Counterfactual Cross-Validation:
Stable Model Selection Procedure for Causal Inference Models

Yuta Saito [1] **Shota Yasui** [2]

Abstract

We study the model selection problem in *conditional average treatment effect* (CATE) prediction. Unlike previous works on this topic, we focus on preserving the rank order of the performance of candidate CATE predictors to enable accurate and stable model selection. To this end, we analyze the model performance ranking problem and formulate guidelines to obtain a better evaluation metric. We then propose a novel metric that can identify the ranking of the performance of CATE predictors with high confidence. Empirical evaluations demonstrate that our metric outperforms existing metrics in both model selection and hyperparameter tuning tasks.

1. Introduction

Predicting conditional average treatment effect (CATE) for certain actions is essential for optimizing metrics of interest in various domains. In digital marketing, incrementality is becoming increasingly important as a performance metric (Diemert et al., 2018). For instance, for a given product, users who will be shown its ads should be chosen based on CATE. It can help avoid showing ads to a user who will buy that product even without seeing the ads. There can be significant applications of CATE prediction in the healthcare segment as well (Alaa & van der Schaar, 2017). This is because, for pursuing an optimal precision medicine, we need to know which treatments will be beneficial or harmful for a particular patient.

To achieve high-accuracy CATE prediction, one has to address the fundamental problem of causal inference, which is that both treated and untreated outcomes can never be observed simultaneously from the same unit (Holland, 1986). Hence we are unable to observe a causal effect and to use it as label to train prediction models. Most previous studies related to the CATE prediction focused on developing methods that can address this fundamental problem and achieve high prediction accuracy (Yoon et al., 2018; Yao et al., 2018; Louizos et al., 2017; Shalit et al., 2017; Du et al., 2019; Alaa & Schaar, 2018).

In model evaluation and selection, the fundamental problem of causal inference poses an additional critical challenge. Because labels are not observed directly, we are unable to calculate loss metrics such as *mean squared error* (MSE). Therefore, data-driven validation procedures such as cross-validation are not directly applicable to model selection and hyperparameter tuning of CATE prediction models. This makes it challenging to identify the suitable model and appropriate hyperparameter values that should be used when applying CATE prediction to real-world problems.

Several prior studies tackle the model evaluation problem in CATE prediction. (Gutierrez & Gérardy, 2017) proposed using the inverse probability weighting (IPW) outcome as the pseudo-label for the true CATE for the calculation of an evaluation metric. (Schuler et al., 2018) used the loss function of R-learner (Nie & Wager, 2017) for the evaluation. (Alaa & Van Der Schaar, 2019) used influence functions to obtain a more efficient estimator for the loss. These works are mainly focused on improving the accuracy of estimating the evaluation metric of interest.

Unlike previous works, we focus on choosing the best model or hyperparameters from potential candidates. For this purpose, ***we only need to know the rank order of the performance of candidate predictors***, which is easier than directly estimating the true performance. To achieve this, we first theoretically analyze the problem of ranking the true performance of CATE predictors and identify the conditions that an ideal metric should satisfy. Building on the analysis, we propose a novel evaluation procedure that preserves the true performance ranking of candidate predictors and minimizes the upper bound of the finite sample uncertainty in model selection. Through empirical evaluations, we demonstrate that the proposed metric performs better than existing heuristic metrics in model selection and hyperparameter tuning of CATE predictors.

[1] Tokyo Institute of Technology, [2] CyberAgent, Inc. Correspondence to: Yuta Saito <saito.y.bj@m.titech.ac.jp>, Shota Yasui <yasui_shota@cyberagent.co.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

2. Related Work

CATE prediction has been extensively studied by combining causal inference and machine learning techniques aiming for the best possible personalization of interventions. State-of-the-art approaches are constructed by utilizing the adversarial generative model, Gaussian process, deep neural networks, and latent variable models (Yoon et al., 2018; Alaa & Schaar, 2018; Louizos et al., 2017; Alaa & van der Schaar, 2017; Hassanpour & Greiner, 2019; 2020; Shi et al., 2019; Bica et al., 2020; Yao et al., 2020). Among the diverse methods that predict CATE from observational data, the approach that is most related to this work is the method based on representation learning (Bengio et al., 2013; Johansson et al., 2020). All methods based on representation learning attempt to map the original feature vectors into the desirable latent representation space so that it eliminates selection biases. Balancing neural network (Johansson et al., 2016) is the most basic method that uses discrepancy distance (Mansour et al., 2009), a domain discrepancy measure in unsupervised domain adaptation for the regularization term. Counterfactual regression (Shalit et al., 2017) minimizes the upper bound of the ground-truth loss for the CATE by utilizing an integral probability metric (Sriperumbudur et al., 2012). In addition to these, methods that obtain a latent representation by preserving a pairwise similarity (Yao et al., 2018; 2019) or by applying adversarial learning (Du et al., 2019) have been proposed.

The prediction methods stated above have provided promising results on standard benchmark datasets. However, previous studies have evaluated such CATE predictors by using synthetic datasets or simple heuristic metrics such as policy risk (Yoon et al., 2018; Shalit et al., 2017; Yao et al., 2018). However, these evaluations do not give a definitive answer about which models would actually be best suited for a given real-world dataset (Alaa & Van Der Schaar, 2019; Setoguchi et al., 2008). Therefore, to bridge the gap between CATE prediction and applications, developing a reliable evaluation metric is critical.

There are only a few studies directly tackling the evaluation problem of CATE prediction models. (Schuler et al., 2018) conducted an extensive survey of several heuristic metrics and provided experimental comparisons. In particular, they introduced inverse probability weighting (IPW) validation, which utilizes an unbiased estimator for the true CATE as an alternative to the true causal effects, and τ-risk, which is based on a loss function of R-learner (Nie & Wager, 2017). In addition, they showed that these metrics empirically outperformed another naive metric, μ-risk, which estimates predictive risk separately for treated and control outcomes using only factual samples. In contrast, (Rolling & Yang, 2014) proposed a propensity matching-based metric called TECV and showed its consistency to the true ranking of the performance of CATE prediction models. However, they did not analyze the uncertainty of the metric, such as its asymptotic variance. It was also empirically outperformed by IPW validation (Schuler et al., 2018). Nonetheless, (Alaa & Van Der Schaar, 2019) improved heuristic plug-in metrics by introducing a meta-estimation technique using influence functions in a theoretically sophisticated manner. Our proposed metric can be further improved by an estimation method based on influence functions.

All the existing metrics aim to estimate the true metric of interest directly, or they do not consider the uncertainty in model selection. However, to conduct accurate model selection and hyperparameter tuning, it is essential to rank model performance accurately, although the aforementioned metrics do not always guarantee the preservation of such rankings. Moreover, analysis of the uncertainty of the model evaluation is necessary, especially in domains in which the size of the validation datasets might be small (e.g., education or public health). Therefore, in contrast to previous works, we investigate a method to accurately preserve the rank order of performance of the candidate predictors while also analyzing the finite sample uncertainty in model selection.

3. Setup

We denote $X \in \mathcal{X} \subseteq \mathbb{R}^d$ as a d-dimensional feature vector and $T \in \mathcal{T} = \{0, 1\}$ as a binary treatment assignment indicator. When an individual i receives treatment, then $T_i = 1$, otherwise, $T_i = 0$. We follow the *potential outcome framework* (Rosenbaum & Rubin, 1983; Rubin, 2005; Imbens & Rubin, 2015) and assume that there exist two potential outcomes denoted as $Y(0), Y(1) \in \mathcal{Y} \subseteq \mathbb{R}$ for each individual. $Y(0)$ is a potential outcome associated with $T = 0$, and $Y(1)$ is associated with $T = 1$. Note that each individual receives only one treatment and reveals the outcome value for the received treatment. We use $p(X, T, Y(0), Y(1))$, or simply p, to denote the joint probability distribution of these random variables.

We formally define the *conditional average treatment effect* (CATE) for a given feature vector $x \in \mathcal{X}$ as:

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x].$$

In addition, we use some notations to represent parameters of p. First, we define the expected potential outcomes conditioned on a feature vector $x \in \mathcal{X}$ as:

$$m_t(x) := \mathbb{E}_{Y(t)}[Y(t) \mid X = x], \; \forall t \in \{0, 1\}.$$

Next, we define the *propensity score* as:

$$e(x) := \mathbb{P}\left(T = 1 \mid X = x\right).$$

This parameter is widely used to estimate treatment effects from observational data (Rosenbaum & Rubin, 1983; Rubin, 1974; Imbens & Rubin, 2015).

Throughout the paper, we make the following standard assumptions in causal inference:

***Assumption* 1.** *(Unconfoundedness) Potential outcomes $(Y(0), Y(1))$ are independent of the treatment assignment indicator T conditioned on the feature vector X, i.e.,*

$$Y(0), Y(1) \perp\!\!\!\perp T \mid X.$$

***Assumption* 2.** *(Overlap) For any $x \in \mathcal{X}$, the true propensity score is strictly between 0 and 1, i.e., $0 < e(x) < 1$.*

***Assumption* 3.** *(Consistency) Observed outcome Y is represented using the potential outcomes and treatment assignment indicator as follows:*

$$Y = TY(1) + (1-T)Y(0).$$

Under these assumptions, the CATE is identifiable from observational data, i.e., $\tau(x) = \mathbb{E}[Y \mid X = x, T = 1] - \mathbb{E}[Y \mid X = x, T = 0]$.

Furthermore, we define some essential notations following (Shalit et al., 2017).

***Definiton* 1.** *(Representation Function) $\Phi : \mathcal{X} \rightarrow \mathcal{R}$ is a representation function and $\mathcal{R}$ is called the representation space. We assume that Φ is a twice differentiable, one-to-one function. Moreover, $p_t^{\Phi} := p(r|t = 1)$ and $p_{1-t}^{\Phi} :- p(r|t = 0)$ are feature distributions for the treated and for the controlled induced over the representation space. We also have $\Psi : \mathcal{R} \rightarrow \mathcal{X}$ as the inverse of Φ, where $\Psi(\Phi(x)) = x, \forall x \in \mathcal{X}$.*

***Definiton* 2.** *(Factual and Counterfactual Loss Functions) Let $h : \mathcal{R} \times \mathcal{T} \rightarrow \mathcal{Y}$ be a hypothesis, $w : \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ be a weighting function, and $L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_{\geq 0}$ be a loss function. In addition, the expected loss for the unit and treatment pair $(x, t) \in \mathcal{X} \times \mathcal{T}$ is denoted as:*

$$\ell_{h,\Phi}^{w}(x, t) := \mathbb{E}_{Y(t)} \left[w(x) L(Y(t), h(\Phi(x), t)) \mid X = x \right].$$

where we use the squared loss: $L(y, y') = (y - y')^2$, hereinafter. Then, the expected factual and counterfactual losses of a combination of a hypothesis h and a representation function Φ are defined as:

$$\epsilon_F^w(h, \Phi) := \int_{\mathcal{X} \times \mathcal{T}} \ell_{h,\Phi}^{w}(x, t) p(x, t) dx dt,$$
$$\epsilon_{CF}^w(h, \Phi) := \int_{\mathcal{X} \times \mathcal{T}} \ell_{h,\Phi}^{w}(x, t) p(x, 1-t) dx dt.$$

where F and CF stand for factual and counterfactual, respectively. Further, the expected factual and counterfactual losses on the treated ($t = 1$) and on the controlled ($t = 0$) are represented as:

$$\epsilon_{F_1}^w(h, \Phi) := \int_{\mathcal{X}} \ell_{h,\Phi}^{w}(x, t = 1) p_1(x) dx,$$
$$\epsilon_{F_0}^w(h, \Phi) := \int_{\mathcal{X}} \ell_{h,\Phi}^{w}(x, t = 0) p_0(x) dx,$$
$$\epsilon_{CF_1}^w(h, \Phi) := \int_{\mathcal{X}} \ell_{h,\Phi}^{w}(x, t = 0) p_1(x) dx,$$
$$\epsilon_{CF_0}^w(h, \Phi) := \int_{\mathcal{X}} \ell_{h,\Phi}^{w}(x, t = 1) p_0(x) dx.$$

where $p_t(x) := p(x \mid T = t)$.

By the definition of the conditional probability, the following equations hold for factual and counterfactual losses:

$$\epsilon_F^w(h, \Phi) = \pi_1 \cdot \epsilon_{F_1}^w(h, \Phi) + \pi_0 \cdot \epsilon_{F_0}^w(h, \Phi),$$
$$\epsilon_{CF}^w(h, \Phi) = \pi_1 \cdot \epsilon_{CF_1}^w(h, \Phi) + \pi_0 \cdot \epsilon_{CF_0}^w(h, \Phi),$$

where $\pi_t := \mathbb{P}(T = t)$.

We also define a class of metrics between probability distributions (Sriperumbudur et al., 2012).

***Definiton* 3.** *(Integral Probability Metric) For two probability density functions defined over a space $\mathcal{S} \subseteq \mathbb{R}^d$ and for a family of functions $G := \{g : \mathcal{S} \rightarrow \mathbb{R}\}$, the IPM between the two density functions p and q is defined as:*

$$\mathrm{IPM}_G(p, q) := \sup_{g \in G} \left| \int_{\mathcal{S}} g(s) \left(p(s) - q(s) \right) ds \right|.$$

Function families G can be the family of bounded continuous functions, the family of 1-Lipschitz functions, and the unit-ball of functions in a universal reproducing Hilbert kernel space.

3.1. Evaluation of CATE prediction models

In previous studies (Gutierrez & Gérardy, 2017; Schuler et al., 2018; Alaa & Van Der Schaar, 2019), the evaluation of a CATE predictor $\hat{\tau}(\cdot)$ has been formulated as accurately estimating the following ground-truth performance metric from a size n of i.i.d observational validation dataset $\mathcal{V} = \{X_i, T_i, Y_i\}_{i=1}^n$:

$$\begin{aligned} \mathcal{R}_{true}(\hat{\tau}) &:= \mathbb{E}_X \left[L \left(\tau(X), \hat{\tau}(X) \right) \right] \\ &= \mathbb{E}_X \left[\left(\tau(X) - \hat{\tau}(X) \right)^2 \right], \end{aligned} \quad (1)$$

where $\mathcal{R}_{true}(\hat{\tau})$ is the true performance metric of $\hat{\tau}(\cdot)$[1].

This approach is intuitive and ideal. However, realizations of the true CATE are never observable, and thus, accurate performance estimation is difficult. Moreover, estimating

[1] Eq. (1) is also termed as the *expected precision in estimation of heterogeneous effect* (PEHE).

the true metric values is not always necessary to conduct valid model selection or hyperparameter tuning. It may be possible to obtain a better evaluation metric under an objective specific to selection and tuning. Thus, we take a different approach from previous works and aim to construct a performance estimator $\widehat{\mathcal{R}}\left(\hat{\tau}\right)$ satisfying the following condition:

$$\mathcal{R}_{true}\left(\hat{\tau}\right) \leq \mathcal{R}_{true}\left(\hat{\tau}'\right) \Rightarrow \widehat{\mathcal{R}}\left(\hat{\tau}\right) \leq \widehat{\mathcal{R}}\left(\hat{\tau}'\right), \; \forall \hat{\tau}, \hat{\tau}' \in \mathcal{M}. \tag{2}$$

where $\mathcal{M} = \{\hat{\tau}_1, ..., \hat{\tau}_{|\mathcal{M}|}\}$ is a set of candidate CATE predictors.

An estimator satisfying Eq. (2) gives an accurate ranking of candidate predictors by the ground-truth metric, and we can identify the best model among $\mathcal{M}$ using such an estimator. Our goal is to construct a sophisticated method to obtain a performance estimator that can help achieve the condition described in Eq. (2) to enable accurate model selection of CATE predictors.

4. Method

To achieve our goal, we consider the following feasible estimator for the ground-truth performance metric:

$$\widehat{\mathcal{R}}\left(\hat{\tau}\right) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \hat{\tau}\left(X_i\right)\right)^2 \tag{3}$$

where $\tilde{\tau}\left(\cdot\right)$ is the *plug-in tau* and is calculated by using validation set. We consider the estimator in the form of Eq. (3), because it can be applied to estimating the performance of a method directly predicting CATE such as R-learner (Nie & Wager, 2017) and doubly robust learner (Foster & Syrgkanis, 2019).

Under our formulation, we aim to answer the following research question: ***What is the best plug-in tau to rank the performance of given candidate CATE predictors from an observational validation dataset?***

To address this question, in Section 4.1, we theoretically analyze the performance estimator in the form of Eq. (3) and identify the conditions that an ideal *plug-in tau* should satisfy. Then, in Section 4.2, we propose a method to obtain a *plug-in tau* that results in an accurate ranking of the true performance of candidate CATE predictors.

4.1. What is the good *plug-in tau*?

First, the following proposition states that a *plug-in tau* that is unbiased for the true CATE provides a desirable property of the resulting performance estimator.

***Proposition* 1.** *Suppose that a given plug-in tau is an unbiased estimator for the true CATE (i.e.,* $\mathbb{E}\left[\tilde{\tau}\left(X, T, Y\right) \mid X\right] = \tau(X)$*), then, the expectation of the performance estimator* $\widehat{\mathcal{R}}$ *is decomposed into the true performance metric and the MSE of the given plug-in tau:*

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left(\hat{\tau}\right)\right] = \mathcal{R}_{true}\left(\hat{\tau}\right) + \underbrace{\mathbb{E}\left[\left(\tau(X) - \tilde{\tau}(X, T, Y)\right)^2\right]}_{\text{independent of } \hat{\tau}} \tag{4}$$

See Appendix A.1 for the proof.

The first term of RHS of Eq. (4) is the true performance metric, and the second term is independent of the given predictor. Therefore, the expectations of the performance estimators preserve the difference between the true metric values as follows:

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left(\hat{\tau}_1\right)\right] - \mathbb{E}\left[\widehat{\mathcal{R}}\left(\hat{\tau}_2\right)\right] = \mathcal{R}_{true}\left(\hat{\tau}_1\right) - \mathcal{R}_{true}\left(\hat{\tau}_2\right)$$

where $\hat{\tau}_1$, $\hat{\tau}_2 \in \mathcal{M}$ are arbitrary candidate predictors. This property is desirable, because the predictor that has the smallest expected value of $\widehat{\mathcal{R}}$ among candidate predictors also has the smallest value of $\mathcal{R}_{true}$ among them; one can expect to select the best predictor among a set of candidates.

However, the expectation of the performance estimator is incalculable, because we can use only a finite sample validation set. This motivates us to consider the finite sample uncertainty of the performance estimator. The empirical version of the performance estimator can be decomposed as

$$\begin{aligned}
&\widehat{\mathcal{R}}\left(\hat{\tau}\right) \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \hat{\tau}(X_i))^2}_{\text{converges to } \mathcal{R}_{true}(\hat{\tau})} \\
&\quad \underbrace{- \frac{2}{n}\sum_{i=1}^{n}\left(\hat{\tau}\left(X_i\right) - \tau\left(X_i\right)\right)\left(\tilde{\tau}\left(X_i, T_i, Y_i\right) - \tau\left(X_i\right)\right)}_{\mathcal{W}\text{:source of uncertainty}} \\
&\quad \underbrace{+ \frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \tilde{\tau}(X_i, T_i, Y_i))^2}_{\text{independent of } \hat{\tau}}.
\end{aligned} \tag{5}$$

In the RHS of Eq. (5), $\mathcal{W}$ is critical to the uncertainty and is controllable by $\tilde{\tau}$. Thus, we try to minimize the variance of $\mathcal{W}$ with the aim of minimizing the uncertainty in model selection. The following theorem upper bounds the variance of $\mathcal{W}$.

***Theorem* 2.** *Suppose that the plug-in tau is unbiased for the CATE and the output of the plug-in tau for an instance is independent of that of other instances. Then, we have the upper bound of the variance of* $\mathcal{W}$ *as follows:*

$$\mathbb{V}\left(\mathcal{W}\right) \leq 4C_{\max}n^{-1}\,\mathbb{E}_X\left[\mathbb{V}\left(\tilde{\tau}(X, T, Y) \mid X\right)\right], \tag{6}$$

where $C_{\max} = \max_{i\in[n]}(\tau(x_i) - \hat{\tau}(x_i))^2$. *See Appendix A.5 for the proof.*

In Eq. (6), the expected conditional variance of $\tilde{\tau}$ is controllable by the construction of the *plug-in tau*. Thus, a *plug-in tau* satisfying the following condition is desirable to construct a stable performance estimator:

$$\min_{\tilde{\tau}\in\Theta} \mathbb{E}_X \left[\mathbb{V}\left(\tilde{\tau}(X,T,Y) \mid X\right)\right], \quad \text{s.t. } \mathbb{E}[\tilde{\tau}(X,T,Y) \mid X] = \tau(X). \tag{7}$$

where Θ is a pre-defined class of *plag-in tau*.

A performance estimator using a *plug-in tau* that achieves Eq. (7) is expected to preserve the difference of the true performance metric and to minimize the upper bound of the finite sample uncertainty term $\mathcal{W}$ in Eq. (5).

4.2. Obtaining *plug-in tau*

Next, we present a method to obtain a desirable *plug-in tau* inspired by the *doubly robust* (DR) estimator in causal inference and *counterfactual regression* (CFR) in CATE prediction (Bang & Robins, 2005; Dudík et al., 2011; Shalit et al., 2017). The proposed procedure is designed to preserve unbiasedness of *plug-in tau* using the DR estimator and to minimize its expected conditional variance with the power of CFR. Thus, the idea of combining the DR estimator and CFR is a key to better satisfy Eq. (7). Subsequently, we formally describe the resulting model selection procedure, *counterfactual cross-validation* (CF-CV).

First, we define a class of *plug-in tau* building on the DR estimator.

***Definiton* 4.** *The doubly robust plug-in tau for a given data* (X, T, Y) *is defined as follows:*

$$\tilde{\tau}_{DR}(X,T,Y;f_t) := \frac{T-e(X)}{e(X)(1-e(X))}(Y - f_T(X)) + f_1(X) - f_0(X), \tag{8}$$

where $f_t : \mathcal{X} \to \mathcal{Y}$ *is an arbitrary regression function.*

We rely on the class of the DR estimator for constructing the *plug-in tau*, because we can design the regression function for a variety of purposes. For example, the *more robust doubly robust* estimator utilizes a weighted squared loss to derive the regression function to minimize the variance of the resulting policy value estimator (Farajtabar et al., 2018). In contrast, we can utilize the regression function to minimize the upper bound of the finite sample uncertainty in model selection. These objectives cannot be achieved with model-free estimators such as the IPW estimator.

Note that our proposed *plug-in tau* cannot be used for the CATE prediction task because only the feature vectors are available while making predictions. In contrast, the treatment assignment and the observed outcome are unavailable. Thus, the *plug-in tau* in the form of Eq. (8) is specialized for the evaluation of CATE predictors.

First, the *plug-in tau* in the form of Eq. (8) is unbiased against the true CATE as follows.

***Proposition* 3.** *Given true propensity scores and a regression function, the proposed plug-in tau is unbiased against the true CATE, i.e.,*

$$\mathbb{E}\left[\tilde{\tau}_{DR}(X,T,Y;f_t) \mid X\right] = \tau(X).$$

See Appendix A.3 for the proof.

Next, to consider the condition in Eq. (7), we state the expected conditional variance of the *plug-in tau*.

***Proposition* 4.** *Given true propensity scores and a regression function, the expected conditional variance of the proposed plug-in tau can be represented as:*

$$\mathbb{E}_X\left[\mathbb{V}\left(\tilde{\tau}_{DR}(X,T,Y;f_t) \mid X\right)\right] = \zeta + \mathbb{E}_X\left[\{\sum_{t\in\mathcal{T}} \sqrt{w_t(X)}(f_t(X) - m_t(X))\}^2\right], \tag{9}$$

where

$$w_t(X) := \frac{t(1-2e(X)) + e(X)^2}{e(X)(1-e(X))},$$

$$\zeta := \mathbb{E}_X\left[\sum_{t\in\mathcal{T}} \frac{e(X) + t(1-2e(X))}{e(X)(1-e(X))}(Y(t) - m_t(X))^2\right].$$

See Appendix A.4 for the proof.

In Eq. (9), ζ is independent of f. Thus, we can pursue the minimization of the expected conditional variance of $\tilde{\tau}_{DR}$ by training f with the following procedure:

$$\min_{f\in\mathcal{F}} \mathbb{E}_X\left[\{\sum_{t\in\mathcal{T}} \sqrt{w_t(X)}(f_t(X) - m_t(X))\}^2\right], \tag{10}$$

where $\mathcal{F}$ is a class of regression functions. A problem is that the direct minimization of Eq. (10) is infeasible, because $m_0(x)$ or $m_1(x)$ is always counterfactual. Therefore, we derive the upper bound of the second term of Eq. (9) using only observable variables.

***Theorem* 5.** *Let* G *be a family of functions* $g : \mathcal{R} \to \mathcal{Y}$ *and suppose that, for any given* $t \in \mathcal{T}$ *and* $w : \mathcal{X} \times \mathcal{T} \to \mathcal{R}_{\geq 0}$, *there exists a positive constant* B_Φ *such that the per-unit expected loss functions obey* $\frac{1}{B_\Phi} \cdot \ell^{w}_{h,\Phi}(\Psi(r), t) \in G$ *where* Ψ *is the inverse image of* Φ. *Then, the following inequality holds:*

$$\begin{aligned}\mathbb{E}_X[\{\sum_{t\in\mathcal{T}} \sqrt{w_t(X)}(f_t(X) - m_t(X))\}^2] &\leq 2\big(\epsilon^{w_1}_{F_1}(h,\Phi) + \epsilon^{w_0}_{F_0}(h,\Phi) \\ &\quad + B_\Phi \mathrm{IPM}_G\left(p^\Phi_t, p^\Phi_{1-t}\right) - 2\sigma^2\big),\end{aligned} \tag{11}$$

```
Algorithm 1 Counterfactual Cross-Validation (CF-CV)
Require: A set of candidate CATE predictors $\mathcal{M} = \{\hat{\tau}_1, ..., \hat{\tau}_{|\mathcal{M}|}\}$; an observational validation dataset $\mathcal{V} = \{X_i, T_i, Y\}_{i=1}^n$; and a trade-off hyperparameter $\alpha$.
    Train $f(X, T)$ by minimizing Eq. (12) using $\mathcal{V}$.
    Estimate the propensity score (if needed).
    Calculate the plug-in tau $\tilde{\tau}_{DR}$ of samples in $\mathcal{V}$.
    Estimate performance of candidate predictors in $\mathcal{M}$ based on the performance estimator $\widehat{\mathcal{R}}$ and $\tilde{\tau}_{DR}$.
Ensure: A selected predictor: $\hat{\tau}^* = \arg\min_{\hat{\tau}\in\mathcal{M}} \widehat{\mathcal{R}}(\hat{\tau})$.
```

where $\sigma^2 := \min_{(t,t')\in\mathcal{T}^2}\{\sigma^2_{t,w_t}(p_{t'})\}$, *and*

$$\sigma^2_{t,w}(p_{t'}) := \int_{\mathcal{X}\times\mathcal{Y}} w(x)(Y(t) - m_t(x))^2 p(Y(t)|x)p_{t'}(x)dY(t)dx.$$

See Appendix A.7 for the proof.

Eq. (11) consists of factual losses and an IPM on the representation space, and thus can be estimated from observed samples. The intuition is that the counterfactual losses can be upper bounded by the sum of weighted factual losses and IPM between distributions of the treated and the controlled. Therefore, we can optimize the upper bound of the expected conditional variance of $\tilde{\tau}_{DR}$ using only factual samples in a manner similar to CFR (Shalit et al., 2017). Thus, we build on the CFR's structure and define our regression function as $f_t(x) = h(\Phi(x), t)$. We then consider the following empirical approximation of Eq. (11) as a loss to derive a hypothesis h and representation function Φ:

$$h, \Phi = \min_{h,\Phi} \underbrace{\sum_{i=1}^{n} \frac{w_t(x_i)}{n} \cdot L\left(h\left(\Phi\left(x_i\right), t_i\right), y_i\right)}_{\textit{empirical weighted risk}} + \underbrace{\alpha \mathrm{IPM}_G\left(\left\{\Phi\left(x_i\right)\right\}_{i:t_i=0}, \left\{\Phi\left(x_i\right)\right\}_{i:t_i=1}\right)}_{\textit{distributional distance}}. \tag{12}$$

We use parameterized deep neural networks for $\Phi(x)$ and $h(\Phi(x), t)$ and train them in an end-to-end manner using the Adam optimizer (Kingma & Ba, 2014). α is a trade-off hyperparameter which is a replacement of the incomputable factor B_Φ. We use the *Wasserstein distance* (Shalit et al., 2017; Cuturi & Doucet, 2014) as IPM_G in the experiments.

The derived *plug-in tau* is unbiased for the true metric and minimizes the upper bound of the controllable term in its expected conditional variance in Eq. (11), enabling the accurate and stable model selection of CATE predictors. Algorithm 1 summarizes the resulting model selection procedure.

5. Experiments

We compare our proposed evaluation procedure and other existing heuristics using a standard semi-synthetic dataset.[2]

5.1. Experimental Setup

We used the Infant Health Development Program (IHDP) dataset provided by (Hill, 2011). The original data is obtained from a randomized study of the impact on educational and follow-up interventions on cognitive development of children (Hill, 2011; Shalit et al., 2017; Alaa & Schaar, 2018; Yao et al., 2018). This is a standard semi-synthetic dataset of 747 children with 25 features and has been widely used to evaluate CATE prediction models (Shalit et al., 2017; Yoon et al., 2018; Alaa & Schaar, 2018; Yao et al., 2018; Johansson et al., 2020). To enable evaluation with the ground-truth CATE, the outcome of this dataset was synthesized by applying several different stochastic models on the observed features. Moreover, to introduce confounding, a biased subset of the treatment group was removed. Note that we did not use real-world causal inference datasets such as *jobs* and *twins* (Yoon et al., 2018; Shalit et al., 2017), because they do not contain the ground-truths for the true CATE and consequently are unable to perform the *evaluation of evaluation metrics*.

We compared the following evaluation metrics in model selection and hyperparameter tuning tasks:

(i) IPW validation (Gutierrez & Gérardy, 2017; Schuler et al., 2018): This metric utilizes the following performance estimator:

$$\hat{\mathcal{R}}_{IPW}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n} \left(\tilde{\tau}_{IPW}(X_i, T_i, Y_i) - \hat{\tau}(X_i)\right)^2$$

where

$$\tilde{\tau}_{IPW}(X_i, T_i, Y_i) = \frac{T_i}{e(X_i)}Y_i - \frac{1-T_i}{1-e(X_i)}Y_i$$

is used as a *plug-in-tau* that satisfies the unbiasedness for the CATE.

(ii) Plug-in validation: This uses predicted values of potential outcomes by an arbitrary machine learning algorithm as the *plug-in tau* of the performance estimator in Eq. (3).

$$\hat{\mathcal{R}}_{\textit{plug-in}}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n} \left(\left(\tilde{\tau}_i^{(1)} - \tilde{\tau}_i^{(0)}\right) - \hat{\tau}(X_i)\right)^2$$

where $\tilde{\tau}_i^{(1)}$ and $\tilde{\tau}_i^{(0)}$ are predictions for the potential outcomes. We used CFR (Shalit et al., 2017) to construct $\tilde{\tau}^{(1)}(\cdot)$ and $\tilde{\tau}^{(0)}(\cdot)$ to ensure a fair comparison.

[2] The code for reproducing the experimetns is availabel at https://github.com/usaito/counterfactual-cv

Table 1. Comparison of Model Selection and Hyperparameter Tuning Performance of Alternative Evaluation Metrics.

	Rank Correlation		Regret		NRMSE	
Methods	**Mean** ±**StdErr**	**Worst-Case**	**Mean** ±**StdErr**	**Worst-Case**	**Mean** ±**StdErr**	**Worst-Case**
IPW	0.195 ±0.039	-0.749	1.032 ±0.100	6.779	0.336 ±0.013	0.737
τ-risk	0.312 ±0.030	-0.553	1.392 ±0.130	7.884	0.324 ±0.013	0.700
Plug-in	0.914 ±0.006	0.591	0.073 ±0.012	0.780	0.257 ±0.010	0.490
CF-CV (ours)	**0.921** ±**0.005**	**0.666**	**0.066** ±**0.012**	**0.562**	**0.256** ±**0.009**	**0.483**

Notes: Mean with standard errors (StdErr), and worst-case performance of the compared evaluation metrics over 100 realizations are reported. The **red fonts** represent the best performance in each performance measures.

(iii) τ-risk (Schuler et al., 2018): This metric is derived from the loss function of R-learner in (Nie & Wager, 2017) and is defined as follows:

$$\hat{\mathcal{R}}_{\tau}(\hat{\tau}) = \frac{1}{n}\sum_{i=1}^{n}\left((Y_i - m(X_i)) - (T_i - e(X_i))\hat{\tau}(X_i)\right)^2$$

where $m(\cdot)$ is the expectation of observed outcome $\mathbb{E}[Y|X]$. We used gradient boosting regressor (GBR) implemented in *scikit-learn* to estimate this parameter.

(iv) Counterfactual Cross-Validation: This is our proposed metric, which relies on the *plug-in tau* in Eq. (8). The hyperparameter tuning procedure to derive the regression function f can be found in Appendix B.1.

We used logistic regression to estimate the propensity score for CF-CV and IPW validation, because the true propensity score is generally unknown in real-world situations. For plug-in validation and CF-CV, we used the μ-risk (Schuler et al., 2018) as a data-driven heuristic to tune hyperparameters of machine learning models to obtain predictions of the potential outcomes or the regression function.

5.2. Model Selection Performance

We first tested the model selection performance.

Experimental Procedure. We followed the experimental procedure in (Schuler et al., 2018); We trained candidate predictors on the training set and made predictions on both validation and test sets. Then, we ranked those predictors based on each evaluation metric on the validation set. Finally, we compare these estimated performances on the validation set and the true performance on the testing set. We conducted the experimental procedure over 100 different realizations with 35/35/30 train/validation/test splits.

Candidate Models. We constructed a set of candidate predictors $\mathcal{M}$ by combining five machine learning algorithms (decision tree, random forest, gradient boosting tree, ridge regressor, and support vector regressor) implemented in *scikit-learn* and five meta-learners (S-learner, X-learner, T-learner, domain adaptation learner, and doubly robust learner) implemented in *EconML*[3]. Thus, we had a set of 25 CATE predictors to select among (i.e., $|\mathcal{M}| = 25$).

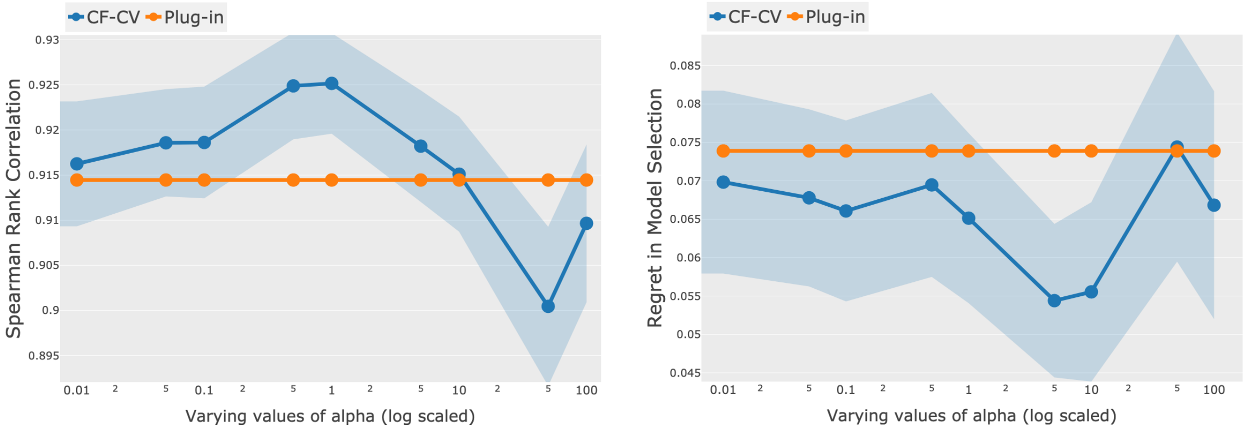
Results. Table 1 reports the mean and worst-case performances over 100 realizations. We evaluated the worst-case model selection performance, because we never know the ground-truth performance of any predictor in the real-world, and stable model selection performance is essential. *Rank correlation* is the Spearman rank correlation between the rankings by the true performance and the estimated metric values. *Regret in model selection* is the difference between the true performance of the selected model and that of the best possible candidate in $\mathcal{M}$, which is defined as:

$$Regret = \frac{\mathcal{R}_{true}\left(\hat{\tau}_{selected}\right) - \mathcal{R}_{true}\left(\hat{\tau}_{best}\right)}{\mathcal{R}_{true}\left(\hat{\tau}_{best}\right)}$$

where $\hat{\tau}_{selected} = \arg\min_{\hat{\tau}\in\mathcal{M}} \widehat{\mathcal{R}}(\hat{\tau})$ is the model selected by $\widehat{\mathcal{R}}$ and $\hat{\tau}_{best} = \arg\min_{\hat{\tau}\in\mathcal{M}} \mathcal{R}_{true}\left(\hat{\tau}\right)$ is the best model in $\mathcal{M}$.

Table 1 shows the effective model selection performance of the proposed CF-CV. In particular, it significantly outperformed the others in terms of the worst-case performance. This result empirically suggests that the proposed metric can stably select a well-performing CATE predictor among potential candidates and is an appropriate choice for real-world situations. The stability of CF-CV could be a result of its variance upper bound minimization property. The improvement of the worst-case performance is essential in many causal inference problems such as personalized medicine, which has a great impact on human lives. Our procedure thus helps avoid deploying poor-performing CATE predictors and enable the safe uses of causal inference in practice. We also evaluated the sensitivity of the proposed metric to changes in the trade-off hyperparameter α. Figure 1 shows the performances of CF-CV with variation of α com-

[3] https://econml.azurewebsites.net/



(a) *Rank correlation* of CF-CV with different values of α

(b) *Regret* of CF-CV with different values of α

Figure 1. Comparing CF-CV with varying α and the plug-in validation. CF-CV (the blue lines) outperforms the plug-in validation (the orange lines) in most cases and demonstrates its robustness to the choice of α.

pared to the performance of the plug-in validation. For the *rank correlation*, CF-CV generally outperformed the plug-in metric with small values of α, although it was slightly outperformed by the plug-in with a larger α. Additionally, CF-CV consistently outperformed the plug-in validation with all values of α in *regret*. These results suggest that the proposed metric is robust to the choice of α.

5.3. Hyperparameter Tuning Performance

Next, we compared the hyperparameter tuning performance.

Tuned Model. We tuned the hyperparameters of the combination of GBR and domain adaptation learner (DAL) implemented in *scikit-learn* and *EconML*, respectively. DAL consists of three base learners including **treated_model**, **controls_model**, and **overall_model**. Thus, we aimed to find the best three sets of hyperparameters of GBR to optimize the resulting CATE prediction model.

Experimental Procedure. We used *Optuna* (Akiba et al., 2019) to tune the CATE predictor and set each metric as its objective function. For each metric, we sought 100 points in the hyperparameter search space[4]. The hyperparameter tuning performance of each metric was evaluated by the true performance of the tuned model on the testing set. We repeated the experimental procedure with 100 different realizations and train/validation/test splits.

Results. Table 1 provides the results of the hyperparameter tuning experiment. We report the mean and worst-case *normalized root-mean-squared-error* (NRMSE) of CATE predictors tuned by each metric defined below[5].

$$\mathit{NRMSE} = \sqrt{\frac{n^{-1}\sum_{i=1}^{n}(\tau(X_i) - \hat{\tau}(X_i))^2}{\hat{\mathbb{V}}(\tau(X))}}$$

where $\{\hat{\tau}(X_i)\}_{i=1}^{n}$ is a set of CATE predictions by $\hat{\tau}(\cdot)$ and $\hat{\mathbb{V}}(\tau)$ is an empirical variance of the ground-truth CATE.

Table 1 shows that our metric improved the worst-case performance by 1.4 % compared to the best baselines. Although the mean NRMSE is almost the same as that with the plug-in validation, the results demonstrate that the proposed metric allows stable hyperparameter tuning of the CATE prediction models.

6. Conclusion

In this work, we studied the model selection problem in CATE prediction. In contrast to previous studies, we aimed to identify the rank order of the true prediction performances of the candidate prediction models. We achieved this by using a modified version of the CFR as a regression function of the DR estimator to minimize the finite sample uncertainty. Empirical evaluations demonstrated the effectiveness and stability of the proposed metric for model selection and hyperparameter tuning of the CATE predictors.

Important future research directions include consideration of situations with hidden confounders, and a possible extension to the *off-policy evaluation* of bandit policies.

[4] The hyperparameter search space is described in Appendix B.2

[5] We used NRMSE, as the potential outcomes of the IHDP dataset have different scales among realizations.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330701.

Alaa, A. and Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138, 2018.

Alaa, A. and Van Der Schaar, M. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pp. 191–201, 2019.

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

Bang, H. and Robins, J. M. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61 (4):962–973, 2005.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Bica, I., Jordon, J., and van der Schaar, M. Estimating the effects of continuous-valued interventions using generative adversarial networks. *arXiv preprint arXiv:2002.12326*, 2020.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. 2014.

Diemert, E., Betlei, A., Renaudin, C., and Amini, M.-R. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London, United Kingdom*, 2018.

Du, X., Sun, L., Duivesteijn, W., Nikolaev, A., and Pechenizkiy, M. Adversarial balancing-based representation learning for causal effect inference with observational data. *arXiv preprint arXiv:1904.13335*, 2019.

Dudík, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 1097–1104. Omnipress, 2011.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1446–1455, 2018.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

Gutierrez, P. and Gérardy, J.-Y. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*, pp. 1–13, 2017.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5880–5887, 2019.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

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

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Imbens, G. W. and Rubin, D. B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029, 2016.

Johansson, F. D., Shalit, U., Kallus, N., and Sontag, D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Mansour, Y., Mohri, M., and Rostamizadeh, A. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*, 2009.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Rolling, C. A. and Yang, Y. Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):749–769, 2014.

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, 1983.

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

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., and Cook, E. F. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6): 546–555, 2008.

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-Volume 70*, pp. 3076–3085. JMLR. org, 2017.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pp. 2503–2513, 2019.

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

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. 2633–2643, 2018.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Ace: Adaptively similarity-preserved representation learning for individual treatment effect estimation. In *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 1432–1437. IEEE, 2019.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

Yoon, J., Jordon, J., and van der Schaar, M. GAN-ITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.