## **Learning Interpretable Models with Causal Guarantees**

## Carolyn Kim 1 Osbert Bastani 2

## **Abstract**

Machine learning has shown much promise in helping improve the quality of medical, legal, and economic decision-making. In these applications, machine learning models must satisfy two important criteria: (i) they must be causal, since the goal is typically to predict individual treatment effects, and (ii) they must be interpretable, so that human decision makers can validate and trust the model predictions. There has recently been much progress along each direction independently, yet the state-of-the-art approaches are fundamentally incompatible. We propose a framework for learning causal interpretable models-from observational data—that can be used to predict individual treatment effects. Our framework can be used with any algorithm for learning interpretable models. Furthermore, we prove an error bound on the treatment effects predicted by our model. Finally, in an experiment on real-world data, we show that the models trained using our framework significantly outperform a number of baselines.

### 1. Introduction

Machine learning is increasingly being used to help inform consequential decisions in healthcare, law, and finance. In these applications, the goal is often to predict the effect of some intervention (called a *treatment effect*)—e.g., the efficacy of a drug on a given patient (Consortium, 2009; Kim et al., 2011; Bastani & Bayati, 2015; Henry et al., 2015), the probability that a defendent in a court case is a flight risk (Kleinberg et al., 2017), or the probability that an applicant will repay a loan (Hardt et al., 2016). There are two important properties that these machine learning models must satisfy: (i) they must be must be causal (Rubin, 2005; Pearl, 2010), and (ii) they must be interpretable.

First, to predict treatment effects, our model must predict outcomes when the world is modified in some way (called a counterfactual outcome). For example, to predict the efficacy of a drug on a patient, we need to know the patient's outcome both when given the drug and when not given the drug. One way to predict counterfactual outcomes is to use randomized controlled experiments (RCTs)—by randomly assigning individuals to treatment and control groups, we can ensure that the model generalizes to predicting counterfactual outcomes. Indeed, RCTs are frequently used to estimate average treatment effects (e.g., whether the drug is effective for the population as a whole). However, they are unsuitable for predicting individual treatment effects (ITEs)—such models have many more parameters, so much more training data is required. <sup>1</sup> Yet, the promise of machine learning is exactly to predict ITEs, which can be used to tailor decisions to specific individuals.

Instead, we consider the more common approach of predicting counterfactual outcomes based on *observational data*. In contrast to RCT data, individuals are selected into treatment and control groups by unknown mechanisms (Rubin, 2005; Shalit et al., 2017). For example, in observational data, sicker patients are more likely to receive drugs. Thus, our model may incorrectly conclude that drugs are ineffective, since individuals who do not take drugs are healthier than those who do. The problem is that supervised learning can only guarantee predictive performance on data that comes from the same distribution as the training data, but counterfactual outcomes do not satisfy this assumption.

To make progress, we have to make assumptions about the distribution of the observational data. Several algorithms along these lines have been proposed, including honest trees (Athey & Imbens, 2016), causal forests (Wager & Athey, 2017), propensity score weighting (Austin, 2011), instrumental variables (Wooldridge, 2015), and causal representations (Johansson et al., 2016; Shalit et al., 2017).

Second, the learned model must be interpretable—i.e., a human domain expert (e.g., a doctor) must be able to validate the model. Interpretability is important since there are often defects in the training data that cause the model to make preventable errors. Indeed, it has been shown that these issues often arise in practice (Caruana et al., 2015; Ribeiro et al., 2016; Bastani et al., 2017). Learning interpretable

<sup>&</sup>lt;sup>1</sup>Stanford University <sup>2</sup>University of Pennsylvania. Correspondence to: Carolyn Kim <ckim@cs.stanford.edu>, Osbert Bastani <obastani@seas.upenn.edu>.

<sup>&</sup>lt;sup>1</sup>Individual treatment effects are also known as heterogeneous treatment effects, or conditional average treatment effects.

models is particularly important when there may be causal issues. In particular, there is often no way to validate the assumptions made by causal learning algorithms. For example, many approaches assume *strong ignorability*, which says that probability of selecting into treatment can be fully predicted from the covariates. However, this assumption often fails in practice (Louizos et al., 2017). Interpretability provides a way for experts to identify causal issues.

Many algorithms have been proposed for learning interpretable models, including decision trees (Breiman, 2017; Bastani et al., 2017), sparse linear models (Tibshirani, 1996; Ustun & Rudin, 2016), generalized additive models (Lou et al., 2012; Caruana et al., 2015), rule lists (Wang & Rudin, 2015; Yang et al., 2017; Angelino et al., 2017), decision sets (Lakkaraju et al., 2016), and programs (Ellis et al., 2015; Verma et al., 2018; Valkov et al., 2018; Ellis et al., 2018).

Thus, while there has been a variety of work on learning causal models and on learning interpretable models, there has been relatively little work on designing algorithms that are capable of achieving both desirable properties.

**Our contributions.** We propose a general framework for learning interpretable models with causal guarantees. In particular, given *any* supervised learning algorithm  $\mathcal{A}$  for learning interpretable models, our framework converts  $\mathcal{A}$  into an algorithm  $\tilde{\mathcal{A}}$  for learning interpretable models  $\hat{\tau}: \mathcal{X} \to \mathcal{Y}$  that predict the ITE  $\tau(x) \in \mathcal{Y}$  of an individual with covariates  $x \in \mathcal{X}$ . Furthermore, we provide guarantees on the performance of the models learned using  $\tilde{\mathcal{A}}$ .

We build on recent work on causal representations (Johansson et al., 2016; Shalit et al., 2017), a general framework for converting any supervised learning algorithm  $\mathcal B$  into an algorithm for learning models that predict ITEs. Their key idea is to first learn a causal representation  $\Phi: \mathcal X \to \mathcal R$ , where  $\mathcal R$  is an embedding space. Intuitively,  $\Phi$  is designed to eliminate the bias from using observational data. In particular, they then use  $\mathcal B$  to train a model  $\hat h: \mathcal R \times \{0,1\} \to \mathcal Y$  on the embedding  $D_{\mathcal R} = \{(\Phi(x),t,y)\}$  of the original dataset  $D=\{(x,t,y)\}$ , where  $t\in\{0,1\}$  is the treatment and  $y\in\mathcal Y$  is the outcome. Finally, assuming strong ignorability, they prove bounds on the error of the following model for predicting ITEs:

$$\hat{\tau}(x) = \hat{h}(\Phi(x), 1) - \hat{h}(\Phi(x), 0)$$

The reason we cannot directly use their approach is that the causal representation  $\Phi$  is uninterpretable. In particular, their approach would use the interpretable learning algorithm  $\mathcal A$  to train an interpretable model  $\hat h:\mathcal R\to\mathcal Y$  on  $D_{\mathbf R}$ . However,  $\hat \tau(x)=\hat h(\Phi(x),1)-\hat h(\Phi(x),0)$  remains uninterpretable since  $\Phi$  is uninterpretable—the problem is that the inputs to h are the uninterpretable features  $\Phi(x)$ .

We propose a solution to this problem inspired by model

compression (Bucilua et al., 2006; Hinton et al., 2015). First, we use (Shalit et al., 2017) to learn an uninterpretable function  $h^*: \mathcal{X} \to \mathcal{Y}$ . We refer to the function  $f^*: \mathcal{X} \to \mathcal{Y}$  defined by  $f^*(x,t) = h^*(\Phi(x),t)$  as the *oracle model*. Then, we use  $\mathcal{A}$  to learn an interpretable model  $\hat{f}: \mathcal{X} \to \mathcal{Y}$  to approximate  $f^*$ —i.e., for some distribution p(x,t) of our choosing,

$$\hat{f} = \mathcal{A}(\{(x_i, t_i, f^*(x_i, t_i)\}),$$
 (1)

where  $(x_i, t_i) \sim p(x, t)$  are i.i.d. samples. Then, we propose to use  $\hat{\tau}(x) = \hat{f}(x, 1) - \hat{f}(x, 0)$  to predict ITEs.

It remains to choose p(x,t) in (1). We make a simple and intuitive choice—namely, the distribution over treatments that would have been induced by running an RCT (which we call the *RCT distribution*), where treatments are randomly assigned and are independent of the covariates x. This choice amounts to using  $f^*$  to label the unobserved counterfactual for each covariate in the original observational dataset, and then running  $\mathcal A$  on the combined dataset to train  $\hat f$ .

Intuitively, since RCTs can be used to predict ITEs, f should have good performance as long as  $f^*$  has good performance and  $\hat{f}$  is a good approximation of  $f^*$  on the RCT distribution. Indeed, under these conditions, we prove a performance guarantee for  $\hat{f}$  analogous to the one available for the causal representations approach (Johansson et al., 2016; Shalit et al., 2017). Finally, in an experimental study, we show how our approach can be used to improve the performance of a wide range of interpretable models.

Related work. There has been prior work proposing the "honest tree" algorithm for learning decision trees for prediting ITEs (Athey & Imbens, 2016). This work builds on the CART algorithm (Breiman, 2017)—in particular, they reduce the bias of CART by using different subsets of the training data to estimate the internal nodes and the leaf nodes. In contrast, our framework can be used to convert *any* interpretable learning algorithm into one for learning models for predicting ITEs. Furthermore, unlike their work, our approach comes with provable performance guarantees. Finally, we show in our experiments that our approach can substantially outperform theirs.

There has also been work using interpretability to identify causal issues in learned predictive models (Caruana et al., 2015; Ribeiro et al., 2016; Bastani et al., 2017). However, there is currently no way to fix these causal issues except by having an expert manually correct the model.

Finally, there has been a wide range of work using an uninterpretable oracle model  $f^*$  to guide the learning of an interpretable model (Lakkaraju et al., 2017; Bastani et al., 2017; Verma et al., 2018; Frosst & Hinton, 2017; Bastani et al., 2018). Our work is the first to leverage this approach in the context of learning causal models.

### 2. Preliminaries

In this section, we give background on causal inference for estimating individual treatment effects (ITEs). Then, we summarize the approach of causal representations proposed in (Johansson et al., 2016; Shalit et al., 2017), as well as a bound they prove on the estimation error for their approach.

**Potential outcomes framework.** We begin by describing the Rubin-Neyman potential outcomes framework (Rubin, 2005). Suppose we have a set of units, and we want to estimate the efficacy of a treatment for a given unit. We assume that each unit is associated with a covariate vector X (e.g., encoding patient-specific characteristics such as their healthcare history). Each unit is either assigned to the control group (denoted T=0) or to the treatment group (denoted T=1). Furthermore, each unit is associated with two potential outcomes—the outcome  $Y_0$  if the unit is assigned to control (i.e., T=0), and the outcome  $Y_1$  if the unit is assigned to treatment (i.e., T=1). The object of interest is the *treatment effect*  $Y_1-Y_0$ , which informs the decision maker whether the unit would experience a better outcome under the treatment or under the control.

For example, units may be patients, and covariates may be patient-specific features such as biomarkers and healthcare history. The treatment may be prescribing a drug to the patient (so the control is not prescribing the drug). Then,  $Y_1$  may be how quickly the patient recovers when prescribed the drug, and  $Y_0$  is how quickly the patient recovers without the drug. Then, the treatment effect measures whether the drug helps the patient recover more quickly. Ideally, the patient would only be given the drug if  $Y_1 - Y_0 > 0$ .

Formally, each unit is associated with a tuple of random variables  $(X, T, Y_0, Y_1)$ . We assume that the covariate vector takes values in  $\mathcal{X} \subseteq \mathbb{R}^d$ , and the potential outcomes take values in  $\mathcal{Y} \subseteq \mathbb{R}$  (of course, the treatment T takes values in  $\{0,1\}$ ). Furthermore, we assume that for each unit, this tuple is drawn i.i.d. from a distribution  $p(x,t,y_0,y_1)$ .

The fundamental challenge in causal inference is that for each unit, we only observe either  $Y_0$  or  $Y_1$ , but never both—in particular, for each unit, we only observe  $(X, T, Y_T)$ .

**Definition 2.1.** The observed outcome  $Y_T$  is the *factual outcome*, and the unobserved outcome  $Y_{1-T}$  is the *counter-factual outcome*.

For example, if we give a patient the drug, we cannot observe what would have happened without the drug.

Thus, we can only estimate the average  $Y_1 - Y_0$  over multiple units. If we average over the entire population, then we obtain average treatment effect (ATE)

$$ATE = \mathbb{E}_n[Y_1 - Y_0].$$

However, the ATE does not yield any information about the efficacy of treatment on an individual unit. Instead, our goal is to estimate the efficacy of a treatment for an individual units based on their covariates.

**Definition 2.2.** The *individual treatment effect (ITE)* is

$$\tau(x) = \mathbb{E}_p[Y_1 - Y_0 \mid X = x]$$

To estimate the ITE, we make the following standard assumption about the treatment assignment mechanism (Johansson et al., 2016; Shalit et al., 2017).

**Assumption 2.3.** We assume that the treatment assignment is *strongly ignorable*, i.e.,

$$(Y_1, Y_0) \perp \!\!\! \perp T \mid X.$$

For example, this assumption eliminates the possibility that we only observe  $Y_1$  for which  $Y_1 > Y_0$ . We also make the standard assumption that each unit has a nonzero probability of being assigned to each the control and the treatment.

**Assumption 2.4.** We assume that for all  $x \in \mathcal{X}$ ,

$$0 < \mathbb{P}_{n}(T = 1 \mid X = x) < 1.$$

For example, this assumption eliminates the possibility that we never get observations of  $Y_1$  for a particular x.

Our goal is to obtain an estimate  $\hat{\tau}(x)$  of the ITE  $\tau(x)$ . A natural metric is our accuracy for predicting  $\tau(x)$  for a unit chosen at random from distribution p.

**Definition 2.5.** The expected *precision in estimation of heterogenous effect (PEHE)* (Hill, 2011) is

$$\epsilon_{\text{PEHE}}(\hat{\tau}) = \int_{\mathcal{X}} (\hat{\tau}(x) - \tau(x))^2 p(x) dx.$$
 (2)

Causal representations. Now, we describe the *causal representations* approach to estimating  $\tau(x)$  (Johansson et al., 2016; Shalit et al., 2017). Suppose that we have observational data  $\{(x_i, t_i, y_{t_i, i})\}_{i=1}^N$  that we want to use to estimate  $\tau(x)$ . One way to do so is by estimating

$$m_1(x) = \mathbb{E}_p[Y_1 \mid x]$$
  

$$m_0(x) = \mathbb{E}_p[Y_0 \mid x],$$

and then using  $\hat{\tau}(x) = \hat{m}_1(x) - \hat{m}_0(x)$ . Naïvely, we can use supervised learning to fit one model  $\hat{f}_0$  to predict  $Y_0$  on samples for which  $t_i = 0$ , yielding an estimate  $\hat{f}_0(x) \approx m_0(x)$ , and a second model  $\hat{f}_1$  to predict  $Y_1$  on samples for which  $t_i = 1$ , yielding an estimate  $\hat{f}_1(x) \approx m_1(x)$ .

This approach corresponds to fitting  $\hat{f}_0(x)$  on samples  $(x, y_0)$  from  $p(x, y_0 \mid T = 0)$ , and fitting  $\hat{f}_1(x)$  on samples  $(x, y_1)$  from  $p(x, y_1 \mid T = 1)$ . However, when evaluating

the PEHE, we are also concerned with the errors of  $\hat{f}_0(x)$  and  $\hat{f}_1(x)$  on the *counterfactual* distributions  $p(x,y_0 \mid T=1)$  and  $p(x,y_1 \mid T=0)$ , respectively—i.e., when fitting  $\hat{f}_0$ , we also need samples  $(x,y_0) \sim p(x,y_0 \mid T=1)$ , and when fitting  $\hat{f}_1$ , we also need samples  $(x,y_1) \sim p(x,y_1 \mid T=0)$ . Otherwise, our estimate  $\hat{\tau}(x)$  may be poor.

Thus, the error  $\epsilon_{\text{PEHE}}$  contains a term that comes from the discrepancy between the factual and counterfactual distributions. More precisely, by strong ignorability,

$$p(x, y_0 \mid T = 0)$$
=  $p(y_0 \mid X = x, T = 0) \cdot p(x \mid T = 0)$   
=  $p(y_0 \mid X = x, T = 1) \cdot p(x \mid T = 0)$ .

Comparing this with

$$p(x, y_0 \mid T = 1) = p(y_0 \mid X = x, T = 1) \cdot p(x \mid T = 1),$$

we observe that the difference between these factual and counterfactual distributions are captured by the difference in the distributions  $p(x \mid T = 0)$  and  $p(x \mid T = 1)$ .

**Definition 2.6.** The distribution of control units is  $p^{t=1}(x) = p(x \mid T=0)$ , and the distribution of treated units is  $p^{t=0}(x) = p(x \mid T=1)$ .

For this source of error to be small, we need  $p^{t=1}(x)$  to be similar to  $p^{t=0}(x)$ . However, for observational data, unlike RCT data, these distributions are given to us, and are not ones that we can choose.

As proposed by (Johansson et al., 2016; Shalit et al., 2017), one solution is to split the prediction problem into two steps: (i) learn a representation  $\Phi: \mathcal{X} \to \mathcal{R}$  for some embedding space  $\mathcal{R} \subseteq \mathbb{R}^{\ell}$ , and (ii) fit a predictive model on  $\mathcal{R}$  rather than on  $\mathcal{X}$ . Then, we can bound the error coming from the discrepancy between  $p^{t=1}(x)$  and  $p^{t=0}(x)$  by the discrepancy between  $\Phi(X) \mid T = 0$  and  $\Phi(X) \mid T = 1$ .

**Assumption 2.7.** The representation  $\Phi$  is a twice-differentiable, one-to-one function. Without loss of generality, we assume that  $\mathcal R$  is the image of  $\mathcal X$  under  $\Phi$ , so that we can define an inverse  $\Phi^{-1}:\mathcal R\to\mathcal X$ .

Next, we define the distributions on  $\mathcal R$  induced by the distributions of treated units and of control units.

**Definition 2.8.** For  $r\in\mathcal{R}$ , define  $p_{\Phi}^{t=0}(r)$  to be the density at r of  $\Phi(X)\mid T=1$ , and define  $p_{\Phi}^{t=1}(r)$  to be the density of  $\Phi(X)\mid T=0$ .

In other words,  $p_{\Phi}^{t=0}(r)$  is the distribution of treated units on  $\mathcal R$  induced by  $\Phi$ , and  $p_{\Phi}^{t=1}$  is the distribution of control units on  $\mathcal R$  induced by  $\Phi$ .

We can now combine the estimates of  $m_1(x)$ ,  $m_0(x)$  into a single function. In particular, consider hypotheses of the

form  $f: \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ , where we estimate  $m_1(x)$  by f(x,1) and  $m_0(x)$  by f(x,0). We are interested in the case where f is derived from an estimator  $h: \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ .

**Definition 2.9.** Given a representation  $\Phi: \mathcal{X} \to \mathbb{R}$ , we say a hypothesis f factors through  $\Phi$  if there exists  $h: \mathcal{R} \times \{0,1\} \to \mathcal{Y}$  such that  $f(x,t) = h(\Phi(x),t)$ .

Then, we consider the following estimate of  $\tau(x)$ :

**Definition 2.10.** The treatment effect estimate of the hypothesis f for a unit with covariate x is

$$\hat{\tau}_f(x) = f(x,1) - f(x,0).$$

We let  $\epsilon_{\text{PEHE}}(f) = \epsilon_{\text{PEHE}}(\hat{\tau}_f)$ . When f factors through a representation  $\Phi$ —i.e.,  $f(x,t) = h(\Phi(x),t)$ —we let  $\epsilon_{\text{PEHE}}(h,\Phi) = \epsilon_{\text{PEHE}}(f)$ .

**Bound on causal error.** Our goal is to bound  $\epsilon_{\text{PEHE}}(h,\Phi)$ . We describe a bound on  $\epsilon_{\text{PEHE}}$  proven in (Shalit et al., 2017) for approaches to estimating the ITE  $\hat{\tau}(x)$  based on causal representations. We have two derived loss functions, one corresponding to the factual loss  $\epsilon_{\text{F}}$  and another corresponding to the counterfactual loss  $\epsilon_{\text{CF}}$ .

**Definition 2.11.** Given  $h, \Phi$ , the expected loss for the unit and treatment pair (x, t) is

$$l_{h,\Phi}(x,t) = \int_{\mathcal{Y}} (Y_t - h(\Phi(x),t))^2 p(Y_t \mid x) dY_t,$$

and the expected factual and counterfactual losses of  $h,\Phi$  are

$$\epsilon_{\mathrm{F}}(h,\Phi) = \int_{\mathcal{X} \times \{0,1\}} l_{h,\Phi}(x,t) p(x,t) dx dt$$

$$\epsilon_{\mathrm{CF}}(h,\Phi) = \int_{\mathcal{X} \times \{0,1\}} l_{h,\Phi}(x,t) p(x,1-t) dx dt.$$

We break up the factual loss  $\epsilon_F(h, \Phi)$  into two parts based on the following definition.

**Definition 2.12.** The expected factual treated and control losses are

$$\epsilon_{\mathcal{F}}^{t=0}(h,\Phi) = \int_{\mathcal{X}} l_{h,\Phi}(x,1) p^{t=0}(x) dx$$

$$\epsilon_{\mathcal{F}}^{t=1}(h,\Phi) = \int_{\mathcal{X}} l_{h,\Phi}(x,0) p^{t=1}(x) dx.$$

It follows immediately that

$$\epsilon_{\mathcal{F}}(h, \Phi) = \mathbb{P}_p[T = 1] \cdot \epsilon_{\mathcal{F}}^{t=1}(h, \Phi) + \mathbb{P}_p[T = 0] \cdot \epsilon_{\mathcal{F}}^{t=0}(h, \Phi).$$

<sup>&</sup>lt;sup>2</sup>We assume that we are using the squared loss.

One term in the bound on  $\epsilon_{\text{PEHE}}(h, \Phi)$  from (Shalit et al., 2017) quantifies the quality of  $\Phi$ , through the discrepancy between two distributions  $p_{\Phi}^{t=1}(r)$  and  $p_{\Phi}^{t=0}(r)$ . We use the following metric to measure this discrepancy:

**Definition 2.13.** Suppose we have two probability distributions p and q on  $S \subseteq \mathbb{R}^d$ . Given a family of functions  $G \subseteq \{g : S \to \mathbb{R}\}$ , we have

$$IPM_G(p,q) = \sup_{g \in G} \left| \int_{\mathcal{S}} g(s)(p(s) - q(s)) ds \right|$$

To obtain guarantees, we require the following assumption on the function family G:

**Assumption 2.14.** The family  $G \subseteq \{g : \mathcal{R} \to \mathbb{R}\}$  satisfies

$$\frac{1}{B_{\Phi}} \cdot l_{h,\Phi}(\Phi^{-1}(r),0), \ \frac{1}{B_{\Phi}} \cdot l_{h,\Phi}(\Phi^{-1}(r),1) \in G.$$

for some  $B_{\Phi} > 0$ .

Then, one desirable property of the representation  $\Phi$  is for  $\mathrm{IPM}_G(p_\Phi^{t=0}, p_\Phi^{t=1})$  to be small. The other term in the bound on the error  $\epsilon_{\mathrm{PEHE}}$  comes from the variances of  $Y_0, Y_1$ .

**Definition 2.15.** Given a distribution p(x,t) on  $\mathcal{X} \times \{0,1\}$ , we denote the *counterfactual density* of p by  $\tilde{p}$ , defined by  $\tilde{p}(x,t) = p(x,1-t)$ .

**Definition 2.16.** Given a distribution p(x,t) on  $\mathcal{X} \times \{0,1\}$ , the expected variances of  $Y_0$  and  $Y_1$  with respect to p are

$$\sigma_{Y_1}^2(p) = \int_{\mathcal{X} \times \mathcal{Y}} (Y_1 - m_1(x))^2 p(Y_1|x) p(x, 1) dY_1 dx$$
  
$$\sigma_{Y_0}^2(p) = \int_{\mathcal{X} \times \mathcal{Y}} (Y_0 - m_0(x))^2 p(Y_0|x) p(x, 0) dY_0 dx.$$

Furthermore, we let

$$\begin{split} \sigma_{Y_T}^2(p) &= \sigma_{Y_1}^2(p) + \sigma_{Y_0}^2(p) \\ \sigma_{Y}^2(p) &= \min\{\sigma_{Y_T}^2(p), \sigma_{Y_T}^2(\tilde{p})\}, \end{split}$$

We have the following bound on  $\epsilon_{PEHE}$  (Shalit et al., 2017):

**Theorem 2.17.** For any  $f: \mathcal{X} \times \{0,1\} \to \mathcal{Y}$  factored as  $f(x,t) = h(\Phi(x),t)$  for some  $h: \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ ,

$$\begin{split} \frac{1}{2} \epsilon_{\text{PEHE}}(h, \Phi) &\leq \epsilon_{\text{CF}}(h, \Phi) + \epsilon_{\text{F}}(h, \phi) - 2\sigma_Y^2(p) \\ &\leq \epsilon_{\text{F}}^{t=1}(h, \Phi) + \epsilon_{\text{F}}^{t=0}(h, \Phi) - 2\sigma_Y^2(p) \\ &+ B_{\Phi} \cdot \text{IPM}_G(p_{\Phi}^{t=0}, p_{\Phi}^{t=1}). \end{split}$$

This theorem shows that the error  $\epsilon_{\text{PEHE}}(f)$  of our estimate of  $\tau(x)$  can be bounded by two terms. The first term

$$\epsilon_{\mathrm{F}}^{t=1}(h,\Phi) + \epsilon_{\mathrm{F}}^{t=0}(h,\Phi) - 2\sigma_{\mathrm{Y}}^{2}(p)$$

**Algorithm 1** Learning interpretable models with causal guarantees.

input Factual observations 
$$D_{\mathrm{F}} = \{(x_i, t_i, y_{t_i, i})\}_{i=1}^N$$

$$f^* \leftarrow \mathrm{LearnCR}(D_{\mathrm{R}})$$

$$D_0 \leftarrow \{(x_i, t_i)\} \cup \{(x_i, 1 - t_i)\}$$

$$D_{f^*} \leftarrow \{(x, t, f^*(x, t)) \mid (x, t) \in D_0\}$$

$$\hat{f} \leftarrow \mathcal{A}(D_{f^*})$$
output  $\hat{f}$ 

captures the error due to the test error of f on the observational dataset. The second term

$$B_{\Phi} \cdot \mathrm{IPM}_G(p_{\Phi}^{t=0}, p_{\Phi}^{t=1})$$

captures the error due to the mismatch between the distributions  $p_{\Phi}^{t=0}$  of treated units and  $p_{\Phi}^{t=1}$  of control units in the embedding space.

# 3. Interpretable Models for Individual Treatment Effect Estimation

Our learning framework can convert any algorithm for learning interpretable models in the supervised setting into an algorithm for learning interpretable models to predict individual treatment effects. Recall that the key issue with applying the causal representations approach is that we cannot simply train an interpretable model  $h: \mathcal{R} \to \mathcal{Y}$  on the causal representation  $\Phi(x) \in \mathcal{R}$ —in particular, the representation function  $\Phi$  is uninterpretable, so the composed model  $f(x,t) = h(\Phi(x),t)$  is uninterpretable.

**Learning algorithm.** We propose an approach where we first train an uninterpretable *oracle model*  $f^*$  using the causal representation approach, and then train an interpretable model  $\hat{f}: \mathcal{X} \times \{0,1\} \to \mathcal{Y}$  to approximate  $f^*$ . In particular, we prove that using our approach, as long as  $\hat{f}$  closely approximates  $f^*$ , we can obtain a bound on the error of  $\hat{f}$  analogous to Theorem 2.17.

Let  $\mathcal{M}\subseteq\{f:\mathcal{X}\times\{0,1\}\to\mathcal{Y}\}$  be the space of interpretable models considered by  $\mathcal{A}$ . Given observations  $\{(x_i,t_i,y_i)\}_{i=1}^N$  from the distribution of  $(X,T,Y_T)$ , our goal is to learn an interpretable model  $f\in\mathcal{M}$  for which we can provide causal guarantees. Let

$$\mathcal{D} = \bigcup_{n=1}^{\infty} \prod_{i=1}^{n} (\mathcal{X} \times \{0, 1\} \times \mathcal{Y})$$

be the set of datasets of any finite size (i.e., of size n for  $n \in \mathbb{N}$ ). Suppose we have a learning algorithm  $\mathcal{A}: \mathcal{D} \to \mathcal{M}$  for interpretable models—i.e., given a dataset  $D = \{(x_i, t_i, y_i)\}_{i=1}^N \in \mathcal{D}$ , then  $\mathcal{A}$  (usually approximately)

solves the supervised learning problem

$$\mathcal{A}(D) = \underset{f \in \mathcal{M}}{\operatorname{arg\,min}} \sum_{i=1}^{N} (f(x_i, t_i) - y_i)^2. \tag{3}$$

We use  $\hat{f} = \mathcal{A}(D)$  to denote the model returned by  $\mathcal{A}$ .

In addition, suppose we also have an oracle model  $f^*: \mathcal{X} \times \{0,1\} \to \mathcal{Y}$  that is not interpretable (so  $f^* \notin \mathcal{M}$ ), but whose associated estimate  $\hat{\tau}_f(x)$  of  $\tau(x)$  is good. We assume that  $f^*$  is learned using the causal representation approach described in Section 2—in particular, that it factors as  $f^*(x,t) = h^*(\Phi(x),t)$ .

Our approach is to train  $\hat{f}$  to approximate  $f^*$ —i.e.,  $\hat{f} = \mathcal{A}(D_{f^*})$ , where  $D_{f^*} = \{(x_i, t_i, f^*(x_i, t_i))\}_{i=1}^{N'}$  for some set  $D_0 = \{(x_i, t_i)\}_{i=1}^{N'}$  of covariate-treatment pairs. The key question is how to choose  $D_0$  so that  $\hat{f}$  produces a good estimate of  $\tau(x)$ —i.e., so that  $\epsilon_{\text{PEHE}}$  is small.

Intuitively, when we have control over the treatment assignment—e.g., in a randomized controlled trial (RCT)—a good distribution to use is to uniformly randomly assign treatments. In particular, consider the following distribution: **Definition 3.1.** Given a distribution p(x) on  $\mathcal{X}$ , the *RCT distribution*  $q_p(x,t)$  derived from p is the distribution on  $\mathcal{X} \times \{0,1\}$  defined by

$$\mathbb{P}_{q_n}[T=0] = \mathbb{P}_{q_n}[T=1] = 1/2.$$

and

$$q_p(x|T=0) = q_p(x|T=1) = p(x).$$

In other words, the random variables (X,T) have joint distribution  $q_p$  if X is distributed as p(x) and T = Bernoulli(1/2) is independent from X.

Letting p(x) be the empirical distribution over covariates  $x \in \mathcal{X}$ , we show below that  $q_p$  is a good candidate for  $D_0$ . In particular, with this choice, we can prove a bound on  $\epsilon_{\text{PEHE}}$  analogous to Theorem 2.17.

Given an observational dataset  $D_{\rm F}$ , our algorithm (shown in Algorithm 1) first uses the causal representations approach to learn an oracle model  $f^*$  based on  $D_{\rm F}$  that has provable guarantees on  $\epsilon_{\rm PEHE}$  (the subroutine LearnCR). Then, our algorithm constructs the distribution  $D_0 = q_p$ , where p is the empirical distribution of covariates in  $D_{\rm F}$ . Next, our algorithm uses  $f^*$  to label the points in  $D_0$ , producing a dataset  $D_{f^*}$ ; this step amounts to using  $f^*$  to label the unobserved counterfactual for each covariate  $x_i$  in  $D_{\rm F}$ . Finally, our algorithm runs the interpretable learning algorithm  $\mathcal A$  on the training set  $D_{f^*}$ , and returns the result  $\hat f = \mathcal A(D_{f^*})$ .

**Bound on causal error.** We prove that as long as  $\hat{f} \in \mathcal{M}$  is close to  $f^*$  on the distribution  $q_p(x,t)$ , where p is the true covariate distribution, then  $\epsilon_{\text{PEHE}}(\hat{f})$  is small.

**Definition 3.2.** The *relative error* of f to  $f^*$  is

$$\begin{split} \epsilon_{f,f^*} &= \mathbb{E}_{q_p}[(f(x,t) - f^*(x,t))^2] \\ &= \int_{\mathcal{X} \times \{0,1\}} (f(x,t) - f^*(x,t))^2 q_p(x,t) dx dt. \end{split}$$

In other words,  $\epsilon_{f,f^*}$  captures the test error of f relative to the oracle model  $f^*$ . Now, we can bound the generalization error by a combination of  $\epsilon_{f,f^*}$  and the bound on  $\epsilon_{\text{PEHE}}(f^*)$ .

**Theorem 3.3.** For any function  $f: \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ , and any function  $f^*: \mathcal{X} \times \{0,1\} \to \mathcal{Y}$  factored as  $f^*(x,t) = h^*(\Phi(x),t)$  for some  $h^*: \mathcal{R} \times \{0,1\} \to \mathcal{Y}$ , we have

$$\frac{1}{4} \epsilon_{\text{PEHE}}(f) \leq 2\epsilon_{f,f^*} + \epsilon_{\text{F}}(f^*) + \epsilon_{\text{CF}}(f^*) - 2\sigma_Y^2(p) 
\leq 2\epsilon_{f,f^*} + \epsilon_{\text{F}}^{t=0}(h, \Phi) + \epsilon_{\text{F}}^{t=1}(h, \Phi) 
+ B_{\Phi} \cdot \text{IPM}_G(p_{\Phi}^{t=1}, p_{\Phi}^{t=0}) - 2\sigma_Y^2(p).$$

We give a proof in Appendix A. Our bound has three terms—the first term  $8\epsilon_{f,f^*}$  captures the test error of f relative to  $f^*$ . The second two terms are from Theorem 2.17—the second term is the test error of  $f^*$  on the observational dataset, and the third term captures the error due to the mismatch between the distributions  $p_\Phi^{t=0}$  of treated units and  $p_\Phi^{t=1}$  of control units in the latent representation.

While the bound in Theorem 3.3 is stated according to the exact error of  $\hat{f}$  with respect to  $f^*$ , it can be straightforwardly converted to a finite sample bound using standard assumptions—e.g., that the model family  $\mathcal{M}$  has finite Rademacher complexity (Bartlett & Mendelson, 2002) and that  $\mathcal{A}$  solves (3) exactly. The other terms can similarly be converted into finite-sample bounds (Shalit et al., 2017).

Finally, note that we can estimate  $\epsilon_{f,f^*}$  on a held-out test set  $(D_{\rm F})_{\rm test}$  of observational data—it is simply the loss of f on the dataset constructed from  $D_{f^*,\rm test}$  constructed from  $(D_{\rm F})_{\rm test}$  the same way Algorithm 1 constructs  $D_{f^*}$  from  $D_{\rm F}$ . As discussed in (Shalit et al., 2017), the remaining terms in the bound can similarly be estimated on  $(D_{\rm F})_{\rm test}$ . Thus, we can obtain an test set estimate of the bound in Theorem 3.3.

## 4. Experiments

Evaluating the performance of causal models is a challenging task, since ground truth data on individual treatment effects (ITEs) is difficult to obtain. Following previous work (Shalit et al., 2017), we evaluate our framework on the IHDP (Hill, 2011) and Jobs (LaLonde, 1986) datasets.

**IHDP dataset.** We use a dataset for causal inference evaluation based on the Infant Health and Development Program, from (Hill, 2011) and preprocessed by (Shalit et al., 2017)

Model	IHDP				Jobs			
	$\sqrt{\epsilon_{ m PEHE}}$		$\epsilon_{ ext{ATE}}$		$R_{\mathrm{POL}}$		$\epsilon_{ ext{ATT}}$	
	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline
CFR-Net	_	$0.926 \pm 0.02$	_	$0.271 \pm 0.01$	_	$0.235 \pm 0.02$	-	$0.086 \pm 0.03$
CART (depth 6)	$3.668 \pm 0.17$	$4.305 \pm 0.20$	$0.485 \pm 0.03$	$0.679 \pm 0.04$	$0.241 \pm 0.01$	$0.271 \pm 0.02$	$0.086 \pm 0.03$	$0.067 \pm 0.02$
CART (depth 5)	$3.824 \pm 0.18$	$4.436 \pm 0.21$	$0.492 \pm 0.02$	$0.725 \pm 0.05$	$0.241 \pm 0.01$	$0.280 \pm 0.02$	$0.086 \pm 0.03$	$0.069 \pm 0.03$
CART (depth 4)	$4.086 \pm 0.19$	$4.605 \pm 0.22$	$0.530 \pm 0.03$	$0.717 \pm 0.05$	$0.241 \pm 0.01$	$0.281 \pm 0.02$	$0.086 \pm 0.03$	$0.064 \pm 0.0$
CART (depth 3)	$4.462 \pm 0.21$	$4.930 \pm 0.23$	$0.585 \pm 0.03$	$0.795 \pm 0.05$	$0.241 \pm 0.01$	$0.285 \pm 0.02$	$0.086 \pm 0.03$	$0.067 \pm 0.02$
Honest Tree (depth 6)	$3.694 \pm 0.17$	$4.086 \pm 0.19$	$0.481 \pm 0.02$	$0.483 \pm 0.03$	$0.235 \pm 0.02$	$0.223 \pm 0.01$	$0.086 \pm 0.03$	$0.073 \pm 0.03$
Honest Tree (depth 5)	$3.760 \pm 0.17$	$4.098 \pm 0.19$	$0.488 \pm 0.02$	$0.486 \pm 0.03$	$0.235 \pm 0.02$	$0.216 \pm 0.01$	$0.086 \pm 0.03$	$0.074 \pm 0.0$
Honest Tree (depth 4)	$3.875 \pm 0.18$	$4.128 \pm 0.19$	$0.498 \pm 0.02$	$0.488 \pm 0.03$	$0.235 \pm 0.02$	$0.223 \pm 0.02$	$0.086 \pm 0.03$	$0.084 \pm 0.0$
Honest Tree (depth 3)	$4.090 \pm 0.19$	$4.237 \pm 0.20$	$0.535 \pm 0.03$	$0.498 \pm 0.03$	$0.235 \pm 0.02$	$0.236 \pm 0.01$	$0.086 \pm 0.03$	$0.080 \pm 0.0$
LASSO	$5.725 \pm 0.26$	$5.777 \pm 0.26$	$0.671 \pm 0.04$	$0.942 \pm 0.05$	$0.235 \pm 0.02$	$0.226 \pm 0.02$	$0.086 \pm 0.03$	$0.080 \pm 0.0$
Kernel Ridge	$2.077 \pm 0.09$	$3.190 \pm 0.14$	$0.361 \pm 0.02$	$0.562 \pm 0.02$	$0.235 \pm 0.02$	$0.234 \pm 0.02$	$0.086 \pm 0.03$	$0.077 \pm 0.0$
GBM	$1.845 \pm 0.09$	$2.799 \pm 0.14$	$0.352 \pm 0.02$	$0.453 \pm 0.03$	$0.241 \pm 0.01$	$0.223 \pm 0.02$	$0.086 \pm 0.03$	$0.080 \pm 0.0$
Random Forest	$2.905 \pm 0.14$	$3.653 \pm 0.19$	$0.439 \pm 0.02$	$0.621 \pm 0.04$	$0.241 \pm 0.01$	$0.239 \pm 0.01$	$0.086 \pm 0.03$	$0.073 \pm 0.0$

Table 1. We show results comparing our approach to a baseline estimator for a number of model families on the IHDP and Jobs datasets. For each value, we show the mean  $\pm$  the standard error. We bold the better of the two values between ours and the baseline.

using the NPCI package (Hill, 2016). The dataset has 747 units (139 treated, 708 control) and 25 covariates of children and their mothers. This dataset contains 1000 realizations of the outcomes with 63/27/10 train/validation/test splits. The outcomes in this dataset are simulated—i.e., we have ground truth values of the ITE for each unit. Using this ground truth, we can obtain a test set estimate  $\hat{\epsilon}_{\text{PEHE}}(f)$  of the error in the predicted ITE. Then, we report the mean and standard errors of  $\sqrt{\hat{\epsilon}_{\text{PEHE}}(f)}$ , as well as the absolute error in the average treatment effect (ATE)

$$\epsilon_{\text{ATE}} = \left| \frac{1}{n} \sum_{i=1}^{n} (\hat{\tau}(x_i) - \tau(x_i)) \right|$$

over the 1000 realizations. Our primary metric of interest is  $\sqrt{\hat{\epsilon}_{\mathrm{PEHE}}(f)}$ , which measures predictive accuracy of ITEs, whereas  $\epsilon_{\mathrm{ATE}}$  measures predictive accuracy of the ATE.

Jobs dataset. We use the Jobs dataset from (Shalit et al., 2017) based on (LaLonde, 1986), where the binary outcome is employment (versus unemployment). This dataset (3212 individuals) is a combination of data from a randomized trial (297 treated and 425 control) and data from an observational study (2490 control). A difficulty with the Jobs dataset is that we do not have ground truth on the ITEs. Instead, we use a metric based proposed in (Shalit et al., 2017), which evaluates a policy  $\pi_f$  that makes treatment decisions based on the predictions of f. In particular, recall that f(x,t) is the predicted outcome for a unit with covariates x and treatment t. We consider the policy  $\pi_f$  that assigns this unit to treatment if the predicted treatment effect is positive—i.e., if f(x,1) > f(x,0). Then, the *policy risk* 

$$R_{\text{POL}}(\pi_f) = 1 - \mathbb{E}[Y_1 | \pi_f(x) = 1] \cdot \mathbb{P}_p[\pi_f = 1]$$
$$- \mathbb{E}[Y_0 | \pi_f(x) = 0] \cdot \mathbb{P}_p[\pi_f = 0]$$

measures the quality of outcomes on average over the test population. For any predictor f, we can estimate  $R_{POL}(\pi_f)$ 

on the randomized subset of the Jobs data as follows:

$$\hat{R}_{POL}(\pi_f) = 1 - \mathbb{E}[Y_1 | \pi_f(x) = 1, T = 1] \cdot \mathbb{P}_{\hat{p}}[\pi_f = 1] - \mathbb{E}[Y_0 | \pi_f(x) = 0, T = 0] \cdot \mathbb{P}_{\hat{p}}[\pi_f = 0].$$

We also use the randomized subset to estimate the "ground truth" effect. In particular, let T, E, C be the set of units in the treated subgroup, the randomized study, and in the control subgroup, respectively (note that  $T \subset E$ ). We report the treatment effect on the treated by

$$ATT = |T|^{-1} \sum_{i \in T} y_i - |C \cap E|^{-1} \sum_{i \in C \cap E} y_i$$

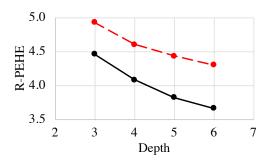
and use as one metric

$$\epsilon_{\text{ATT}} = |\text{ATT}| = |T|^{-1} \left| \sum_{i \in T} (f(x_i, 1) - f(x_i, 0)) \right|.$$

We report the mean and standard error of  $\hat{R}_{POL}(\pi_f)$  and  $\epsilon_{ATT}$  over 10 outcomes with 56/24/20 train/validation/test splits. For this study, our primary outcome of interest is the  $R_{POL}$ , since it to some degree measures the predictive accuracy of ITEs; in contrast, similar to  $\epsilon_{ATE}$ ,  $\epsilon_{ATT}$  measures the predictive accuracy of a population average effect.

**Oracle model.** For  $f^*$ , we train a CFR-net from (Shalit et al., 2017), which has 3 fully connected exponential-linear layers for each the representation function  $\Phi$  and for the prediction function  $h^*$ , with layer sizes 100 for all layers used for Jobs and 200 and 100 for the representation and hypothesis layers for IHDP. For IHDP, we used mean squared loss; for Jobs, we use logistic loss.

**Interpretable models.** We evaluate the performance of our approach on a variety of models with a range of interpretability: CART trees (Breiman, 2017), honest trees (Athey & Imbens, 2016), LASSO regression (Tibshirani, 1996), kernel ridge regression (Murphy, 2012), gradient boosted models (GBMs) (Friedman, 2001), and random



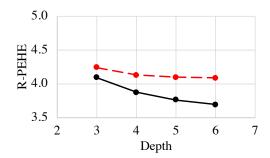


Figure 1. Performance (in terms of  $\sqrt{\epsilon_{\text{PEHE}}}$ ) of CART (left) and honest trees (right) using our approach (black, solid) and the baseline approach (red, dashed) on the IHDP dataset, as a function of the depth of the decision tree.

forests (Breiman, 2001). For each model family, we train one model using our approach, and a baseline model using only the observational data for training.

Of these models, only honest trees are designed to handle causality; however, their focus is on obtaining unbiased estimates rather than low-variance estimates. In particular, they split the dataset into two, using the first part to estimate splits and the second to estimate values at the leaf nodes. This approach ensures that the estimates at the leaf nodes are unbiased, but also greatly increases variance since they are only using half the data at each point.

**Results.** We show results in Table 1. Note that we run CART and honest trees with different maximum depths; Figure 1 shows how  $\sqrt{\epsilon_{\rm PEHE}}$  scales with depth on IHDP.

**Discussion.** On the IHPD dataset, our approach uniformly outperforms the baseline approach in terms of  $\sqrt{\epsilon_{\rm PEHE}}$ , which measures performance on predicting ITEs. Even on predicting ATEs, our approach mostly outperforms the baseline; the only exception are honest trees, which are interpretable models tailored towards estimating treatment effects. As we discussed before, honest trees are focused on reducing bias at the expense of increased variance. Otherwise, we observe the usual trends—more complex models (e.g., GBMs and random forests) outperform more interpretable models (LASSO, CART, honest trees).

On the Jobs dataset, our performance was more mixed. Our approach significantly benefited CART in terms of  $R_{\rm POL}$ , as well as honest trees of depth 3. However, for the remaining models (including honest trees of depth  $\geq 4$ ), the baseline approach outperformed ours.

The problem is that the oracle model CFR-Net did not perform as well as even some of the simpler models—indeed, the baseline honest tree of depth 5 was the best performing model on the dataset. In particular, we were unable to replicate the results of (Shalit et al., 2017), despite using their available code and obtaining the original

train/validation/test splits from the authors. The gap in our performance ( $R_{\rm POL}=0.235$ ) relative to ones reported in (Shalit et al., 2017) ( $R_{\rm POL}=0.21$ ) is not very large; however, even in their results, a number of baseline models perform very similarly (or even better) than CFR-Net.

As a consequence, many of the models trained using our approach achieved performance equal to that of CFR-Net—in particular, since we are training our models using labels provided by CFR-Net as ground truth, we cannot expect to do better than than their performance (i.e.,  $R_{\rm POL}=0.235$  and  $\epsilon_{\rm ATT}=0.086$ ). Furthermore, CFR-Net appears to have learned a relatively simple function, since LASSO and kernel ridge regression both performed exactly as well CFR-Net when trained to imitate it; similarly, none of the CART and honest trees trained to imitate CFR-Net grew beyond depth 3.

In summary, while our approach proved less useful for the Jobs dataset, where simple models already perform as well as (or better than) more expressive models, our results on the IHDP dataset clearly demonstrate the potential for our approach to substantially improve the performance of interpretable learning algorithms used to predict ITEs.

#### 5. Conclusion

We have proposed a general framework for learning interpretable models with causal guarantees. A number of directions remain for future work. Most importantly, as with previous work, our approach makes the strong ignorability assumption. The predominant approach to avoiding this assumption is to use instrumental variables. Incorporating these ideas with the instrumental variables framework could enable causal guarantees without strong ignorability.

### References

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., and Rudin, C. Learning certifiably optimal rule lists. In *Proceedings of the 23rd ACM SIGKDD International* 

- Conference on Knowledge Discovery and Data Mining, pp. 35–44. ACM, 2017.
- Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.
- Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.
- Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.
- Bastani, H. and Bayati, M. Online decision-making with high-dimensional covariates. 2015.
- Bastani, O., Kim, C., and Bastani, H. Interpreting blackbox models via model extraction. *arXiv* preprint *arXiv*:1705.08504, 2017.
- Bastani, O., Pu, Y., and Solar-Lezama, A. Verifiable reinforcement learning via policy extraction. In *NIPS*, 2018.
- Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.
- Breiman, L. Classification and regression trees. Routledge, 2017.
- Bucilua, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541. ACM, 2006.
- Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730. ACM, 2015.
- Consortium, I. W. P. Estimation of the warfarin dose with clinical and pharmacogenetic data. *New England Journal* of Medicine, 360(8):753–764, 2009.
- Ellis, K., Solar-Lezama, A., and Tenenbaum, J. Unsupervised learning by program synthesis. In *Advances in neural information processing systems*, pp. 973–981, 2015.
- Ellis, K., Ritchie, D., Solar-Lezama, A., and Tenenbaum, J. Learning to infer graphics programs from hand-drawn images. In *Advances in Neural Information Processing Systems*, pp. 6062–6071, 2018.
- Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

- Frosst, N. and Hinton, G. Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*, 2017.
- Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.
- Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299): 299ra122–299ra122, 2015.
- Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 2011.
- Hill, J. L. Npci: Non-parametrics for causal inference. https://github.com/vdorie/npci, 2016.
- Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.
- Kim, E. S., Herbst, R. S., Wistuba, I. I., Lee, J. J., Blumenschein, G. R., Tsao, A., Stewart, D. J., Hicks, M. E., Erasmus, J., Gupta, S., et al. The battle trial: personalizing therapy for lung cancer. *Cancer discovery*, 2011.
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293, 2017.
- Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD in*ternational conference on knowledge discovery and data mining, pp. 1675–1684. ACM, 2016.
- Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154, 2017.
- LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pp. 604–620, 1986.
- Lou, Y., Caruana, R., and Gehrke, J. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–158. ACM, 2012.

- Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latentvariable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.
- Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.
- Pearl, J. Causal inference. In *Causality: Objectives and Assessment*, pp. 39–58, 2010.
- Ribeiro, M. T., Singh, S., and Guestrin, C. Model-agnostic interpretability of machine learning. In *KDD*, 2016.
- Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 2005.
- Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
- Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* (*Methodological*), pp. 267–288, 1996.
- Ustun, B. and Rudin, C. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.
- Valkov, L., Chaudhari, D., Srivastava, A., Sutton, C., and Chaudhuri, S. Houdini: Lifelong learning as program synthesis. In *Advances in Neural Information Processing Systems*, pp. 8701–8712, 2018.
- Verma, A., Murali, V., Singh, R., Kohli, P., and Chaudhuri, S. Programmatically interpretable reinforcement learning. In *ICML*, 2018.
- Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.
- Wang, F. and Rudin, C. Falling rule lists. In *Artificial Intelligence and Statistics*, pp. 1013–1022, 2015.
- Wooldridge, J. M. *Introductory econometrics: A modern approach*. Nelson Education, 2015.
- Yang, H., Rudin, C., and Seltzer, M. Scalable bayesian rule lists. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3921–3930. JMLR. org, 2017.

## A. Proof of Theorem 3.3

From the proof of Theorem 1 in (Shalit et al., 2017), we have

$$\epsilon_{\text{PEHE}}(f) \le 2 \int_{\mathcal{X}} (f(x,t) - m_t(x))^2 p(x,t) dx dt + 2 \int_{\mathcal{X}} (f(x,t) - m_t(x))^2 p(x,1-t) dx dt. \tag{4}$$

Then, we have

$$\int_{\mathcal{X}\times\{0,1\}} (f(x,t) - m_t(x))^2 p(x,t) dx dt 
= \int_{\mathcal{X}\times\{0,1\}} ((f(x,t) - f^*(x,t)) + (f^*(x,t) - m_t(x)))^2 p(x,t) dx dt 
\leq 2 \int_{\mathcal{X}\times\{0,1\}} (f(x,t) - f^*(x,t))^2 p(x,t) dx dt + 2 \int_{\mathcal{X}\times\{0,1\}} (f^*(x,t) - m_t(x))^2 p(x,t) dx dt 
= 2 \int_{\mathcal{X}\times\{0,1\}} (f(x,t) - f^*(x,t))^2 p(x,t) dx dt + 2(\epsilon_{\mathcal{F}}(f^*) - \sigma_{Y_T}^2(p)),$$
(5)

where equation (5) follows from Lemma A5 in (Shalit et al., 2017). Similarly,

$$\begin{split} \int_{\mathcal{X}\times\{0,1\}} (f(x,t)-m_t(x))^2 p(x,1-t) dx dt \\ &= \int_{\mathcal{X}\times\{0,1\}} ((f(x,t)-f^*(x,t)) + (f^*(x,t)-m_t(x)))^2 p(x,1-t) dx dt \\ &\leq 2 \int_{\mathcal{X}\times\{0,1\}} (f(x,t)-f^*(x,t))^2 p(x,1-t) dx dt + 2 \int_{\mathcal{X}\times\{0,1\}} (f^*(x,t)-m_t(x))^2 p(x,1-t) dx dt \\ &= 2 \int_{\mathcal{X}\times\{0,1\}} (f(x,t)-f^*(x,t))^2 p(x,1-t) dx dt + 2 (\epsilon_{\mathrm{CF}}(f^*)-\sigma_{Y_T}^2(\tilde{p})). \end{split}$$

Plugging this in equation 4, we obtain

$$\epsilon_{\text{PEHE}}(f) \leq 4 \int_{\mathcal{X} \times \{0,1\}} (f(x,t) - m_t(x))^2 (p(x,t) + p(x,1-t)) dx dt + 4(\epsilon_{\text{F}}(f^*) - \sigma_{Y_T}^2(p)) + 4(\epsilon_{\text{CF}}(f^*) - \sigma_{Y_T}^2(\tilde{p}))$$

$$= 4 \int_{\mathcal{X} \times \{0,1\}} (f(x,t) - m_t(x))^2 (2q_p(x,t)) dx dt + 4(\epsilon_{\text{F}}(f^*) - \sigma_{Y_T}^2(p)) + 4(\epsilon_{\text{CF}}(f^*) - \sigma_{Y_t}^2(\tilde{p}))$$

$$= 8\epsilon_{f,f^*} + 4(\epsilon_{\text{F}}(f^*) - \sigma_{Y_T}^2(p)) + 4(\epsilon_{\text{CF}}(f^*) - \sigma_{Y_T}^2(\tilde{p})).$$

The result follows from the definition of  $\sigma_Y^2(p)$  and Theorem 2.17 (i.e., Theorem 1 of (Shalit et al., 2017)).