Budgeted Heterogeneous Treatment Effect Estimation

Tian Qin [1] **Tian-Zuo Wang** [1] **Zhi-Hua Zhou** [1]

Abstract

Heterogeneous treatment effect (HTE) estimation is receiving increasing interest due to its important applications in fields such as healthcare, economics, and education. Current HTE estimation methods generally assume the existence of abundant observational data, though the acquisition of such data can be costly. In some real scenarios, it is easy to access the pre-treatment covariates and treatment assignments, but expensive to obtain the factual outcomes. To make HTE estimation more practical, in this paper, we examine the problem of estimating HTEs with a budget constraint on observational data, aiming to obtain accurate HTE estimates with limited costs. By deriving an informative generalization bound and connecting to active learning, we propose an effective and efficient method which is validated both theoretically and empirically.

1. Introduction

Treatment effect estimation is of great importance in many applications, such as advertising (Sun et al., 2015; Wang et al., 2015), recommendation (Schnabel et al., 2016), and healthcare (Shalit, 2019). It aims to estimate the effects of a treatment T (e.g., a kind of drug) on an outcome Y (e.g., whether the patient recovers) based on some pre-treatment variables X (e.g., demographic characteristics). Among different levels of treatment effects, we focus on heterogeneous treatment effects (HTEs), which captures treatment effects at a subgroup level and therefore is useful for personalizing treatment plans.

When the treatment is binary (e.g., taking a drug or not), the group of individuals that receives the treatment is called the *treated group*, and others the *control group*. Ideally, analysts can design and conduct randomized controlled trials (RCTs) to estimate treatment effects. In RCTs, the treatment assignment is randomized, so the estimates can be readily obtained. However, RCTs are often costly to run, and sometimes they are even unethical or illegal. To overcome this issue, researchers resort to observational data, which is usually easy to access. Observational data mainly consists of past pre-treatment variables, treatments, and outcomes. Since the generating process of observational data is uncontrolled, there exists a gap between the treated distribution and the control distribution, which hinders accurate estimation of treatment effects. In light of this, methods that try to balance the two distributions were proposed and have achieved impressive performance (Johansson et al., 2016; Shalit et al., 2017; Kallus, 2020).

In some scenarios, however, observational data can be expensive to collect. E.g., to fight against COVID-19, each individual can choose to receive free injection of vaccines provided by the government. Now, the medical department needs to collect factual data such as the antibody level, by providing expensive medical tests, to evaluate the effectiveness of the vaccines. In this case, the amount of potential observational data is huge, but the cost of collecting all factuals is unacceptable. It would be better to obtain accurate HTE estimates by using only a limited amount of factual observational data. Therefore, we focus on estimating HTE with a budget constraint on observational data.

In this problem, given a pool of observational data *without* factual outcomes, we allow an algorithm to select a limited subset of it and pay some costs to obtain the corresponding outcomes, then use them as training data to estimate HTEs. To obtain as accurate estimates as possible using limited training data, we bring in active learning, which is a classical machine learning paradigm. The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to select the data from which it learns (Settles, 2012), which fits in our setting and thus can be helpful.

However, unlike supervised learning tasks, true HTEs are not easily available due to unobserved counterfactuals. As a result, it is impractical to directly apply active learning that mainly focuses on supervised learning tasks. In light of this, we need a criterion that measures the goodness of estimates and also guides the selection of training data. Therefore, we derive a generalization bound for HTE using the concept of

[1]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. Correspondence to: Zhi-Hua Zhou <zhouzh@lamda.nju.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

core-sets (Tsang et al., 2005; Sener & Savarese, 2018). The bound takes into account the representativeness of selected training data, along with empirical regression losses and the discrepancy between the treated and control distribution. Inspired by the theoretical result, we propose a method called QHTE (**Q**uery-based **H**eterogeneous **T**reatment **E**ffect estimation), which tries to minimize the generalization bound using a neural network. Specifically, QHTE alternates between two steps. The first step is to minimizing the empirical losses and distribution discrepancy on current training data. The second step is to actively selecting the potentially most beneficial data to query their outcomes to augment the training set. In this way, QHTE can obtain accurate estimates with very limited yet representative training data.

The main contributions of this paper are threefold. First, we resolve the problem of estimating HTEs with an observational data budget, which is a practical problem but lacks relevant studies. Second, we derive an informative generalization error bound for HTE, which connects the problem with active learning and is instructive for the selection of training data. Finally, we propose a theory-guided method QHTE based on the theoretical analysis. Experiments across three datasets show that our method outperforms baselines given a fixed observational data budget.

The remainder of the paper is structured as follows. Section 2 introduces related work. Section 3 presents background knowledge about HTE estimation and formulates the budgeted HTE estimation problem. Section 4 and 5 describes the generalization bound and the proposed QHTE method, respectively. Section 6 presents empirical evaluation results. Finally, we conclude in Section 7.

2. Related Work

Due to the flexibility and predictive ability of machine learning models, there has been considerable interest in bringing machine learning techniques into HTE estimation, including methods based on Bayesian additive regression trees (Hill, 2011; Hahn et al., 2020), random forests (Wager & Athey, 2018; Athey et al., 2019), neural networks (Shalit et al., 2017; Yao et al., 2018; Yoon et al., 2018), etc. There are also some meta-algorithms (Künzel et al., 2019; Nie & Wager, 2020) that can take advantage of any supervised learning or statistical regression methods to estimate HTEs.

Among various work on HTE estimation, Shalit et al. (2017) proposed a generalization bound for HTE estimation, which is expressed in terms of factual loss and the discrepancy between the treated and control distribution. By minimizing the upper bound, a neural network maps the treated and control data into a common vector space, where the discrepancy between two groups of data is reduced and thus better estimates can be obtained. Our work refers to their generalization bound, but differs in that we consider a new problem of estimating HTEs with a budget, hence our bound is not described in terms of fully available observational data but of selected core-sets, which informs us to actively query the potentially most beneficial data to augment the training set.

Active learning attempts to achieve high accuracy using as few labeled instances as possible by asking queries in the form of unlabeled instances to be labeled by an oracle (e.g., a human annotator) (Settles, 2012). The key difference between active learning methods is the querying criterion. Effective and practical criteria includes informativeness (Roy & McCallum, 2001; Guo & Schuurmans, 2007), representativeness (Yu et al., 2006; Chattopadhyay et al., 2013), and both (Huang et al., 2014; Wang & Ye, 2015), etc.

There are studies considering budgets on various resources that are allowed to be used in the learning process. For example, storage budget (Zhou et al., 2009; Hou et al., 2017), query budget (Huang et al., 2017), and cost budget (Wang & Zhou, 2016), etc. These methods can control the resource consumption before learning, and thus are of great importance in real applications such as online recommendation, edge computing, and stream computing.

The work most relevant to ours is Deng et al. (2011) and Sundin et al. (2019), both of which utilize the concept of active learning to inform decision-making. Our work is very different from theirs. First, their work focuses on decision-making, which only cares about the signs of HTEs rather than exact values. Second, we only query for factual outcomes, which is more practical in real applications. Instead, Deng et al. (2011) selected individuals to conduct experiments to obtain new data, and Sundin et al. (2019) queried for counterfactuals, which are difficult to obtain. Moreover, we start from a generalization bound and propose a theoretically sound while effective and efficient method.

3. Problem Setup

HTE estimation aims to measure the effect of treatment $t \in \mathcal{T}$ on the outcome $y \in \mathcal{Y}$ of a specific subgroup described by $x \in \mathcal{X}$. In this work, we focus on the binary treatment case where $\mathcal{T} = \{0, 1\}$ and let the bounded set $\mathcal{Y}$ denote the set of possible outcomes. A unit is treated if $t = 1$ and controlled if $t = 0$. Under the potential outcome framework (Neyman, 1923; Rubin, 1974), HTE is also known as the conditional average treatment effect, and is defined as:

$$\tau(x) \triangleq \mathbb{E}[Y_1 - Y_0 \mid x],$$

where Y_t denotes the potential outcome for treatment t, i.e., the value that Y would obtain had x received treatment t. The challenge of this task lies in that in real applications, we can only observe at most one factual outcome y_t for a unit, but never the counterfactual outcome y_{1-t}.

Currently, most HTE estimation methods implicitly assume that there exists plenty of fully observable data, which is described as a dataset of (x, t, y) triplets. In contrast to this assumption, we assume that only a dataset of (x, t) pairs is available, and the request for factual outcomes is associated with a cost, which may be related to the covariates and treatments. E.g., in the scenario described in Section 1, some extra medical tests might be necessary for some people considering their health conditions and thus more costs are needed to obtain the factuals. Under this setting, we seek to achieve accurate HTE estimates within a query cost budget.

We formulate the above problem as follows. Let $D = \{(x_i, t_i)\}_{i=1}^m$ denote the training data, where each (x_i, t_i) pair is collected from an observational study. Let $L = \{(x_i, t_i, y_i)\}_{i=1}^m$ denote the corresponding set of full observational data with factual outcomes, which was assumed to be available in previous studies when there was no concern about the query costs. Whereas in this paper, we do not assume it is readily available. Instead, we try to explore it in a cost-effective way. We further write $D = D_0 \cup D_1$, where $D_t = \{(x_i, t_i) \mid (x_i, t_i) \in D \wedge t_i = t\}$, and $L = L_0 \cup L_1$ analogously. Also, we denote the size of D_0 and D_1 by m_0 and m_1, respectively. Let $c : \mathcal{X} \times \mathcal{T} \to \mathbb{R}_+$ be the cost function, meaning that in order to obtain y_i from L, we need to pay a cost of $c(x_i, t_i)$ for all $(x_i, t_i) \in D$. Let $S \subseteq L$ denote a set of data with factual outcomes, and $h_S(x)$ denote the HTE estimate for unit x output by a learning algorithm trained on data S. We can measure the quality of $h_S(x)$ using the expected Precision in Estimation of Heterogeneous Effect (PEHE) loss (Hill, 2011):

$$\epsilon_{\text{PEHE}}(h_S) \triangleq \int_{\mathcal{X}} \left(h_S(x) - \tau(x)\right)^2 p(x)dx.$$

Assume that we have a budget of $B \in \mathbb{R}_+$, then estimating HTE with a budget constraint can be formalized as:

$$\begin{aligned} \min_{S \subseteq L} \quad & \epsilon_{\text{PEHE}}(h_S) \\ \text{s.t.} \quad & \sum_{(x,t,y) \in S} c(x, t) \leq B. \end{aligned} \tag{1}$$

The difficulty of the above optimization problem lies in that we do not have the full observational data L, and cannot access h_S without using S to train a learning algorithm. A practical algorithm for this problem should be able to determine which factual outcomes to query so that the cost does not exceed B and training on them leads to small PEHE loss *before* actually obtaining those outcomes.

For simplicity, in the rest of the paper, we consider the cost function $c(x, t) = 1$ for all $x \in \mathcal{X}$ and $t \in \mathcal{T}$. Then B is equivalent to the maximum number of queries, i.e., the maximum number of training samples that are associated with factual outcomes. The problem in (1) becomes:

$$\begin{aligned} \min_{S \subseteq L} \quad & \epsilon_{\text{PEHE}}(h_S) \\ \text{s.t.} \quad & |S| \leq B. \end{aligned} \tag{2}$$

We make the following common assumptions under the potential outcome framework (Imbens & Rubin, 2015).

Assumption 1 (SUTVA)**.** *The potential outcomes for any unit do not vary with the treatment assigned to other units. For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

Assumption 2 (Consistency)**.** *The potential outcome of treatment t equals the observed outcome if the actual treatment received is t.*

Assumption 3 (Strong ignorability)**.** *Given pre-treatment covariates X, the potential outcome variables are independent of treatment assignment, i.e., $\{Y_0, Y_1\} \perp\!\!\!\perp T \mid X$. For any X, the probability to receive each treatment is positive, i.e., $0 < P(T = t \mid X = x) < 1, \forall t \in \mathcal{T}$ and $x \in \mathcal{X}$.*

With above assumptions, we have $\tau(x) = \mathbb{E}[Y \mid x, T = 1] - \mathbb{E}[Y \mid x, T = 0]$, which only consists of quantities that can be estimated from data, thus HTE is identifiable.

4. Theory

In this section, we derive a generalization bound that is useful to inform an algorithm to select which data points to query. We first present the generalization bound on HTE estimation by Shalit et al. (2017), then generalize their result by introducing the concept of core-sets.

Shalit et al. (2017) discussed one-to-one representation functions of the form $\Phi : \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a representation space. Their method first maps $x \in \mathcal{X}$ to a new space $\mathcal{Z}$ and then estimates HTE by predicting two potential outcomes in $\mathcal{Z}$ and differentiating them. The idea is that the distribution of training data becomes more similar in $\mathcal{Z}$ and thus benefits the estimation. In the following analysis, as in Shalit et al. (2017), we assume a joint distribution $p(x, t, y_0, y_1)$ over $\mathcal{X} \times \mathcal{T} \times \mathcal{Y} \times \mathcal{Y}$. For simplicity, we write $\Phi(S) = \{(\Phi(x), t) \mid (x, t) \in S\}$ for $S \subseteq D$, and $\Phi(S) = \{(\Phi(x), t, y) \mid (x, t, y) \in S\}$ for $S \subseteq L$.

Definition 1. Let $\Phi : \mathcal{X} \to \mathcal{Z}$ be a representation function, $f : \mathcal{Z} \times \{0, 1\} \to \mathcal{Y}$ be a hypothesis predicting the outcome of unit x given treatment t using the mapped covariates $\Phi(x)$. Let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denote a loss function. The expected factual *treated* and *control* losses with respect to

Φ and f are:

$$\epsilon_1(f,\Phi) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} l(y, f(\Phi(x),1))p(x,y \mid T=1)dxdy,$$
$$\epsilon_0(f,\Phi) \triangleq \int_{\mathcal{X}\times\mathcal{Y}} l(y, f(\Phi(x),0))p(x,y \mid T=0)dxdy.$$

Definition 2 (IPM)**.** For two probability density functions p, q defined over $\mathcal{Z} \subseteq \mathbb{R}^d$, and for a function family G of functions $g : \mathcal{Z} \to \mathbb{R}$, the *Integral Probability Metric* is

$$\mathrm{IPM}_{\mathrm{G}}(p,q) \triangleq \sup_{g\in\mathrm{G}} \left| \int_{\mathcal{Z}} g(z)(p(z)-q(z))dz \right|.$$

It can be seen that IPM measures the distance between two distributions. For rich enough function families G, such as the family of 1-Lipschitz functions (Sriperumbudur et al., 2012) and the unit-ball of functions in a reproducing kernel Hilbert space (Gretton et al., 2012), IPM is a true metric over the corresponding set of probabilities. In the rest of the paper, we consider an estimate for HTE in the form of $h(x) = f(\Phi(x),1) - f(\Phi(x),0)$. Next is the generalization bound for $h(x)$ derived in Shalit et al. (2017).

Proposition 1 (Shalit et al. (2017))**.** *Let $\Phi : \mathcal{X} \to \mathcal{Z}$ be a one-to-one representation function. Let $f : \mathcal{Z} \times \{0,1\} \to \mathcal{Y}$ be a hypothesis. Let* G *be a family of functions $g : \mathcal{Z} \to \mathcal{Y}$. Assume that $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is the ℓ_2 loss and that there exists a constant $C_\Phi > 0$, such that for fixed $t \in \{0,1\}$, $\frac{1}{C_\Phi} \cdot \ell_{f,\Phi}(x,t) \in \mathrm{G}$, where $\ell_{f,\Phi}(x,t) \triangleq \int_{\mathcal{Y}} l\left(y_t, f(\Phi(x),t)\right) p\left(y_t \mid x\right) dy_t$. Let $h(x) = f(\Phi(x),1) - f(\Phi(x),0)$ be an estimate. Then,*

$$\epsilon_{\mathrm{PEHE}}(h) \leq 2\left(\epsilon_1(f,\Phi) + \epsilon_0(f,\Phi)\right) + 2C_\Phi \cdot \mathrm{IPM}_{\mathrm{G}}\left(p_1^\Phi, p_0^\Phi\right) - C_Y,$$

where $p_t^\Phi = p(\Phi(x) \mid t)$ and C_Y is a constant related to the expected variance of outcomes.

Proposition 1 bounds the PEHE loss of $h(x)$ with standard regression generalization error on treated and control data and an IPM term measuring the dissimilarity between the treated and control distribution. This bound indicates that we should uniformly sample B data points in D to query their outcomes to solve the problem in (2) since in that way the selected data best mimics the underlying observational data distribution, which is exactly the distribution that ϵ_t is defined on. However, different training data can lead to various gains in prediction performance, which is not utilized by uniform random sampling.

In light of this, we need to guide the selection of query targets in a more refined manner. We bring in the concept of core-sets (Tsang et al., 2005), which can be seen as a representative subset of a dataset. Sener & Savarese (2018) adopted core-sets to formulate an active querying criterion for convolutional neural networks in image classification. However, their approach only applies to classification tasks, and assumes training error is zero, which is reasonable in classification but unrealistic in regression tasks. We derive a new generalization bound for HTE using the concept of core-sets and without the zero training error assumption, which can be used as an effective querying criterion.

Definition 3 (r-cover and core-set)**.** A set Z is a *r-cover* of a set U if

$$U \subseteq \bigcup_{z\in Z} \{u \mid u \in U \wedge \|u - z\| \leq r\}.$$

A r-cover Z is a *core-set* if all of its elements are from the original set U. Note that in the rest of the paper, if the elements of U are (x,t,y) triplets, we ignore the existence of y when calculating the norm $\|\cdot\|$.

Definition 4 (Augmented r-cover)**.** Let set Z be a r-cover of a set U, the *augmented r-cover* w.r.t. Z and U is a multiset and denoted as Z^A. Z^A is constructed by the following procedure: initialize Z^A with $\emptyset$, then for each $u \in U$, randomly choose an element from $\{z \mid z \in Z \wedge \forall z' \in Z, \|u - z\| \leq \|u - z'\|\}$ to join Z^A.

All data points from the original set are covered by a set of balls with radius r centered at each element of the r-cover. Therefore, a r-cover of a dataset can be treated as a representative subset of the whole dataset. An augmented r-cover simply uses nearest neighbors in the r-cover to construct a set that is similar to the original one. Intuitively, a learning algorithm that performs well on an augmented r-cover with a small covering radius r will probably generalize to the original full dataset. We next give some definitions, then state a theorem that bounds the difference between the expected loss on training data and the expected loss on an augmented r-cover, which matches the intuition.

Definition 5. Let $\Phi : \mathcal{X} \to \mathcal{Z}$ be a representation function, $f : \mathcal{Z} \times \{0,1\} \to \mathcal{Y}$ be a hypothesis predicting potential outcomes, $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function. Given $S \subseteq D$, the expected loss on $Z = \Phi(S)$ is

$$\epsilon_Z(f) \triangleq \frac{1}{|Z|} \sum_{(z,t)\in Z} \int_{\mathcal{Y}} p(y \mid z,t) l(y, f(z,t)) dy.$$

Definition 6. Under the conditions of Definition 5, given $S \subseteq L$, the empirical loss on $Z = \Phi(S)$ is

$$\hat{\epsilon}_Z(f) \triangleq \frac{1}{|Z|} \sum_{(z,t,y)\in Z} l(y, f(z,t)).$$

It can be seen that $\epsilon_Z(f)$ measures the expected loss on a mapped empirical distribution induced by Z, while $\hat{\epsilon}_Z(f)$ is the standard empirical regression loss on Z.

Theorem 2. *Let $\Phi : \mathcal{X} \to \mathcal{Z}$ be a one-to-one representation function, $f : \mathcal{Z} \times \{t\} \to \mathcal{Y}$ be a hypothesis predicting the outcomes for treatment t. Assume that the ℓ_2 loss function $l(y, f(z,t)) : \mathcal{Z} \to \mathbb{R}_+$ is λ_l-Lipschitz continuous for fixed y, t and upper bounded by M, i.e., $l(y, y') \leq M$ for all $y, y' \in \mathcal{Y}$. Assume the conditional probability density function $p(y \mid z) : \mathcal{Z} \to \mathbb{R}_+$ is λ_p-Lipschitz for fixed y. Then, given a r-cover Z_t of $\Phi(D_t)$, We have*

$$\epsilon_{\Phi(D_t)}(f) \leq \epsilon_{Z_t^A}(f) + r\left(\lambda_l + \lambda_p \frac{M^{\frac{3}{2}}}{3}\right). \tag{3}$$

Proof. The key is to utilize the covering property of a r-cover. For every data point in $\Phi(D_t)$, there is at least an element in Z_t^A such that the distance between them does not exceed r. Then the difference of the losses on $\Phi(D_t)$ and Z_t^A can be bounded using the Lipschitzness assumptions. The complete proof is in the appendix. □

Remark. Theorem 2 bounds the difference between the expected loss on training data and the expected loss on an augmented r-cover with a term of r, which confirms that an augmented r-cover with a small r can be effectively used as a surrogate of the whole dataset for training. Note that the bound does not involve the size of the r-cover, so it is possible to represent the whole dataset with a small r-cover.

Lemma 3. *Under the conditions of Theorem 2, and $l(y, y')$ being λ-Lipschitz for any fixed $y' \in \mathcal{Y}$, let Z_t be a core-set of $\Phi(L_t)$ for $t \in \{0, 1\}$ with covering radius r. Let $\mathcal{H}$ be a set of mappings from $\mathcal{Z} \times \{t\}$ to $\mathcal{Y}$, and $f \in \mathcal{H}$. Then,*

$$\epsilon_t(f, \Phi) \leq \hat{\epsilon}_{Z_t^A}(f) + r \cdot C_M + 2\lambda \mathfrak{R}_{m_t}(\mathcal{H}) + 3M\sqrt{\frac{\ln \frac{3}{\delta}}{2m_t}}$$

holds with probability at least $1 - \delta$, where $C_M = \lambda_l + \lambda_p \frac{M^{\frac{3}{2}}}{3}$ and $\mathfrak{R}_{m_t}(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$.

Proof. The proof mainly utilizes Hoeffding's inequality and some results from Mohri et al. (2018). The complete proof is in the appendix. □

Theorem 4. *Under the conditions of Proposition 1 and Lemma 3, with probability at least $1 - \delta$,*

$$\begin{aligned}\epsilon_{\text{PEHE}}(h) \leq 2 \sum_{t\in\{0,1\}} \left(\hat{\epsilon}_{Z_t^A}(f) + 2\lambda \mathfrak{R}_{m_t}(\mathcal{H})\right) \\ + 4r \cdot C_M + 2C_\Phi \cdot \text{IPM}_\text{G}\left(p_1^\Phi, p_0^\Phi\right) + C\end{aligned}$$

holds, where $C = 6M\left(\sqrt{\frac{\ln \frac{6}{\delta}}{2m_0}} + \sqrt{\frac{\ln \frac{6}{\delta}}{2m_1}}\right) - C_Y$.

Proof. The theorem follows by bounding the two ϵ_t terms in Proposition 1 with the inequality in Lemma 3 and substituting δ with $\frac{\delta}{2}$. □

Remark. Theorem 4 provides an upper bound for PEHE loss, which mainly consists of the covering radius r, empirical regression losses on the augmented r-cover, IPM term between treated and control distribution, and model complexity terms, all of which can be empirically estimated or approximated. The upper bound decreases as the covering radius r gets small. Therefore, unlike Proposition 1, Theorem 4 instructs us to query data points that form a core-set of the observational data with a small covering radius. Note that a set is a core-set of itself with a covering radius of 0, so the result in Shalit et al. (2017) is a special case of ours.

5. The Proposed Method

Based on the theoretical analysis in Section 4, we propose a method called QHTE (Query-based Heterogeneous Treatment Effect estimation), which actively selects data points to query to minimize the upper bound in Theorem 4.

We write the following optimization problem:

$$\begin{aligned}\min_{f,\Phi,Z_0,Z_1} \quad & \sum_{t\in\{0,1\}} \sum_{(x,t,y)\in Z_t^A} \frac{1}{m_t} \cdot l(y, f(\Phi(x), t)) \\ & + \alpha \cdot \|f\| + \beta \cdot r + \gamma \cdot \text{IPM}_\text{G}(\hat{p}_1^\Phi, \hat{p}_0^\Phi) \\ \text{s.t.} \quad & Z_t \subseteq \Phi(L_t), \quad t = 0, 1 \\ & |Z_0| + |Z_1| \leq B,\end{aligned} \tag{4}$$

where r is the largest covering radius of Z_0 and Z_1, α, β, γ are tunable hyperparameters and $\hat{p}_t^\Phi$ is the empirical distribution induced by $\Phi(D_t)$. It is a difficult optimization problem involving the joint optimization over four variables, two of which can have infinite dimensionality. Also, we can obtain factual outcomes from L_t only through explicit queries, which means during the entire optimization process, only B labeled data points can be utilized. To overcome such difficulty, we propose to sequentially add elements to Z_t by alternating between optimizing the objective function over f and Φ, and selecting data points to query. We first elaborate how to determine which points to join Z_t and obtain corresponding factual outcomes, then explicate how to optimize over f and Φ, and present the full method.

Let r_t denote the covering radius of core-set Z_t, the objective function requires us to minimize $r = \max\{r_0, r_1\}$. Since we do not count the contribution of y when deciding a r-cover, finding a core-set of $\Phi(L_t)$ is equivalent to finding a core-set of $\Phi(D_t)$, which does not requires factual outcomes. Formally, this core-set finding problem is:

$$\begin{aligned}\min_{Z_0,Z_1} \quad & \max\left\{\max_{z'\in\Phi(D_t)} \min_{z\in Z_t} \|z - z'\|\right\}_{t\in\{0,1\}} \\ \text{s.t.} \quad & Z_t \subseteq \Phi(D_t), \quad t = 0, 1 \\ & |Z_0| + |Z_1| \leq B.\end{aligned} \tag{5}$$

The special case of (5), where one set $\Phi(D_t)$ is empty, is

known as the k-center problem in theoretical computer science, and is NP-hard (Vazirani, 2003). So the problem in (5) is at least as hard as NP-hard problems, and could not be solved within polynomial time if P $\neq$ NP. However, the k-center problem has an efficient approximation algorithm that greedily selects the point farthest from the current core-set and achieves an approximation ratio of 2 (Vazirani, 2003), i.e., the returned covering radius $r \leq 2 \cdot \mathrm{OPT}$, where OPT is the minimum covering radius. Inspired by that, we propose Algorithm 1 for the problem in (5), where $\mathrm{dist}(s, Z) = \min_{z \in Z} \|s - z\|$. The input Z_t is a placeholder for later use. For now, we just set Z_t to an empty set. We prove Theorem 5 to validate the algorithm.

Algorithm 1 CoreSet

Input: Initial core-sets: Z_0, Z_1; candidate sets: S_0, S_1; size constraint: B

target_size $\leftarrow |Z_0| + |Z_1| + B$
for $t \in \{0, 1\}$ **do**
  if $Z_t = \emptyset$ **then**
    Initialize Z_t with a random element in S_t
  end if
end for
while $|Z_0| + |Z_1| <$ target_size **do**
  $a \leftarrow \arg\max_{a \in S_0} \mathrm{dist}(a, Z_0)$
  $b \leftarrow \arg\max_{b \in S_1} \mathrm{dist}(b, Z_1)$
  if $\mathrm{dist}(a, Z_0) > \mathrm{dist}(b, Z_1)$ **then**
    $Z_0 \leftarrow Z_0 \cup \{a\}$
  else
    $Z_1 \leftarrow Z_1 \cup \{b\}$
  end if
end while

Output: Z_0, Z_1

Theorem 5. *Let r^* be the optimal objective value of the problem in (5), r be the maximum covering radius of the two core-sets returned by Algorithm 1, then $r \leq 2 \cdot r^*$.*

Proof. The proof is based on the observation that the output Z_t is identical to the k centers returned by the greedy algorithm for the k-center problem with $k = |Z_t|$. The complete proof is in the appendix. □

Therefore, once f and Φ are fixed, feasible Z_0 and Z_1 can be efficiently found. We now consider minimizing the objective function in (4) with fixed Z_t. Note that r is uniquely determined by Z_t, so the $\beta \cdot r$ term is a constant when optimizing over f and Φ. We parameterize both f and Φ with neural networks, and adopt the CFR (CounterFactual Regression) network architecture and training procedure described in Shalit et al. (2017). The upper half of Figure 1 shows the general architecture. The mapping Φ is parameterized by the beginning representation part and is shared by both groups of data. The outcome prediction function f is trained with separate heads using corresponding groups of data. A simplified training process is in Algorithm 2. Note that for better sample utilization, the gradient of the IPM term at line 6 in Algorithm 2 can also be calculated using a combination of labeled and unlabeled data.

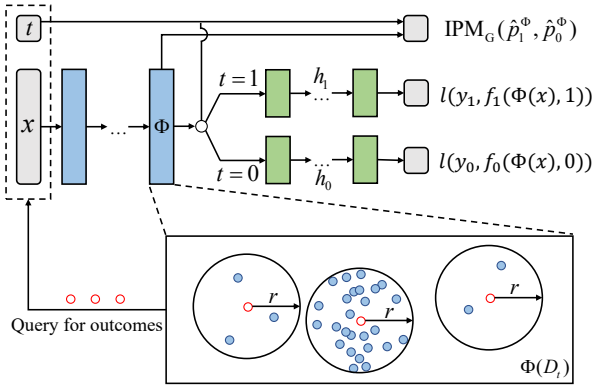

Figure 1. Network architecture and querying strategy.

Algorithm 2 CFR

Input: Factual data: $Z = \{(x_i, t_i, y_i)\}_i$; learning rate: η hyperparameters: α, γ

Initialize a neural network as in Figure 1
$\mathbf{W} \leftarrow$ parameters of Φ
$\mathbf{V} \leftarrow$ parameters of f
while not converged **do**
  Sample mini-batch $\{i_1, i_2, \ldots, i_k\} \subseteq [|Z|]$
  $g_1 \leftarrow \nabla_{\mathbf{W}} \mathrm{IPM}_G(\{\Phi(x_{i_j})\}_{t_{i_j}=0}, \{\Phi(x_{i_j})\}_{t_{i_j}=1})$
  $g_2 \leftarrow \nabla_{\mathbf{W}} \frac{1}{k} \sum_j l(y_{i_j}, f(\Phi(x_{i_j}), t_{i_j}))/m_{t_{i_j}}$
  $g_3 \leftarrow \nabla_{\mathbf{V}} \frac{1}{k} \sum_j l(y_{i_j}, f(\Phi(x_{i_j}), t_{i_j}))/m_{t_{i_j}}$
  $\mathbf{W}, \mathbf{V} \leftarrow \mathbf{W} - \eta(\gamma g_1 + g_2), \mathbf{V} - \eta(g_3 + 2\alpha \mathbf{V})$
end while

Output: Φ, f

The full QHTE method is in Algorithm 3. An illustration of the process is in Figure 1. We add a parameter b to control the number of queries at each optimizing step. During each iteration, QHTE first finds the core-sets for mapped treated and controlled data, then queries an oracle to obtain the corresponding factual outcomes (e.g., performing medical tests to a patient to obtain her health condition), and re-train the neural network with augmented core-sets. Note that training on an augmented core-set can be efficiently done by training on a weighted core-set, which contains much less training samples. By performing a coordinate descent style optimizing procedure – alternating between finding core-sets and optimizing a neural network, QHTE actively selects representative core-set data points. Based on very limited yet representative training data, QHTE balances the

distribution between the treated and controlled group, and meanwhile preserves the predictive ability for unseen test data. Implementation details can be found in Section 6 and the appendix.

Algorithm 3 QHTE

Input: Training data: $D_0 = \{(x_i, 0)\}_i$, $D_1 = \{(x_j, 1)\}_j$; query budget: B; query batch size: b; oracle: $\mathcal{O}$ learning rate: η; hyperparameters: α, γ

Initialize Φ with an identity mapping
$U_0 \leftarrow \emptyset, U_1 \leftarrow \emptyset$ ▷ core-sets without factual outcomes
$Z_0 \leftarrow \emptyset, Z_1 \leftarrow \emptyset$ ▷ core-sets with factual outcomes
for $i \in [\frac{B}{b}]$ **do**
  $U_0', U_1' \leftarrow \text{CoreSet}(U_0, U_1, \Phi(D_0), \Phi(D_1), b)$
  $Z_0 \leftarrow Z_0 \cup \mathcal{O}(U_0' \setminus U_0), Z_1 \leftarrow Z_1 \cup \mathcal{O}(U_1' \setminus U_1)$
  $Z_0^A \leftarrow \text{Augment}(Z_0, \Phi(D_0))$
  $Z_1^A \leftarrow \text{Augment}(Z_1, \Phi(D_1))$
  $\Phi, f \leftarrow \text{CFR}(Z_0^A \cup Z_1^A, \eta, \alpha, \gamma)$
  $U_0, U_1 \leftarrow U_0', U_1'$
end for
$h \leftarrow f(\Phi(\cdot), 1) - f(\Phi(\cdot), 0)$

Output: h

6. Experiments

Because of the nature of the HTE estimation problem, we rarely have access to ground truth in real-world data. In order to evaluate the proposed method, we conduct experiments on three semi-synthetic datasets. Since the setting discussed in this paper is new in the field of causal inference, there are no published methods to be compared. Proposition 1 inspires a successful HTE estimation method (Shalit et al., 2017), which also suggests random sampling under the proposed setting as discussed in Section 4. Therefore, we follow the implications of Proposition 1 and compare QHTE with other HTE estimation methods accompanied by a random querying strategy. We first describe three datasets, then experimental settings, last the results and analysis.

IHDP. This is a common benchmark dataset introduced by Hill (2011). It is from the Infant Health and Development Program (IHDP), in which the covariates come from a randomized experiment studying the effects of specialist home visits on future cognitive test scores. An imbalanced observational dataset is created by removing all children with non-white mothers in the treated group. The dataset consists of 747 units, 139 of which are treated and 608 are controlled, and 25 covariates measuring the children and their mothers. The treatment assignments and pre-treatment covariates are from the experimental data, and the outcomes are generated from the response surface B setting described in Hill (2011). We average over 1,000 realizations of the outcomes with 63/27/10 train/validation/test splits.

ACIC. The datasets were developed for 2016 Atlantic Causal Inference Conference competition (Dorie et al., 2019). They consist of 58 variables and 4,802 individuals. The treatments, factual outcomes, and counterfactual outcomes are all generated by simulation, and selection bias is created as in the IHDP dataset. We randomly choose one dataset from them and average over 100 realizations with 63/27/10 train/validation/test splits.

IBM causal inference benchmark. This is a semi-synthetic dataset created in Shimoni et al. (2018). It uses the cohort of 100,000 samples in Linked Births and Infant Deaths Database (LBIDD) and comprises 177 covariates. We find that the HTEs in this dataset is a fixed constant, which does not exhibit heterogeneity of treatment effects, and it is comprised of datasets that only use a small portion of data. In order to examine the performance of QHTE with large scale data, we simulate the treatments and outcomes for all 100,000 units. Specifically, we create selection bias using $t \mid x \sim \text{Bern}((1 + \exp(-(w^T x + b)))^{-1})$, where $w \sim \mathcal{U}((-0.1, 0.1)^{177 \times 1}$ and $b \sim \mathcal{N}(0, 0.1)$ as in Yoon et al. (2018). The outcomes are simulated based on Hill (2011), we set $Y_0 \sim \mathcal{N}(\exp(u^T(x + 0.5)/10), 0.1)$, $Y_1 \sim \mathcal{N}(u^T x - v, 0.1)$, where u is a vector of regression coefficients $(0, 0.1, 0.2, 0.3, 0.4)$ randomly sampled with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, v is set to make the average treatment effect on the treated group or on the control group equals 4 with probability 0.5 respectively in each simulation. We average over 100 realizations with 63/27/10 train/validation/test splits.

Implementation. We implement QHTE based on CFR (Shalit et al., 2017). We use the same set of hyperparameters for QHTE across three datasets. Specifically, QHTE uses 3 layers to parameterize the representation mapping function Φ, and 3 layers for the outcome prediction function f. Layer sizes are 200 for each of the first 3 layers, and 100 for others. All but the output layer use ReLU (Rectified Linear Unit) (Agarap, 2018) as activation functions, and use batch normalization (Ioffe & Szegedy, 2015) to facilitate training. We use stochastic gradient descent with an initial learning rate of 0.001 and a batch size of 100 to train the network. The learning rate decays with a factor of 0.1 when the validation error plateaus. The family of 1-Lipschitz functions is used in the IPM term, which makes the IPM term the Wasserstein distance (Villani, 2008). We approximate it with the Sinkhorn-Knopp matrix scaling algorithm (Cuturi, 2013). We set $\alpha = 1 \times 10^{-4}$ and $\gamma = 1$.

Baselines. We compare our method with 10 baselines using random querying strategy: Ordinary Least Squares with treatment as a feature (OLS-1), OLS with separate regressors for each treatment (OLS-2), k-Nearest Neighbor (k-NN), Propensity Score Matching with logistic regression (PSM) (Rosenbaum & Rubin, 1983), Bayesian Additive

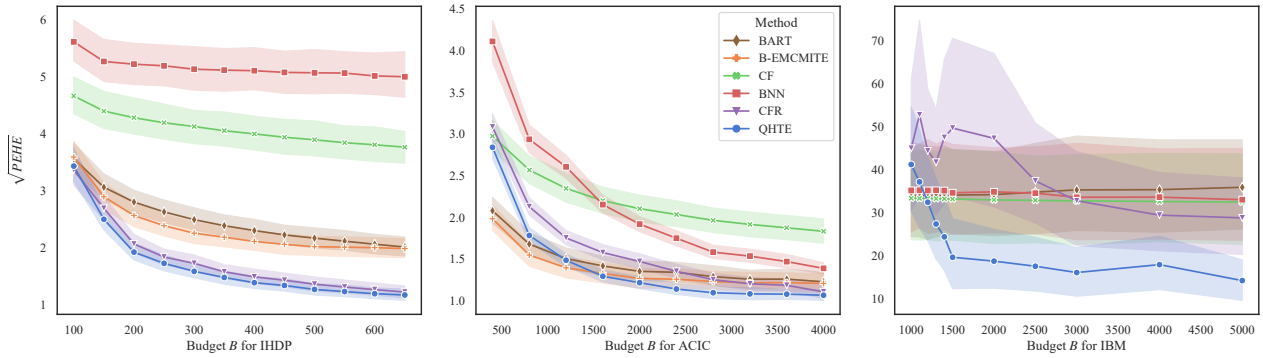

Figure 2. Results on the test sets of three datasets. Lower is better. The bound represents the 95% confidence interval generated by bootstrap sampling. Only 6 superior methods are drawn to avoid clutter.

Regression Trees (BART) (Chipman et al., 2010), Random Forest (RF) (Breiman, 2001), Causal Forest (CF) (Wager & Athey, 2018; Athey et al., 2019), Balancing Neural Network (BNN) (Johansson et al., 2016), Treatment-Agnostic Representation Network (TARNet) (Shalit et al., 2017) as well as CounterFactual Regression with Wasserstein metric (CFR) (Shalit et al., 2017), along with a BART-based method using uncertainty-based querying strategy named B-EMCMITE (Puha et al., 2020).

Table 1. $\sqrt{\text{PEHE}}$ loss on the test set of ACIC. The first row means the budget B. B-EMC stands for B-EMCMITE.

METHOD	500	1,500	2,500	3,500
OLS1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1
OLS2	2.5±0.1	1.6±0.1	1.5±0.1	1.4±0.1
k-NN	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2
PSM	4.9±0.1	4.8±0.1	4.8±0.1	4.8±0.1
BART	2.1±0.1	1.5±0.1	1.3±0.1	1.2±0.1
B-EMC	**2.0±0.1**	**1.4±0.1**	1.3±0.1	1.2±0.1
RF	2.4±0.1	1.9±0.0	1.7±0.0	1.6±0.0
CF	3.0±0.1	2.3±0.1	2.0±0.1	1.9±0.1
BNN	4.1±0.1	2.4±0.1	1.8±0.0	1.5±0.0
TARNET	3.6±0.1	1.9±0.1	1.6±0.0	1.4±0.0
CFR	3.1±0.1	1.7±0.1	1.4±0.0	1.2±0.0
QHTE	2.8±0.1	**1.4±0.0**	**1.1±0.0**	**1.1±0.0**

Results. We compare QHTE with baselines given different budget constraints. A part of the results is shown in Figure 2, where the horizontal axis is the budget and the vertical axis is the square root of the PEHE loss on test data. We also list the performance of all methods under some budgets in Table 1 and 2. Full results can be found in the appendix. The results show that on all three datasets, QHTE outperforms other methods, especially on the IBM dataset. Besides, on the ACIC dataset, the average loss of QHTE is smaller than that of CFR at a significance level of 5% in 17 out of 21 settings. On the IBM dataset, the average loss of QHTE is significantly smaller in 8 out of 11 settings.

Table 2. $\sqrt{\text{PEHE}}$ loss on the test set of IBM.

METHOD	1,500	2,000	3,000	5,000
OLS1	33.7±5.4	33.7±5.4	33.7±5.4	33.7±5.4
OLS2	34.4±5.9	34.5±5.5	33.8±5.1	33.8±5.5
k-NN	35.1±5.4	35.1±5.4	35.1±5.4	35.0±5.4
PSM	34.6±5.5	34.2±5.5	34.6±5.5	34.7±5.5
BART	34.1±5.5	34.2±5.4	35.3±5.4	35.9±5.4
B-EMC	34.0±5.6	34.2±5.6	35.1±5.4	35.0±5.5
RF	31.4±5.3	31.2±5.3	30.9±5.2	30.0±5.2
CF	33.1±5.4	32.9±5.4	32.7±5.4	32.4±5.4
BNN	34.6±5.4	34.8±5.4	33.5±5.5	33.0±5.4
TARNET	58.7±9.9	48.6±9.1	30.3±5.3	29.6±5.2
CFR	49.7±10.0	47.3±8.9	32.7±5.6	28.8±4.9
QHTE	**19.6±4.2**	**18.7±3.5**	**16.0±3.1**	**14.1±2.4**

Analysis. Because the implementation of QHTE is based on CFR, their performance are similar when they use identical training data ($B = m$). However, we care more about the case where $B < m$, and that is where QHTE significantly outperforms baselines. On the IHDP dataset, QHTE is only slightly better than CFR. We argue that because the training set of IHDP only comprises 672 units and 25 covariates, the core-sets selected by Algorithm 1 are not significantly better than those randomly selected. The right two subfigures in Figure 2 further verifies this argument since QHTE performs significantly better than baselines on ACIC (4,321 units) and IBM (90,000 units). With a large potentially available dataset, the core-sets selected by QHTE can be more representative, therefore it can achieve competitive performance with much less training data than baselines.

7. Conclusion

In this paper, we examine the HTE estimation problem with a budget constraint on observational data. We connect it

with active learning, and give an informative error bound based on the concept of core-sets. Our bound mainly consists of a covering radius term, the empirical loss, and the distributional discrepancy between the treated and control data. Based on this result, we propose QHTE to minimize the upper bound with a coordinate descent style optimization procedure, with which QHTE succeeds in actively querying factual outcomes that are more likely to benefit HTE estimation. We apply this theory-guided approach to three datasets with increasing scale, showing that our method achieves the best performance, and that it requires much less training data to achieve the same performance as baselines.

Acknowledgements

This research was supported by the NSFC (61921006) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

The authors would like to thank Jia-Wei Shan for feedback on drafts of the paper, and thank the anonymous reviewers for their helpful comments.

References

Agarap, A. F. Deep learning using rectified linear units (ReLU). *CoRR*, abs/1803.08375, 2018.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Chattopadhyay, R., Wang, Z., Fan, W., Davidson, I., Panchanathan, S., and Ye, J. Batch mode active sampling based on marginal probability distribution matching. *ACM Transactions on Knowledge Discovery from Data*, 7(3):13:1–13:25, 2013.

Chipman, H. A., George, E. I., and McCulloch, R. E. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266–298, 03 2010.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

Deng, K., Pineau, J., and Murphy, S. A. Active learning for developing personalized treatment. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pp. 161–168, 2011.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, (1):43–68, 02 2019.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Guo, Y. and Schuurmans, D. Discriminative batch mode active learning. In *Advances in Neural Information Processing Systems*, pp. 593–600, 2007.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 15(3):965–1056, 2020.

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

Hou, B.-J., Zhang, L., and Zhou, Z.-H. Storage fit learning with unlabeled data. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1844–1850, 2017.

Huang, S.-J., Jin, R., and Zhou, Z.-H. Active learning by querying informative and representative examples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(10):1936–1949, 2014.

Huang, S.-J., Chen, J.-L., Mu, X., and Zhou, Z.-H. Cost-effective active learning from diverse labelers. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 1879–1885, 2017.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

Johansson, F. D., Shalit, U., and Sontag, D. A. Learning representations for counterfactual inference. In *Proceedings of the 33nd International Conference on Machine Learning*, pp. 3020–3029, 2016.

Kallus, N. DeepMatch: Balancing deep covariate representations for causal inference using adversarial training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5067–5077, 2020.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, 2nd edition, 2018.

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

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 09 2020.

Puha, Z., Kaptein, M., and Lemmens, A. Batch mode active learning for individual treatment effect estimation. In *2020 International Conference on Data Mining Workshops*, pp. 859–866, 2020.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

Roy, N. and McCallum, A. Toward optimal active learning through sampling estimation of error reduction. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 441–448, 2001.

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

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1670–1679, 2016.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Settles, B. *Active Learning*. Morgan & Claypool Publishers, 2012.

Shalit, U. Can we learn individual-level treatment policies from clinical data? *Biostatistics*, 21(2):359–362, 2019.

Shalit, U., Johansson, F. D., and Sontag, D. A. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3076–3085, 2017.

Shimoni, Y., Yanover, C., Karavani, E., and Goldschmidt, Y. Benchmarking framework for performance-evaluation of causal inference analysis. *CoRR*, abs/1802.05046, 2018.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Sun, W., Wang, P., Yin, D., Yang, J., and Chang, Y. Causal inference via sparse additive models with application to online advertising. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 297–303, 2015.

Sundin, I., Schulam, P., Siivola, E., Vehtari, A., Saria, S., and Kaski, S. Active learning for decision-making from imbalanced observational data. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6046–6055, 2019.

Tsang, I. W., Kwok, J. T., and Cheung, P.-M. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

Vazirani, V. V. *Approximation Algorithms*. Springer-Verlag, 2003.

Villani, C. *Optimal Transport: Old and New*. Springer-Verlag, 2008.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Wang, L. and Zhou, Z.-H. Cost-saving effect of crowdsourcing learning. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*, pp. 2111–2117, 2016.

Wang, P., Sun, W., Yin, D., Yang, J., and Chang, Y. Robust tree-based causal inference for complex ad effectiveness analysis. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, pp. 67–76, 2015.

Wang, Z. and Ye, J. Querying discriminative and representative samples for batch mode active learning. *ACM Transactions on Knowledge Discovery from Data*, 9(3): 17:1–17:23, 2015.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2638–2648, 2018.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In *Proceedings of the 23nd International Conference on Machine Learning*, pp. 1081–1088, 2006.

Zhou, Z.-H., Ng, M. K., She, Q.-Q., and Jiang, Y. Budget semi-supervised learning. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 588–595, 2009.