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

Anonymous Author(s)
Affiliation
Address
email

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, but when measuring the outcome for an individual is costly (e.g. biopsy) a sample efficient strategy for acquiring outcomes is required. Deep Bayesian active learning provides a framework for efficient data acquisition by selecting points with high uncertainty. However, naive application of existing methods selects training data that is biased toward regions where the treatment effect cannot be identified because there is non-overlapping support between the treated and control populations. To maximize sample efficiency for learning personalized treatment effects, we introduce new acquisition functions grounded in information theory that bias data acquisition towards regions where overlap is satisfied, by combining insights from deep Bayesian active learning and causal inference. 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 [23]? How will a user's question be answered by a search recommendation [22]? Insight into these questions can be gained by learning about personalized treatment effects. Estimating personalized treatment effects from observational data is essential in situations where 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 had the individual been treated and the outcome had they not been treated. But, it is impossible to observe both outcomes for an individual, so the difference must instead be computed between populations. Therefore, in the common setting of binary treatments, data is partitioned into the *treatment group* (individuals that received the treatment) and the *control group* (individuals who did not). The personalized treatment effect is then given by the expected difference 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).

Increasingly, pre-treatment covariates are being assembled from high-dimensional, heterogeneous measurements such as medical images and electronic health records [29]. Deep learning methods have been shown capable of learning personalized treatment effects from such data [27, 28, 13].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

However, a key 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, often 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 [4]. In active learning for treatment effects [5, 30], a model is trained on available labeled data consisting of covariates, assigned treatments, and acquired outcomes. The model predictions are then used to select the most informative examples from a set of data consisting of only covariates and treatment indicators. Outcomes are then acquired, e.g. by performing a biopsy, for the selected patients and the model is retrained and evaluated. This process is repeated 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 bias that results in divergence between the distribution of the acquired training data and the distribution of the pool set data [7]. In the context of learning causal effects, such bias can have important positive and negative consequences. For example, while random acquisition active learning results in an unbiased sample of the training data, 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 acquisitions for which we cannot identify the treatment effect, which could in turn lead to uninformed, potentially harmful, personalized recommendations.

To gain insight into how biasing the acquisition of training data can be beneficial for learning treatment effects, consider a key difference between experimental and observational data: the treatment assignment mechanism is not available for observational data. This means that there may be unobserved variables that affect treatment assignment (an untestable condition), but also that the relative proportion of individuals treated to those controlled can vary across different sub-populations of the data. The later point is illustrated in fig. 1, where there are relatively equal proportions of treated and controlled examples for data in region 3, but 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* [24]. This is illustrated in fig. 1 by region 1, where there are only controlled examples, and by region 5, where there are only treated examples. In the language of causal inference, the necessity of seeing both treated and untreated examples for each sub-population corresponds to satisfaction of the overlap (or positivity) assumption (see 2.3). Regions 2 and 4 of fig. 1 are interesting as while either the treated or control group are underrepresented, there may still be sufficient coverage to learn treatment effects.

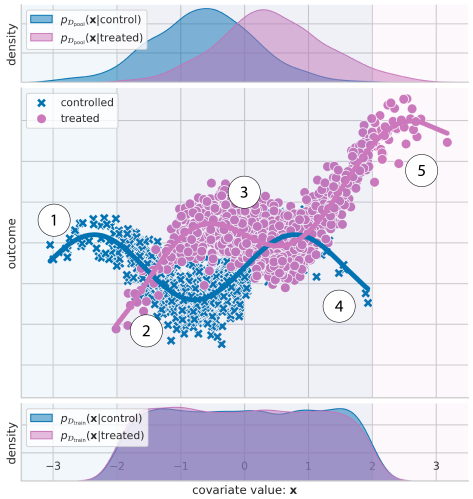


Figure 1: Observational data. Top: data density of treatment (right) and control (left) groups. Middle: observed outcome response for treatment (circles) and control (x's) groups. Bottom: data density for active learned training set after a number of acquisition steps.

We propose that the acquisition of unlabeled data should focus on exploring all regions with sufficient overlap, but not areas with no overlap. 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 an epistemic uncertainty aware method for active learning of personalized treatment effects from high dimensional observational data. We demonstrate the performance of the proposed acquisition strategies on a 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 [21, 26], 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 [26]:

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 [25, 31, 27] and non-parametric estimators [10, 34, 2, 8] 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.

Bayesian inference over the model parameters $\boldsymbol{\omega}$ treated as stochastic instances of the random variable $\boldsymbol{\Omega} \in \mathcal{W}$ has been shown by Jesson et al. [12] to yield models capable of quantifying when assumption 2.3 (overlap) does not hold, or 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 $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x})$. Epistemic uncertainty is a measure of "disagreement" between the functions at a given value $\mathbf{x}$ [16]. Jesson et al. [12] 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 in the CATE.

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 $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\mathrm{train}})$. It is further assumed that an oracle exists to provide outcomes y for any data point in $\mathcal{D}_{\mathrm{pool}}$. After model training, a batch of data $\{\mathbf{x}_i^*\}_{i=1}^b$ is selected from $\mathcal{D}_{\mathrm{pool}}$ using an acquisition function a according to the informativeness of the batch.

We depart from the standard active learning setting and include the treatment: for active learning of treatment effects, we define $\mathcal{D}_{\mathrm{pool}} = \{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^{n_{\mathrm{pool}}}$, a labeled training set $\mathcal{D}_{\mathrm{train}} = \{\mathbf{x}_i, \mathrm{t}_i, \mathrm{y}_i\}_{i=1}^{n_{\mathrm{train}}}$, and a predictive model with likelihood $p_{\boldsymbol{\omega}}(\mathrm{y} \mid \mathbf{x}, \mathrm{t})$ parameterized by $\boldsymbol{\omega} \sim p(\boldsymbol{\Omega} \mid \mathcal{D}_{\mathrm{train}})$. The acquisition function takes as input $\mathcal{D}_{\mathrm{pool}}$ and returns a batch of data $\{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^b$ which are labelled using an oracle and added to $\mathcal{D}_{\mathrm{train}}$. We are specifically examining the case when there is only access to the observed treatments $\{\mathrm{t}_i\}_{i=1}^{n_{\mathrm{pool}}}$: 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 [6, 15]. Epistemic

(or model) uncertainty, arises from uncertainty in the model parameters. This is for example caused by the model not having seen similar data points before, and therefore is unclear what the correct label would be. In this paper, we focus on using epistemic uncertainty to identify informative points to acquire the label for.

Bayesian Active Learning by Disagreement (BALD) [11] 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. For discrete outcomes, where Bernoulli or Categorical likelihoods are used, this is a straightforward quantity to calculate.

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:

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

which is the joint mutual information between the set $\{\mathrm{Y}_i\}$ and the model parameters [18]. This 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 [18]. Estimating the joint mutual information is computationally expensive, as the evaluation of 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 soft-BALD, which involves importance weighted sampling across $\mathcal{D}_{\text{pool}}$ with the individual importance weights given by BALD [17]. This acquisition function is computationally more efficient, and performs competitively with BatchBALD. In this paper, we use soft-BALD for batch acquisition. 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, analyze how they bias the acquisition of training data, and 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:

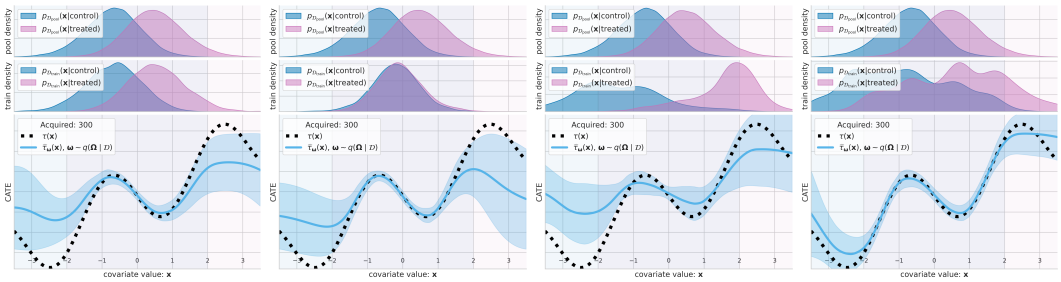
$$a(\mathcal{D}_{\text{pool}}, p(\boldsymbol{\Omega} \mid \mathcal{D}_{\text{train}})) = \underset{\{\mathbf{x}_i, \mathrm{t}_i\}_{i=1}^b \subseteq \mathcal{D}_{\text{pool}}}{\arg\max} \; \mathrm{I}(\bullet \mid \{\mathbf{x}_i, \mathrm{t}_i\}, \mathcal{D}_{\text{train}}), \tag{3}$$

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 naive acquisition functions bias the training data, and the effect on the estimated CATE function.

To motivate Causal–BALD, we first look at a set of naive acquisition functions. The simplest function selects data points uniformly at random from $\mathcal{D}_{\text{pool}}$ and adds them to $\mathcal{D}_{\text{train}}$. In fig. 2a we have acquired 300 such examples from a synthetic dataset and trained a deep-kernel Gaussian process [32] on those labelled examples. Comparing the top two panes, 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 pane we see that while the CATE estimator is accurate (and certain) near the modes of $\mathcal{D}_{\text{pool}}$, it becomes less accurate as we move to lower density regions. In this way random acquisition 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.



(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.

Definition 3.1. *Propensity based 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 acquisition towards the modes of $\mathcal{D}_{\text{pool}}$.

The goal of BALD is to acquire data $(\mathbf{x}, \mathrm{t})$ that maximally reduce 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*

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

Building off the result in [12], we show how the LHS measure about the *unobservable potential outcomes* can be estimated by the variance over $\mathbf{\Omega}$ of the *identifiable difference in expected outcomes* in Theorem 1 of the appendix. A similar result has been proposed for non-parametric models [3]. Intuitively, this measure represents the information gain for $\mathbf{\Omega}$ if we could 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: labels for the random variable $\mathrm{Y}^1 - \mathrm{Y}^0$ are never observed. As a result, it 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 label given the observed 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 do not have full overlap, it leads to uninformative acquisitions.

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

Definition 3.3. *μBALD*

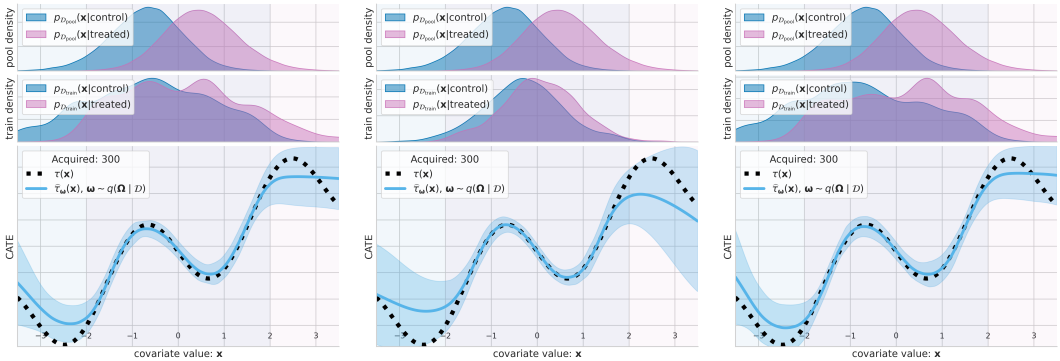
$$\mathrm{I}(\mathrm{Y}^{\mathrm{t}}; \mathbf{\Omega} \mid \mathbf{x}, \mathrm{t}, \mathcal{D}_{\text{train}}) \approx \underset{\boldsymbol{\omega} \sim p(\mathbf{\Omega} \mid \mathcal{D}_{\text{train}})}{\mathrm{Var}} \left(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}, \mathrm{t})\right). \tag{6}$$

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 datapoint $(\mathbf{x}, \mathrm{t})$ and the data we have trained on $\mathcal{D}_{\text{train}}$. Proof for these results are given in Theorem 2 of the appendix.

μBALD only contains observable quantities; however, it does not take into account 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, acquisition would not be 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 take into account both factors when choosing a new point to acquire for model training.



(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.

The most straightforward way 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*

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

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

Alternatively we can take an information theoretic approach to combining 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 $\widehat{\mu}^{\mathrm{t}}_{\boldsymbol{\Omega}} \in \mathbb{R}$ corresponding to the expected outcome conditioned on t. Further, let $\widehat{\tau}_{\boldsymbol{\omega}}(\mathbf{x})$ be an instance of the random variable $\widehat{\tau}_{\boldsymbol{\Omega}} = \widehat{\mu}^{1}_{\boldsymbol{\Omega}} - \widehat{\mu}^{0}_{\boldsymbol{\Omega}}$ corresponding to the CATE. Then,

Definition 3.5. *ρBALD*

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

This measure represents the information gain for the CATE $\tau_{\boldsymbol{\Omega}}$ if we observe the outcome Y for a datapoint $(\mathbf{x}, \mathrm{t})$ and the data we have trained on $\mathcal{D}_{\text{train}}$. Proof for this result is given 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 will bias acquisition towards examples for which we are certain about counterfactual outcome, and 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 is an interesting 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 in a similar manner to propensity acquisition in fig. 2b. Further, we see that the CATE estimator is more accurate over the support of the data. However, there is a shortcoming of ρ–BALD. 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 $\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2)) = \mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}'_2))$. That is, for each point we are as uncertain about the conditional expectation given the observed treatment as we would be given the counterfactual treatment. Further, let $\mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}'_1)) = \mathrm{Cov}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2), \widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}'_2))$ and $\mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_1, \mathrm{t}_1)) > \mathrm{Var}_{\boldsymbol{\omega}}(\widehat{\mu}_{\boldsymbol{\omega}}(\mathbf{x}_2, \mathrm{t}_2))$. In this scenario ρ–BALD would rank these two points equally, but in practice it may be preferable 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. When naively acquiring multiple points per acquisition step, this method biases training data to the modes of $\mathcal{D}_{\text{pool}}$.

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 acquisition of training data examples is more uniformly distributed over the support of the pool data where overlap is satisfied. Furthermore, the accuracy of the CATE estimator is highest over that region.

4 Related Work

Deng et al. [5] 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 than ours – we assume that suggesting treatments are too risky or even potentially lethal and instead we acquire patients for the purpose of revealing their outcome (e.g. by having a biopsy). Additionally, although their method uses the predictive uncertainty to identify which patients to recruit, it does not disentangle the sources of uncertainty and as such treatments with high outcome variance will be recruited as well. Closer to our proposal is the work from Sundin et al. [30], where the authors propose the use of 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 modeling uncertainty they do not work well on high dimensional input spaces. In this work, we use Neural network based methods to obtain uncertainty: Deep Ensembles [19] and DUE [32], a Deep Kernel Learning based GP, which are shown to work well even on high dimensional inputs. Additionally, the authors assume that noisy observations about the counterfactual treatments are available at training time, 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 [19] and DUE [32] (more details about the adaptation is provided in Appendix B).

5 Experiments

In this section we evaluate our acquisition objectives on synthetic and semi-synthetic datasets.

5.1 Models

Our objectives rely on methods that are capable of modeling the uncertainty and handling high-dimensional data modalities. We evaluate two models, Deep Ensembles [19] and DUE [32].

Deep Ensembles. We use Deep Ensembles [19] to disentangle *epistemic* and *aleatoric* uncertainty. Each ensemble component corresponds to a Neural Network parameterising a mixture of Gaussian distributions $q(\mathrm{Y}|\mathbf{x}, \mathbf{t}; \boldsymbol{\omega}) = \sum_{i=1}^{K} \alpha_i(\mathbf{x}, \mathbf{t}; \boldsymbol{\omega})\mathcal{N}(\mathrm{Y}; \mu_i(\mathbf{x}, \mathbf{t}; \boldsymbol{\omega}), \Sigma_i(\mathbf{x}, \mathbf{t}; \boldsymbol{\omega}))$, where $\alpha_i(\cdot; \boldsymbol{\omega})$ are the mixing coefficients, and $\mu_i(\cdot; \boldsymbol{\omega})$ and $\Sigma_i(\cdot; \boldsymbol{\omega})$ the sufficient statistics of the Gaussian distributions. In practice we find that an ensemble of size 5 with 5 mixture components is sufficient. In Appendix F we present the detailed architecture choices and hyper parameters.

DUE. DUE [32] is an instance of Deep Kernel Learning [33], where a deep feature extractor is used to transform the inputs over which a Gaussian process' (GP) kernel is defined. In particular, DUE uses a variational inducing point approximation [9] and a constrained feature extractor which contains residual connections and spectral normalisation to enable reliable uncertainty. It was previously shown to obtain SotA results on IHDP [32]. 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 our outputs 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 hyper parameters in Appendix F.

5.2 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 combination of the train and pool dataset which we then use acquire points based on their propensity score. Please note that this is a valid assumption as training a propensity model does not require outcomes. $\boldsymbol{\gamma}$ **(S-type error rate) [30].** This acquisition function is the S-type error rate based method proposed by Sundin et al. [30]. 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 $\gamma(x) = \text{probit}^{-1}(-\frac{|\mathbf{E}_{p(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}[\tau]|}{\sqrt{\text{Var}(\tau|\mathbf{x},\mathcal{D}_{\text{train}})}})$ 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.

5.3 Datasets

Table 1: Summary of active learning setup per dataset.

Dataset	Warm up size	Acquisition size	Number of Acquisitions
Synthetic	0	100	100
IHDP	100	10	38 (max)
HCMNIST	0	25	20

Starting from the hypothesis that different objectives can target different types of imbalances and overlap ratios we construct a synthetic dataset [14] demonstrating the different biases. Additionally, we study the performance of our acquisition functions on the IHDP dataset [10, 27], a standard benchmark in causal treatment effect literature, and finally we demonstrate that our method is suitable for high dimensional datasets on CMNIST [13], an MNIST [20] based dataset adapted for causal treatment effect studies.

Synthetic Data. We use the one-dimensional simulated dataset introduced by Kallus et al. [14] to study different dataset biases. Figure1 illustrates the response surface and we provide more details in C.1. Treatments are sampled from a Binomial distribution. In contrast to [14], we sample covariates X from a Gaussian distribution $X \sim \mathcal{N}(0, 1)$. We set γ and u to zero as we are not interested in the hidden confounder setting and we apply additive Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$.

IHDP Data. Infant Health and Development Program (IHDP) is a semi-synthetic dataset [10, 27] commonly used in literature to study the performance of causal effect estimation methods. The dataset consists of 747 cases, out of which 139 are assigned in treatment group and 608 in control. Each unit is represented by 25 covariates describing different aspects of the infants and their mothers.

CMNIST Data. Following the setup from [13], we use a simulated dataset based on MNIST [20]. Each digit d is mapped to a latent feature $\phi(d) \in [-2, 2]$ which is then used in the outcome model as defined in the synthetic dataset [14]. Extended description is provided in C.2

5.4 Experimental Results

For each of the acquisition objective, dataset and model we present the mean and standard error of empirical square root of precision in estimation of heterogenous effect (PEHE) $\sqrt{\epsilon_{PEHE}} = \sqrt{\frac{1}{N}\sum_x (\hat{\tau}(x) - \tau(x))^2}$. We summarize in table 1 the active learning setup per dataset.

Table 2: $\sqrt{\epsilon_{PEHE}} \downarrow$ on IHDP dataset after 10 acquisitions. Our proposed methods consistently outperform random acquisition independently of the model. We used 100 seeds for DUE experiments and 50 seeds for the Deep Ensemble as they are computationally more expensive.

	DUE	Deep Ensembles
random	2.09 ± 0.27	2.56 ± 0.32
$\mu\rho$BALD	1.58 ± 0.20	1.59 ± 0.13
$\mu\pi$BALD	$\mathbf{1.48} \pm 0.18$	1.81 ± 0.24

In fig. 4, we see that epistemic uncertainty aware methods, $\mu\rho$BALD and $\mu\pi$BALD outperform the baselines, random, propensity and S-Type error rate ($\boldsymbol{\gamma}$). As we analysed in section 3, this is expected as our acquisition objectives target the type of uncertainty that can be reduced – that is the epistemic uncertainty for which we have overlap between treatment and control. Additionally, $\mu\rho$BALD shows superior performance over the other objectives in the high dimensional dataset CMNIST verifying our qualitative analysis in Figure 3c.

To test the hypothesis that our method is model independent we tested the acquisition objectives in DUE [32], a Deep Kernel Learning method, suitable for assessing uncertainty in high dimensional datasets. In Table 2 we observe that the BALD objectives significantly outperform random baseline.

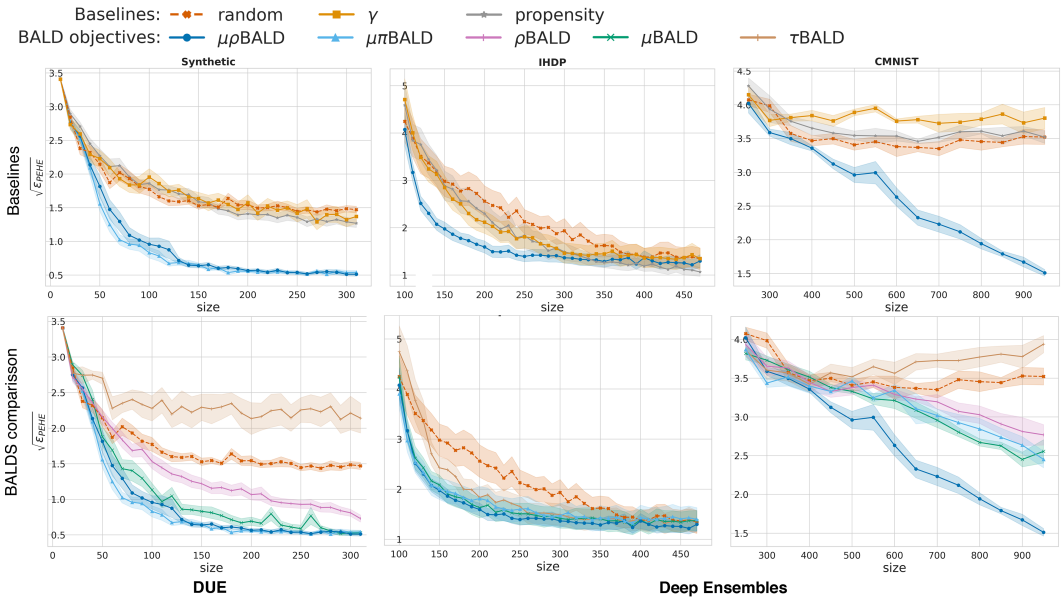


Figure 4: $\sqrt{\epsilon_{PEHE}}$ performance (shaded standard error) for Deep Ensembles based models. **(left to right) synthetic** (20 seeds), **IHDP** (50 seeds) and **CMNIST** (5 seeds) dataset results. For the synthetic dataset we use DUE model and for IHDP and CMNIST, Deep Ensemble method. We present results for both DUE and Deep Ensembles in the Appendix D. 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.

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 [11]. We derive our proposed method from an information theoretic perspective and compared with various acquisition functions that do not take into consideration epistemic uncertainty (like random or propensity based) or they target uncertainties that cannot be reduced in the observational setting (i.e. when we do not have access to counterfactual observations). We show that our methods significantly outperform the baselines while also studying the various properties of each of our proposed objectives in both a quantitative and a qualitative analysis, potentially impacting areas like healthcare where sample efficiency in acquisition of new examples imply improved safety and reductions in costs.

7 Broader Impact

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

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] 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.

[3] 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.

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

[5] 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.

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

[7] 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`.

[8] Zijun Gao and Yanjun Han. Minimax optimal nonparametric estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2002.06471*, 2020.

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

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

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

[12] 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.

[13] Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *arXiv preprint arXiv:2103.04850*, 2021.

[14] 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.

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

[16] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf`.

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

[18] Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

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

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

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

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

[24] 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.

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

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

[27] 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.

[28] Claudia Shi, David M Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *arXiv preprint arXiv:1906.02120*, 2019.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

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

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] [14]
   (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]