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

Yuta Saito
Tokyo Institute of Technology
saito.y.bj@m.titech.ac.jp

Shota Yasui
CyberAgent, Inc.
yasui_shota@cyberagent.co.jp

Abstract

What is the most effective way to select the best causal model among potential candidates? In this paper, we propose a method to effectively select the best conditional average treatment effect (CATE) predictors from a set of candidates using only an observational validation set. When conducting a model selection or tuning hyperparameters, we are interested in choosing the best model or hyperparameter value. Thus, we focus on accurately preserving the rank order of the CATE prediction performance of causal model candidates. For this purpose, we propose a new model selection procedure that preserves the true ranking of the model performance and minimizes the upper bound of the finite sample uncertainty in model selection. Consistent with the theoretical properties, empirical evaluations demonstrate that our proposed method is more likely to select the best model and set of hyperparameters for both model selection and hyperparameter tuning tasks.

1 Introduction

Predicting individual-level treatment effects (ITEs) for certain actions is essential for optimizing metrics of interest in various domains. In digital marketing, for instance, incrementality is becoming increasingly important for performance metrics [1]. For this purpose, users that will be shown ads for a given product should be chosen based on the ITE to avoid showing ads to a user who will already buy that product. Healthcare is also an important application of ITE prediction [2]. For precision medicine, we need to know which treatments will be more beneficial or harmful for a given patient. The fundamental problem of causal inference is that we never observe both treated and untreated outcomes from the same unit simultaneously. This is widely known as the central problem of causal inference [3]. Because of this situation, we are unable to observe a causal effect itself and to use causal effects as labels in the prediction model. Most previous studies related to the topic of ITE focus on prediction methods using observational data [4, 5, 6, 7, 8, 9]. 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 applicable in model selection and hyperparameter tuning. In other words, we are unsure which model and which hyperparameter values should be used when applying causal effect predictors to real-world problems.

There are only a few studies that tackle this problem. [10] proposed using the inverse probability weighting (IPW) outcome as the pseudo-label for the true ITE in the calculation of an evaluation metric. [11] used the loss function of R-learner [12] as the evaluation metric. [13] used influence functions to obtain a more efficient estimator for the loss. These works are mainly interested in more accurately and efficiently estimating the evaluation metric of interest.

In this work, we focus on the problem of choosing the best conditional average treatment effect (CATE) predictors among a set of candidates using only an observational validation set. Typically, we are interested in choosing the best model or hyperparameters from potential candidates. In such a situation, **we only need to know the rank order of the value of the loss for those candidates**. To achieve this objective, in this work, we propose a new model selection procedure for causal inference models. The proposed metric is theoretically proven to preserve the ranking of candidate predictors and minimize the upper bound of the finite sample uncertainty in model selection. To demonstrate the practical significance of our evaluation metric, we conducted two experiments. In those experiments, our proposed method accurately and stably finds the best performance model among candidates and the best sets of hyperparameters for a given causal inference model compared with other existing heuristic metrics.

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, and latent variable models [2, 4, 6, 9]. Among the diverse methods that predict the CATE from observational data, the approach that is most related to this work is the method based on representation learning [14]. All the methods based on representation learning attempt to map the original feature vectors into the desirable latent representation space such that it eliminates selection biases. Balancing Neural Network [15] is the most basic method and uses discrepancy distance [16], a domain discrepancy measure in unsupervised domain adaptation, for the regularization term. CounterFactual Regression [7] minimizes the upper bound of the ground-truth loss for the CATE by utilizing an integral probability metric (IPM) [17]. In addition to these, methods that obtain a latent representation by preserving a local similarity [5] or by applying adversarial learning [8] have been proposed.

The prediction methods stated above have provided promising results on standard benchmark datasets; however, the evaluation of such CATE predictors has been conducted by using synthetic datasets or simple heuristic metrics such as policy risk, in previous studies [4, 5, 7]. These evaluations do not guarantee which models would actually be best on a given real-world dataset [13, 18]. Therefore, to bridge the gap between causal inference and applications, developing a reliable evaluation metric is critical.

There are only a few studies directly tackling the evaluation problem of CATE prediction models. [11] 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 [12]. In addition, they showed that these metrics empirically outperformed another naive metric, μ-risk, where predictive risk is estimated separately for treated and control outcomes using factual samples only. In contrast, [19] 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, and the TECV metric was empirically outperformed by IPW validation [11]. Nontheless, [13] 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 so far aim to estimate the true metric of interest (e.g., MSE for the true CATE) accurately, or they do not consider the uncertainty in model selection. However, to conduct accurate model selection and hyperparameter tuning, it is critical to rank model performance accurately, although the metrics above do not always guarantee the preservation of such rankings. Moreover, analysis of the uncertainty of the model evaluation is essential, especially in domains where the size of validation datasets might be small (e.g., education or public health). Therefore, in contrast to previous works, we investigate a way to construct a metric that accurately preserves the ranking of candidate predictors while also analyzing the uncertainty with respect to the ranking performance.

3 Problem Setting and Notation

In this section, we introduce notation and formulate the evaluation of CATE prediction models.

3.1 Notation

We denote $X \in \mathcal{X} \subset \mathbb{R}^d$ as the 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, $T_i = 0$. Here, we follow the potential outcome framework [20, 21, 22] and assume that there exist two potential outcomes denoted as $(Y^{(0)}, Y^{(1)}) \in \mathcal{Y} \times \mathcal{Y}$ 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 CATE for an individual with a feature vector $x \in \mathcal{X}$ as

$$\tau(x) = \mathbb{E}[Y^{(1)} - Y^{(0)} \mid X = x], \tag{1}$$

which is the conditional expectation of the ITE $Y^{(1)} - Y^{(0)}$. In addition, we use some notation to represent parameters of p. First, the conditional expectations of potential outcomes are:

$$m^{(k)}(x) = \mathbb{E}[Y^{(k)} \mid X = x], \ \forall k \in \{0, 1\}. \tag{2}$$

Next, we define the conditional probability of treatment assignment as:

$$e(x) = \mathbb{P}\left(T = 1 \mid X = x\right) \tag{3}$$

This parameter is called the propensity score in causal inference and is widely used to estimate treatment effects from observational data [20, 22, 23].

Throughout this 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 feature vector X, i.e.,*

$$(Y^{(0)}, Y^{(1)}) \perp T \mid X \tag{4}$$

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

$$\mathbb{P}\left(e(X) \in (0, 1)\right) = 1 \tag{5}$$

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

$$Y^{obs} = TY^{(1)} + (1 - T)Y^{(0)} \tag{6}$$

Under these assumptions, the CATE is identifiable from observational data (i.e., $\tau(X) = \mathbb{E}[Y^{\text{obs}} \mid X, T = 1] - \mathbb{E}[Y^{\text{obs}} \mid X, T = 0]$) [7].

Furthermore, we define some critical notation following [7].

***Definiton* 1.** *(Representation Function) $\Phi : \mathcal{X} \to \mathcal{R}$ 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, $p_\Phi(r|t = 1)$ and $p_\Phi(r|t = 0)$ are feature distributions for the treated and controlled outcomes induced over the representation space.*

***Definiton* 2.** *(Factual and Counterfactual Loss Functions) Let $h : \mathcal{R} \times \mathcal{T} \to \mathcal{Y}$ be a hypothesis, $w : \mathcal{X} \to \mathbb{R}_{\geq 0}$ be a weighting function, and $L : \mathcal{Y} \times \mathcal{Y} \to \mathcal{R}$ 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:*

$$\ell_{h,\Phi,w}(x,t) = \int_{\mathcal{Y}} w(x) L(Y^{(t)}, h(\Phi(x), t)) p(Y^{(t)}|x) dY^{(t)}$$

Then, the expected factual and counterfactual losses of a combination of a hypothesis h and a representation function Φ are defined as:

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

Further, the expected factual and counterfactual losses on the treated ($t = 1$) and on the controlled ($t = 0$) are represented as:

$$\epsilon_F^{t=1}(h, \Phi, w) = \int_{\mathcal{X}} \ell_{h,\Phi,w}(x, t = 1) p(x|t = 1) dx$$
$$\epsilon_F^{t=0}(h, \Phi, w) = \int_{\mathcal{X}} \ell_{h,\Phi,w}(x, t = 0) p(x|t = 0) dx$$
$$\epsilon_{CF}^{t=1}(h, \Phi, w) = \int_{\mathcal{X}} \ell_{h,\Phi,w}(x, t = 1) p(x|t = 0) dx$$
$$\epsilon_{CF}^{t=0}(h, \Phi, w) = \int_{\mathcal{X}} \ell_{h,\Phi,w}(x, t = 0) p(x|t = 1) dx$$

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

$$\epsilon_F(h, \Phi, w) = u \cdot \epsilon_F^{t=1}(h, \Phi, w) + (1-u) \cdot \epsilon_F^{t=0}(h, \Phi, w)$$
$$\epsilon_{CF}(h, \Phi, w) = (1-u) \cdot \epsilon_{CF}^{t=1}(h, \Phi, w) + u \cdot \epsilon_{CF}^{t=0}(h, \Phi, w)$$

where $u = P(t = 1)$.

We also define a class of metrics between probability distributions [17].

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

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

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.

***Definiton* 4.** *(Weighted Variance of Potential Outcomes) The weighted expected variance of a potential outcome $Y^{(t)}$ with respect to a joint probability distribution $p(x,t)$ is defined as:*

$$\sigma^2_{t,w}(p(x\,|\,t)) = \int_{\mathcal{X}\times\mathcal{Y}} w(x)(m^{(t)}(x) - Y^{(t)})^2 p(Y^{(t)}, x|\,t)dY^{(t)}dx$$

3.2 Evaluating CATE prediction models

In previous studies [10, 11, 13], the evaluation of a CATE predictor $\widehat{\tau}(\cdot)$ has been formulated as accurately estimating the following metric from observational validation dataset $\mathcal{V} = \{X_i, T_i, Y_i^{obs}\}$ as:

$$\mathcal{R}_{\text{true}}\left(\widehat{\tau}\right) = \mathbb{E}_X\left[\left(\tau(X) - \widehat{\tau}(X)\right)^2\right] \tag{7}$$

Here, $\mathcal{R}_{\text{true}}$ is the true performance metric for a CATE predictor $\widehat{\tau}(\cdot)$[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 the true metric values is not always necessary to conduct valid model selection or hyperparameter tuning of causal models. It may be possible to construct a better 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 $\widehat{\mathcal{R}}\left(\widehat{\tau}\right)$ satisfying the following condition:

$$\mathcal{R}_{\text{true}}\left(\widehat{\tau}\right) \leq \mathcal{R}_{\text{true}}\left(\widehat{\tau}'\right) \Rightarrow \widehat{\mathcal{R}}\left(\widehat{\tau}\right) \leq \widehat{\mathcal{R}}\left(\widehat{\tau}'\right),\ \forall\,\widehat{\tau}, \widehat{\tau}' \in \mathcal{M}. \tag{8}$$

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

An estimator satisfying Eq. (8) gives accurate **ranking** of candidate predictors by the true metric values, and we can select the best model among $\mathcal{M}$ using the estimator. The main focus of this paper is to propose a theoretically sophisticated way to construct a performance estimator $\widehat{\mathcal{R}}$ that achieves the condition described in Eq. (8) as much as possible.

4 Method

To achieve our goal of interest, we consider the following feasible estimator of the performance metric:

$$\widehat{\mathcal{R}}\left(\widehat{\tau}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{\tau}\left(X_i, T_i, Y_i^{\text{obs}}\right) - \widehat{\tau}\left(X_i\right)\right)^2 \tag{9}$$

where $\tilde{\tau}\left(\cdot\right)$ is called **the plug-in tau** and is constructed from validation set $\mathcal{V}$. We consider the estimator for the true risk as represented in Eq. (9) because it can be applied to estimating the performance of a predictor, directly predicting CATE such as R-learner [12], Domain Adaptation Learner, or Doubly Robust Learner [24].

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

In the following subsections, we theoretically analyze the performance estimator as represented in the form of Eq. (9) and propose a plug-in tau that gives an accurate **ranking** of candidate CATE prediction models by the true performance metric.

[1]Several papers have called $\mathcal{R}_{\text{true}}$ the expected Precision in Estimation of Heterogeneous Effect (PEHE).

4.1 Theoretical Analysis of the Performance Estimator

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

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

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left(\widehat{\tau}\right)\right] = \mathcal{R}_{true}\left(\widehat{\tau}\right) + \underbrace{\mathbb{E}\left[\left(\tau(X) - \tilde{\tau}(X, T, Y^{obs})\right)^2\right]}_{\text{independent of } \widehat{\tau}} \tag{10}$$

The first term of RHS of Eq. (10) 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:

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

where $\widehat{\tau}_1$, $\widehat{\tau}_2 \in \mathcal{M}$ 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 $\mathcal{R}_{\text{true}}$ 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 dataset. This motivates us to consider the finite sample uncertainty of the performance estimator. Here, the empirical version of the performance estimator can be decomposed as

$$\begin{aligned}\widehat{\mathcal{R}}\left(\widehat{\tau}\right) = & \frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \widehat{\tau}(X_i))^2 \\ & \underbrace{- \frac{2}{n}\sum_{i=1}^{n}\left(\widehat{\tau}\left(X_i\right) - \tau\left(X_i\right)\right)\left(\tilde{\tau}\left(X_i, T_i, Y_i^{\text{obs}}\right) - \tau\left(X_i\right)\right)}_{\mathcal{W}\text{:source of uncertainty}} \\ & + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) - \tilde{\tau}(X_i, T_i, Y_i^{\text{obs}}))^2}_{\text{independent of } \widehat{\tau}}\end{aligned} \tag{11}$$

In Eq. (11), the second term of RHS ($\mathcal{W}$) is critical to the uncertainty and is controllable by the plug-in tau. Thus, we consider the plug-in tau minimizing the variance of $\mathcal{W}$ to construct the performance estimator. The following theorem states the upper bound of the variance of $\mathcal{W}$.

***Theorem* 2.** *Assume that* $(\tau(X) - \widehat{\tau}(X))^2$ *is upper bounded by a positive constant* C *for a given predictor. Additionally, 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:*

$$\mathbb{V}\left(\mathcal{W}\right) \leq \frac{4C}{n}\left(\mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}(X, T, Y^{obs}) \mid X\right)\right] + \mathbb{V}\left(\tau(X)\right)\right) \tag{12}$$

In Eq. (12), the expectation of the conditional variance of the plug-in tau is the only controllable term 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:

$$\min \ \mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}(X,T,Y^{\text{obs}}) \mid X\right)\right] \tag{13}$$

$$\text{s.t. } \mathbb{E}[\tilde{\tau}(X)] = \tau(X). \tag{14}$$

A performance estimator using a plug-in tau that satisfies the conditions above 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. (11).

4.2 Proposed Plug-in Tau

In this subsection, we propose a new plug-in tau inspired by the doubly robust (DR) estimator in causal inference and counterfactual regression (CFR) in CATE prediction [7, 25, 26]. The proposed plug-in tau is designed to preserve unbiasedness using the DR estimator, and to minimize its own variance using the power of CFR. Thus, the idea of combining the DR estimator and CFR is critical to better satisfy the conditions in Eq. (13) and Eq. (14). Subsequently, we formally describe the resulting model selection procedure, called counterfactual cross-validation (CF-CV).

Here we define the proposed plug-in tau based on the DR estimator.

Definiton 5. *Let $f(\cdot,\cdot) : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ be a hypothesis predicting potential outcomes, called a regression function, where $f(x,t) = h(\Phi(x),t)$. Then, the plug-in tau for a given data (X,T,Y^{obs}) is defined as follows:*

$$\begin{aligned}&\tilde{\tau}_{DR}(X,T,Y^{obs}) \\ &= \frac{T}{e(X)}\left(Y^{obs} - f(X,1)\right) - \frac{1-T}{1-e(X)}\left(Y^{obs} - f(X,0)\right) + \left(f(X,1) - f(X,0)\right)\end{aligned} \tag{15}$$

We rely on the class of DR estimators for constructing the plug-in tau because we can design the regression function f in Eq. (15) for a variety of objectives. For example, a more robust doubly robust estimator [27] utilizes a weighted squared loss to derive the regression function to minimize the variance of the resulting policy value estimator. For the proposed plug-in tau, we utilize the regression function to minimize the upper bound of the finite sample uncertainty in model selection. Those 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 simply because, when making predictions, only the feature vectors are available; the treatment assignment and the observed outcome are unavailable. Thus, the plug-in tau in the form of Eq. (15) is specialized for the **evaluation** of CATE predictors.

First, the plug-in tau in the form of Eq. (15) is proved to be unbiased against the true CATE, thus satisfying Eq. (14).

Proposition 3. *Given true propensity scores, the proposed plug-in tau is unbiased against the true CATE:*

$$\mathbb{E}\left[\tilde{\tau}_{DR} | X\right] = \tau(X) \tag{16}$$

Algorithm 1 Counterfactual Cross-Validation (CF-CV)

Input: A set of candidate CATE predictors $\mathcal{M} = \{\widehat{\tau}_1, ..., \widehat{\tau}_{|\mathcal{M}|}\}$; an observational validation dataset $\mathcal{V} = \{X_i, T_i, Y^{\mathrm{obs}}\}_{i=1}^n$; and a trade-off hyperparameter α.

Output: A selected CATE predictor $\widehat{\tau}^* \in \mathcal{M}$.

1: Train a function $f(X, T)$ by minimizing Eq. (19) using validation set $\mathcal{V}$.
2: Estimate the propensity score (if needed).
3: Derive the plug-in tau values $\tilde{\tau}_{DR}$ for $\mathcal{V}$.
4: Estimate performance of candidate predictors in $\mathcal{M}$ based on the performance estimator $\widehat{\mathcal{R}}$ in Eq. (9).
5: Select a predictor minimizing the performance estimator among $\mathcal{M}$, i.e., $\widehat{\tau}^* = \arg\min_{\widehat{\tau}\in\mathcal{M}} \widehat{\mathcal{R}}(\widehat{\tau})$

Next, to consider the condition in Eq. (13), we state the expectation of the conditional variance of the DR plug-in tau.

***Proposition* 4.** *Given true propensity scores, the expectation of the conditional variance of the proposed plug-in tau can be represented as*

$$\mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}_{DR} \,|\, X\right)\right] = \mathbb{E}\left[\left(\xi^{(1)}\right)^2\right] + \mathbb{E}\left[\left(\xi^{(0)}\right)^2\right] + \mathcal{Z} \tag{17}$$

where $\xi^{(1)} = \frac{T}{e(x)}(Y^{(1)} - m^{(1)}(x))$, $\frac{1-T}{1-e(x)}(Y^{(0)} - m^{(0)}(x))$, $w^{(t)}(x) = \frac{t(1-2e(x))+e(x)^2}{e(x)(1-e(x))}$, *and*

$$\mathcal{Z} = \int_{\mathcal{X}} \left(\sqrt{w^{(1)}(x)}\left(f(x,1) - m^{(1)}(x)\right) + \sqrt{w^{(0)}(x)}\left(f(x,0) - m^{(0)}(x)\right)\right)^2 p(x)dx$$

To find a plug-in tau that satisfies the variance condition in Eq. (13), we aim to train a hypothesis f minimizing the variance derived in Eq. (17). In the variance, $\xi^{(1)}$ and $\xi^{(0)}$ are independent of f, and are thus uncontrollable. On the other hand, the third term of RHS of Eq. (17) is dependent on both $m^{(0)}(x)$ and $m^{(1)}(x)$. However, either $m^{(0)}(x)$ or $m^{(1)}(x)$ is always counterfactual, and thus, the direct minimization of $\mathcal{Z}$ is infeasible.

Therefore, to find the appropriate hypothesis f that minimizes $\mathcal{Z}$ from an observational validation set, we derive the upper bound of $\mathcal{Z}$ depending only on observable variables.

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

$$\begin{aligned}\mathcal{Z} \leq\ & 2\left(\epsilon_F^{t=1}\left(h, \Phi, w^{(1)}\right) + \epsilon_F^{t=0}\left(h, \Phi, w^{(0)}\right) - 2\sigma^2\right) \\ &+ 2B_\Phi\left((1-u) \cdot \mathrm{IPM}_G\left(p_{\Phi,w^{(1)}}^{t=1}(r), p_{\Phi,w^{(1)}}^{t=0}(r)\right) + u \cdot \mathrm{IPM}_G\left(p_{\Phi,w^{(0)}}^{t=1}(r), p_{\Phi,w^{(0)}}^{t=0}(r)\right)\right)\end{aligned} \tag{18}$$

where $\sigma^2 = \min_{(t,t')\in\mathcal{T}^2}\{\sigma^2_{t,w^{(t)}}\left(p(x \,|\, t')\right)\}$, $p_{\Phi,w}^t(r) = w\left(\Psi(r)\right) \cdot p^t(r)$.

Eq. (18) in Theorem 5 consists of factual losses and an IPM on the representation space and thus can be estimated from finite samples. From the theoretical implications above, the loss function to derive a hypothesis h and a representation function Φ is:

$$h, \Phi = \min_{h,\Phi} \frac{1}{n}\sum_{i=1}^{n} w^{(t)}(x_i) \cdot L\left(h\left(\Phi\left(x_i\right), t_i\right), y_i\right) + \alpha \cdot \mathrm{IPM}_{\mathrm{G}}\left(\left\{\Phi\left(x_i\right)\right\}_{i:t_i=0}, \left\{\Phi\left(x_i\right)\right\}_{i:t_i=1}\right) \tag{19}$$

where α is a trade-off hyperparameter, and we use the squared loss as the loss function L in the experiments.

The derived plug-in tau is unbiased for the true metric and minimizes the upper bound of the variance in Eq. (18). A summary of the resulting model selection procedure called *counterfactual cross-validation* is given in Algorithm 1.

5 Experiments

In this section, we compare our proposed performance estimator and the other baselines using a standard semi-synthetic dataset.

5.1 Basic Experimental Setups

5.1.1 Dataset

We used the Infant Health Development Program (IHDP) dataset provided by [28]. The IHDP is an interventional program aimed to improve the health of premature infants [9, 28]. This is a standard semi-synthetic dataset containing 747 children with 25 features and has been widely used to evaluate CATE prediction models [4, 7, 9]. A detailed description of this dataset can be found in Section 5.1 of [7]. We used the simulated outcome implemented in the *EconML* package.

5.1.2 Baseline & Proposed Metrics

We compared the following evaluation metrics:

- IPW validation [10, 11]: This metric utilizes the following form of performance estimator:

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

  where $\tilde{\tau}_{IPW}(X_i, T_i, Y_i^{obs}) = \frac{T_i}{e(X_i)}Y_i^{obs} - \frac{1-T_i}{1-e(X_i)}Y_i^{obs}$ is used as a plug-in-tau that satisfies the requirement of unbiasedness for the CATE.

- 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. (9).

$$\frac{1}{n}\sum_{i=1}^{n}\left((\tilde{\tau}_i^{(1)} - \tilde{\tau}_i^{(0)}) - \hat{\tau}(X_i)\right)^2$$

  where $\tilde{\tau}_i^{(1)}$ and $\tilde{\tau}_i^{(0)}$ are predictions for potential outcomes. We used Counterfactual Regression [7] to construct the plug-in tau $\tilde{\tau}^{(1)}(\cdot)$ and $\tilde{\tau}^{(0)}(\cdot)$ to ensure a fair comparison.

- τ-risk [11]: This metric is derived from the loss function of R-learner in [12] and is defined as follows:

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

  where $m(\cdot)$ is the expectation of observed outcome $\mathbb{E}[Y^{\text{obs}}|X]$. We used Gradient Boosting Regressor to estimate this parameter.

- Counterfactual Cross-Validation: This is our proposed metric, which relies on the plug-in tau in Eq. (15). The hyperparameter to derive the regression function f can be found in Section B.1 of the Supplementary Material (e.g., the model class $\mathcal{G}$).

We used logistic regression to estimate the propensity score for our proposed plug-in tau and for IPW validation. This is because the true propensity score is generally unknown in real-world situations.

Table 1: Mean with standard errors (SE), median, and worst-case performance of the compared evaluation metrics over 100 realizations are reported. The results show that the proposed CF-CV outperformed the other baselines with respect to both model selection and hyperparameter tuning performance in most cases.

	Rank Correlation (in model selection)			Regret (in model selection)			NRMSE (in hyperparameter tuning)		
	Mean (SE)	Median	Worst-Case	Mean (SE)	Median	Worst-Case	Mean (SE)	Median	Worst-Case
IPW	0.195 (0.039)	0.233	-0.749	1.032 (0.100)	0.766	6.779	0.336 (0.013)	0.325	0.737
τ-risk	0.312 (0.030)	0.306	-0.553	1.392 (0.130)	1.164	7.884	0.324 (0.013)	0.332	0.700
Plug-in	0.914 (0.006)	0.934	0.591	0.073 (0.012)	0.017	0.780	0.257 (0.010)	0.267	0.490
CF-CV	**0.921 (0.005)**	**0.935**	**0.666**	**0.066 (0.012)**	**0.009**	**0.562**	**0.256 (0.009)**	**0.252**	**0.483**

5.2 Comparison on Model Selection Performance

We first compared the model selection performance of our CF-CV with the other baselines.

5.2.1 Experimental Procedure

We follow the experimental procedure in [11]; We first trained candidate predictors on the training set, and then made predictions on both validation and test sets by pre-trained predictors. Then, we ranked those predictors by using each metric on the observational 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.

5.2.2 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 with RBF kernel) and five meta-learners (S-Learner, X-Learner, T-Learner, Domain Adaptation Learner, and Doubly Robust Learner) as implemented in *EconML* package[2]. Thus, we had a set of 25 CATE predictors to select among (i.e., $|\mathcal{M}| = 25$).

5.2.3 Results

Table 1 reports the mean, median, and worst-case performance over 100 realizations. We evaluated the worst-case model selection performance of each metric because, in real-world causal inference problems, we never know the ground-truth performance of any predictor, and stable model selection performance is essential. *Rank Correlation* is the Spearman rank correlation between the ranking by the true performance

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

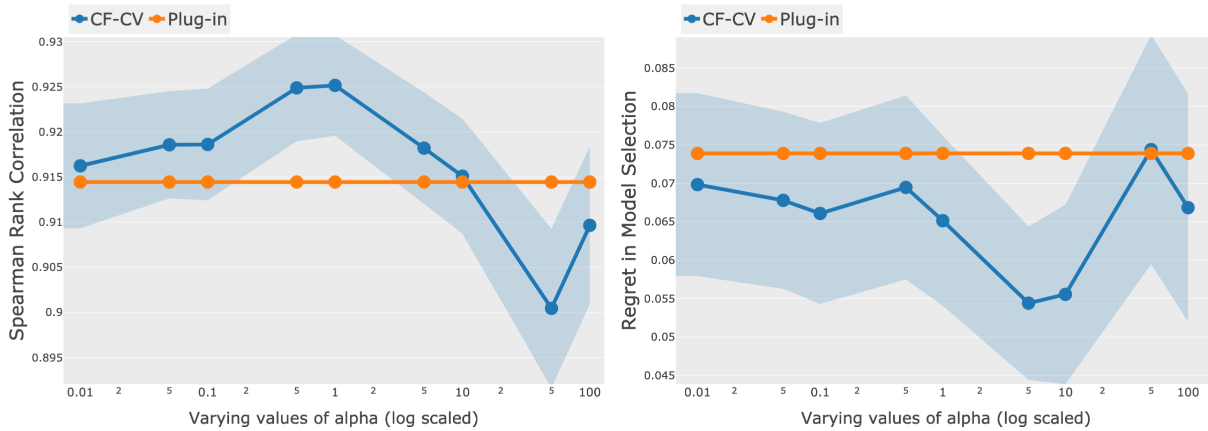

Figure 1: (Left) Rank correlation of CF-CV with different values of α in Eq. (19). It outperforms the pulg-in with small values of α and shows a stable ranking performance with variation in α. (Right) Regret of CF-CV with different values of α in Eq. (19). It consistently outperforms the plug-in in all settings and demonstrates robustness to the choice of α.

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 one in $\mathcal{M}$, and is defined as follows:

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

where $\hat{\tau}_{\text{selected}} = \arg\min_{\hat{\tau}\in\mathcal{M}} \widehat{\mathcal{R}}(\hat{\tau})$ is the model selected by metrics and $\hat{\tau}_{\text{best}} = \arg\min_{\hat{\tau}\in\mathcal{M}} \mathcal{R}_{\text{true}}\left(\hat{\tau}\right)$ 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 with respect to worst-case performance. This result empirically suggests that the proposed metric can stably select a better predictor among potential candidates, and is an appropriate choice for real-world situations. We also evaluated the sensitivity of the proposed metric to changes of the trade-off hyperparameter α in Eq. (19)[3]. Figure 1 shows the performances of CF-CV with variation of α compared to the performance of the plug-in. 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 larger values of α. Additionally, the proposed metric consistently outperformed the plug-in metric with all values of α. These results suggest that the proposed metric can assist practitioners to stably select better causal inference models using only observation validation data.

5.3 Comparison on Hyperparameter Tuning Performance

Next, we compared the hyperparameter tuning performance of our CF-CV with the other baseline metrics.

5.3.1 Tuned Model

We tuned the hyperparameters of the combination of Gradient Boosting Regressor and Domain Adaptation Learner as implemented in *scikit-learn* and *EconML*, respectively. Domain Adaptation Learner consists

[3]Tested values were $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$.

of three base learners including **treated_model**, **controls_model**, and **overall_model**. Thus, we aimed to find the best three sets of hyperparameters of Gradient Boosting Regressor to optimize Domain Adaptation Learner.

5.3.2 Experimental Procedure

We used *Optuna* software [29] to tune the CATE predictor and set each metric as the objective function of *Optuna*. 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 conducted the experimental procedure over the same 100 realizations, using 35/35/30 train/validation/test splits as the model selection experiment.

5.3.3 Results

Table 1 provides the results of the hyperparameter tuning experiment. We report the mean, median, and worst-case *normalized root-mean-squared-error (NRMSE)* of CATE predictors tuned by each metric defined below. We used NRMSE because potential outcomes of the IHDP dataset have different scales among realizations.

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

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

Table 1 shows that our metric improved the median NRMSE by 5.6 % and the worst-case performance by 1.4 % compared to the best baselines. Although the mean of NRMSE is almost the same with the plug-in metric, the results demonstrate that the proposed metric allows one to conduct stable hyperparameter tuning of CATE prediction models.

6 Conclusion

In this paper, we explored the evaluation problem of CATE prediction models. In contrast to previous studies, the proposed plug-in tau preserves the rankings of the true prediction performances of candidate models and minimizes the upper bound of the finite sample uncertainty in model evaluation. We achieved this by using a modified version of CFR as a regression function of the DR estimator to minimize the finite sample uncertainty. Empirical evaluations using the IHDP dataset demonstrated the effectiveness and stability of the proposed metric for evaluating causal inference models.

Important future research directions include consideration of situations with hidden confounders, and a theoretical analysis of the selection of hyperparameters for the regression function f used in the proposed plug-in tau.

Acknowledgements. The work was done during YS's internship at CyberAgent, Inc., AI Lab. The authors would like to thank Yusuke Narita for his helpful comments and discussions.

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

Bibliography

[1] Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *Proceedings of the AdKDD and TargetAd Workshop, KDD, London, United Kingdom*, 2018.

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

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

[4] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

[5] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pages 2633–2643, 2018.

[6] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017.

[7] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

[8] Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *arXiv preprint arXiv:1904.13335*, 2019.

[9] Ahmed Alaa and Mihaela Schaar. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pages 129–138, 2018.

[10] Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*, pages 1–13, 2017.

[11] Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

[12] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

[13] Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pages 191–201, 2019.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[15] Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029, 2016.

[16] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. In *22nd Conference on Learning Theory, COLT 2009*, 2009.

[17] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[18] Soko Setoguchi, Sebastian Schneeweiss, M Alan Brookhart, Robert J Glynn, and E Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

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

[20] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

[22] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

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

[24] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

[25] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[26] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104. Omnipress, 2011.

[27] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1446–1455, 2018.

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

[29] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2623–2631, New York, NY, USA, 2019. ACM.

Supplementary Materials

For all the proofs below, we denote $\tau(X_i)$, $\widehat{\tau}(X_i)$, and $\tilde{\tau}(X_i, T_i, Y_i^{\text{obs}})$ as τ_i, $\widehat{\tau}_i$, and $\tilde{\tau}_i$ for simplicity.

A Proof of Proposition 1

Proof. First, the following equality holds:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{R}}(\widehat{\tau})\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i - \widehat{\tau}_i)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[(\tilde{\tau}_i - \tau_i + \tau_i - \widehat{\tau}_i)^2\right] \\
&= \mathbb{E}\left[(\tilde{\tau}_i - \tau(X))^2\right] - \frac{2}{n}\sum_{i=1}^{n}\underbrace{\mathbb{E}\left[(\widehat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)\right]}_{(a)} + \underbrace{\mathbb{E}\left[(\widehat{\tau}(X) - \tau(X))^2\right]}_{\mathcal{R}_{\text{true}}}
\end{aligned}
$$

Then, we have,

$$
\begin{aligned}
(a) &= \mathbb{E}\left[(\widehat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(\widehat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i) \mid X\right]\right] \\
&= \mathbb{E}[\mathbb{E}\left[(\widehat{\tau}_i - \tau_i) \mid X\right]\mathbb{E}\left[(\tilde{\tau}_i - \tau_i) \mid X\right]] = 0 \because \text{unbiasedness of } \tilde{\tau}_i.
\end{aligned}
$$

Thus, we obtain,

$$
\mathbb{E}\left[\widehat{\mathcal{R}}\left(\widehat{\tau}\right)\right] = \mathcal{R}_{\text{true}}\left(\widehat{\tau}\right) + \mathbb{E}\left[(\tau(X) - \tilde{\tau}(X))^2\right]
$$

□

B Derivation of Eq. (11)

Proof. Following the same procedure as in the proof of Proposition 1,

$$
\begin{aligned}
\widehat{\mathcal{R}}(\widehat{\tau}) &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i - \widehat{\tau}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i - \tau_i + \tau_i - \widehat{\tau}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}_i - \tau_i)^2 - \frac{2}{n}\sum_{i=1}^{n}(\widehat{\tau}_i - \tau_i)(\tilde{\tau}_i - \tau_i) + \frac{1}{n}\sum_{i=1}^{n}(\widehat{\tau}_i - \tau_i)^2
\end{aligned}
$$

□

C Proof of Proposition 3

Proof. We define the DR oracle as:

$$\begin{aligned}\tilde{\tau}_{DR}(X,T,Y) &= \frac{T}{e(X)}\left(Y^{\text{obs}} - f(X,1)\right) - \frac{1-T}{1-e(X)}\left(Y^{\text{obs}} - f(X,0)\right) + (f(X,1) - f(X,0)) \\ &= \tilde{\tau}_{DR}^{(1)}(X,T,Y) - \tilde{\tau}_{DR}^{(0)}(X,T,Y)\end{aligned}$$

where

$$\begin{aligned}\tilde{\tau}_{DR}^{(1)}(X,T,Y) &= \frac{T}{e(X)}\left(Y^{\text{obs}} - f(X,1)\right) + f(X,1) \\ \tilde{\tau}_{DR}^{(0)}(X,T,Y) &= \frac{1-T}{1-e(X)}\left(Y^{\text{obs}} - f(X,0)\right) + f(X,0)\end{aligned}$$

Then, the expectation of $\tilde{\tau}_{DR}^{(1)}$ is:

$$\begin{aligned}\mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \,|\, X\right] &= \mathbb{E}\left[\frac{T}{e(X)}\left(Y^{(1)} - f(X,1)\right) + f(X,1) \,|\, X\right] \\ &= \mathbb{E}\left[\frac{T}{e(X)} \,|\, X\right]\mathbb{E}\left[\left(Y^{(1)} - f(X,1)\right) \,|\, X\right] + f(X,1) \quad \because \text{Unconfoundedness} \\ &= \mathbb{E}\left[Y^{(1)} \,|\, X\right]\end{aligned}$$

We also have $\mathbb{E}\left[\tilde{\tau}_{DR}^{(0)} \,|\, X\right] = \mathbb{E}\left[Y^{(0)} \,|\, X\right]$ is the same way. Thus,

$$\mathbb{E}\left[\tilde{\tau}_{DR} \,|\, X\right] = \mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} - \tilde{\tau}_{DR}^{(0)} \,|\, X\right] = \mathbb{E}\left[Y^{(1)} - Y^{(0)} \,|\, X\right] = \tau(X)$$

□

D Proof of Proposition 4

Proof. The second moment of $\tilde{\tau}_{DR}^{(1)}$ is

$$\begin{aligned}\mathbb{E}\left[\left(\tilde{\tau}_{DR}^{(1)}\right)^2 \,|\, X\right] &= \mathbb{E}\left[\left(\frac{T}{e(X)}\left(Y^{(1)} - f(X,1)\right) + f(X,1)\right)^2 \,|\, X\right] \\ &= \mathbb{E}\left[\left(\left(1 - \frac{T}{e(X)}\right)\left(f(X,1) - Y^{(1)}\right) + Y^{(1)}\right)^2 \,|\, X\right] \\ &= \mathbb{E}\left[\left(\xi^{(1)}\right)^2 \,|\, X\right] + \left(m^{(1)}(X)\right)^2 + \frac{1-e(X)}{e(X)}\left(f(X,1) - m^{(1)}(X)\right)^2\end{aligned}$$

We also have the second moment of $\tilde{\tau}_{DR}^{(0)}$ in the same manner as follows:

$$\mathbb{E}\left[\left(\tilde{\tau}_{DR}^{(0)}\right)^2 \,|\, X\right] = \mathbb{E}\left[\left(\xi^{(0)}\right)^2 \,|\, X\right] + \left(m^{(0)}(X)\right)^2 + \frac{e(X)}{1-e(X)}\left(f(X,0) - m^{(0)}(X)\right)^2$$

Thus, by using the result of Proposition 2, we obtain

$$\begin{aligned}\mathbb{V}\left(\tilde{\tau}_{DR}^{(1)} \,|\, X\right) &= \mathbb{E}\left[\left(\tilde{\tau}_{DR}^{(1)}\right)^2 \,|\, X\right] - \left(\mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \,|\, X\right]\right)^2 \\ &= \mathbb{E}\left[\left(\xi^{(1)}\right)^2 \,|\, X\right] + \frac{1-e(X)}{e(X)}\left(f(X,1) - m^{(1)}(X)\right)^2\end{aligned}$$

$$\mathbb{V}\left(\tilde{\tau}_{DR}^{(0)} \,|\, X\right) = \mathbb{E}\left[\left(\tilde{\tau}_{DR}^{(0)}\right)^2 \,|\, X\right] - \left(\mathbb{E}\left[\tilde{\tau}_{DR}^{(0)} \,|\, X\right]\right)^2 = \mathbb{E}\left[\left(\xi^{(0)}\right)^2 \,|\, X\right] + \frac{e(X)}{1-e(X)}\left(f(X,0) - m^{(0)}(X)\right)^2$$

In addition, from Lemma 3.

$$Cov\left(\tilde{\tau}_{DR}^{(1)}, \tilde{\tau}_{DR}^{(0)} \,|\, X\right) = -\left(f(X,1) - m^{(1)}(X)\right)\left(f(X,0) - m^{(0)}(X)\right)$$

Therefore,

$$\begin{aligned}\mathbb{V}\left(\tilde{\tau}_{DR} \,|\, X\right) &= \mathbb{V}\left(\tilde{\tau}_{DR}^{(1)} - \tilde{\tau}_{DR}^{(0)} \,|\, X\right) \\ &= \mathbb{V}\left(\tilde{\tau}_{DR}^{(1)} \,|\, X\right) - 2Cov\left(\tilde{\tau}_{DR}^{(1)}, \tilde{\tau}_{DR}^{(0)} \,|\, X\right) + \mathbb{V}\left(\tilde{\tau}_{DR}^{(0)} \,|\, X\right) \\ &= \mathbb{E}\left[\left(\xi^{(1)}\right)^2 \,|\, X\right] + \mathbb{E}\left[\left(\xi^{(0)}\right)^2 \,|\, X\right] + \frac{1-e(X)}{e(X)}\left(f(X,1) - m^{(1)}(X)\right)^2 \\ &+ \frac{e(X)}{1-e(X)}\left(f(X,0) - m^{(0)}(X)\right)^2 + 2\left(f(X,1) - m^{(1)}(X)\right)\left(f(X,0) - m^{(0)}(X)\right) \\ &= \mathbb{E}\left[\left(\xi^{(1)}\right)^2 \,|\, X\right] + \mathbb{E}\left[\left(\xi^{(0)}\right)^2 \,|\, X\right] \\ &\quad + \left(\sqrt{\frac{1-e(X)}{e(X)}}\left(f(X,1) - m^{(1)}(X)\right) + \sqrt{\frac{e(X)}{1-e(X)}}\left(f(X,0) - m^{(0)}(X)\right)\right)^2 \\ &= \mathbb{E}\left[\left(\xi^{(1)}\right)^2 \,|\, X\right] + \mathbb{E}\left[\left(\xi^{(0)}\right)^2 \,|\, X\right] + \mathcal{Z}\end{aligned}$$

$\square$

E Proof of Theorem 2

Proof.

$$
\begin{aligned}
\mathbb{V}\left(\frac{2}{n}\sum_{i=1}^{n}(\widehat{\tau}_i-\tau_i)(\tilde{\tau}_i-\tau_i)\right) &= \frac{4}{n^2}\mathbb{V}\left(\sum_{i=1}^{n}(\widehat{\tau}_i-\tau_i)(\tilde{\tau}_i-\tau_i)\right)\\
&= \frac{4}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n}(\widehat{\tau}_i-\tau_i)(\tilde{\tau}_i-\tau_i)\right)^2\right] \ \because (a)=0\\
&= \frac{4}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(\widehat{\tau}_i-\tau_i)^2(\tilde{\tau}_i-\tau_i)^2\right]\\
&\because \mathbb{E}[(\widehat{\tau}_i-\tau_i)(\tilde{\tau}_i-\tau_i)(\widehat{\tau}_j-\tau_j)(\tilde{\tau}_j-\tau_j)]=0, \forall i,j(i\neq j)\\
&\leq \frac{4C}{n}\mathbb{E}\left[(\tilde{\tau}_i-\tau_i)^2\right]\\
&= \frac{4C}{n}\mathbb{V}\left(\tilde{\tau}_i\right) \ \because \mathbb{E}[\tilde{\tau}_i\,|\,X]=\tau_i\\
&= \frac{4C}{n}\left(\mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}_i\,|\,X\right)\right]+\mathbb{V}\left[\mathbb{E}\left(\tilde{\tau}_i\,|\,X\right)\right]\right) \ \because \text{law of total variance}\\
&= \frac{4C}{n}\left(\mathbb{E}\left[\mathbb{V}\left(\tilde{\tau}_i\,|\,X\right)\right]+\mathbb{V}\left(\tau_i\right)\right) \ \because \mathbb{E}[\tilde{\tau}_i\,|\,X]=\tau_i
\end{aligned}
$$

□

F Proofs of Technical Lemmas

First, we prove some lemmas.

Lemma 6. *(Similar to Lemma A.4 of [7]) Let $\Phi:\mathcal{X}\to\mathcal{R}$ be an invertible representation with Ψ its inverse. Let G be a family of functions $g:\mathcal{R}\to\mathbb{R}$ and $h:\mathcal{R}\times\mathcal{T}\to\mathcal{Y}$ be a hypothesis. Assume that, for any given $t\in\mathcal{T}$, there exists a constant $B_\Phi>0$, such that $\frac{1}{B_\Phi}\cdot\ell_{h,\Phi}(\Psi(r),t)\in$ G. Then we have:*

$$
\epsilon_{CF}^{t}\left(h,\Phi,w\right)\leq\epsilon_{F}^{t}\left(h,\Phi,w\right)+B_\Phi\cdot\mathrm{IPM}_G\left(p_{\Phi,w}^{t=1}(r),\, p_{\Phi,w}^{t=0}(r)\right)
$$

Proof.

$$
\begin{aligned}
\epsilon_{CF}^{t}\left(h,\Phi,w\right)-\epsilon_{F}^{t}\left(h,\Phi,w\right) &= \int_{\mathcal{X}} w(x)\ell_{h,\Phi}(x,t)\left(p^{1-t}(x)-p^{t}(x)\right)dx\\
&= \int_{\mathcal{R}} w(\Psi(r))\ell_{h,\Phi}(\Psi(r),t)\left(p_\Phi^{1-t}(r)-p_\Phi^{t}(r)\right)dr\\
&= B_\Phi\cdot\int_{\mathcal{R}}\frac{\ell_{h,\Phi}(\Psi(r),t)}{B_\Phi}\left(w(\Psi(r))p_\Phi^{1-t}(r)-w(\Psi(r))p_\Phi^{t}(r)\right)dr\\
&\leq B_\Phi\cdot\sup_{g\in G}\left|\int_{\mathcal{R}} g(r)\left(w(\Psi(r))p_\Phi^{1-t}(r)-w(\Psi(r))p_\Phi^{t}(r)\right)dr\right|\\
&= B_\Phi\cdot IPM_G\left(w(\Psi(r))p_\Phi^{1-t}(r),\, w(\Psi(r))p_\Phi^{t}(r)\right)
\end{aligned}
$$

□

Lemma 7. *(Similar to Lemma A.5 of [7]) For any function $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ and conditional probability distribution $p^t(x)$ over the space $\mathcal{X} \times \mathcal{T}$, the following equalities hold:*

$$\int_{\mathcal{X}} w(x) \left(f(x,t) - m^{(t)}(x)\right)^2 p^t(x)dx = \epsilon_F^t(f,w) - \sigma_{t,w}^2 \left(p^t(x)\right)$$
$$\int_{\mathcal{X}} w(x) \left(f(x,t) - m^{(t)}(x)\right)^2 p(x \,|\, 1-t)dx = \epsilon_{CF}^t(f,w) - \sigma_{t,w}^2 \left(p^{1-t}(x)\right)$$

Proof.

$$\begin{aligned}
\epsilon_F^t(f,w) &= \int_{\mathcal{X}\times\mathcal{Y}} w(x) \left(f(x,t) - Y^{(t)}\right)^2 p\left(Y^{(t)}|x\right) p(x \,|\, t) dY^{(t)} dx \\
&= \int_{\mathcal{X}} w(x) \left(f(x,t) - m^{(t)}(x)\right)^2 p(x \,|\, t)dx \\
&- 2\int_{\mathcal{X}\times\mathcal{Y}} w(x) \left(f(x,t) - m^{(t)}(x)\right) \left(Y^{(t)} - m^{(t)}(x)\right) p\left(Y^{(t)}, x \,|\, t\right) dY^{(t)} dx \\
&+ \int_{\mathcal{X}\times\mathcal{Y}} w(x) \left(Y^{(t)} - m^{(t)}(x)\right)^2 p\left(Y^{(t)}, x \,|\, t\right) dY^{(t)} dx \\
&= \int_{\mathcal{X}} w(x) \left(f(x,t) - m^{(t)}(x)\right)^2 p(x \,|\, t)dx + \sigma_{t,w}^2 \left(p(x \,|\, t)\right)
\end{aligned}$$

Thus, we have,

$$\int_{\mathcal{X}} w(x) \left(f(x,t) - m^{(t)}(x)\right)^2 p^t(x)dx = \epsilon_F^t(f,w) - \sigma_{t,w}^2 \left(p^t(x)\right)$$

We can derive the analogous equality for counterfactual losses in the same manner. □

Lemma 8. *The covariance of $\tilde{\tau}_{DR}^{(1)}$ and $\tilde{\tau}_{DR}^{(0)}$ is:*

$$Cov\left(\tilde{\tau}_{DR}^{(1)}, \tilde{\tau}_{DR}^{(0)} \,|\, X\right) = -\left(f(X,1) - m^{(1)}(X)\right)\left(f(X,0) - m^{(0)}(X)\right)$$

Proof.

$$\begin{aligned}
Cov\left(\tilde{\tau}_{DR}^{(1)}, \tilde{\tau}_{DR}^{(0)} \,|\, X\right) &= \mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \cdot \tilde{\tau}_{DR}^{(0)} \,|\, X\right] - \mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \,|\, X\right] \cdot \mathbb{E}\left[\tilde{\tau}_{DR}^{(0)} \,|\, X\right] \\
&= \mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \cdot \tilde{\tau}_{DR}^{(0)} \,|\, X\right] - m^{(1)}(X) \cdot m^{(0)}(X)
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \cdot \tilde{\tau}_{DR}^{(0)} \,|\, X\right] &= f(X,1)f(X,0) + f(X,1)\mathbb{E}\left[\frac{1-T}{1-e(X)}\left(Y^{(0)} - f(X,0)\right) \,|\, X\right] \\
&\quad + f(X,0)\mathbb{E}\left[\frac{T}{e(X)}\left(Y^{(1)} - f(X,1)\right) \,|\, X\right] \\
&= f(X,1)f(X,0) + f(X,1)(m^{(0)}(X) - f(X,0)) + f(X,0)(m^{(1)}(X) - f(X,1))
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\tilde{\tau}_{DR}^{(1)} \cdot \tilde{\tau}_{DR}^{(0)} \,|\, X\right] - m^{(1)}(X)m^{(0)}(X) \\
&= f(X,1)f(X,0) + f(X,1)(m^{(0)}(X) - f(X,0)) \\
&\quad + f(X,0)(m^{(1)}(X) - f(X,1)) - m^{(1)}(X)m^{(0)}(X) \\
&= -\left(f(X,1) - m^{(1)}(X)\right)\left(f(X,0) - m^{(0)}(X)\right)
\end{aligned}$$

□

G Proof of Theorem 5

Proof.

$$
\begin{aligned}
\mathcal{Z} &\le 2\int_{\mathcal{X}}\Big(w^{(1)}(x)\left(f(x,1)-m^{(1)}(x)\right)^2 + w^{(0)}(x)\left(f(x,0)-m^{(0)}(x)\right)^2\Big)p(x)\,dx \;\;\because (x+y)^2 \le 2(x^2+y^2) \\
&= 2u\int_{\mathcal{X}} w^{(1)}(x)\left(f(x,1)-m^{(1)}(x)\right)^2 p^{t=1}(x)\,dx = +2(1-u)\int_{\mathcal{X}} w^{(1)}(x)\left(f(x,1)-m^{(1)}(x)\right)^2 p^{t=0}(x)\,dx \\
&+ 2u\int_{\mathcal{X}} w^{(0)}(x)\left(f(x,0)-m^{(0)}(x)\right)^2 p^{t=1}(x)\,dx + 2(1-u)\int_{\mathcal{X}} w^{(0)}(x)\left(f(x,0)-m^{(0)}(x)\right)^2 p^{t=0}(x)\,dx \\
&= 2u\left(\epsilon_F^{t=1}(f,w^{(1)}) - \sigma^2_{t=1,w^{(1)}}\left(p^{t=1}(x)\right)\right) + 2(1-u)\left(\epsilon_{CF}^{t=1}(f,w^{(1)}) - \sigma^2_{t=1,w^{(1)}}\left(p^{t=0}(x)\right)\right) \\
&+ 2(1-u)\left(\epsilon_F^{t=0}(f,w^{(0)}) - \sigma^2_{t=0,w^{(0)}}\left(p^{t=0}(x)\right)\right) + 2u\left(\epsilon_{CF}^{t=0}(f,w^{(0)}) - \sigma^2_{t=0,w^{(0)}}\left(p^{t=1}(x)\right)\right) \;\;\because \text{Lemma 2} \\
&\le 2\epsilon_F^{t=1}(f,w^{(1)}) + 2\epsilon_F^{t=0}(f,w^{(0)}) \\
&+ 2B_\Phi\left((1-u)\cdot \mathrm{IPM}_G\left(p^{t=1}_{\Phi,w^{(1)}}(r),\, p^{t=0}_{\Phi,w^{(1)}}(r)\right) + u\cdot \mathrm{IPM}_G\left(p^{t=1}_{\Phi,w^{(0)}}(r),\, p^{t=0}_{\Phi,w^{(0)}}(r)\right)\right) - 4\sigma^2 \;\;\because \text{Lemma 1}
\end{aligned}
$$

where $\sigma^2 = \min_{(t,t')\in\mathcal{T}^2}\{\sigma^2_{t,w^{(t)}}\left(p(x\,|\,t')\right)\},\; p^t_{\Phi,w}(r) = w\left(\Psi(r)\right)\cdot p^t(r).$ □

H Detailed Experimental Settings

H.1 Model Selection Experiment in Section 5.2

The weighted counterfactual regression model used as a regression function of our proposed metric has some hyperparameters itself. To tune the hyperparameters of this model, we used the simple μ-risk as described in [11]. Table 2 describes the hyperparameter search spaces and the resulting set of hyperparameters for the weighted counterfactual regression .

Table 2: Hyperparameter search spaces and the selected values of the hyperparameters for the weighted counterfactual regression in our proposed CFCV. A set of hyperparameters optimzed the μ-risk [11] was selected for the weighted CFR.

Hyperparameters	Search spaces	Selected values
Num. of hidden layers for h and Φ in Eq. (19)	$\{1, 2, 3\}$	3
Dim. of hidden layers for h and Φ in Eq. (19)	$\{20, 50, 100\}$	100
trade-off parameter α in Eq. (19)	$[0.01, 100]$	0.356
learning_rate	$[0.0001, 0.01]$	4.292×10^{-4}
batch_size	256 (fixed)	256 (fixed)
dropout rate	0.2 (fixed)	0.2 (fixed)

H.2 Hyperparameter Tuning Experiment in Section 5.3

Table 3 provides the hyperparameter search space of the Gradient Boosting Regressor used in the hyperparameter tuning experiment in Section 5.3.

Table 3: Hyperparameter search space for Gradient Boosting Regressors.

Hyperparameters	Search spaces
n_estimators	100 (fixed)
max_depth	$[1, 20]$
min_samples_leaf	$[1, 20]$
learning_rate	$[10^{-5}, 10^{-1}]$
subsample	$\{0.1, 0.2, \dots 1.0\}$