efficiency. Consider two examples in $\mathcal{D}_{\text{pool}}$, $(\mathbf{x}_1, \mathrm{t}_1)$ and $(\mathbf{x}_2, \mathrm{t}_2)$ where $\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1)) = \mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1'))$ and $\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2)) = \mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2'))$: for each point, we are as uncertain about the conditional expectation given the factual treatment as we are uncertain given the counterfactual treatment. Further, let $\mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1')) = \mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2'))$. Finally, let $\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1)) > \mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2))$: we are more uncertain about the conditional expectation given the factual treatment for data point $(\mathbf{x}_1, \mathrm{t}_1)$ than we are for data point $(\mathbf{x}_2, \mathrm{t}_2)$. Under these three conditions, ρ–BALD would rank these two points equally, and so this method would bias training data to the modes of $\mathcal{D}_{\text{pool}}$ when $(\mathbf{x}_2, \mathrm{t}_2)$ is more frequent than $(\mathbf{x}_1, \mathrm{t}_1)$. In practice, it may be more data-efficient to choose $(\mathbf{x}_1, \mathrm{t}_1)$ over $(\mathbf{x}_2, \mathrm{t}_2)$ as it would more likely be a point as yet unseen by the model.

To combine the positive attributes of μ–BALD and ρ–BALD, while mitigating their shortcomings, we introduce $\mu\rho$BALD.

Definition 3.6. *$\mu\rho$BALD*

$$\mathrm{I}(\mu\rho \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \equiv \underset{\boldsymbol{\omega}}{\mathrm{Var}}\left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})\right) \frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}))}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}. \tag{9}$$

Here, we scale Equation 8, which has equivalent expression $\frac{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x}))}{\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t}'))}$ by our measure for μBALD such that in the cases where the ratio may be equal, there is a preference for data points the current model is more uncertain about. We can see in Fig. 3c that the training data acquisition is distributed more uniformly over the support of the pool data where the overlap assumption is satisfied. Furthermore, the accuracy of the CATE estimator is highest over that region.

4 Related Work

Deng et al. [10] propose the use of Active Learning for recruiting patients to assign treatments that will reduce the uncertainty of an Individual Treatment Effect model. However, their setting is different from ours – we assume that suggesting treatments are too risky or even potentially lethal. Instead, we acquire patients to reveal their outcome (e.g., by having a biopsy). Additionally, although their method uses predictive uncertainty to identify which patients to recruit, it does not disentangle the sources of uncertainty; therefore, it will also recruit patients with high outcome variance. Closer to our proposal is the work from Sundin et al. [45]. They propose using a Gaussian process (GP) to model the individual treatment effect and use the expected information gain over the S-type error rate, defined as the error in predicting the sign of the CATE, as their acquisition function. Although GPs are suitable for quantifying uncertainty, they do not work well on high-dimensional input spaces. In this work, we use Neural network methods to obtain uncertainty: Deep Ensembles [28] and DUE [47], a Deep Kernel Learning GP, both of which work well even on high dimensional inputs. Additionally, the authors assume that noisy observations about the counterfactual treatments are available at training time where we make no such assumptions. We compare to this in our experiment by limiting the access to counterfactual observations (γ baseline) and adapting it to Deep Ensembles [28] and DUE [47] (we provide more details about the adaptation in Appendix B.1). Recent work by Qin et al. [39] looks at budgeted heterogeneous effect estimation but does not factor weak or limited overlap into their acquisition function.

5 Experiments

In this section, we evaluate our acquisition objectives on synthetic and semi-synthetic datasets. Code to reproduce these experiments is available at https://github.com/anndvision/causal-bald.

Datasets Starting from the hypothesis that different objectives can target different types of imbalances and degrees overlap, we construct a **synthetic** dataset [22] demonstrating the various biases. We depict this dataset graphically in Fig. 1. We use this dataset primarily for illustrative purposes. By design, we have constructed a primary data mode and have regions of weak or no overlap. Additionally, we study the performance of our acquisition functions on the **IHDP** dataset [17, 42], which is a standard benchmark in causal treatment effect literature. Finally, we demonstrate that our method is suitable for high dimensional, large-sample datasets on **CMNIST** [21], an MNIST [29] based dataset adapted for causal treatment effect studies. In Fig. 4, we see that CMNIST is an adaptation of the synthetic dataset. Model inputs are MNIST digits and assigned treatments, and the response surfaces are generated based on a projection of the digits onto a latent 1-dimensional manifold. The observed digits are high-dimensional proxies for the confounding covariate ϕ. Detailed descriptions of each dataset are available in Appendix C.

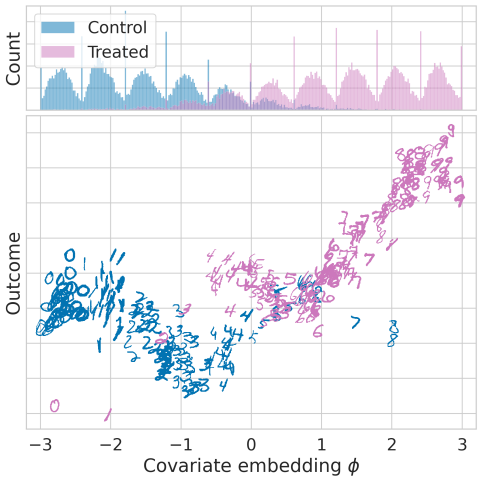


Figure 4: Visualizing CMNIST dataset. Model inputs are MNIST digits and assigned treatments. The MNIST digits are high-dimensional proxies for the latent confounding covariate ϕ. Digits are projected onto ϕ by ordering them first by image intensity and then by digit class (0 - 9). Methods must be able to implicitly learn this non-linear mapping in order to predict the conditional expected outcomes.

Model Our objectives rely on methods that are capable of modeling uncertainty and handling high-dimensional data modalities. DUE [47] is an instance of Deep Kernel Learning [48] that uses a deep feature extractor to transform the inputs and defines a Gaussian process (GP) kernel over the extracted feature representation. In particular, DUE uses a variational inducing point approximation [16] and a constrained feature extractor that contains residual connections and spectral normalization to enable reliable uncertainty. Due obtains SotA results on IHDP [47]. In DUE, we distinguish between the model parameters θ and the variational parameters $\boldsymbol{\omega}$, and we are Bayesian only over the $\boldsymbol{\omega}$ parameters. Since DUE is a GP, we obtain a full Gaussian posterior over outcomes from which we can use the mean and covariance directly. When necessary, sampling is very efficient and only requires a single forward pass in the deep model. We describe all hyperparameters in Appendix F.

Baselines We compare against the following baselines: **Random.** This acquisition function selects points uniformly at random. **Propensity.** An acquisition function based on the propensity score (Eq. 4). We train a propensity model on the pool data, which we then use to acquire points based on their propensity score. Please note that this is a valid assumption as training a propensity model does not require outcomes. **γ (S-type error rate) [45].** This acquisition function is the S-type error rate based method proposed by Sundin et al. [45]. We have adapted the acquisition function to use with Bayesian Deep Neural Networks. The objective is defined as $\mathrm{I}(\boldsymbol{\gamma}; \boldsymbol{\Omega} \mid \mathbf{x}, \mathcal{D}_{\text{train}})$, where $\gamma(x) = \text{probit}^{-1}\left(-\frac{|\mathbf{E}_{p(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}[\tau]|}{\sqrt{\text{Var}(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}}\right)$ and $\text{probit}^{-1}(\cdot)$ is the cumulative distribution function of normal distribution. In contrast to the original formulation, we do not assume access to counterfactual observations at training time.

5.1 Experimental Results

For each of the acquisition objectives, dataset, and model we present the mean and standard error of empirical square root of precision in estimation of heterogenous effect (PEHE) [2]. We summarize

[2] $\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{N}\sum_x (\hat{\tau}(x) - \tau(x))^2}$

Table 1: Summary of active learning parameters for each dataset.

Dataset	Warm-up size	Acquisition size	Acquisition steps	Pool Size	Valid Size
Synthetic	10	10	30	10k	1k
IHDP	100	10	38	471	201
CMNIST	250	50	55	35k	15k

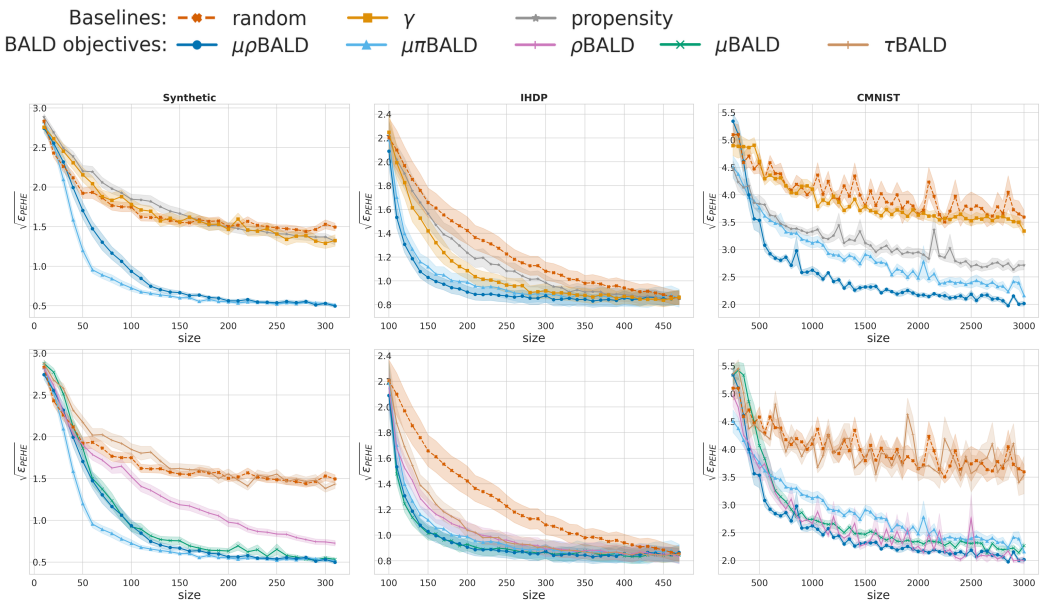


Figure 5: $\sqrt{\epsilon_{PEHE}}$ performance (shaded standard error) for DUE models. **(left to right) synthetic** (40 seeds), and **IHDP** (200 seeds). We observe that BALD objectives outperform the **random**, γ and **propensity** acquisition functions significantly, suggesting that epistemic uncertainty aware methods that target reducible uncertainty can be more sample efficient.

each active learning setup in Table 1. The *warm up size* is the number of examples in the initial pool dataset. *Acquisition size* is the number of examples labeled at each acquisition step. *Acquisition steps* is the number of times we query a batch of labels. *Pool size* is the number of examples in the pool dataset. Finally, *valid size* is the number of examples used for model selection when optimizing the model at each acquisition step.

In Fig. 5, we see that epistemic uncertainty aware $\mu\rho$BALD outperforms the baselines, random, propensity, and S-Type error rate (γ). As analyzed in section 3, we expect this improvement as our acquisition objectives target reducible uncertainty – that is, epistemic uncertainty when there is overlap between treatment and control. Additionally, $\mu\rho$BALDshows superior performance over the other objectives in the high dimensional dataset CMNIST verifying our qualitative analysis in Figure 3c.

Each of (μBALD, ρBALD, $\mu\pi$BALD, and $\mu\rho$BALD) outperform the baseline methods on these tasks. Of note, the performance ρBALD improves as the dimensionality of the covariates increases. In contrast, the performance of the propensity score-based $\mu\pi$BALD worsens as the dimensionality of the covariates increases. Propensity score estimation is known to be a problem in high-dimensions [12]. We see that both μBALD and $\mu\rho$BALD perform consistently as dimensionality increases, with $\mu\rho$BALD showing a statistically significant improvement over μBALD on two of the three tasks. These improvements indicate that $\mu\rho$BALD is more robust for data with high-dimensional covariates than $\mu\pi$BALD ; moreover, $\mu\rho$BALD does not need an additional propensity score model.

6 Conclusion

We have introduced a new acquisition function for active learning of individual-level causal-treatment effects from high dimensional observational data, based on Bayesian Active Learning by Disagreement [18]. We derive our proposed method from an information-theoretic perspective and compare it with acquisition strategies that either do not consider epistemic uncertainty (i.e., random or propensity-based) or target irreducible uncertainty in the observational setting (i.e., when we do not have access to counterfactual observations). We show that our methods significantly outperform baselines, while also studying the various properties of each of our proposed objectives in both quantitative and qualitative analyses, potentially impacting areas like healthcare where sample efficiency in the acquisition of new examples implies improved safety and reductions in costs.

7 Broader Impact

Active Learning for learning treatment effects from observational data is highly actionable research, and there are several sectors where our research can have an impact. Take, for example, a hospital that needs to decide whom to treat, based on some model. To these choices, the decision-maker needs to have a confident and accurate treatment effect prediction model. However, improving the performance of such a model requires data from patients, which might be costly and perhaps even unethical to acquire. With this work, we assume that we cannot assign new treatments to patients but only perform biopsies or questionnaires post-treatment to reveal the outcome. We believe that this is an impactful and realistic scenario that will directly benefit from our proposal. However, our method can also impact fields like computational-advertisement, where the goal is to learn a model to predict the captivate the attention of users, or policymaking where a government wants to decide how to intervene for beneficial or malicious reasons.

Active learning inherently biases the acquisition of training data. We attempt to show how this can be beneficial or detrimental for learning treatment effects under different acquisition functions under the unconfoundedness assumption. We are not making guarantees on the overall unbiasedness of our methods. We guarantee only that our results are conditional on the unconfoundedness assumption. Unobserved confounding can result in the biased sampling of training data concerning the hidden confounding variable. This bias can result in performance inequality between groups and a biased estimate of the unconfounded CATE function. Further, the models' uncertainty estimates are not informative of when this may occur.

An anonymous reviewer writes, "one risk is that the method could, e.g., lead to a biased, non-representative sampling in terms of ethnicity and other protected attributes - particularly, if the unconfoundedness assumption this work is based on is blindly trusted" Recent work by Andrus et al. [5] does a great job of highlighting the difficulties practitioners face when accounting for algorithmic bias across protected attributes. In such cases, model uncertainty is not enough to identify non-representative sampling concerning the protected attribute. They report about a practitioner's use of structural causal models in concert with domain expert feedback as a means to inform clients of potential sources of bias. Perhaps such methodology could be used with causal sensitivity analysis for CATE [22, 50, 21] as a way to model beliefs about the protected attribute without observing it.

Impactful avenues for future work include relaxing the unconfoundedness assumption, incorporating beliefs about hidden confounding into the acquisition function. Furthermore, in addition to uncovering the outcome, we think it is interesting to revisit the active treatment assignment problem. On the active learning side, exploring more rigorous batch acquisition methods could yield improvements over the current stochastic sampling estimation we use. Finally, in this work, we assume access to a validation set, which may not be available in practice, so exploration of active acquisition of validation data will also have an impact [27].

Acknowledgments and Disclosure of Funding

We would like to thank OATML group members and all anonymous reviewers for sharing their valuable feedback and insights. PT and AK are supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems (grant reference EP/L015897/1). JvA is grateful for funding by the EPSRC (grant reference EP/N509711/1) and by Google-DeepMind. U.S. was partially supported by the Israel Science Foundation (grant No. 1950/19).

References

[1] Jason Abrevaya, Yu-Chin Hsu, and Robert P Lieli. Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, 33(4):485–505, 2015.

[2] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

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

[4] Ahmed M Alaa and Mihaela van der Schaar. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12 (5):1031–1046, 2018.

[5] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 249–260, 2021.

[6] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/bergstra13.html.

[7] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *International conference on machine learning*, pages 448–456. PMLR, 2016.

[8] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[9] Alexander D'Amour and Alexander Franks. Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*, 2021.

[10] Kun Deng, Joelle Pineau, and Susan Murphy. Active learning for personalizing treatment. In *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL)*, pages 32–39. IEEE, 2011.

[11] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[12] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654, 2021.

[13] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=JiYq3eqTKY.

[14] Zijun Gao and Yanjun Han. Minimax optimal nonparametric estimation of heterogeneous treatment effects. In *Advances in Neural Information Processing Systems*, 2020.

[15] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

[16] James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

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

[18] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *stat*, 1050:24, 2011.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[20] Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, pages 4829–4838. PMLR, 2021.

[22] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2281–2290. PMLR, 2019.

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30:5574–5584, 2017.

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[25] Andreas Kirsch. PowerEvaluationBALD: Efficient evaluation-oriented deep (bayesian) active learning with stochastic acquisition functions. *arXiv preprint arXiv:2101.03552*, 2021.

[26] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32:7026–7037, 2019.

[27] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. *arXiv preprint arXiv:2103.05331*, 2021.

[28] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[29] Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[30] Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

[31] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[32] Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A distributed framework for emerging ai applications, 2018.

[33] Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

[34] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

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

[37] Caroline Criado Perez. *Invisible women: Exposing data bias in a world designed for men*. Random House, 2019.

[38] Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

[39] Tian Qin, Tian-Zuo Wang, and Zhi-Hua Zhou. Budgeted heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, pages 8693–8702. PMLR, 2021.

[40] James M Robins, Miguel Angel Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):551, 2000.

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

[42] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

[43] Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 32:2507–2517, 2019.

[44] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 12(3):e1001779, 2015.

[45] Iiris Sundin, Peter Schulam, Eero Siivola, Aki Vehtari, Suchi Saria, and Samuel Kaski. Active learning for decision-making from imbalanced observational data. In *International Conference on Machine Learning*, pages 6046–6055. PMLR, 2019.

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

[47] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. Improving deterministic uncertainty estimation in deep learning for classification and regression. *arXiv preprint arXiv:2102.11409*, 2021.

[48] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.

[49] Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

[50] Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Section 1
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In Section 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] In Section 1
   (b) Did you include complete proofs of all theoretical results? [Yes] In Section 3 and Appendix A.2, A.3
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In supplemental material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Appendix sections C and F
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In Appendix section E
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] [22]
   (b) Did you mention the license of the assets? [No] All datasets and assets are open source and their license is retrievable from the citation.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Described in Appendix C and code to generate is provided
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] In 5 IHDP section
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]