Appendix: Budgeted Heterogeneous Treatment Effect Estimation

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

This is the appendix for *Budgeted Heterogeneous Treatment Effect Estimation*. We provide proofs of the theorems in Appendix A and full experimental results in Appendix B.

Appendix A

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. Let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ denote a loss function. The expected factual treated and control losses w.r.t. Φ and f are

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

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

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

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

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 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 $\ell_{f,\Phi}(x,t) \triangleq \int_{\mathcal{Y}} l\left(f(\Phi(x),t),y\right) p\left(y \mid x,t\right) dy$. Let $h(x) = f(\Phi(x),1) - f(\Phi(x),0)$ be an estimate for HTE. Assume $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is the ℓ_2 loss, we have*

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

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

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 $x \in U$, randomly choose an element from $\{z \mid z \in Z \wedge \forall z' \in Z, \|x-z\| \leq \|x-z'\|\}$ to join Z^A.

[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. Copyright 2021 by the author(s).

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, the empirical loss on $Z = \Phi(S)$ is

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

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*

$$\epsilon_{\Phi(D_t)}(f) \leq \epsilon_{Z_t^A}(f) + r \left(\lambda_l + \lambda_p \frac{M^{\frac{3}{2}}}{3} \right). \tag{1}$$

Proof. We first bound the expected loss on a single data point $(\Phi(x), t)$ in $\Phi(D_t)$. According to Definition 4, for every $(\Phi(x), t) \in \Phi(D_t)$, there exists a point $(z, t) \in Z_t^A$ such that $\|\Phi(x) - z\| \leq r$. We write l_x as an abbreviation for $l(y, f(\Phi(x), t))$. We have

$$\begin{aligned}
&\int_{\mathcal{Y}} p(y \mid x, t) l(y, f(\Phi(x), t)) dy \\
\overset{(a)}{=} &\int_{\mathcal{Y}} p(y \mid \Phi(x), t) l_x dy \\
= &\int_{\mathcal{Y}} \left(p(y \mid \Phi(x), t) - p(y \mid z, t) + p(y \mid z, t) \right) l_x dy \\
\overset{(b)}{\leq} &\, r \cdot \lambda_p \int_{\mathcal{Y}} l_x dy + \int_{\mathcal{Y}} p(y \mid z, t) l_x dy \\
\overset{(c)}{\leq} &\, r \cdot \lambda_p \cdot \frac{M^{\frac{3}{2}}}{3} + \int_{\mathcal{Y}} p(y \mid z, t) l_x dy,
\end{aligned}$$

where we use the one-to-one property of Φ in (a), the Lipschitz property in (b) and the fact that l is the ℓ_2 loss in (c). Also, we have

$$\begin{aligned}
&\int_{\mathcal{Y}} p(y \mid z, t) l_x dy \\
= &\int_{\mathcal{Y}} \left(p(y \mid z, t) \left(l_x - l_z \right) + p(y \mid z, t) l_z \right) dy \\
\leq &\, r \cdot \lambda_l + \int_{\mathcal{Y}} p(y \mid z, t) l_z dy.
\end{aligned}$$

Combine the above inequalities, sum over all $(\Phi(x), t) \in \Phi(D_t)$, and choose the corresponding neighbor in Z_t^A for each element as described in Definition 4, then (1) follows. □

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 with covering radius r of $\Phi(L_t)$ for $t \in \{0, 1\}$. Let $\mathcal{H}$ be a set of mappings from $\mathcal{Z} \times \{t\}$ to $\mathcal{Y}$, and $f \in \mathcal{H}$. Then, we have that*

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

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. With Hoeffding's inequality, we have

$$\Pr\left[\hat{\epsilon}_{\Phi(L_t)}(f) \leq \epsilon_{\Phi(D_t)}(f) + C_\delta\right] \geq 1 - \delta,$$

$$\Pr\left[\epsilon_{Z_t^A}(f) \leq \hat{\epsilon}_{Z_t^A}(f) + C_\delta\right] \geq 1 - \delta,$$

where $C_\delta = M\sqrt{\frac{\ln\frac{1}{\delta}}{2m_t}}$ and $\epsilon_{Z_t^A}$ is defined analogously as in Definition 5. Combined with (1), we have

$$\Pr\left[\hat{\epsilon}_{\Phi(L_t)}(f) \leq \hat{\epsilon}_{Z_t^A}(f) + r \cdot C_M + 2C_\delta\right] \geq 1 - 2\delta.$$

Together with Theorem 11.3 in Mohri et al. (2018) and substitute δ with $\frac{\delta}{3}$, Lemma 3 follows. □

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

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

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 $\frac{\delta}{2}$. □

$$\begin{aligned} \min_{Z_0,Z_1} \quad & \max\left\{\max_{z'\in\Phi(L_t)} \min_{z\in Z_t} \|z - z'\|\right\}_{t\in\{0,1\}} \\ \text{s.t.} \quad & Z_t \subseteq \Phi(L_t), \quad t = 0, 1 \\ & |Z_0| + |Z_1| \leq B. \end{aligned} \tag{2}$$

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

Proof. Let b_t^* denote the size and r_t^* denote the covering radius of the core-set for set S_t given by an optimal solution. We use r_t^g to denote the covering radius returned by the greedy facility location algorithm with constraint b_t^* for S_t, then $r_t^g \leq 2r_t^*$ (Vazirani, 2003). Let r_t denote the covering radius of Z_t, which is returned by Algorithm 1.

(1) If $|Z_0| = b_0^*$ and $|Z_1| = b_1^*$, then $r_t = r_t^g$, thus $r = \max\{r_0, r_1\} = \max\{r_0^g, r_1^g\} \leq \max\{2r_0^*, 2r_1^*\} = 2\max\{r_0^*, r_1^*\} = 2r^*$.

(2) Otherwise, without loss of generality, assume $|Z_0| > b_0^*$. We jump to the point that the algorithm is about to add an element to Z_0 such that the size of Z_0 will become $b_0^* + 1$, and record the current covering radius of Z_t as r_t', then $r_0' = r_0^g$ and $r_0' \geq r_1'$. Also note that the covering radius of Z_t is monotonic non-increasing, we have $r = \max\{r_0, r_1\} \leq \max\{r_0', r_1'\} = r_0' = r_0^g \leq 2r_0^* \leq \max\{2r_0^*, 2r_1^*\} = 2\max\{r_0^*, r_1^*\} = 2r^*$. □

Appendix B

In this section, we present implementation details and full experimental results. The code and data can be found at https://github.com/Qcer17/QHTE.

Implementation. We implement QHTE based on CFR (Shalit et al., 2017). For a fair comparison, we do not use any parameter-searching techniques because other estimation methods usually do not have many hyperparameters like a neural network does. Instead, we use the same set of hyperparameters for QHTE across three datasets. Specifically, QHTE uses 3

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
    $a = \arg\max_{a \in S_0} \mathrm{dist}(a, Z_0)$
    $b = \arg\max_{b \in S_1} \mathrm{dist}(b, Z_1)$
    if $\mathrm{dist}(a, Z_0) > \mathrm{dist}(b, Z_1)$ then
        $Z_0 = Z_0 \cup \{a\}$
    else
        $Z_1 = Z_1 \cup \{b\}$
    end if
end while
Output: $Z_0$, $Z_1$
```

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 ten 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 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). The experiments are built upon some well-developed and open-source packages, including Pedregosa et al. (2011), Kapelner & Bleich (2016) and Tibshirani et al. (2020), and most of the hyperparameters are set to the default ones.

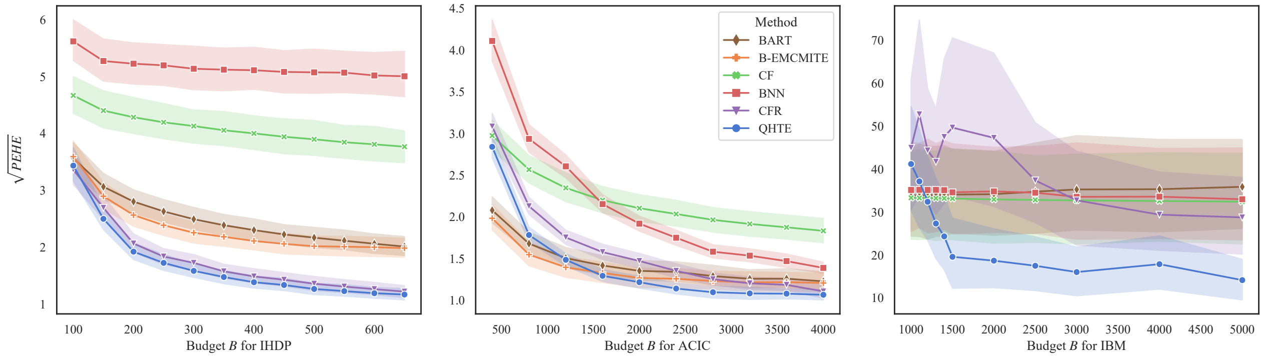

Figure 1. Results on the test sets of three datasets. Lower is better. Only five superior methods are drawn to avoid clutter.

B.1 Results on IHDP

Results on the IHDP dataset are in Table 1-6. Since the number of samples is small in IHDP, making training on IHDP efficient, we run experiments with small budget steps. In total 58 different settings of the budget B, the win/tie/loss counts of QHTE are 43/12/3. Note that in real applications, the budget B is usually small compared to the size of the dataset D, so we care more about the performance of different methods with small B. When $B \leq 500$, the win/tie/loss counts of QHTE are 36/2/3, which further verifies the effectiveness of QHTE.

Table 1. $\sqrt{\text{PEHE}}$ loss on the test set of IHDP. The first row is the budget B.

METHOD	100	110	120	130	140	150	160	170	180	190
OLS1	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6
OLS2	**3.4±0.4**	**3.1±0.3**	3.0±0.3	2.9±0.3	2.8±0.3	2.7±0.3	2.7±0.3	2.6±0.3	2.6±0.3	2.6±0.3
k-NN	5.2±0.5	5.2±0.5	5.1±0.5	5.1±0.5	5.1±0.5	5.1±0.5	5.1±0.5	5.0±0.5	5.0±0.5	5.0±0.5
PSM	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.6±0.6	5.5±0.6	5.5±0.6	5.5±0.6
BART	3.6±0.4	3.4±0.4	3.3±0.4	3.2±0.4	3.1±0.4	3.1±0.4	3.0±0.4	2.9±0.4	2.9±0.4	2.8±0.4
RF	3.6±0.4	3.6±0.4	3.4±0.4	3.4±0.4	3.3±0.4	3.2±0.4	3.2±0.4	3.1±0.4	3.1±0.4	3.0±0.4
CF	4.7±0.5	4.5±0.5	4.5±0.5	4.5±0.5	4.4±0.5	4.4±0.5	4.4±0.5	4.3±0.5	4.3±0.5	4.3±0.5
BNN	5.6±0.6	5.6±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.3±0.6	5.3±0.6	5.3±0.6	5.2±0.6	5.2±0.6
TARNET	3.5±0.5	3.3±0.5	3.1±0.4	2.9±0.4	2.8±0.4	2.7±0.4	2.6±0.4	2.4±0.3	2.4±0.3	2.3±0.3
CFR	**3.4±0.4**	**3.1±0.4**	**2.9±0.4**	**2.7±0.4**	**2.6±0.3**	2.7±0.3	2.4±0.3	2.3±0.3	2.2±0.3	2.1±0.3
QHTE	**3.4±0.5**	**3.1±0.4**	3.0±0.4	2.8±0.4	2.7±0.4	**2.5±0.3**	**2.3±0.3**	**2.1±0.3**	**2.1±0.3**	**2.0±0.3**

Table 2. $\sqrt{\text{PEHE}}$ loss on the test set of IHDP. The first row is the budget B.

METHOD	200	210	220	230	240	250	260	270	280	290
OLS1	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6
OLS2	2.5±0.3	2.5±0.3	2.5±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.3±0.3
k-NN	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5
PSM	5.5±0.6	5.5±0.6	5.6±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6
BART	2.8±0.3	2.8±0.3	2.7±0.3	2.7±0.3	2.7±0.3	2.6±0.3	2.6±0.3	2.6±0.3	2.5±0.3	2.5±0.3
RF	3.0±0.4	3.0±0.4	2.9±0.4	2.9±0.3	2.9±0.3	2.8±0.3	2.8±0.3	2.8±0.3	2.8±0.3	2.7±0.3
CF	4.3±0.5	4.3±0.5	4.2±0.5	4.2±0.5	4.2±0.5	4.2±0.5	4.2±0.5	4.2±0.5	4.2±0.5	4.1±0.5
BNN	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.1±0.6
TARNET	2.2±0.3	2.2±0.3	2.1±0.3	2.1±0.3	2.0±0.3	2.0±0.3	2.0±0.3	1.9±0.3	1.9±0.3	1.9±0.3
CFR	2.1±0.3	2.0±0.3	2.0±0.3	1.9±0.2	1.9±0.2	1.8±0.2	1.8±0.2	1.8±0.2	1.8±0.2	1.8±0.2
QHTE	**1.9±0.2**	**1.9±0.2**	**1.8±0.2**	**1.8±0.2**	**1.7±0.2**	**1.7±0.2**	**1.7±0.2**	**1.7±0.2**	**1.7±0.2**	**1.7±0.2**

Table 3. $\sqrt{\text{PEHE}}$ loss on the test set of IHDP. The first row is the budget B.

Method	300	310	320	330	340	350	360	370	380	390
OLS1	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6
OLS2	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.2±0.3	2.2±0.3	2.2±0.3
k-NN	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	5.0±0.5	4.9±0.5	4.9±0.5	4.9±0.5
PSM	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6
BART	2.5±0.3	2.5±0.3	2.5±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.3±0.3	2.3±0.3
RF	2.7±0.3	2.7±0.3	2.7±0.3	2.7±0.3	2.6±0.3	2.6±0.3	2.6±0.3	2.6±0.3	2.6±0.3	2.5±0.3
CF	4.1±0.5	4.1±0.5	4.1±0.5	4.1±0.5	4.1±0.5	4.1±0.5	4.0±0.5	4.0±0.5	4.0±0.5	4.0±0.5
BNN	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6
TARNet	1.8±0.2	1.8±0.2	1.8±0.2	1.7±0.2	1.7±0.2	1.7±0.2	1.7±0.2	1.7±0.2	1.6±0.2	1.6±0.2
CFR	1.7±0.2	1.7±0.2	1.7±0.2	1.6±0.2	1.6±0.2	1.6±0.2	1.6±0.2	1.5±0.2	1.5±0.2	1.5±0.2
QHTE	**1.6±0.2**	**1.5±0.2**	**1.5±0.2**	**1.5±0.2**	**1.5±0.2**	**1.5±0.2**	**1.5±0.2**	**1.4±0.2**	**1.4±0.2**	**1.4±0.2**

Table 4. $\sqrt{\text{PEHE}}$ loss on the test set of IHDP. The first row is the budget B.

Method	400	410	420	430	440	450	460	470	480	490
OLS1	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6
OLS2	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3
k-NN	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5
PSM	5.5±0.6	5.6±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.6±0.6	5.5±0.6	5.5±0.6	5.5±0.6
BART	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3
RF	2.5±0.3	2.5±0.3	2.5±0.3	2.5±0.3	2.5±0.3	2.5±0.3	2.5±0.3	2.5±0.3	2.4±0.3	2.4±0.3
CF	4.0±0.5	4.0±0.5	4.0±0.5	4.0±0.5	4.0±0.5	3.9±0.5	3.9±0.5	3.9±0.5	3.9±0.5	3.9±0.5
BNN	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6
TARNet	1.6±0.2	1.6±0.2	1.6±0.2	1.6±0.2	1.5±0.2	1.5±0.2	1.5±0.2	1.5±0.2	1.5±0.2	1.5±0.2
CFR	1.5±0.2	1.5±0.2	1.5±0.2	1.5±0.2	1.5±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2
QHTE	**1.4±0.2**	**1.4±0.2**	**1.4±0.2**	**1.4±0.2**	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**

Table 5. $\sqrt{\text{PEHE}}$ loss on the test set of IHDP. The first row is the budget B.

Method	500	510	520	530	540	550	560	570	580	590
OLS1	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6
OLS2	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.2±0.3	2.1±0.3	2.1±0.3	2.1±0.3
k-NN	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5
PSM	5.5±0.6	5.5±0.6	5.6±0.6	5.6±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6
BART	2.2±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3
RF	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.4±0.3	2.3±0.3
CF	3.9±0.5	3.9±0.5	3.9±0.5	3.9±0.5	3.9±0.5	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5
BNN	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.1±0.6	5.0±0.6	5.0±0.6
TARNet	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2	1.4±0.2
CFR	1.4±0.2	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2
QHTE	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**	**1.3±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**

Table 6. $\sqrt{\text{PEHE}}$ loss on the test set of IHDP. The first row is the budget B.

METHOD	600	610	620	630	640	650	660	670
OLS1	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6	5.2±0.6
OLS2	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3	2.1±0.3
k-NN	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5	4.9±0.5
PSM	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.5±0.6	5.6±0.6	5.5±0.6
BART	2.1±0.3	2.0±0.3	2.0±0.3	2.0±0.3	2.0±0.3	2.0±0.3	2.0±0.3	2.0±0.3
RF	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3	2.3±0.3
CF	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5	3.8±0.5
BNN	5.0±0.6	5.0±0.6	5.0±0.6	5.0±0.6	5.0±0.6	5.0±0.6	5.0±0.6	5.0±0.6
TARNET	1.4±0.2	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2	1.3±0.2
CFR	1.3±0.2	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**
QHTE	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**	**1.2±0.2**

B.2 Results on ACIC

Results on the ACIC dataset are in Table 7, 8, and 9. We run experiments on ACIC with 21 different settings of the budget B. The win/tie/loss counts of QHTE are 13/4/4.

Table 7. $\sqrt{\text{PEHE}}$ loss on the test set of ACIC. The first row is the budget B.

METHOD	200	500	600	800	1,000	1,200	1,500
OLS1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1
OLS2	4.2±0.2	2.5±0.1	2.1±0.1	1.9±0.1	1.8±0.1	1.7±0.1	1.6±0.1
k-NN	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2
PSM	4.9±0.1	4.9±0.1	4.9±0.1	4.9±0.1	4.8±0.1	4.8±0.1	4.8±0.1
BART	**2.7±0.1**	**2.1±0.1**	**1.8±0.1**	**1.7±0.1**	**1.6±0.1**	**1.5±0.1**	1.5±0.1
RF	2.8±0.1	2.4±0.1	2.2±0.1	2.1±0.1	2.0±0.0	1.9±0.0	1.9±0.0
CF	3.4±0.1	3.0±0.1	2.7±0.1	2.6±0.1	2.4±0.1	2.3±0.1	2.3±0.1
BNN	4.8±0.1	4.1±0.1	3.5±0.1	2.9±0.1	2.9±0.1	2.6±0.1	2.4±0.1
TARNET	5.8±0.2	3.6±0.1	2.8±0.1	2.4±0.1	2.2±0.1	2.1±0.1	1.9±0.1
CFR	4.2±0.1	3.1±0.1	2.4±0.1	2.1±0.1	1.9±0.0	1.8±0.0	1.7±0.1
QHTE	4.3±0.2	2.8±0.1	2.2±0.1	1.8±0.0	**1.6±0.0**	**1.5±0.0**	**1.4±0.0**

Table 8. $\sqrt{\text{PEHE}}$ loss on the test set of ACIC. The first row is the budget B.

METHOD	1,600	1,800	2,000	2,200	2,500	2,600	2,800
OLS1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1
OLS2	1.6±0.1	1.5±0.1	1.5±0.1	1.5±0.1	1.5±0.1	1.4±0.1	1.4±0.1
k-NN	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2
PSM	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1
BART	1.4±0.1	1.4±0.1	1.4±0.1	1.3±0.1	1.3±0.1	1.3±0.1	1.3±0.1
RF	1.8±0.0	1.8±0.0	1.8±0.0	1.8±0.0	1.7±0.0	1.7±0.0	1.7±0.0
CF	2.2±0.1	2.2±0.1	2.1±0.1	2.1±0.1	2.0±0.1	2.0±0.1	2.0±0.1
BNN	2.2±0.1	2.0±0.1	1.9±0.0	1.8±0.0	1.8±0.0	1.7±0.0	1.6±0.0
TARNET	1.8±0.1	1.7±0.0	1.6±0.1	1.7±0.1	1.6±0.0	1.5±0.1	1.5±0.0
CFR	1.6±0.0	1.5±0.0	1.5±0.0	1.4±0.0	1.4±0.0	1.3±0.0	1.3±0.0
QHTE	**1.3±0.0**	**1.2±0.0**	**1.2±0.0**	**1.2±0.0**	**1.1±0.0**	**1.1±0.0**	**1.1±0.0**

Table 9. $\sqrt{\text{PEHE}}$ loss on the test set of ACIC. The first row is the budget B.

METHOD	3,000	3,200	3,500	3,600	3,800	4,000	4,200
OLS1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1	3.9±0.1
OLS2	1.4±0.1	1.4±0.1	1.4±0.1	1.4±0.1	1.4±0.1	1.3±0.1	1.3±0.1
k-NN	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2	5.6±0.2
PSM	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1	4.8±0.1
BART	1.3±0.1	1.3±0.1	1.2±0.1	1.3±0.1	1.2±0.1	1.2±0.1	1.2±0.1
RF	1.7±0.0	1.7±0.0	1.6±0.0	1.6±0.0	1.6±0.0	1.6±0.0	1.6±0.0
CF	1.9±0.1	1.9±0.1	1.9±0.1	1.9±0.1	1.9±0.1	1.8±0.1	1.8±0.1
BNN	1.6±0.0	1.5±0.0	1.5±0.0	1.5±0.0	1.4±0.0	1.4±0.0	1.4±0.0
TARNET	1.5±0.0	1.4±0.0	1.4±0.0	1.4±0.0	1.4±0.0	1.3±0.0	1.3±0.0
CFR	1.2±0.0	1.2±0.0	1.2±0.0	1.2±0.0	1.2±0.0	**1.1±0.0**	**1.1±0.0**
QHTE	**1.1±0.0**	**1.1±0.0**	**1.1±0.0**	**1.1±0.0**	**1.1±0.0**	**1.1±0.0**	**1.1±0.0**

B.3 Results on IBM

Results on the IBM dataset are in Table 10 and 11. Since the size of the IBM dataset is large, methods like k-NN, PSM, and BART do not scale well on such a dataset and take lots of running time and memory. So we run experiments with 11 different settings where the budget B is less than 5,000. The win/tie/loss counts of QHTE are 8/0/3. Note that methods other than QHTE perform much worse than QHTE and does not improve much with more data, which confirms that the core-sets selected by our algorithm are more representative for the whole dataset, and training on core-sets is effective for minimizing the generalization error.

Table 10. $\sqrt{\text{PEHE}}$ loss on the test set of IBM. The first row is the budget B.

METHOD	1,000	1,100	1,200	1,300	1,400	1,500
OLS1	33.8±5.4	33.7±5.4	33.8±5.4	33.8±5.4	33.7±5.4	33.7±5.4
OLS2	34.8±6.0	34.8±6.0	34.8±5.9	34.8±5.8	34.5±5.9	34.4±5.9
k-NN	35.2±5.4	35.2±5.4	35.2±5.4	35.2±5.4	35.1±5.4	35.1±5.4
PSM	34.4±5.5	34.6±5.5	34.8±5.5	34.5±5.5	34.5±5.5	34.6±5.5
BART	33.8±5.4	33.8±5.4	33.9±5.4	34.0±5.4	34.0±5.4	34.1±5.5
RF	**31.7±5.3**	**31.6±5.3**	**31.6±5.3**	31.5±5.3	31.7±5.3	31.4±5.3
CF	33.3±5.4	33.3±5.4	33.3±5.4	33.2±5.4	33.2±5.4	33.1±5.4
BNN	35.1±5.4	35.2±5.4	35.1±5.4	35.2±5.4	35.1±5.4	34.6±5.4
TARNET	52.7±7.4	46.7±6.5	49.4±7.5	47.1±8.1	48.4±8.6	58.7±9.9
CFR	45.0±7.5	52.8±11.1	44.3±7.0	41.7±6.5	47.5±9.1	49.7±10.0
QHTE	41.2±6.3	37.1±5.9	32.4±5.2	**27.3±4.9**	**24.3±4.6**	**19.6±4.2**

Table 11. $\sqrt{\text{PEHE}}$ loss on the test set of IBM. The first row is the budget B.

METHOD	2,000	2,500	3,000	4,000	5,000
OLS1	33.7±5.4	33.7±5.4	33.7±5.4	33.8±5.4	33.7±5.4
OLS2	34.5±5.5	34.2±5.4	33.8±5.1	33.7±5.1	33.8±5.5
k-NN	35.1±5.4	35.1±5.4	35.1±5.4	35.0±5.4	35.0±5.4
PSM	34.2±5.5	34.7±5.5	34.6±5.5	34.5±5.5	34.7±5.5
BART	34.2±5.4	34.7±5.4	35.3±5.4	35.3±5.4	35.9±5.4
RF	31.2±5.3	31.1±5.2	30.9±5.2	30.6±5.2	30.0±5.2
CF	32.9±5.4	32.8±5.4	32.7±5.4	32.6±5.4	32.4±5.4
BNN	34.8±5.4	34.5±5.4	33.5±5.5	33.6±5.4	33.0±5.4
TARNET	48.6±9.1	41.2±6.9	30.3±5.3	31.0±5.2	29.6±5.2
CFR	47.3±8.9	37.4±5.9	32.7±5.6	29.4±4.9	28.8±4.9
QHTE	**18.7±3.5**	**17.5±3.0**	**16.0±3.1**	**17.9±3.3**	**14.1±2.4**

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.

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.

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.

Kapelner, A. and Bleich, J. bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, 70(4):1–40, 2016. doi: 10.18637/jss.v070.i04.

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

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

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.

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.

Tibshirani, J., Athey, S., and Wager, S. *grf: Generalized Random Forests*, 2020. URL `https://CRAN.R-project.org/package=grf`. R package version 1.2.0.

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.