Estimating Potential Outcome Distributions with Collaborating Causal Networks

Anonymous authors
Paper under double-blind review

Abstract

Traditional causal inference approaches leverage observational study data to determine an individual's outcome change due to a potential treatment, known as the Conditional Average Treatment Effect (CATE). However, CATE is the comparison on the first moment and might be insufficient in reflecting the full picture of treatment effects. As an alternative, estimating the full potential outcome distributions could provide greater insights. However, existing methods in estimating the potential outcome distributions often impose assumptions about the treatment effect outcome distributions. Here, we propose Collaborating Causal Networks (CCN), a novel methodology which goes beyond the estimation of CATE alone by learning the *full potential outcome distribution*. Our proposed method facilitates estimation of the utility of each possible treatment and permits individual-specific variation in utility functions (e.g., risk tolerance variability). CCN not only extends outcome estimation beyond traditional risk difference, but also enables a more comprehensive decision making process through definition of flexible comparisons. Under standard causal inference assumptions, we show that CCN learns distributions that asymptotically capture the correct potential outcome distributions. Furthermore, we propose an adjustment approach that is empirically effective in alleviating sample imbalance between treatment groups in observational studies. Finally, we evaluate CCN's performance in multiple extensive experiments. We demonstrate CCN learns improved distribution estimates compared to existing Bayesian and deep generative methods as well as improved decisions on a variety of utility functions.

1 Introduction

Personalized medicine requires estimating how an individual's intrinsic biological characteristics trigger unique and heterogeneous responses to treatments (Yazdani & Boerwinkle, 2015). Under the causal inference potential outcome framework (Imbens & Rubin, 2015), these treatment effects are characterized by the difference between the conditional expectations of potential outcomes under different administered treatments, known as the Conditional Average Treatment Effect (CATE). Since only the outcome for the assigned treatment is observed, estimating CATE often requires inferring the missing potential outcomes (Ding & Li, 2018). Recently, machine learning approaches have been applied to estimation of CATEs by extending approaches such as Random Forests (Wager & Athey, 2018) and creating bespoke neural network frameworks (Shalit et al., 2017; Shi et al., 2019).

However, CATEs do not necessarily align with optimal choices. In a decision theoretic framework, the optimal decision maximizes the expected utility function $U(\gamma)$ over the distribution of outcomes (Joyce, 1999). CATE is a special case with an identity utility function $U(\gamma) = \gamma$, but more general utility functions require alternative estimation approaches. One approach to learn a decision maker is to optimize the function with respect to transformed outcomes in accordance to the utility function as the objective function, known as policy learning (Kallus & Zhou, 2018; Qian & Murphy, 2011). However, traditional policy learning methods require the utility function to be pre-specified, whereas utility functions typically vary between individuals Pennings & Smidts (2003). Defining proper losses with respects to complex decision criteria could be challenging from both a theoretical as well as a computational standpoint. Moreover, if we also want to account for an individual's specific needs or to have a trained model generalize to new individuals

with different utilities, traditional policy learning methods fall short. Previous efforts to estimate potential outcome distributions include Bayesian Additive Regression Trees (BART) (Chipman et al., 2010; Hill, 2011), variational methods (Louizos et al., 2017), generalized additive models with location, shape and scale (GAMLSS) (Hohberg et al., 2020), and techniques based on adversarial networks (Yoon et al., 2018; Ge et al., 2020). Empirically, these techniques often impose explicit or implicit assumptions about the outcome distributions (e.g., Gaussian errors), which may not match or align with the true data generating mechanism.

To address these issues, we propose a novel neural network-based approach, Collaborating Causal Networks (CCN), to estimate the full potential outcome distributions, in turn providing flexible personalization and valuable insights with regards to specific individuals. CCN modifies the structure of Collaborating Networks (CN) (Zhou et al., 2021) to create a new causal framework that flexibly represents distributions. Under standard causal inference assumptions, we prove that CCN asymptotically captures conditional potential outcome distributions. We also introduce an adjustment method that alleviates effects of imbalance between treatment groups, thus addressing a common confound that impairs model generalization. Empirically, this adjustment method improves point estimates, distribution estimates, and decision-making.

We summarize the main contributions of our work below:

1. We propose the Collaborating Causal Networks (CCN) to estimate full potential outcome distributions.
2. We theoretically prove the asymptotic properties of CCN.
3. We propose an adjustment scheme that combines domain invariant, propensity-specific information and propensity stratification to alleviate the effects of treatment group imbalance.
4. We propose and evaluate personalized utilities in decision making in causal inference, which is practically more meaningful and addresses distinct user needs.
5. We empirically show that CCN improves conditional decisions in a potential outcomes framework.

2 Problem Statement

2.1 Notations

We define the covariates as $X \in \mathcal{X} \subset \mathbb{R}^p$. We assume a binary treatment condition and each unit is assigned a treatment $T \in \{0, 1\}$. We choose the binary setup for clarity, and the multi-class case could be constructed under the same framework. We let $Y(0) \in \mathbb{R}^1$ and $Y(1) \in \mathbb{R}^1$ represent the continuous potential outcomes under the two treatments; $Y(T)$ is the observed outcome. We use lowercase letters with subscript i to denote observations on each subject: $\{y_i(1), y_i(0), t_i, y_i(t_i), x_i\}$. Previously, a common goal is to estimate the CATE, $\tau(x_i) = \mathbb{E}[Y(1)|X = x_i] - \mathbb{E}[Y(0)|X = x_i]$. In some of the existing literature, this is also referred to as the Individual Treatment Effect (ITE), (Yao et al., 2018; Shalit et al., 2017). Due to the existence of unmeasured features in practice, Vegetabile (2021) points out that $\tau(x_i)$ is still an average, but taken over individuals with same observed features. We adopt this rigor in definition and also describe the main objective of this paper as to address the conditional causal inference.

Additionally, we wish to study a wider range of objectives beyond CATE. Specifically, our goal is to use the incomplete data to infer the distributions on both potential outcomes, $p(Y(0)|X)$ and $p(Y(1)|X)$. Successful estimation of these distributions enables us to explore personalized needs through the introduction of utility functions $U(\gamma)$ (Dehejia, 2005).

2.2 Utility Functions

We can vary the structure of utility functions to adapt to different levels of comparisons. They may be described by four categories:

(i) unified utility: $\mathbb{E}_{\gamma \sim p(Y(1)|X)}[U(\gamma)] - \mathbb{E}_{\gamma \sim p(Y(0)|X)}[U(\gamma)]$;

(ii) treatment-specific utility: $\mathbb{E}_{\gamma \sim p(Y(1)|X)}[U_1(\gamma)] - \mathbb{E}_{\gamma \sim p(Y(0)|X)}[U_0(\gamma)]$;

(iii) utility related to inherent feature: $\mathbb{E}_{\gamma\sim p(Y(1)|X)}[U_1(\gamma, X)] - \mathbb{E}_{\gamma\sim p(Y(0)|X)}[U_0(\gamma, X)]$;

(iv) utility for personalized needs: $\mathbb{E}_{\gamma\sim p(Y(1)|X)}[U_{1,i}(\gamma, X)] - \mathbb{E}_{\gamma\sim p(Y(0)|X)}[U_{0,i}(\gamma, X)]$.

From (i) to (iv), the scope of comparison is broadened. (iv) may or may not depend on X. A major distinction between (iii) and (iv) is that, for subject with same feature X, (iv) allows the utility to differ in accordance to personal preferences, which better aligns with the concept of personalized medicine. We note that $U(\gamma) = \gamma$ in (i) makes the objective CATE. Therefore, we view the incorporation of utility as forming a more general framework.

(i), (ii) and (iii) may still be addressed and estimated directly by transforming outcomes in accordance to their defined utilities (Kennedy, 2020). One shortcoming is that utility functions are often nonlinear in measured outcomes in practice (Pennings & Smidts, 2003), and naively transforming the outcome such as for threshold utility, $U(\gamma) = 1_{\gamma>C}$ could result in loss of information. On the contrary, estimating distributions not only enables us to study (iv) but also retains all the information in the system during training.

Using estimated distributions to evaluate utility functions can be viewed as a two-step procedure or a plug-in estimator, which might be less efficient than direct optimization in some cases (Bickel & Ritov, 2003). However, its flexibility and adaptability in studying various types of utilities makes it advantageous for practical purposes (Grunewalder, 2018).

2.3 Causal Assumptions

Like most causal methods, CCN relies on the standard strong ignorability and consistency assumptions (Rosenbaum & Rubin, 1983; Hernan & Robins, 2020), to estimate the potential outcome distributions when each datum only observes a single treatment outcome. They consist of three sub-assumptions:

Assumption 1 (Positivity or overlap)**.** *$\forall X \in \mathcal{X} \subset \mathbb{R}^p$, the probability of assignment to any treatment group is bounded away from zero: $0 < Pr(T = 1|X) < 1$.*

Assumption 2 (Consistency)**.** *The observed outcome given a specific treatment is equal to its potential outcome: $Y|T, X = Y(T)|T, X$.*

Assumption 3 (Ignorability or Unconfoundedness)**.** *The potential outcomes are jointly independent of the treatment assignment conditional on X: $[Y(0), Y(1)] \perp T|X$.*

3 Collaborating Causal Networks

The CCN approach approximates the conditional distributions $Y(0)|X$ and $Y(1)|X$. It uses a two-function framework based on the Collaborating Networks (CN) method (Zhou et al., 2021). We choose to extend CN to the causal setting because it automatically adapts to different distribution families, including non-Gaussian distributions. We first give an overview of CN, then present CCN, and finally introduce our adjustment strategies. Proofs of all theoretical claims are in Appendix A.

3.1 Overview of Collaborating Networks

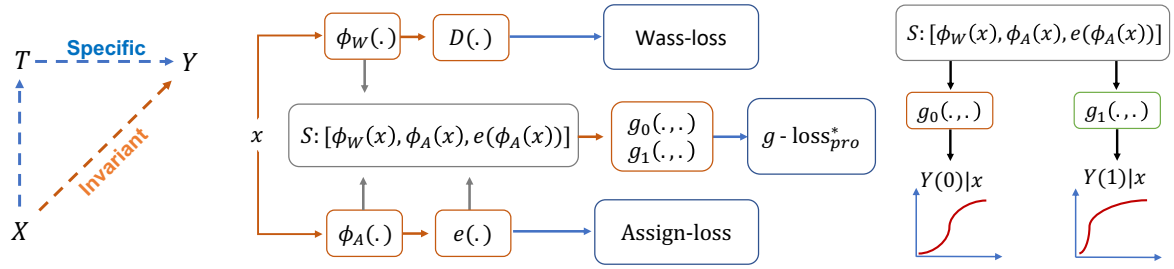
CN estimates the conditional distribution, $Y|X$, with two neural networks: a network $g(Y, X)$ to approximate the conditional CDF, $Pr(Y < y|X)$, and a network $f(q, X)$ to approximate its inverse. Information sharing is enforced by the fact that the CDF and its inverse are an identity mapping for any quantile q: $g(f(q, x), x) = q$. The networks form a *collaborative* scheme with their respective losses,

$$\text{g-loss} : \mathbb{E}_{q,y,x}\left[\ell(1_{y<f(q,x)}, g(f(q, x), x))\right], \tag{1}$$

$$\text{f-loss} : \mathbb{E}_{q,x}\left[(q - g(f(q, x), x))^2\right]. \tag{2}$$

The quantile q is randomly sampled (e.g., $q \sim Unif(0, 1)$). $\ell(\cdot, \cdot)$ represents the binary cross-entropy loss. The parameters for f and g are only updated with their respective losses.

When trained with equation 1 and equation 2, a fixed point of the optimization is at the true conditional CDF and its inverse (Zhou et al., 2021). In this framework, $g(\cdot)$ is the main function which only relies on the



(a) Causal Diagram (b) FCCN Training (c) Infer Distribution

Figure 1: 1(a) depicts how two sources of information could impact Y. 1(b) visualizes the FCCN network. 1(c) depicts how the trained g_0 and g_1 functions can be used to sketch the underlying CDFs of $Y(0)|X$ and $Y(1)|X$.

auxiliary function $f(\cdot)$ to extensively search and cover the full outcome space to attain optimality. On the other hand, $f(\cdot)$'s optimality depends on an optimal $g(\cdot)$, as it acquires information solely through inverting $g(\cdot)$. Zhou et al. (2021) shows that $f(\cdot)$ can be replaced by other space searching tools, including prefixed uniform distributions, which may result in a minor performance loss in favor of ease of optimization. This ease of optimization can be especially beneficial when combining CN with other regularization terms.

Thus, we focus only on extending $g(\cdot)$ and g-loss for causal inference. We replace $f(q, x)$ with a variable z as a general form of a space searching tool, such as a *uniform distribution covering the range of the observed outcomes*. We simplify the g-loss to

$$\text{g-loss} : \mathbb{E}_{y,x,z}\left[\ell(1_{y<z}, g(z,x))\right]. \tag{3}$$

In practice, equation 3 is replaced with an empirical approximation.

3.2 Causal Inference Formulation

Following the taxonomy of Künzel et al. (2019), CN can be extended to an "S-learner," where the treatment label is included as an additional covariate and thus is more scalable for multiple treatment groups, or to a "T-learner," where the outcome under each treatment arm is estimated separately.

Below, we give the T-learner extension of the CN. Its S-learner counterpart could be formulated similarly. Based on equation 3, we define a network for each group $g_0(\cdot)$ and $g_1(\cdot)$, with the corresponding loss as combining the losses of two groups.

$$\text{g-loss}^* \quad = \quad \mathbb{E}_{y(t=0),x,z}\left[\ell(1_{y(t=0)<z}, g_0(z,x))\right] + \mathbb{E}_{y(t=1),x,z}\left[\ell(1_{y(t=1)<z}, g_1(z,x))\right]. \tag{4}$$

We call this framework the Collaborating Causal Networks (CCN). Under Assumptions 1, 2 and 3, CCN's fixed point solution and consistency still hold regardless of the treatment group imbalance.

To summarize, Assumption 2 connects the conditional distribution, $Y|X,T$, to the potential outcome distribution $Y(T)|X,T$ in each treatment space: $p(X|T=0)$ and $p(X|T=1)$. Assumption 1 guarantees that despite the treatment group imbalance: $p(X|T=1) \neq p(X|T=1) \neq p(x)$, each treatment space can still has sufficient coverage of the full space $p(x)$ given enough samples. Lastly, Assumption 3 generalizes the potential outcome distributions from each space $Y(T)|X,T$ to the full space $Y(T)|X$. Given our assumptions, we state:

Proposition 1 (Optimal solution for g_0 and g_1)**.** *When the support of the outcomes is a subset of the support of z, or as z covers the whole outcome space, the functions g_0 and g_1 that minimize g-loss* are optimal when they are equivalent to the conditional CDF of $Y(0)|X=x$ and $Y(1)|X=x$, $\forall x$ such that $p(x)>0$.*

Proposition 2 (Consistency of g_0 and g_1)**.** *Assume the ground truth CDF functions for $T \in \{0,1\}$ is Lipschitz continuous with respect to both the feature X and the potential outcomes $\{Y(0), Y(1)\}$, and the support of the outcomes is a subset of the support of z. Denote the ground truth functions as g_0^* and g_1^*. As $n \to \infty$, the*

finite sample estimators g_0^n and g_1^n have the following consistency property: $d(g_0^n, g_0^*) \to_P 0; d(g_1^n, g_1^*) \to_P 0$ under some metrics d, such as the $\mathbb{L}_1$ norm.

Taken together, these propositions state that the CDF estimators g_0 and g_1 inherit the large sample properties from CN for estimating potential outcome distributions. Detailed proofs are provided in Appendix A. Though the consistency is only claimed for the true functions from Lipschitz family, we regard it to be a reasonable assumption, as the Lipschitz family can provide good approximation to a wide range of functions (Sohrab, 2003), especially considering the boundedness of CDFs between $(0, 1)$. In practice, our approximation functions are represented by generic neural networks, which allow for additional flexibility.

3.3 Adjustment for Treatment Group Imbalance

One obstacle in observational studies is the treatment group imbalance, where the distributions of the covariate spaces $p(x|T = 0)$ and $p(x|T = 1)$ differ significantly. It creates two major issues for causal predictions: generalization over different treatment spaces and confounding effects. In the asymptotic regime (Proposition 2), this imbalance is less problematic since overlap (Assumption 1) ensures all regions with positive density will eventually be densely covered with samples. For finite samples, this imbalance hurts inference. Thus, we pair our method with an adjustment scheme to address this challenge.

We encode the covariates into a space that facilitates the generalization between spaces and adjusts for confounding effects simultaneously. The new space is represented as $S = [\phi_W(X), \phi_A(X), e(\phi_A(X))]$. Two neural networks, $\phi_W(\cdot) : \mathbb{R}^p \to \mathbb{R}^{q_W}$ and $\phi_A(\cdot) : \mathbb{R}^p \to \mathbb{R}^{q_A}$ transform the input $X \in \mathbb{R}^p$ into q_W- and q_A-dimensional latent spaces. In Figure 1(a), they correspond to two sources of information that possibly impact the outcome. One is invariant between the treatment groups (domain invariant) and the other is specific to the differences between the treatment spaces (domain specific) (Shalit et al., 2017; Ben-David et al., 2010). The invariant component $\phi_W(\cdot)$ finds a more balanced representation between spaces that benefits generalization, while the specific component $\phi_A(\cdot)$ controls for confounding effects through learning the treatment assignment mechanism. Additionally, a neural network $e(\cdot) : \mathbb{R}^{q_A} \to [0, 1]$ uses the output of $\phi_A(\cdot)$ to predict the propensity of treatment assignment, $Pr(T = 1|X = x)$. While $e(\phi_A(X))$ seems redundant given $\phi_A(X)$, including $e(\phi_A(X))$ in the covariate space encourages *propensity score stratification* (Hahn et al., 2020).

Domain-invariance is encouraged in the space $\phi_W(\cdot)$ through a penalty on the Wasserstein distance (Wass-loss) between the two treatment arms, and $[\phi_A(\cdot), e(\cdot)]$ is encouraged through a cross-entropy loss on the assigned treatment labels (Assign-loss), as detailed in Sections 3.3.1 and 3.3.2, respectively. If both approaches were encouraged by a single network, they would simply compete with each other. We denote the g-loss* trained on this representation space as g-loss$^*_{pro}$. The representations $[\phi_W(\cdot), \phi_A(\cdot), e(\cdot))]$ with their respective losses are incorporated as regularization terms. This framework is sketched in Figure 1, and the full loss can be expressed as the sum of g-loss$^*_{pro}$,

$$L(g_0, g_1, \phi_W, \phi_A, e) = \text{g-loss}^*_{pro}(g_0, g_1, \phi_W, \phi_A, e) + \alpha\text{Wass-loss}(\phi_W) + \beta\text{Assign-loss}(\phi_A, e). \tag{5}$$

The tuning parameters α and β adjust the relative importance of the losses during learning. Empirically, we have found performance to be robust to selection of these hyperparameter values.. We call this full adjustment CCN (FCCN).

3.3.1 Wass-loss to Alleviate the Covariate Space Imbalance (Domain Invariant)

The introduction of Wass-loss is motivated by CounterFactual Regression (CFR) implemented with the Wasserstein distance (Shalit et al., 2017). CFR is a causal estimator based on representation learning $\phi_W(\cdot) : \mathbb{R}^p \to \mathbb{R}^{q_W}$. The goal is to find latent representations where $p(\phi_W(x)|T = 1)$ and $p(\phi_W(x)|T = 0)$ are more balanced or domain invariant than the original space. We use the Wasserstein-1 distance, which represents the total "work" required transform one distribution to another (Vallender, 1974). Through the Kantorovich-Rubinstein duality Villani (2008), this distribution distance is,

$$W(\mathbb{P}_a, \mathbb{P}_b) = \sup_{\|D\|_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}_a}[D(x)] - \mathbb{E}_{x\sim\mathbb{P}_b}[D(x)].$$

$\|D\|_L \leq 1$ represents the family of 1-Lipschitz functions. We approximate this distance by adopting the approach of Arjovsky et al. (2017). This turns into the following min-max regime,

$$\text{Wass-loss} : \max_{D} \min_{\phi_W} \mathbb{E}_t[(-1)^t \mathbb{E}_{x \sim p(x|t)}[D(\phi_W(x)]].$$

$D(\cdot)$ is parameterized by a small neural network, and the Lipschitz constraint on D is enforced through weight clipping. The Wass-loss penalizes differences in the latent space between the treatment and control group, which improves generalization between groups.

3.3.2 Assign-loss and propensity stratification for confounding effects (Domain Specific)

The introduction of Assign-loss is inspired by Dragonnet (Shi et al., 2019), a deep learning method that learns a latent representation for treatment assignment mechanism to control for confounding effects. It is defined as a binary cross-entropy loss on predicting the treatment assignment label with $e(\phi_A(x))$,

$$\text{Assign-loss} : -\mathbb{E}_{t,x}\left[t \log(e(\phi_A(x)) + (1-t)\log(1 - e(\phi_A(x))\right].$$

We additionally incorporate the estimated propensity $e(\phi_A(x))$ directly into our covariate space. Using propensity scores in the predictive model is a form of continuous stratification to reduce the bias of estimation by facilitating information sharing within sample strata created by $e(\phi_A(x))$ (Hahn et al., 2020).

4 Related Work

CATE estimation. A common approach to estimate CATEs with machine learning is matching, which identifies pairs of similar individuals (Rubin, 1973; Rosenbaum & Rubin, 1983; Li & Fu, 2017; Schwab et al., 2018). This idea motivates many tree-based methods that identify similar individuals within automatically-identified regions of the covariate space (Liaw & Wiener, 2002; Zhang & Lu, 2012; Athey & Imbens, 2016; Wager & Athey, 2018).

Deep learning methods are also common to predict CATEs. As previously mentioned, these include networks with additional loss terms to encourage a treatment-invariant space (Johansson et al., 2020; 2016; Du et al., 2019) and networks that explicitly encode treatment propensity information (Shi et al., 2019). Representation learning can be combined with weighting strategies to enforce covariate balance (Assaad et al., 2021; Hassanpour & Greiner, 2019; 2020a). Despite some of the resemblance in learning the feature embedding, this manuscript performs extensive ablation study to support and justify that each of our adjustment components leads to increased robustness. Additionally, these methods largely focus on estimating only the CATEs (with some exceptions noted below), which may be insufficient to reflect the full picture of different treatment regimes (Park et al., 2021). In contrast, CCN estimates full distributions to assess the utility and confidence of a decision.

Potential outcome distribution sketching. Bayesian methods have been used to estimate outcome distributions, including methods such as Gaussian Processes (Alaa & van der Schaar, 2017), Bayesian dropout (Alaa et al., 2017), and Bayesian Additive Regression Trees (BART) (Chipman et al., 2010). BART has gained popularity in recent years and has been the focus of further modifications, including variations to account for regions with poor overlap (Hahn et al., 2020). However, Bayesian methods can suffer under model mis-specification (Walker, 2013), such as mismatch between the assumed and true outcome distributions. Bayesian methods have also been integrated with deep learning, such as the Causal Effect Variational Autoencoder (CEVAE) and its extensions (Louizos et al., 2017; Jesson et al., 2020); hybrid architectures are sometimes adopted to account for certain types of missing data mechanisms (Hassanpour & Greiner, 2020b).

Frequentist approaches can achieve flexible representations of distributions. A well-known adaptation is the Generalized Additive Model with Location, Scale and Shift (GAMLSS), which estimates the parameters for a baseline distribution with up to three transformations given a specific distribution family (Briseño Sanchez et al., 2020; Hohberg et al., 2020). The CDF may also be estimated nonparametrically by adapting density estimation methods such as nearest neighbors (Shen, 2019), which is less reliable in areas of treatment group imbalance due to sparse samples. GAN-inspired methods, including GANITE (Yoon et al., 2018), can also

learn non-Gaussian outcome distributions. There is emerging literature on conformal prediction in treatment effect estimation (Lei & Candès, 2020; Chernozhukov et al., 2021). However, conformal prediction only learns a specific level of coverage and its coverage probabilities are proven for marginal rather than conditional distributions.

Policy learning and utility functions. A key purpose of estimating the conditional causal effect is to serve decision making. A common strategy called policy learning is to express the policy as a function of the feature space and learn the policy to optimize a pre-defined utility (Kallus & Zhou, 2018; Qian & Murphy, 2011; Bertsimas et al., 2017; Beygelzimer & Langford, 2009). This involves transforming observed outcomes in accordance to the pre-defined utility as the new objective for optimization. Traditionally, utilities studied in policy learning are linear transformation of the potential outcomes, which can be described as the difference between the benefit and cost (Athey & Wager, 2021). However, when we are presented with threshold based utilities, transforming the observed outcomes may be subject to information loss according to the Data Processing Inequality (Beaudry & Renner, 2012) (e.g., binarization of a continuous variable greatly reduces information). Additionally, policy learning can only study one pre-specified utility at a time, whereas a trained potential outcome distribution estimator can serve as a one-stop shop to explore a myriad of personal utility functions.

5 Experiments

We follow established literature and use semi-synthetic scenarios to assess conditional causal effects. First, we use the Infant Health and Development Program (IHDP) (Hill, 2011), where the outcome of each subject is simulated under a standard Gaussian distribution with a heterogeneous treatment effect. This first situation describes an ideal scenario for many methods, including BART. The second example is based on a field experiment in India studying the impact of education (EDU). In this case, we synthesize each individual outcome with heterogeneous effect and variability using a non-Gaussian distribution. The semi-synthetic procedures are briefly outlined below with full details in Appendix B. We additionally provide evaluations on a number of different synthetic outcome distributions to compare methods under different scenarios, including multi-modal distributions and many different distribution families.

We include our base approach, CCN, and its adjusted version, FCCN. We compare it to existing approaches that estimate potential outcome distributions, including Bayesian approaches (CEVAE (Louizos et al., 2017), BART (Hill, 2011)), a frequentist approach, GAMLSS (Hohberg et al., 2020), and a GAN-based approach, GANITE (Yoon et al., 2018). Causal Forests (CF) (Wager & Athey, 2018) is benchmarked for non-distribution metrics as a popular recent CATE-only method. GAMLSS's flexibility and strength in estimating distributions is dependent on a close match to the true distribution families, which is rarely known in practice. However, we evaluate GAMLSS where it is provided the closest possible distribution, meaning that GAMLSS is provided *more information than any other method.* We also benchmark the proposed approaches against policy learning approaches on decision-making metrics. To fully understand the impact of the various adjustments in FCCN compared to CCN, we run ablation studies and evaluate the performance over a suite of hyperparameters for tuning the adjustment.

Detailed specifications for all models are given in Appendix C. We use a standard neural network architecture for CCN, but detail an alternative structure that enforces a monotonic constraint in Appendix D. *Code to replicate all experiments has been included, which will be released with an MIT license if accepted.*

5.1 Metrics

We evaluate mean estimates via Precision in Estimation of Heterogenous Effect (PEHE) and the full distribution by estimating the log-likelihood (LL) of the potential outcomes. LL is regarded as the key metric since it evaluates full distributions. In addition, we evaluate how well each method makes decisions by the Area Under the Curve (AUC) for chosen utility functions to show that improved distributional estimates lead to improved decisions.

Precision in Estimation of Heterogeneous Effect (PEHE): We adopt the definition in Hill (2011). For unit i and the covariates x_i, we have the CATE and its estimates as $\tau(x_i) = E[Y(1)|X = x_i] - E[Y(0)|X = x_i]$ and $\hat{\tau}(x_i)$. PEHE measures their distance: $\text{PEHE} = \sqrt{\sum_{i=1}^{N}[(\hat{\tau}(x_i) - \tau(x_i)]^2/N}$.

Log Likelihood (LL): Log likelihood measures how well each method models potential outcome distributions. It is normally based on evaluating the PDF functions at the observed points. However, closed-form distributions are not directly available for sampling based approaches such as the variants of CCN and GANITE. Instead, we approximate the LL by calculating the probability on a neighborhood of the realized outcome y, $B_{y,\epsilon} = (y - \epsilon, y + \epsilon)$, where ϵ is a small positive value. $\text{LL} = \sum_{t=0}^{1}\sum_{i=1}^{N} \log(\hat{Pr}[Y_i(t) \in B_{y_i(t),\epsilon}|X_i = x_i])/2N$.

Asymptotically, the true distribution dominates in this evaluation, and this can be shown under the criterion of Kullback–Leibler divergence (Kullback & Leibler, 1951) as $N \to \infty$ and $\epsilon \to 0$. We define $\epsilon = 0.5$ for IHDP and $\epsilon = 0.2$ for EDU to adjust for the scale of the outcomes.

Area Under the Curve (AUC): As mentioned in 2.2, personalized decisions can be described by the quantity: $\tau(x_i, U_{0,i}, U_{1,i}) = \mathbb{E}_{\gamma\sim p(Y(1)|X=x_i)}[U_{0,i}(\gamma, x_i)] - \mathbb{E}_{\gamma\sim p(Y(0)|X=x_i)}[U_{1,i}(\gamma, x_i)]$. The optimal decision is based on the sign of this contrast $1_{\tau(x_i,U_{0,i},U_{1,i})}$, which is regarded as the true label. Then, using the estimated contrast $\hat{\tau}(x_i, U_{0,i}, U_{1,i})$ as the decision score, we can estimate the AUC by varying the decision threshold.

5.2 IHDP

The Infant Health and Development Program (IHDP) describes a randomized experiment and is converted to an observational study by removing a nonrandom portion from the treatment group. We use the response surface B in Hill (2011) for heterogeneous treatment effect. The study consists of 747 subjects (139 in the treated group) with 19 binary and 6 continuous variables ($x_i \in \mathbb{R}^{25}$). We utilize 100 replications of the data for out-of-sample evaluation by following the simulation process of Shalit et al. (2017).

The quantitative results are summarized in Table 1. Overall, CCN outperforms other competing methods in both mean and distribution metrics. The advantages of combining the adjustment strategy is evident as FCCN improves over CCN by a large, clear margin. It is worth noting that LL calculated under the ground truth model is -1.41, demonstrating that FCCN is highly effective in capturing the true distributions.

Table 1: Quantitative results on IHDP. Each metric's mean and standard error are reported. FCCN outperforms CCN in all metrics with statistical significance. *GANITE is only used for estimating the CATE as it is relatively challenging to optimize for this small dataset according to Yoon et al. (2018).

Metrics/Method	CCN	FCCN	GANITE	CEVAE	BART	GAMLSS	CF
PEHE	1.59 ± .16	**1.13 ± .14**	2.40 ± .40	2.60 ± .10	2.23 ± .33	3.00 ± .39	3.52 ± .57
LL	-1.78± .02	**-1.62± .02**	*	-2.82 ± .08	-1.99 ± .08	-2.34 ± .13	NA
AUC (Linear)	.925 ± .011	**.942 ± .010**	.723 ± .017	.523 ± .008	.923 ± .009	.930 ± .10	.896 ± .009
AUC (Threshold)	.913 ± .011	**.935 ± .010**	*	.564 ± .010	.917 ± .009	.925 ± .10	NA

We evaluate two sets of utility functions: a linear utility $U_0(\gamma) = \gamma, U_1(\gamma) = \gamma - 4$ and a non-linear utility with $U_{0,i}(\gamma, x_i) = 1_{\gamma>E[Y(0)|X=x_i]}$ and $U_{1,i}(\gamma, x_i) = 1_{\gamma>(E[Y(0)_i|X=x_i]+4)}$, as the ATE for surface B is 4 (Hill, 2011). Table 1 shows AUC (Linear) and AUC (Non-Linear) corresponding to two utilities, demonstrating that CCN's improved distribution estimates contribute to more accurate decisions, despite the fact that a homoskedastic Gaussian distribution is well matched to the specifications in BART and GAMLSS.

The evaluations of all these metrics can be regarded as a plug-in process for distribution estimators. We notice that PEHEs of these distribution methods may not be as competitive as some of the state-of-the-art CATE estimators (Shalit et al., 2017; Hassanpour & Greiner, 2020a), as optimizing the mean is simpler and has an exact match with this metric. The strategy of estimating distributions can flexibly be adapted to various comparisons with trained models.

Comparison to Policy Learning. We next compare the proposed approaches to a policy learning approach, specifically policytree (Sverdrup et al., 2021), with full details in Appendix F. Policy learning is limited to pre-specified utility functions, so we set up two scenarios: one with a linear utility ($U_0(\gamma) = \gamma$, $U_1(\gamma) = \gamma - 4$), and one with a threshold (binary) utility ($U_0(\gamma) = 1_{\gamma > E[Y(0)]}$, $U_1(\gamma) = 1_{\gamma > E[Y(1)]}$). Since policytree only outputs its predicted optimal treatment, we compare on accuracy (predicted vs true optimal treatment). On the IHDP dataset, FCCN performs well on both with 88.6% and 87.72% accuracy, respectively. However, policytree's accuracy drops from 76.9% to 57.6% when we switch to the threshold utility, signifying how much information is lost from the system by binarizing the outcomes.

5.3 EDU

The EDU dataset is based on a randomized field experiment in India between 2011 and 2012 (Banerji et al., 2017; 2019). The experiment studies whether providing a mother with adult education benefits their children's learning. We define the binary treatment as whether a mother receives adult education and the continuous outcome as the difference between the final and the baseline test scores. After the preprocessing described in Appendix B, the sample size is 8,627 with 18 continuous covariates and 14 binary covariates, $x_i \in \mathbb{R}^{32}$.

We create a semi-synthetic case over the two potential outcomes by the following procedures. We first train two neural networks, $f_{\hat{y}_0}(\cdot), f_{\hat{y}_1}(\cdot)$, on the observed outcomes for the control and treatment groups. The uncertainty model for the control and treatment group are based on a Gaussian distribution and an exponential distribution, respectfully, which helps showcase that CCN and FCCN can automatically adapt to different distribution families. We represent s_i as an indicator of whether the mother has received any previous education, as we hypothesize the variability is higher for the mothers not educated previously. Then the potential outcomes are synthesized as,

$$Y(0)_i|x_i \sim f_{\hat{y}_0}(x_i) + (2 - s_i)N(0, .5^2); \quad Y(1)_i|x_i \sim f_{\hat{y}_1}(x_i) + (2 - s_i)Exp(2).$$

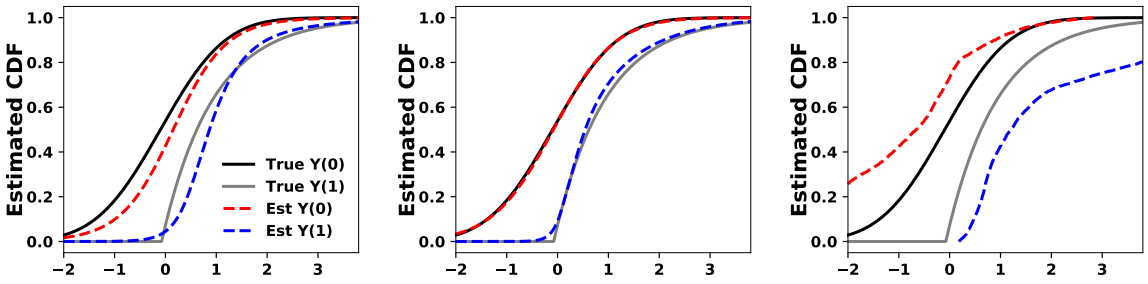
The treatment group imbalance comes from two aspects. One is from a treatment assignment model with propensity $Pr(T_i = 1|x_i) = 1/[1 + exp(-x_i^T\beta)]$ where we assign large coefficients in β to add imbalance. The other is from truncation, as we remove well-balanced subjects with estimated propensities in the range of $0.3 < Pr(T_i = 1|x_i) < 0.7$. We keep 1,000 samples for evaluation and use the rest for training. The full procedure is repeated 10 times for variability assessment.

The utility function is customized for each subject to mimic personalized decisions. For subject i, $U_{0,i}(\gamma, x_i) = I(\gamma > v_i)$, and $U_{1,i}(\gamma, x_i) = I(\gamma > v_i + 1 - s_i)$ where $v_i \sim U(0, 1.5)$. The interpretation of this utility is that different mothers have different expectations of their children's improvements with threshold v_i. For the mothers without previous education, their expectations are higher by 1. This design coincides with the expectation that the education should have a positive effect on outcomes in exchange for a finite cost, and we would only invest in the intervention for a positive return. Results from this experiment are summarized in Table 2. Both CCN-based methods flexibly model different distributions, with FCCN providing clear improvement over CCN.

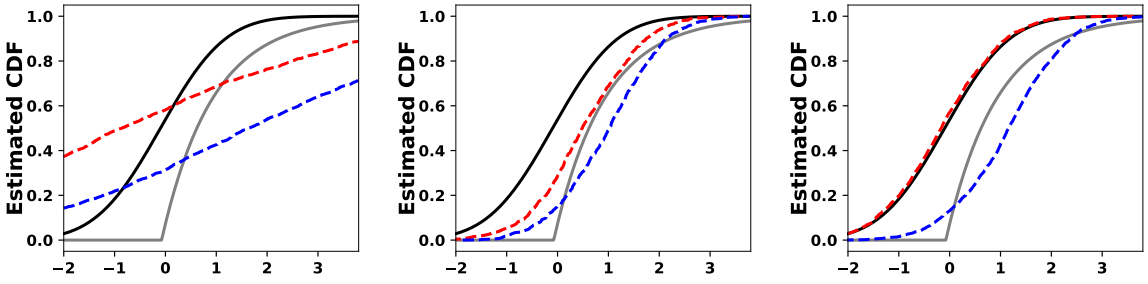
Table 2: Quantitative results on the EDU dataset. FCCN outperforms CCN in all metrics with statistical significance.

Metrics/Method	CCN	FCCN	GANITE	CEVAE	BART	GAMLSS	CF
PEHE	.392 ± .049	**.296 ± .042**	1.253 ± .181	1.911 ± .351	.534 ± .042	.314 ± .053	1.022 ± .051
LL	-2.178 ± .024	**-2.125± .022**	-5.092 ± .596	-3.558 ± .055	-2.443 ± .063	-2.250 ± .025	NA
AUC	.933 ± .026	**.953 ± .014**	.760 ± .053	.622 ± .039	.906 ± .015	.941 ± .010	NA

Making an optimal decision is highly dependent on how close a method's estimated distribution aligns with the true values and all relevant heterogeneity. Thus, the AUCs follow their respective LLs with the CCN based methods performing better. CEVAE has two facets of misspecification that hurt performance: (*i*) its homogeneous Gaussian error, whereas the real outcomes come from heteroskedastic Gaussian or exponential distributions, and (*ii*) it decodes the continuous covariates into a Gaussian distribution. Hence, CEVAE captures the marginal distributions well but does not provide helpful personalized suggestions.

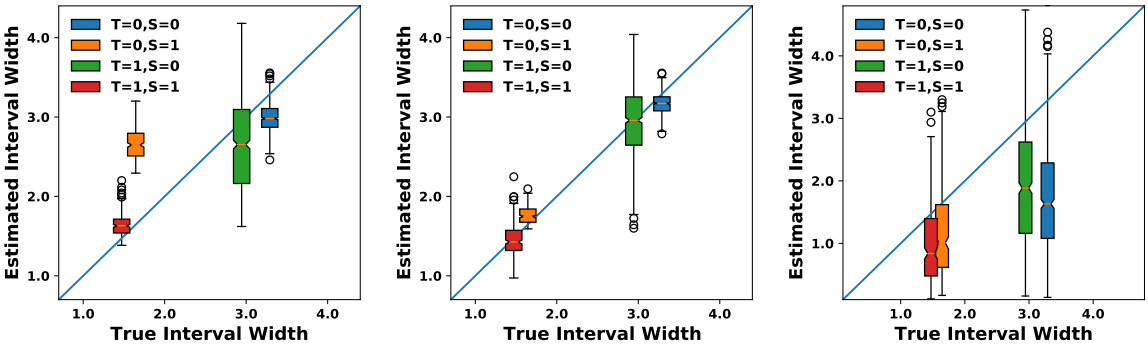


(a) CCN (b) FCCN (c) GANITE



(d) CEVAE (e) BART (f) GAMLSS

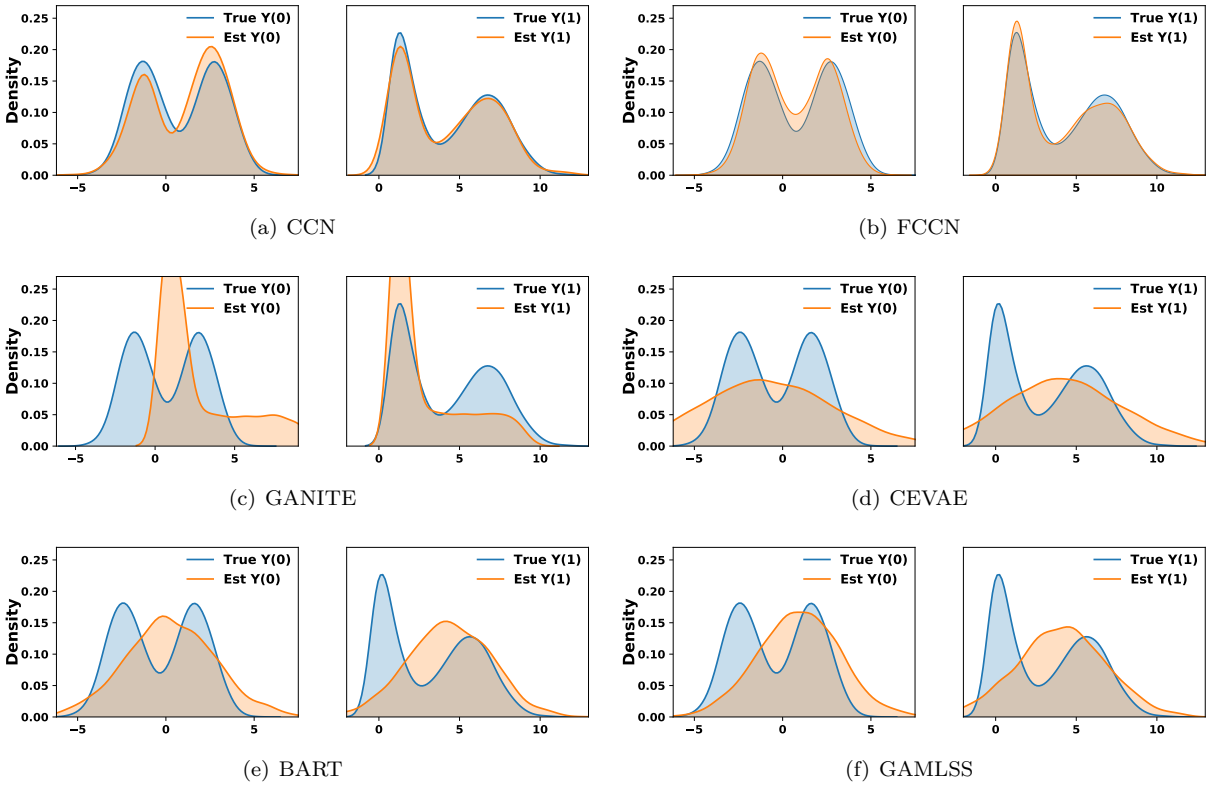
Figure 2: Visualization of a random sample of EDU. FCCN closely follows the theoretical curve and the rest diverge.



(a) CCN (b) FCCN (c) GANITE

Figure 3: The estimated versus true 90% interval widths given the four combinations of T and S. GANITE reflects the main trend of how uncertainties change with T and S. FCCN can clearly discern the four scenarios by aligning its estimated interval widths and the true interval widths. Additional visualizations are in Appendix J.

Next, we randomly draw a data sample and compare the estimated CDFs against the true CDFs in Figure 2 (see Figure S1 for additional samples). We find that CCN-based approaches are capable of faithfully recovering the true conditional CDFs, whereas the other methods have gaps in their estimation. GAMLSS is accurate on the control group but not the treatment group. This is partly due to gamlss package (Stasinopoulos et al., 2021) not supporting the exponential distribution with location shift, so skewed normal is chosen as the closest reasonable substitute. Overall, GAMLSS is flexible but requires precise specification on a case-by-case basis, whereas CCN can robustly use the same approach. In our experiments for GAMLSS, we must choose very close distributions and limit the uncertainty to the relevant variables or the package does not converge.



(a) CCN (b) FCCN

(c) GANITE (d) CEVAE

(e) BART (f) GAMLSS

Figure 4: Visualization of each method's estimated density of the potential outcomes. Variants of CCN outperform other methods with a clear margin. GANITE is aware of certain mixture here. As its objective makes a trade-off between GAN loss and supervised loss, it is not guaranteed to approximate the true distributions.

Lastly, we assess whether the heteroskedasticity of the outcomes is captured. The combination of $S = 0, 1$ and $T = 0, 1$ produces four uncertainty models. We visualize the predictive 90% interval widths in Figure 3. FCCN captures the bimodal nature of the interval widths. In contrast, GANITE only captures a small fraction of the difference between the low and high variance cases. Both CEVAE (Louizos et al., 2017) and BART (Hill, 2011) fail to capture the heteroskedasticity and are not shown. For GAMLSS, we explicitly feed its uncertainty model with only T and S for it to converge effectively. It does not produce variability in interval widths. In summary, the true interval widths for four combinations are 1.47, 1.64, 2.94 and 3.29, while GAMLSS reports 0.92, 1.52, 3.37 and 3.37. Overall, CCN and its variants produce higher quality ranges.

5.4 Additional Comparisons and Properties

In this section, we include several additional experiments to reflect other properties of CCN and its adjustment to demonstrate their advantages in estimating the potential outcome distributions.

5.4.1 Estimating Multimodal Distributions

A fundamental reason that we choose to extend CN to estimate potential outcome distributions is its adaptability to different outcome families. We demonstrate this with an example from a multi-modal distribution. Many of other methods, including GAMLSS, BART, and CEVAE, cannot capture it with their existing structure, whereas CCN, FCCN, and methods like GANITE can. To succinctly summarize this experiment, only CCN and FCCN naturally adjust to the multi-modal space, as shown briefly in Figure

4. GANITE is aware of the mixtures but does weight them well. The other algorithms do not capture the mixture model. The data generating process is described in Appendix H.

5.4.2 Additional Distribution Tests

Next, we sketch out how different methods work on a variety of outcome distributions, including Gumbel, Gamma, and Weibull distributions, with full details in Appendix G. The results in Table 3 are qualitatively similar to the previously presented semi-synthetic cases, where CCN straightforwardly adapts to these distributions and FCCN provides additional improvements. In fact, FCCN even slightly outperforms GAMLSS even when *GAMLSS is provided the true outcome distribution*. When GAMLSS is given a flexible but not perfectly matched outcome distribution, it does not come close to the CCN approach.

Table 3: The estimated LL under different simulated distributions.[1] and [2] represent fitting GAMLSS with the true family and heterosekdastic Gaussian, respectively.

	True Value	CCN	FCCN	BART	GAMLSS[1]	GAMLSS[2]
Gumbel	-2.87	-3.67 ± .02	**-3.56 ± .02**	-3.92 ± .06	-3.67 ± .02	-3.90 ± .05
Gamma	-3.17	-3.83 ± .04	**-3.74 ± .06**	-3.97 ± .02	-3.77 ± .02	-3.95 ± .02
Weibull	-2.87	-3.41 ± .02	**-3.32 ± .03**	-3.86 ± .12	**-3.32 ± .04**	-3.71 ± .09

5.4.3 Sample Size and Convergence

As suggested in Proposition 2, CCN can asymptotically estimate the optimal value given large sample size. We create an example with the logistic distribution to visualize it. The data generating process is described in Appendix I.

We vary the input data size and compare all methods on log-likelihood in Figure 5. We note that CCN and its variants can all approach the optimal value. All adjusted versions of CCN present faster convergence rates with FCCN dominating the curve. Gaussian methods (CEVAE, BART) provide more stable approximations in smaller samples. Due to distributional mismatch, the optimal value can not be attained by those methods. Though GAMLSS has the correct family specification, it is restrained by the flexibility of the additive model to approach the optimal value. GANITE progresses slower and needs over 15,000 samples to generate competitive results.

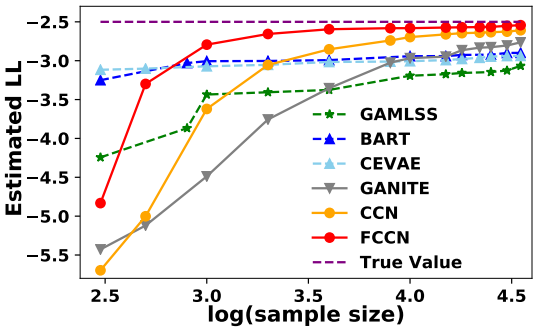
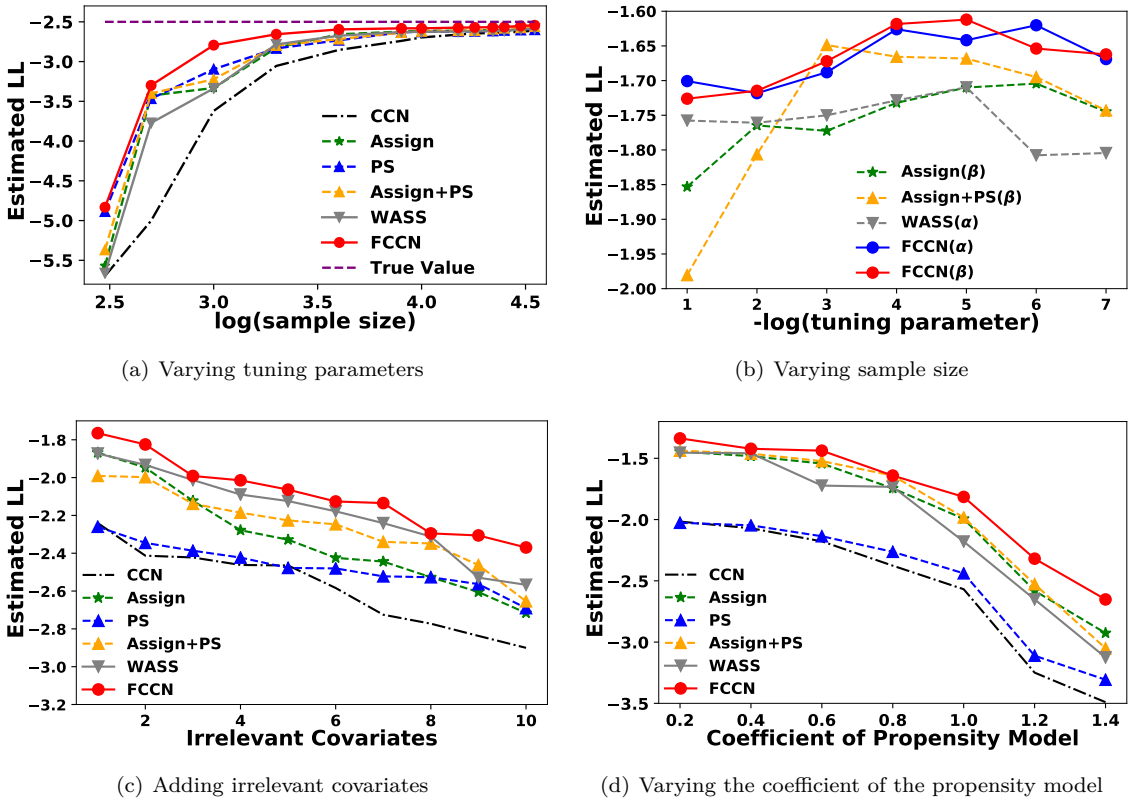


Figure 5: Sample size and convergence rate.

5.4.4 Ablation Study

We include three components to account for the treatment group imbalance in our adjustment scheme. They are the Assign-loss (Assignment), the Wass-loss (Wass), and the propensity stratification (PS). Below, we inspect how CCN empirically benefits from each component. Additional details and visualizations can be found in Appendix J.

Table 4 and 5 summarize the results on the evaluating metrics in both the IHDP and EDU datasets. Overall, variants of CCN with adjustment more accurately estimate the potential outcomes. However, the aspects on how these components contribute vary by their attributes. The propensity score stratification mainly facilitates information sharing between strata (Lunceford & Davidian, 2004), hence it excels in the IHDP dataset where the imbalance is caused by removing a specific subset from the treatment group. However, on the conditional level, the stratification can be regarded as a form of aggregation, which might hinder the precision. Hence, we do not see much gain in distribution based metrics or personalized decisions by solely including the propensity. The Assign-loss overcomes the confounding effect by extracting representation relevant to confounding effects, and we observe that it effectively increases the model performance in all metrics. The combination of Assign-loss and propensity stratification displays the merits of these two approaches. The



(a) Varying tuning parameters

(b) Varying sample size

(c) Adding irrelevant covariates

(d) Varying the coefficient of the propensity model

Figure 6: The estimated LL in different scenarios. 6(a) demonstrates the differences in convergence speed of CCN and its variants with increased sample size. 6(b) depicts the performance of CCN's variants given different values of tuning parameters. 6(c) and 6(d) exemplify that adding noises or treatment group imbalance worsens the model performance. Under different circumstances, FCCN remains the most robust.

Wass-loss finds a representation that balances the treatment and control groups and improves both point and distribution estimates in the two datasets. Nevertheless, it does not account for the domain specific information (Shi et al., 2019). FCCN wins in all cases by a clear margin.

Table 4: Quantitative results on the IHDP dataset regarding different variants of CCN.

Metrics/Method	CCN	Wass	Assign	PS	Assign+PS	FCCN
PEHE	1.59 ± .16	1.32 ± .17	1.42 ± .25	1.15 ± .10	1.22 ± .15	**1.13 ± .14**
LL	-1.78 ± .02	-1.65± .02	-1.64± .03	-1.67 ± .15	-1.65 ± .02	**-1.62 ± .02**
AUC (Linear)	.925 ± .011	.938 ± .010	.940 ± .010	.918 ± .012	.940 ± .010	**.942 ± .010**
AUC (Threshold)	.913 ± .011	.932 ± .011	.932 ± .011	.911 ± .012	.934 ± .011	**.935 ± .010**

Table 5: Quantitative results on the EDU dataset regarding different variants of CCN.

Metrics/Method	CCN	Wass	Assign	PS	Assign+PS	FCCN
PEHE	.392 ± .049	.324 ± .046	.343 ± .041	.400 ± .052	.339 ± .039	**.296 ± .042**
LL	-2.178 ± .024	-2.128± .020	-2.132± .023	-2.171 ± .024	-2.129 ± .029	**-2.125 ± .022**
AUC	.933 ± .026	.951 ± .013	.946 ± .018	.932 ± .022	.952 ± .019	**.953 ± .014**

Moreover, we study different adjustment components in different scenarios in Figure 6. Figure 6(a) shows that CCN and its variants all asymptotically approach the optimal value with FCCN being the quickest. Figure 6(b) is made by varying the value of α or β. It suggests that too large or small tuning parameters are more likely to hurt the models with only single adjustment component, while FCCN is more robust against it.

Figure 6(c) describes a case where irrelevant dimensions with standard Gaussian distributions are added to the covariate space. We observe that adding noise worsens the performance. The Assign-loss in this case is more likely to overfit the propensity model with extra covariates containing noises only. Hence, adjustment through WASS-loss is preferable. In Figure 6(d), we vary the propensity model by changing its coefficient. Larger values represent less balanced space and larger confounding effects. In this setup, the Assign-loss is slightly better when the imbalance is more extreme, as a balanced representation becomes more challenging to obtain. Among them, FCCN is the most robust due to simultaneously considering the domain invariant and specific information.

6 Discussion

Here, we propose Collaborating Causal Networks (CCN): a novel framework to estimate conditional potential outcome *distributions*, backed by theoretical proofs and paired with an adjustment method that rectifies treatment group imbalance. Our experiments demonstrate that CCN automatically adapts to a variety of outcomes, including exponential family distributions and multi-modal distributions. Empirically, we show that CCN is effective in inferring *full potential outcomes*. Additionally, incorporation of our proposed adjustment technique in FCCN is relatively robust with regard to treatment group imbalance in semi-synthetic and synthetic experiments. We note that improving distribution estimates leads to improved decision-making even without *a priori* access to utility functions by comparing to policy learning. In all of our evaluations, FCCN meets or exceeds the current state-of-the-art methodology for potential outcomes distribution estimation, and asymptotically approaches our theoretical claims.

References

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

Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. Deep counterfactual networks with propensity-dropout. *Proceedings of International Conference on Machine Learning*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. *International Conference on Machine Learning*, 2017.

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. *Artificial Intelligence and Statistics*, 2021.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Rukmini Banerji, James Berry, and Marc Shotland. The impact of maternal literacy and participation programs: Evidence from a randomized evaluation in india. *American Economic Journal: Applied Economics*, 9(4):303–37, 2017.

Rukmini Banerji, James Berry, and Marc Shotland. *The impact of mother literacy and participation programs on child learning: evidence from a randomized evaluation in India*, 2019. URL `https://doi.org/10.7910/DVN/WEOLSW`.

Normand J Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *Quantum Information & Computation*, 12(5-6):432–441, 2012.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010.

Dimitris Bertsimas, Nathan Kallus, Alexander M Weinstein, and Ying Daisy Zhuo. Personalized diabetes management using electronic medical records. *Diabetes Care*, 40(2):210–217, 2017.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 129–138, 2009.

Peter J Bickel and Ya'acov Ritov. Nonparametric estimators which can be "plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003.

Guillermo Briseño Sanchez, Maike Hohberg, Andreas Groll, and Thomas Kneib. Flexible instrumental variable distributional regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4): 1553–1574, 2020.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, pp. 1–44, 2021.

Hugh Chipman and Robert McCulloch. *BayesTree: Bayesian Additive Regression Trees*, 2016. URL `https://CRAN.R-project.org/package=BayesTree`. R package version 1.4.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayesian ensemble learning. In *Advances in Neural Information Processing Systems*, pp. 265–272, 2007.

Hugh A Chipman, Edward I George, and Robert E McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

Rajeev H Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125(1-2):141–173, 2005.

Peng Ding and Fan Li. Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237, 2018.

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.

Qiyang Ge, Xuelin Huang, Shenying Fang, Shicheng Guo, Yuanyuan Liu, Wei Lin, and Momiao Xiong. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in Genetics*, 11, 2020.

Steffen Grunewalder. Plug-in estimators for conditional expectations and probabilities. In *International Conference on Artificial Intelligence and Statistics*, pp. 1513–1521, 2018.

P Richard Hahn, Jared S Murray, and Carlos Carvalho. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Jun Han and Claudio Moraga. The influence of the sigmoid function parameters on the speed of back-propagation learning. In *International Workshop on Artificial Neural Networks*, pp. 195–201. Springer, 1995.

Negar Hassanpour and Russell Greiner. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020a.

Negar Hassanpour and Russell Greiner. Variational auto-encoder architectures that excel at causal inference. *Advances in Neural Information Processing Systems*, 2020b.

Miguel A Hernan and James M Robins. *Causal inference*. Boca Raton: Chapman & Hall/CRC, 2020.

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

Maike Hohberg, Peter Pütz, and Thomas Kneib. Treatment effects beyond the mean using distributional regression: Methods and guidance. *PloS one*, 15(2):e0226514, 2020.

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

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33, 2020.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

James M Joyce. *The foundations of causal decision theory*. Cambridge University Press, 1999.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *Advances in Neural Information Processing Systems*, 31, 2018.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

S Kullback and RA Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

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

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *arXiv preprint arXiv:2006.06138*, 2020.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. *Advances in Neural Information Processing Systems*, pp. 929–939, 2017.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

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

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960, 2004.

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. *CoRR*, abs/2102.08208, 2021. URL `https://arxiv.org/abs/2102.08208`.

Joost ME Pennings and Ale Smidts. The shape of utility functions and organizational behavior. *Management Science*, 49(9):1251–1263, 2003.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of Statistics*, 39(2):1180, 2011.

R. A. Rigby and D. M. Stasinopoulos. Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554, 2005.

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.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pp. 159–183, 1973.

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085, 2017.

Shu Shen. Estimation and inference of distributional partial effects: theory and application. *Journal of Business & Economic Statistics*, 37(1):54–66, 2019.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, pp. 2507–2517, 2019.

Houshang H Sohrab. *Basic Real Analysis*, volume 231. Springer, 2003.

Mikis Stasinopoulos, Bob Rigby, Vlasios Voudouris, Calliope Akantziliotou, Marco Enea, and Daniil Kiose. *gamlss: Generalised Additive Models for Location Scale and Shape*, 2021. URL `https://CRAN.R-project.org/package=gamlss`. R package version 5.3-4.

Erik Sverdrup, Ayush Kanodia, Zhengyuan Zhou, Susan Athey, and Stefan Wager. *policytree: Policy Learning via Doubly Robust Empirical Welfare Maximization over Trees*, 2021. URL `https://CRAN.R-project.org/package=policytree`. R package version 1.1.1.

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

SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

Brian G Vegetabile. On the distinction between" conditional average treatment effects"(cate) and" individual treatment effects"(ite) under ignorability assumptions. *ICML 2021 Workshop*, 2021.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

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

Stephen G Walker. Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, 143(10):1621–1633, 2013.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 31, 2018.

A Yazdani and E Boerwinkle. Causal inference in the age of decision medicine. *Journal of Data Mining in Genomics & Proteomics*, 6(1), 2015.

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

Guoyi Zhang and Yan Lu. Bias-corrected random forests in regression. *Journal of Applied Statistics*, 39(1): 151–160, 2012.

Tianhui Zhou, Yitong Li, Yuan Wu, and David Carlson. Estimating uncertainty intervals from collaborating networks. *Journal of Machine Learning Research*, 22(257):1–47, 2021. URL `http://jmlr.org/papers/v22/20-1100.html`.

A Proofs of Propositions 1 and 2

Zhou et al. (2021) assumes that the covariate distributions are the same for training and generalization for CN. In observational studies, the training spaces $p(x|T=1)$ and $p(x|T=0)$ differ from the evaluation space $p(x)$. Hence, the central challenge of migrating CN's properties to CCN is to show its robustness to the covariate space mismatch.

First, we explore the properties of CN and CCN under the presence of covariate space mismatch. Second, we expand on how CCN with standard causal assumptions can overcome the covariate space mismatch in causal setups.

To make this section self complete, we restate the two propositions from the main article. Note that they are both claimed on the full covariate space, $\forall x$ such that $p(x) > 0$.

Proposition 1 (Optimal solution for g_0 and g_1). When the support of the outcomes is a subset of the support of z, or as z covers the whole outcome space, the functions g_0 and g_1 that minimize g-loss* are optimal when they are equivalent to the conditional CDF of $Y(0)|X=x$ and $Y(1)|X=x$, $\forall x$ such that $p(x) > 0$.

Proposition 2 (Consistency of g_0 and g_1). Assume the ground truth CDF functions for $T \in \{0,1\}$ is Lipschitz continuous with respect to both feature X and the potential outcomes, and that the support of the outcomes is a subset of the support of z. Denote the ground truth as g_0^* and g_1^*. As $n \to \infty$, the finite sample estimators g_0^n and g_1^n have the following consistency property: $d(g_0^n, g_0^*) \to_P 0; d(g_1^n, g_1^*) \to_P 0$ under some metric d such as $\mathbb{L}_1$ norm and with the space searching tool z being able to cover the full outcome space.

A.1 Claims from Zhou et al. (2021) with and without Space Mismatch

We first restate two similar propositions in CN under the non-causal setting where there is no space mismatch between the training and the evaluation spaces.

Proposition S1 (Optimal solution for g from Zhou et al. (2021))**.** *Assume that $f(q,x)$ approximates the conditional q^{th} quantile of $Y|X=x$ (inverse CDF, not necessarily perfect). If $f(q,x)$ spans $\mathbb{R}^1$, then a g minimizing equation 1 is optimal when it is equivalent to the conditional CDF, or $Y|X=x \sim g(Y,x), \forall x$ such that $p(x) > 0$.*

Proposition S2 (Consistency of g from Zhou et al. (2021))**.** *Assume the true CDF function g^* satisfies Lipschitz continuity with respect to x and outcome y. As $n \to \infty$, the finite sample estimator g^n has the following consistency property: $d(g^n, g^*) \to_P 0$ under some metric d such as $\mathbb{L}_1$ norm and with f capable of searching the full outcome space.*

Proposition S1 demonstrates that a fixed point solution estimates the correct distributions. Proposition S2 states that the optimal learned function asymptotically estimates the true distributions.

However, they do not answer whether these properties hold on a new space that differs from the original training space. We address this existing limitation by developing Proposition S3 which shows that these properties can still be retained given a certain type of space mismatch. For generality, we define the training space as $p(x)$ and the new space as $p'(x)$.

Proposition S3 (The dependency of CN on $p(x)$)**.** *If the conditional outcome distribution $p(Y|X=x)$ remains invariant between $p(x)$ and $p'(x)$ (covariate shift), and $p(x) > 0 \implies p'(x) > 0$, the solutions in Propositions S1 and the consistency in Proposition S2 also generalize to the new space where $p'(x) > 0$.*

Proof of Proposition S3. With the Propositions S1 and S2, we observe that g estimates the conditional distribution of $Y|X=x$ in the space where $p(x) > 0$. Next we generalize it to a new space $p'(x)$. Given the condition $p(x) > 0 \implies p'(x) > 0$, for any x in evaluation space with $p'(x) > 0$, it is covered in the training space where $p(x) > 0$. From Proposition S1 and S2, we know that for such x, the optimum can be obtained. The covariate shift assumption on the invariance of outcome distributions then guarantees that the optimum of such x in the training space is also the optimum in the new space. Therefore, each point x in the new space with $p'(x) > 0$ can obtain their optimum through training the model on the training space $p(x)$. We claim that the optimum remains invariant when we generalize from $p(x)$ to $p'(x)$. □

A.2 Overcoming Covariate Space Mismatch for CCN

Proposition S3 enables us to extend CN's optimum to new spaces given two conditions: the *covariate shift* and the *space overlap* $p(x) > 0 \implies p'(x) > 0$. Our main task is to show how they hold in the causal settings.

First, we give a weaker version of Propositions 1 and 2 as a direct result from Proposition S1 and S2 without accounting for the mismatch between the training and generalization spaces.

Claim S1 (Potential outcome distributions on each treatment space)**.** *Given all the conditions in Proposition S1 and S2, an optimal solution for g_0 and g_1 in g-loss* in equation 4 are the true CDFs of $Y(0)|X = x, T = 0,\ \forall x$, such that $p(x|T = 0) > 0$; and $Y(1)|X = x, T = 1,\ \forall x$, such that $p(x|T = 1) > 0$. The finite sample estimators g_0^n and g_1^n can also consistently estimate these CDFs.*

Discussion on Claim S1. We only discuss the first part of Claim S1 for $T = 0$ without loss of generality, while optimizing g_0 only involves updating parameters using batches with $t = 0$. The other group can be shown with identical steps.

A direct conclusion from Propositions S1 and S2 is that the optimal g_0 is the fixed point solution, and finite sample estimator g_0^n consistently estimates the true CDF of $Y|X = x, T = 0,\ \forall x$, such that $p(x|T = 0) > 0$. This is claimed for the full conditional distribution $Y|X = x, T = 0$ rather than for the potential outcome $Y(0)$. By virtue of the treatment consistency (Assumption 2), the following two outcomes are identically distributed: $Y|X = x, T = 0 \iff Y(0)|X = x, T = 0$. Therefore, we can successfully establish the estimators for potential outcome distributions, but currently limited to each treatment subspace. □

As mentioned above, we need two conditions to generalize Claim S1 back to the full space with density $p(x)$. The covariate shift has been explicitly described in (Johansson et al., 2020), and can be induced by Assumption 3.

Lemma S1 (Covariate Shift)**.** *Both potential outcome distributions $p(Y(0)|X = x)$ and $p(Y(1)|X = x)$ are independent of the covariate space in which they are located.*

Discussion on Covariate Shift. The ignorability states that $[Y(0), Y(1)] \perp T|X$, which implies that

$$P(Y(0), Y(1)|X, T = 0) = P(Y(0), Y(1)|X, T = 1) = P(Y(0), Y(1)|X).$$

This supports that the potential outcome distributions are invariant to the treatment groups. □

We next show the condition of the space overlap.

Lemma S2 (Positivity relating to the space overlap)**.** *Under Assumption 1, the equivalent condition holds: $p(x) > 0 \iff p(x|T = 0) > 0$, and $p(x) > 0 \iff p(x|T = 1) > 0$.*

Proof of Lemma S2. The positivity in Assumption 1 claims that $\forall x, 0 < Pr(T = 1|x) < 1$. Then for each x where $p(x) > 0$, we can find a constant $1 > C_x > 0$ that satisfies $Pr(T = 1|x) > C_x$.

By Bayes rule, $p(x|T = 1) = Pr(T = 1|x)p(x)/Pr(T = 1) > C_x p(x) > 0$. Then $p(x) > 0 \implies p(x|T = 1) > 0$. From the other direction, if $p(x|T = 1) = Pr(T = 1|x)p(x)/Pr(T = 1) > 0$, each component on the right hand side needs to be positive. Therefore, $p(x|T = 1) > 0 \implies p(x) > 0$. We can use the same procedure to verify the condition for $T = 0$. □

With Lemma S2 and S2 satisfying the conditions in Proposition S3, we acquire that the space mismatch given the standard causal assumptions is the specific type that does not impact the large sample properties of CCN. As a result, Propositions 1 and 2 naturally follow. Thus, under the standard assumptions in causal inference, CCN will capture the full potential outcome distributions. This procedure is not limited to a binary treatment condition and is extendable to multiple treatment setups.

B Semi-synthetic Data Generation

B.1 IHDP

We focus on making our simulation results comparable to other causal methods' published results. The simulation replications for the IHDP data are downloaded directly from `https://github.com/clinicalml/cfrnet`, which are used to generate the results of WASS-CFR (Shalit et al., 2017) and CEVAE (Louizos et al., 2017). The dataset does not contain personally identifiable information or offensive content.

B.2 Education Data

The raw education data are downloaded from the Harvard Dataverse[1], which consist of 33,167 observations and 378 variables. The dataset does not contain personally identifiable information or offensive content. We pre-process the data such as by combining repetitive information, deleting covariates with over 2,5000 missing values. Then we end up with a clean dataset containing 8,627 observations.

The function $f_{\hat{y_1}}(\cdot)$ and $f_{\hat{y_0}}(\cdot)$ are learned from the observed outcomes for the treatment and control groups. They are both designed as single-hidden-layer neural networks with 32 units and sigmoid activation functions (Han & Moraga, 1995). A logistic regression model with coefficient $\beta = [\beta_1, ..., \beta_{28}]$ and propensity score $Pr(T = 1|X = x_i) = \frac{1}{1+exp(-x_i'\beta)}$ are used to generate treatment labels and mimic a observational study. The coefficients are randomly generated as $\beta_i \sim U(-0.8, 0.8)$.

C Detailed Method Implementations

All Python-based methods: CCN, CEVAE and GANITE are run on a single NVIDIA P100 GPU; the R-based methods, BART, CF, and GAMLSS are run on a Intel(R) Xeon(R) Gold 6154 CPU.

CCN and FCCN

The implementation of CCN and all its variants are based on the code base for CN Zhou et al. (2021), which is provided at `https://github.com/thuizhou/Collaborating-Networks` with the MIT license. Functions g_0 and g_1 follow the structures of g in Zhou et al. (2021). We implement the full collaborating structure and find the optimization to be harder when added with regularization terms. Therefore, we fix f as a uniform distribution covering the range of the observed outcomes. It is called the g-only in Zhou et al. (2021), and is easier to optimize with only marginal loss in accuracy.

In FCCN, we introduce two latent representation $\phi_A(\cdot)$ and $\phi_W(\cdot)$. We set their dimensions to 25. They are both parametrized through a neural network with a single hidden layer of 100 units. The Wasserstein distance in Wass-loss is learned through $D(\cdot)$ which is a network with two hidden layers of 100 and 60 units per layer. We adopt the weight clipping strategy with threshold: (-0.01,0.01) to maintain its Lipschitz constraint (Arjovsky et al., 2017). The hyper-parameter α and β are tuned for FCCN according to the log-likelihood calculated upon the observed outcomes. We propose a few candidate values for α and β as: 5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5, as we do not want these values to be too large to overtake the main part of the loss that learns the distribution. Then we do grid search to determine the hyper-parameters. Based on the results on the first few simulations, we fix α=5e-4 and β=1e-5 in IDHP and α=1e-5 and β=5e-3 in EDU. We find this specification to consistently improve the performance over regular CCN.

To assess the potential outcome distributions, and take expectation over a defined utility function, we draw 3,000 samples for each test data point with the learned g_0 and g_1.

The code for CCN and its adjustment will be public on Github with the MIT license when the manuscript is accepted.

BART (Chipman et al., 2010) The implementation of BART uses the R package BayesTree (Chipman & McCulloch, 2016) with GPL (>= 2) license. We use the default setups in its model structure. Chipman et al. (2007) suggest that BART's performance with default prior is already highly competitive and is not

[1] `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/19PPE7`

highly dependent on fine tuning. We set the burn-in iteration to 1,000. We draw 1,000 random samples per individual to access their posterior predicted distributions, as the package stores all the chain information and is not scalable for large data.

CEVAE (Louizos et al., 2017)

The CEVAE is implemented with the publicly available code from `https://github.com/rik-helwegen/CEVAE_pytorch/`. We follow its default structure in defining encoders and decoders. The latent confounder size is 20. The optimizer is based on ADAM with weight decay according to Louizos et al. (2017). We use their recommended learning rate and decay rate in IHDP. In EDU dataset, the learning rate is set to 1e-4, and the decay rate to 1e-3 after tuning. We draw 3,000 posterior samples to access the posterior distributions.

GANITE (Yoon et al., 2018) The implementation of GANITE is based on `https://github.com/jsyoon0823/GANITE` with the MIT license. The model consists of two GANs: one for imputing missing outcomes (counterfactual block) and one for generating the potential outcome distributions (ITE block). Within each block, they have a supervised loss on the observed outcomes to augment the mean estimation of the potential outcomes. We use the recommended specifications in Yoon et al. (2018) to train the IHDP data. In the EDU dataset, the hyper-parameters for the supervised loss are set to α =2 (counterfactual block) and β =1e-3 (ITE block) after tuning.

CF (Wager & Athey, 2018) The implementation of CF uses the R package grf Tibshirani et al. (2020) with GPL-3 license. We specify the argument `tune.parameters="all"` so that all the hyper-parameters are automatically tuned.

GAMLSS (Hohberg et al., 2020) The implementation of GAMLSS uses the R package gamlss Rigby & Stasinopoulos (2005) with GPL-3 license. Since the method uses likelihood to estimate its parameters and often does not converge under complex models, we feed its location, scale and shape models with relevant variables only. In location models, we fit all continuous variables with penalized splines. In scale and shape models, we use relevant variables in their linear forms. The choice is based on a balanced consideration of the representation power and model's stability.

D Enforcing a Monotonicity Constraint

The learned CCN should have a monotonic property that $g(x, z+\epsilon) \geq g(x, z),\ \forall \epsilon \geq 0$. In our experiments, this condition is learned with a standard neural network architecture with our training scheme. We do not see any non-trivial violations of this requirement.

If required, though, this scheme can be enforced by modifying the neural network structure. One way of accomplishing this goal is to use a neural network with the form,

$$g(x, z) = \sum_{j=1}^{J} \text{softmax}(g_x^w(x))_j \sigma(g_x^b(x)_j + \exp(g_x^a(x)_j) z).$$

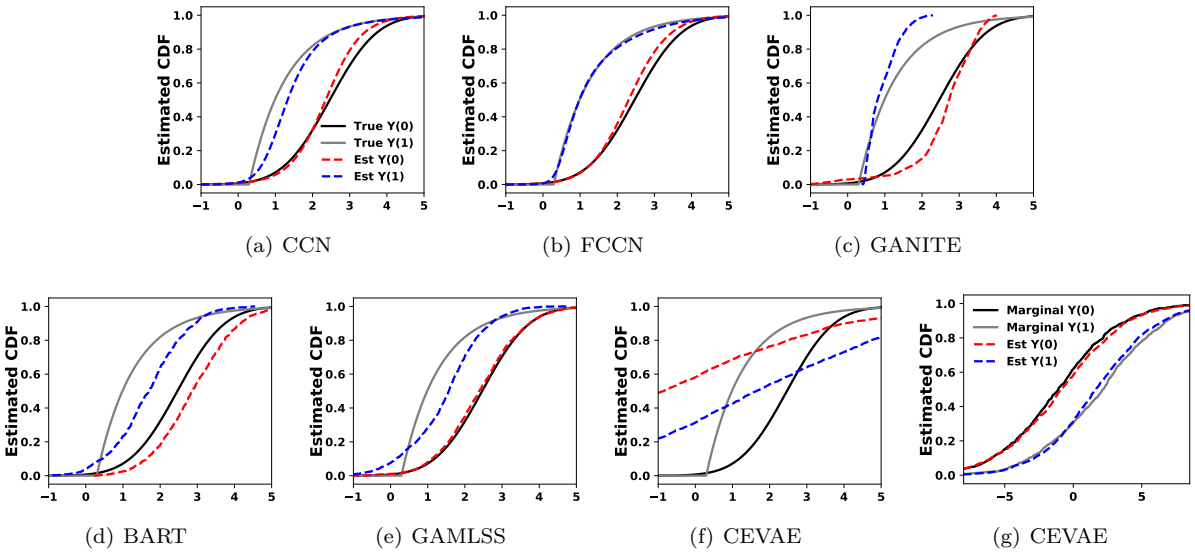
Here, $\sigma(\cdot)$ represents the sigmoid function. g_x^w, g_x^b, and g_x^a are all neural networks that map from the input space to a J-dimensional vector, $\mathbb{R}^p \to \mathbb{R}^J$. In this case, the formulation of the outcome is still highly flexible but becomes an admixture of sigmoid functions. As the multiplier on z is required to be positive, each individual sigmoid function is monotonically increasing as a function of z. Because the weight on each sigmoid is positive, this creates a full monotonic function as a function of z.

We implement this structure and find that it is competitive with a more standard architecture but is more difficult to optimize. As it is not empirically necessary to implement this strategy, we prefer the standard architecture in our implementations.

E Additional CDF Visualizations

We provide additional visualizations for the estimated potential outcome distributions with each method in Figure S1 based on another data point, which agrees with the results visualized in Figure 2. The two

variants of CCN are capable of capturing the main shape of the true CDF curves, including the asymmetry of the exponential distribution, with higher fidelity. GAMLSS is less accurate in the treatment group due to using skewed normal for the exponential distribution with location shifts. GANITE's two-GAN structures are highly reliant on data richness for accurate predictions (Yoon et al., 2018), so it falls short in cases with greater treatment group imbalance. CEVAE captures the overall marginal distribution for the potential outcomes as shown in Figure 1(g), but fails to discern the heterogeneity in conditional distributions in figure S1(f). BART provides reasonable estimates but struggles with the misspecification by its Gaussian form.



(a) CCN (b) FCCN (c) GANITE

(d) BART (e) GAMLSS (f) CEVAE (g) CEVAE

Figure S1: Visualization of each method's estimated CDFs. The two variants of CCN give good distribution estimates. Other methods give less accurate estimates. By comparing the posterior distributions of CEVAE against the conditional distribution and marginal distribution of the ground truth in S1(f) and S1(g), we conclude that CEVAE primarily captures the marginal distribution in this study.

F Policy Learning

In decision making, the core difference between a traditional policy learning method and a distribution learning method is whether the utility is determined in advance. Though a policy learning method can tailor decisions based on different utilities, it is at the cost of fitting a new model towards each proposed utility. Regardless of the inconvenience in computation, we discuss below another shortcoming of traditional policy learning methods. To train a traditional policy learning approach, the first step is often to convert the raw outcome to the observed utility. While this is less problematic for bijective transformations, it might incur information loss if we deal with discretized utilities.

To demonstrate, we compare FCCN to policytree with its published package (Athey & Wager, 2021; Sverdrup et al., 2021) on IHDP. We propose two types of utilities. They are the linear utility, $U_0(\gamma) = \gamma, U_1(\gamma) = \gamma - 4$, and the threshold utility, $U_0(\gamma) = 1_{\gamma > E[Y(0)]}, U_1(\gamma) = 1_{\gamma > E[Y(1)]}$. Since the policytree package only outputs the decision, we use accuracy as the metric. The results are summarized in Table S1. FCCN consistently makes more correct decisions. The information loss in the threshold utility negatively impacts policytree. We note that a threshold utility drastically reduces the available information by converting a continuous scale to a binary scale.

Fundamentally, these two methods are different and they address similar problems from different perspectives. There might be some possibilities that we could combine their merits. Hence, we will consider exploring their interactions more in future work.

Table S1: Comparing the accuracy of policy learning between FCCN and policytree.

Utility/Method	FCCN %	policytree %
Linear	88.60 ± 1.09	76.86 ± 1.03
Threshold	87.72 ± 1.19	57.64 ± .71

G Additional Distribution Tests

To further illustrate CCN's potential to model different types of distributions with high fidelity, we simulate potential outcomes from three extra distributions to assess its adaptability. To compare, we include GAMLSS, BART, CCN and FCCN. The assessment is based on log likelihood (LL) to reflect the closeness to the true distributions. We simulate the covariate spaces and treatment labels with the following procedures:

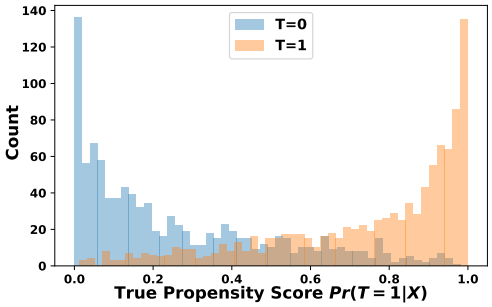


Figure S2: The limited propensity score overlap between two treatment groups. This indicates a severe treatment group imbalance.

Covariates:

$$x_i = (x_{1,i}, \cdots, x_{10,i})^\intercal, \quad x_{j,i} \overset{i.i.d.}{\sim}, \ N(0,1);$$

Treatment assignment:

$$Pr(T_i = 1|x_i) = \frac{1}{1 + exp(-x_i^\intercal \beta)}, \beta = (0.8, ..., 0.8)^\intercal$$

Given the magnitude of β, we have created a covariate space with limited overlap between two treatment groups as shown in Figure S2. With limited sample size, it also helps us evaluate the robustness of our method when positivity in Assumption 1 is possibly violated. Then we specify three scenarios with sufficient nolinearity added to the potential outcome generating processes. We generate 2,000 data points in each case, and the variability assessment is based on 5-fold cross validations.

Gumbel Distribution:

$$Y(0)_i|x_i \sim \text{Gumbel}\left(5\left[\sin(\sum_{j=1}^{10} x_{j,i})\right]^2, 5\left[\cos(\sum_{j=1}^{10} x_{j,i})\right]^2\right); \ Y(1)_i|x_i \sim \text{Gumbel}\left(5\left[\cos(\sum_{j=1}^{10} x_{j,i})\right]^2, 5\left[\sin(\sum_{j=1}^{10} x_{j,i})\right]^2\right).$$

Gamma Distribution:

$$Y(0)_i|x_i \sim \text{Gamma}\left(4\sqrt{\left|\sin(\sum_{j=1}^{5} x_{j,i}) + \cos(\sum_{j=6}^{10} x_{j,i})\right|} + .5, 2\sqrt{\left|\cos(\sum_{j=1}^{5} x_{j,i}) + \sin(\sum_{j=6}^{10} x_{j,i})\right|}\right);$$

$$Y(1)_i|x_i \sim \text{Gamma}\left(4\sqrt{\left|\cos(\sum_{j=1}^{5} x_{j,i}) + \sin(\sum_{j=6}^{10} x_{j,i})\right|} + .52\sqrt{\left|\sin(\sum_{j=1}^{5} x_{j,i}) + \cos(\sum_{j=6}^{10} x_{j,i})\right|}\right).$$

Weibull Distribution:

$$Y(0)_i|x_i \sim \text{Weibull}\left(5\sqrt{\left|\sin(\sum_{j=1}^{5} x_{j,i}) + \cos(\sum_{j=6}^{10} x_{j,i})\right|}, 2\sqrt{\left|\cos(\sum_{j=1}^{5} x_{j,i}) + \sin(\sum_{j=6}^{10} x_{j,i})\right|} + .2\right);$$

$$Y(1)_i|x_i \sim \text{Weibull}\left(5\sqrt{\left|\cos(\sum_{j=1}^{5} x_{j,i}) + \sin(\sum_{j=6}^{10} x_{j,i})\right|}, 2\sqrt{\left|\sin(\sum_{j=1}^{5} x_{j,i}) + \cos(\sum_{j=6}^{10} x_{j,i})\right|} + .2\right).$$

H Estimating Multimodal Distributions

We use the following steps to generate the data from a multimodal distribution.

Covariates: $x_i \overset{i.i.d.}{\sim} N(0,1)$;

Treatment assignment: $Pr(T_i = 1) = 1_{x_i>0}$;

Mixture parameter: $\phi_i \sim i.i.d, \text{Bernoulli}(0.5)$;

Potential outcomes: $Y(0)_i|x_i \sim \phi_i N(-2,1)+(1-\phi_i)N(2,1)+x_i, Y(1)_i|x_i \sim \phi_i N(6,1.5^2)+(1-\phi_i)\text{Exp}(1)+x_i$.

The control group is a mixture of two Gaussian distributions, and the treatment group is a mixture of Gaussian and exponential distributions. The mixture information is not given to any model, and we simply use their original form to approximate the distributions. Each model is trained using 1,600 simulated samples.

I Sample Size and Convergence

We create an example with the logistic distribution to visualize the performance of models with respect to sample size. We simulate 40,000 samples in total and hold out 2,000 for evaluation based on log likelihood (LL) with the following procedures:

Covariates:$x_i = (x_{1,i}, x_{2,i}, x_{3,i})^\intercal, \;\; x_{j,i} \overset{i.i.d.}{\sim} N(0,1)$;

Treatment assignment: $Pr(T_i = 1|x_i) = \dfrac{1}{1+exp(-x_i^\intercal \beta)}, \beta = (2,2,2)^\intercal$

Scale Parameter: $\sigma_i = |x_{1,i} + x_{2,i} + x_{3,i}| + .5$

Location Parameter: $\mu(0)_i = \sin(x_{1,i}\pi + x_{2,i}\pi) + \sin(x_{3,i}\pi),\ \mu(1)_i = \cos(x_{1,i}\pi + x_{2,i}\pi) + \cos(x_{3,i}\pi)$

Potential Outcomes: $Y(0)_i|x_i \sim \text{Logistic}(\mu(0)_i, \sigma_i),\ Y(1)_i|x_i \sim \text{Logistic}(\mu(1)_i, \sigma_i)$

J Ablation Study

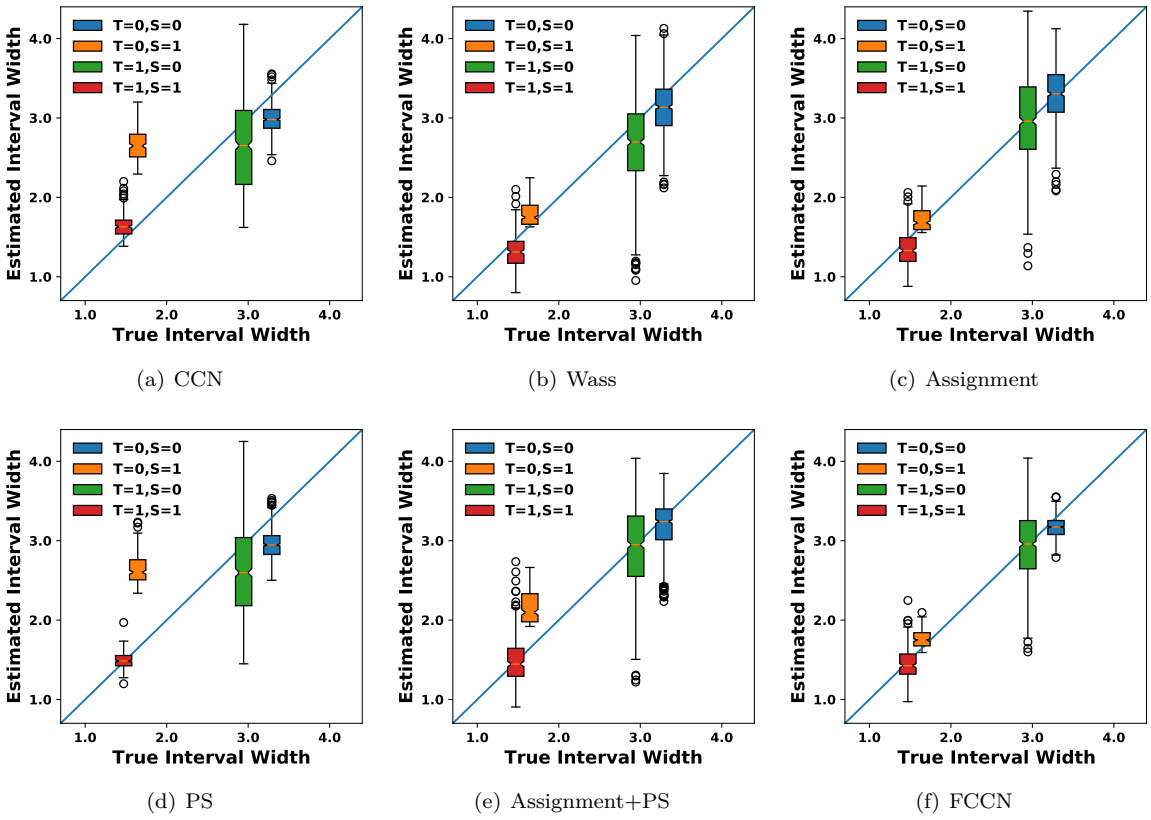
First, we provide an additional visualization in Figure S3 to reflects how well each adjustment can improve the the uncertainty estimates in EDU dataset. FCCN not only captures the heteroskedasticity, but also reduces uncertainty by exhibiting narrower bar in plots.

J.1 An Additional Imbalance Adjustment Study

Then, we give another motivating example to visualize the added robustness with our adjustment scheme. The visualizations in Figure 6(b), 6(c) and 6(d) are also based on this data example. We use the same covariate space and treatment assignment mechanism in Appendix G. However, we posit a nonstandard distribution with its location model as a trigonometric function, and outcome uncertainty model as a heterosekdastic Beta distribution. We generate 2,000 data points in total, with 8/2 split for training and testing. The detailed synthetic procedure for the outcomes is described below:

$$Y(0)_i|x_i \sim \text{Beta}\left(\frac{\sum_{j=1}^{5}|x_{j,1}|}{5}, \frac{\sum_{j=6}^{10}|x_{j,1}|}{5}\right) + \sin\left(\sum_{j=1}^{10} x_{j,i}\right);\ Y(1)_i|x_i \sim \text{Beta}\left(\frac{\sum_{j=6}^{10}|x_{j,1}|}{5}, \frac{\sum_{j=1}^{5}|x_{j,1}|}{5}\right) + \cos\left(\sum_{j=1}^{10} x_{j,i}\right).$$

We visualize the performance of different adjustment schemes in scatter plots where x axis corresponds to each point's true propensity score, a measurement of imbalance. Figure S4 depicts the absolute difference between the inferred CATEs and true CATEs. Lower vertical positions represent lower errors. We observe

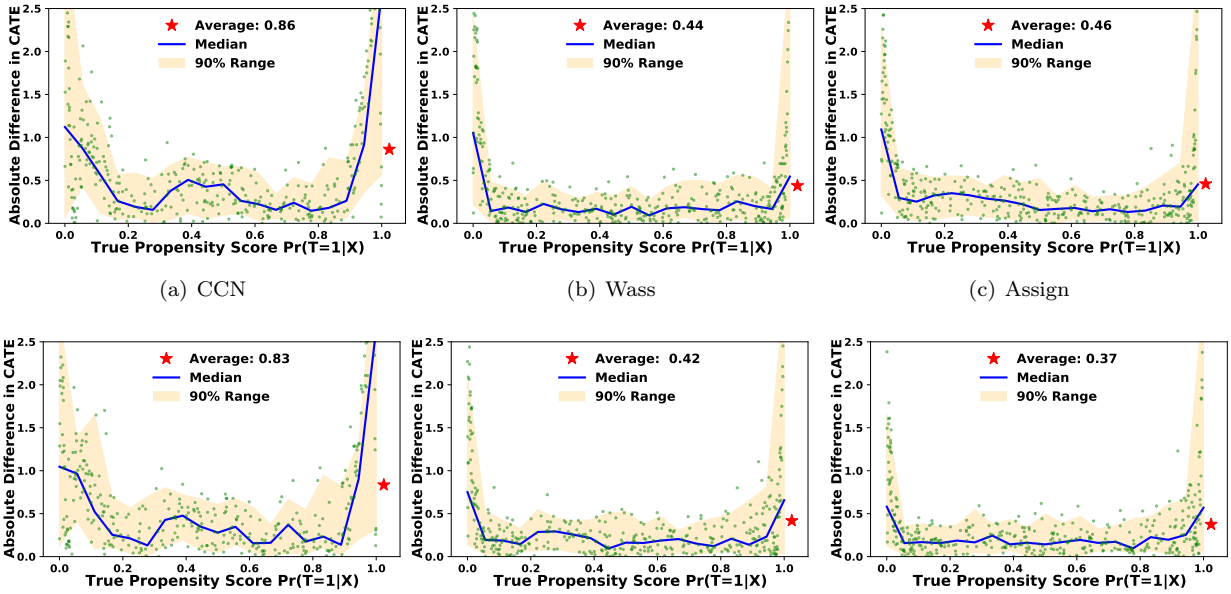


(a) CCN (b) Wass (c) Assignment

(d) PS (e) Assignment+PS (f) FCCN

Figure S3: The true 90 % interval widths versus the estimated 90 % interval widths given the four combinations of T and S with all variants of CCN. FCCN not only captures the heteroskedasticity, but also reduces the uncertainty by exhibiting narrower bar in plots.

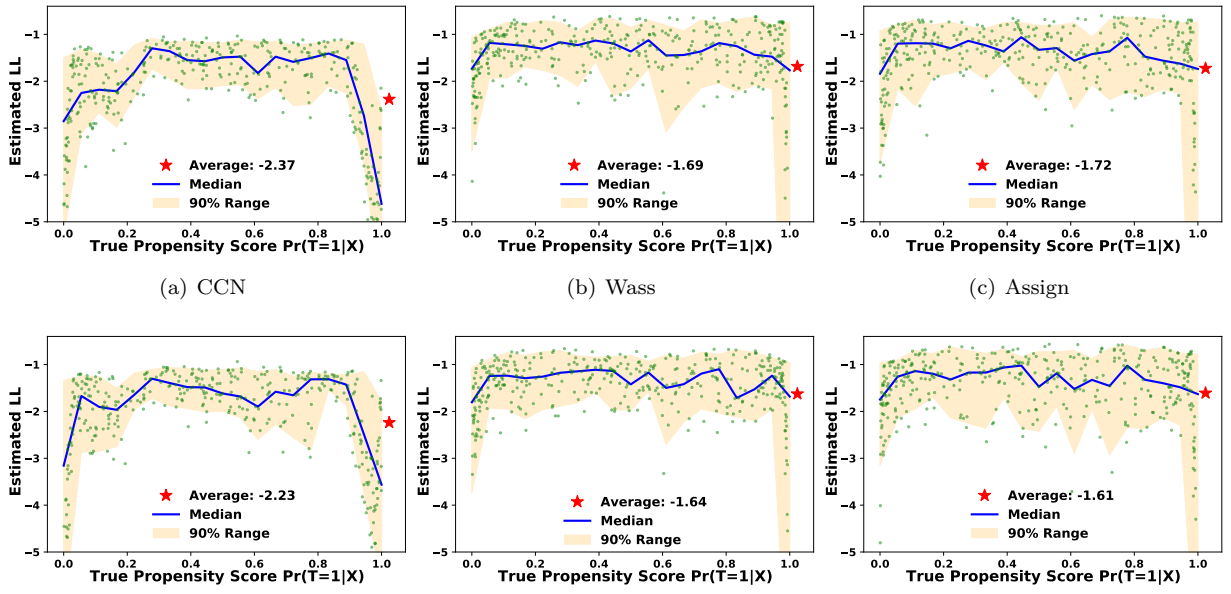
that the performance deteriorates in each method if a point is close to two boundaries (extreme propensity scores), which is the area that generally struggles the most in observational studies. Compared with the baseline CCN, each adjustment scheme by itself lowers the error to some extent. Among them, WASS-CCN, Assign-CCN and FCCN are able to reduce the average error by over 50 %. The propensity stratification (PS) can effectively reduce bias when there is more homogeneity within each stratum. However, severe imbalance in this case only gives homogeneity in the strata where propensity is around 0.5. Hence, PS provides limited benefits. In contrast, WASS-loss and Assign-loss seek new representations to either rectify group level imbalance or exclude confounding effects. They prove to be more effective in the regions of imbalance, which represent the majority of the data points.



(a) CCN (b) Wass (c) Assign

(d) PS (e) Assign+PS (f) FCCN

Figure S4: The scatter plot of the propensity scores (x-axis) versus the absolute difference between the true CATEs and their estimates (y-axis). Among them, WASS-CCN, Assign-CCN and FCCN are able to reduce the average error by over 50%.



(a) CCN (b) Wass (c) Assign

(d) PS (e) Assign+PS (f) FCCN

Figure S5: The scatter plot of the propensity scores (x-axis) versus the estimated log-likelihood (LL) (y-axis). Collectively, the Assign-loss and Wass-loss contribute to making FCCN more robustly estimate distributions than any single component does.