Optimal Policy Adaptation Under Covariate Shift

Xueqing Liu[1] , **Qinwei Yang**[1] , **Zhaoqing Tian**[1] , **Ruocheng Guo**[2] , **Peng Wu***
[1]Beijing Technology and Business University
[2]ByteDance Research
pengwu@btbu.edu.cn

Abstract

Transfer learning of prediction models under covariate shift has been extensively studied, while the corresponding policy learning approaches are rarely discussed. In this paper, we propose principled approaches for learning the optimal policy in the target domain by leveraging two datasets: one with full information from the source domain and the other from the target domain with only covariates. First, in the setting of covariate shift, we formulate the problem from a perspective of causality and present the identifiability assumptions for the reward induced by a given policy. Then, we derive the efficient influence function and the semiparametric efficiency bound for the reward. Based on this, we construct a doubly robust and semiparametric efficient estimator for the reward and then learn the optimal policy by optimizing the estimated reward. Moreover, we theoretically analyze the bias and the generalization error bound for the learned policy. Furthermore, in the presence of both covariate and concept shifts, we propose a novel sensitivity analysis method to evaluate the robustness of the proposed policy learning approach. Extensive experiments demonstrate that the approach not only estimates the reward more accurately but also yields a policy that closely approximates the theoretically optimal policy.

1 Introduction

In many real-world scenarios, labeled data is often scarce due to budget constraints and time-consuming collection processes [Zhuang *et al.*, 2020; Imbens *et al.*, 2024], significantly limiting the generalizability of the resulting models. For example, in medical research, collecting labeled data involves extensive clinical trials and follow-up periods, making it costly and time-consuming [Dahabreh *et al.*, 2020; Hu *et al.*, 2023]. In autonomous driving, obtaining labeled data requires manual annotation of large amounts of sensor data, which is laborious and expensive [Sun *et al.*, 2020]. To address this problem and enhance a model's performance in a target domain without labels, an active area of research is transfer learning. It aims to improve the performance of target learners in the target domain by transferring the knowledge contained in a different but related source domain.

While transfer learning has been extensively studied in the context of prediction models [Wang *et al.*, 2018; Wang *et al.*, 2020; Pesciullesi *et al.*, 2020], how to transfer a policy is still underdeveloped. Policy learning refers to identifying individuals who should receive treatment/intervention based on their characteristics by maximizing rewards [Murphy, 2003]. It has broad applications in recommender systems [Chen and Sun, 2021; Wu *et al.*, 2022], precision medicine [Bertsimas *et al.*, 2017] and reinforcement learning [Liu *et al.*, 2021; Kwan *et al.*, 2023]. Unlike transfer learning for prediction models, policy transfer faces identification challenges due to its counterfactual nature [Athey and Wager, 2021; Li *et al.*, 2023b; Wu *et al.*, 2024c; Yang *et al.*, 2024]. Instead of predicting outcomes based on observed data, policy transfer requires considering what would happen under different actions, making the process more complex.

We aim to learn the optimal policies in the target and entire domains using a dataset from the source domain (source dataset) and a dataset from the target domain (target dataset). The source dataset includes the covariates, treatment, and outcome for each individual, whereas the target dataset contains only the covariates. We assume that the source dataset satisfies the unconfoundedness and overlap assumptions while imposing fewer restrictions on the target dataset. We allow for substantial differences in the covariate distributions between the source and target datasets (referred to as covariate shift), while assuming that the conditional distributions of potential outcomes given covariates are the same.

In this article, we first propose a principled policy learning approach under covariate shift. Specifically, we define the reward and the optimal policy in the target domain using the potential outcome framework in causal inference. Under the widely used assumptions of unconfoundedness and transportability, we establish the identifiability of the reward in the target domain and then derive its efficient influence function and semiparametric efficiency bound. Building on this, we develop a novel estimator for the reward. Theoretical analysis shows that the proposed estimator is doubly robust and achieves the semiparametric efficient bound, that is, it is the optimal regular estimator in terms of asymptotic variance

*corresponding author

[Newey, 1990]. Then we propose to learn the optimal policy by maximizing the estimated reward. We analyze the bias of the estimated reward and the generalization error bound of the learned policy. In addition, we extend the proposed method to learn the optimal policy in the entire domain consisting of the source and target domains by leveraging data from both domains to address distributional discrepancies and ensure robust generalization across heterogeneous environments.

The main contributions are summarized as follows: (1) We propose a principled approach for learning the optimal policy under covariate shift from a perspective of causality, by introducing plausible identifiability assumptions and efficient estimation methods; (2) We provide a comprehensive theoretical analysis of the proposed approach, including the consistency, asymptotic normality, and semiparametric efficiency of the estimator of reward. Additionally, we derive the bias and generalization error bound for the learned policy; (3) We conduct extensive experiments to demonstrate the effectiveness of the proposed policy learning approach.

2 Problem Formulation

2.1 Notation and Setup

Let $A \in \mathcal{A} = \{0, 1\}$ denote the binary indicator for treatment, where $A = 1$ indicates receiving treatment and $A = 0$ indicates not receiving treatment. The random vector $X \in \mathcal{X} \subset \mathbb{R}^p$ represents the p-dimensional covariates measured before treatment, and $Y \in \mathcal{Y} \subset \mathbb{R}$ denotes the outcome of interest. Assume that a larger outcome is preferable. Under the potential outcome framework [Rubin, 1974; Splawa-Neyman, 1990], let $Y(a)$ denote the potential outcome that would be observed if A were set to a for $a \in \mathcal{A}$. By the consistency [Hernán and Robins, 2020], the observed outcome Y satisfies $Y = Y(A) = AY(1) + (1 - A)Y(0)$.

Without loss of generality, we consider a typical scenario involving two datasets: a source dataset and a target dataset, which are representative samples of the source domain and target domain, respectively. Let $G \in \{0, 1\}$ be the indicator for the data source, where $G = 1$ denotes the source domain and $G = 0$ denotes the target domain. The observed data are represented as follows,

$$\mathcal{D}_1 = \{(X_i, A_i, Y_i, G_i = 1) : i = 1, ..., n_1\},$$
$$\mathcal{D}_0 = \{(X_i, G_i = 0) : i = n_1 + 1, ..., n_1 + n_0\},$$

where the source dataset $\mathcal{D}_1$ consists of n_1 individuals, with observed covariates, treatment, and outcome for each individual. The target dataset $\mathcal{D}_0$ contains n_0 individuals, with only covariates for each individual. This is common in real life due to the scarcity of outcome data. For example, in medical research, patient features are observed, but obtaining outcomes requires long-term follow-up [Hu *et al.*, 2023; Imbens *et al.*, 2024]. Let $\mathbb{P}(\cdot|G = 1)$ and $\mathbb{P}(\cdot|G = 0)$ be the distributions of the two datasets respectively. Then $n = n_0 + n_1$ and $q = n_1/n$ represent the probability of an individual belonging to the source population.

2.2 Formulation

We formulate the goal of learning the optimal policy in the target domain. Specifically, let $\pi : \mathcal{X} \to \mathcal{A}$ denote a policy that maps individual covariates $X = x$ to the treatment space $\mathcal{A}$. A policy $\pi(X)$ is a treatment rule that determines whether an individual receives treatment $(A = 1)$ or not $(A = 0)$. For a given policy π applied to the target domain, the average reward is defined as follows

$$R(\pi) = \mathbb{E}[\pi(X)Y(1) + (1 - \pi(X))Y(0)|G = 0]. \quad (1)$$

We aim to learn the optimal policy π^* defined by $\pi^* = \arg\max_{\pi \in \Pi} R(\pi)$, where Π is a pre-specified policy class. For example, $\pi(X)$ can be modeled with a parameter θ using methods such as logistic regression or multilayer perceptron, with each value of θ corresponding to a different policy.

In addition, for a policy $\pi(x)$ applied across the whole domain, the corresponding average reward is defined as

$$V(\pi) = \mathbb{E}[\pi(X)Y(1) + (1 - \pi(X))Y(0)]. \quad (2)$$

There is a subtle difference between $R(\pi)$ and $V(\pi)$. For $R(\pi)$, our focus is on transferring the policy from the source domain to the target domain, and for $V(\pi)$, we aim to generalize the policy from the source domain to the entire domain. In the main text, we focus on learning the policy maximizing $R(\pi)$ to avoid redundancy. We also develop a similar approach to learn the policy maximizing $V(\pi)$ and briefly present it in Section 4.3.

3 Oracle Policy and Identifiability

3.1 Oracle Policy

The optimal policy that maximizes Eq. (1) has an explicit form. Let $\tau(X) = \mathbb{E}[Y(1) - Y(0)|X, G = 0]$ be the conditional average treatment effect (CATE) in the target domain,

$$\begin{aligned} R(\pi) &= \mathbb{E}[\pi(X)\{Y(1) - Y(0)\} + Y(0)|G = 0] \\ &= \mathbb{E}[\pi(X)\tau(X)|G = 0] + \mathbb{E}[Y(0)|G = 0] \end{aligned}$$

where the last equality follows from the law of iterated expectations. Then we have the following conclusion.

Lemma 1. *The oracle policy*

$$\pi_0^*(x) = \arg\max_{\pi} R(\pi) = \begin{cases} 1, & \textit{if } \tau(x) \geq 0 \\ 0, & \textit{if } \tau(x) < 0, \end{cases}$$

where $\max_\pi$ *is taken over all possible policies without constraints, rather than being restricted to* Π.

For an individual characterized by $X = x$ in the target domain, Lemma 1 asserts that the decision to accept treatment $(A = 1)$ should be based on the sign of $\tau(x)$. The oracle policy π_0^* recommends treatment for individuals expected to experience a positive benefit, thereby optimizing the overall reward within the target domain. The target policy π^* equals the oracle policy π_0^* in Lemma 1 if $\pi_0^* \in \Pi$; otherwise, they may not be equal, and their difference is the systematic error induced by limited hypothesis space of Π.

3.2 Identifiability of the Reward

To learn the optimal policy π^*, we first need to address the identifiability problem of $R(\pi)$, as this forms the foundation for policy evaluation. Since the target dataset only contains covariates X, $R(\pi)$ cannot be identified from the target data alone due to the absence of treatment and outcome. To identify $R(\pi)$, it is necessary to borrow information from the source dataset by imposing several assumptions.

Assumption 1. *For all X in the source domain,*
(i) Unconfoundedness: $(Y(1), Y(0)) \perp\!\!\!\perp A \mid X, G = 1$*;*
(ii) Overlap: $0 < e_1(X) \triangleq \mathbb{P}(A = 1|X, G = 1) < 1$*, where* $e_1(X)$ *is the propensity score.*

Assumption 1(i) states that, in the source domain, the treatment is independent of the potential outcomes given the covariates, implying that all confounders affecting both the treatment and outcome are observed. Assumption 1(ii) asserts that any individual characterized by X in the source domain has a positive probability of receiving treatment. Assumption 1 is a standard assumption for identifying causal effects in the source domain [Rosenbaum and Rubin, 1983]. However, Assumption 1 is not enough to identify the causal effects in the target domain. Thus, we further invoke Assumption 2.

Assumption 2 (Transportability)**.** *Suppose that*
(i) $(Y(0), Y(1)) \perp\!\!\!\perp G \mid X$ *for all* X*;*
(ii) $0 < s(X) \triangleq \mathbb{P}(G = 1|X) < 1$ *for all* X *in the source domain, where* $s(X)$ *is the sampling score.*

Assumption 2 is widely adopted in causal effects estimation via data combination in causal inference [Stuart *et al.*, 2011; Hartman *et al.*, 2015; Kern *et al.*, 2016a; Lesko *et al.*, 2017; Buchanan *et al.*, 2018; Li *et al.*, 2022a; Li *et al.*, 2023a; Colnet *et al.*, 2024; Wu *et al.*, 2025; Wu and Mao, 2025]. Assumption 2(ii) indicates that all individuals in the source domain have a positive probability of belonging to the target domain. Assumption 2(i) implies that $\mathbb{E}[Y(a)|X, G = 1] = \mathbb{E}[Y(a)|X, G = 0] = \mathbb{E}[Y(a)|X]$ for $a = 0, 1$, which ensures the transportability of the CATE from the source domain to the target domain and leads to the identifiability of $\tau(X)$, that is,

$$
\begin{aligned}
\tau(X) &= \mathbb{E}[Y(1) - Y(0)|X, G = 1] \\
&= \mathbb{E}[Y|X, A = 1, G = 1] - \mathbb{E}[Y|X, A = 0, G = 1] \\
&\triangleq \mu_1(X) - \mu_0(X),
\end{aligned}
$$

where the third equality follows from Assumption 1. Thus, under Assumptions 1-2, the reward $R(\pi)$ can be identified as

$$
\begin{aligned}
R(\pi) &= \mathbb{E}[\pi(X)\mu_1(X) + (1 - \pi(X))\mu_0(X)|G = 0] \\
&= \mathbb{E}\Big[\frac{1 - G}{1 - q}\{\pi(X)\mu_1(X) + (1 - \pi(X))\mu_0(X)\}\Big] \qquad (3)
\end{aligned}
$$

Assumption 2 allows the presence of covariate shift, i.e., the distribution of X in the source domain may significantly differ from that in the target domain [Gama *et al.*, 2014].

4 Policy Adaptation Under Covariate Shift

In this section, we proposed a method for learning the optimal policy. It consists of two steps: (a) policy evaluation, estimating the reward $R(\pi)$ for a given π, and (b) policy learning, learning the optimal policy based on the estimated reward.

4.1 Estimation of the Reward $R(\pi)$

According to Eq. (3), a direct method for estimating the reward $R(\pi)$ is given as

$$
\begin{aligned}
\hat{R}_{\text{Direct}}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \frac{1 - G_i}{1 - q} \\
\times \{\pi(X_i)\hat{\mu}_1(X_i) + (1 - \pi(X_i))\hat{\mu}_0(X_i)\},
\end{aligned}
$$

where $\hat{\mu}_a(X)$ $(a = 0, 1)$ represents the estimated outcome regression function $\mu_a(X)$. This can be implemented by regressing Y on X using the source dataset with $A = a$. The unbiasedness of the direct estimator $\hat{R}_{\text{Direct}}(\pi)$ depends on the accuracy of $\hat{\mu}_a(X)$. If $\hat{\mu}_a(X)$ is a biased estimator of $\mu_a(X)$, then $\hat{R}_{\text{Direct}}(\pi)$ will also be a biased estimator of $R(\pi)$. Moreover, the generalization performance of the direct method is often poor because $\hat{\mu}_a(X)$ is trained using the source dataset of $A = a$, but is applied to the whole target dataset [Li *et al.*, 2022b; Wu *et al.*, 2024b]. When there is a significant difference in the covariate distributions between the source and target datasets, the direct method suffers from the problem of model extrapolation, resulting in poor practical performance.

In addition to the direct method, one can use the propensity score $e_1(X) = \mathbb{P}(A = 1|X, G = 1)$ and sampling score $s(X) = \mathbb{P}(G = 1|X)$ to construct the inverse probability weighting (IPW) estimator of $R(\pi)$. Note that

$$
R(\pi) = \mathbb{E}\Big[\frac{G}{1 - q}\omega(X)\Big\{\frac{\pi(X)AY}{e_1(X)} + \frac{(1 - \pi(X))(1 - A)Y}{1 - e_1(X)}\Big\}\Big],
$$

with $\omega(X) = (1 - s(X))/s(X)$. Based on it, the IPW estimator of $R(\pi)$ is given as

$$
\begin{aligned}
\hat{R}_{\text{IPW}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{G_i}{1 - q}\frac{\pi(X_i)A_iY_i}{\hat{e}_1(X_i)}\frac{1 - \hat{s}(X_i)}{\hat{s}(X_i)} \\
+ \frac{G_i}{1 - q}\frac{(1 - \pi(X_i))(1 - A_i)Y_i}{1 - \hat{e}_1(X_i)}\frac{1 - \hat{s}(X_i)}{\hat{s}(X_i)}\Big\},
\end{aligned}
$$

where $\hat{e}_1(X)$ and $\hat{s}(X)$ are estimates of $e_1(X)$ and $s(X)$, respectively. The IPW estimator $\hat{R}_{\text{IPW}}(\pi)$ is an unbiased estimator of $R(\pi)$ when $\hat{e}_1(X)$ and $\hat{s}(X)$ are accurate estimators of $e_1(X)$ and $s(X)$, respectively, i.e., $\hat{e}_1(X) = e_1(X)$ and $\hat{s}(X) = s(X)$. However, a limitation of the IPW estimator is its inefficiency, meaning it tends to have a large variance.

The limitations of direct and IPW methods are essentially caused by the insufficiency of utilizing the observed data. The direct method does not leverage the information of data indicator G and treatment A, while the IPW method does not extract the relationship between covariates X and outcome Y. To fully utilize the observed data, we employ the semiparametric efficiency theory [Tsiatis, 2006] to derive the efficient influence function and the efficiency bound of $R(\pi)$. This allows us to obtain the semiparametric efficient estimator of $R(\pi)$. A semiparametric efficient estimator is considered optimal as it reaches the semiparametric efficiency bound, meaning it has the smallest asymptotic variance under several regularity conditions [Newey, 1990].

Theorem 1 (Efficiency Bound of $R(\pi)$)**.** *Under Assumptions 1-2, the efficient influence function of* $R(\pi)$ *is*

$$
\begin{aligned}
\varphi_R = \frac{G}{1 - q}\frac{\pi(X)A\{Y - \mu_1(X)\}}{e_1(X)}\frac{1 - s(X)}{s(X)} \\
+ \frac{G}{1 - q}\frac{(1 - \pi(X))(1 - A)\{Y - \mu_0(X)\}}{1 - e_1(X)}\frac{1 - s(X)}{s(X)} \\
+ \frac{1 - G}{1 - q}\{\pi(X)\mu_1(X) + (1 - \pi(X))\mu_0(X) - R(\pi)\}.
\end{aligned}
$$

The semiparametric efficiency bound of $R(\pi)$ *is Var*(φ_R).

Theorem 1 (See Appendix A.1 for proofs) presents the efficient influence function and semiparametric efficiency bound of $R(\pi)$ under Assumptions 1-2. From Theorem 1, we can construct the semiparametric efficient (SE) estimator of $R(\pi)$, which is given as

$$\hat{R}_{\text{SE}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\Big[\frac{G_i}{1-q}\frac{\pi(X_i)A_i\{Y_i-\hat{\mu}_1(X_i)\}}{\hat{e}_1(X_i)}\frac{1-\hat{s}(X_i)}{\hat{s}(X)} + \frac{G_i}{1-q}\frac{(1-\pi(X_i))(1-A_i)\{Y_i-\hat{\mu}_0(X_i)\}}{1-\hat{e}_1(X_i)}\frac{1-\hat{s}(X_i)}{\hat{s}(X_i)} + \frac{1-G_i}{1-q}\{\pi(X_i)\hat{\mu}_1(X_i)+(1-\pi(X_i))\hat{\mu}_0(X_i)\}\Big].$$

Next, we analyze the theoretical properties of $\hat{R}_{\text{SE}}(\pi)$.

Proposition 1 (Double Robustness of $\hat{R}_{\text{SE}}(\pi)$)**.** *Under Assumptions 1 and 2, $\hat{R}_{SE}(\pi)$ is an unbiased estimator of $R(\pi)$ if one of the following conditions is satisfied:*

(i) $\hat{\mu}_a(x) = \mu_a(x)$, i.e., $\hat{\mu}_a(x)$ estimates $\mu_a(x)$ accurately for $a = 0, 1$.

(ii) $\hat{e}_1(x) = e_1(x)$ and $\hat{s}(x) = s(x)$, i.e., $\hat{e}(x)$ and $\hat{s}(x)$ estimate $e(x)$ and $s(x)$ accurately.

Proposition 1 (See Appendix A.2 for proofs) shows the double robustness of $\hat{R}_{\text{SE}}(\pi)$, i.e., if the outcome regression function $\mu_a(X)$ for $a = 0, 1$ can be estimated accurately, or the propensity score $e_1(X)$ and the sampling score the $s(X)$ can be estimated accurately, $\hat{R}_{\text{SE}}(\pi)$ is unbiased estimator of $R(\pi)$. Compared to the direct method that requires $\hat{\mu}_a(X) = \mu_a(X)$ for unbiasedness, and the IPW method that requires $\hat{e}_1(X) = e_1(X)$ and $\hat{s}(X) = s(X)$ for unbiasedness, double robustness provides more reliable results by mitigating the inductive bias caused by inaccurate models for the nuisance parameters $e_1(X), s(X)$, and $\mu_a(X)$ for $a = 0, 1$.

Theorem 2 (Efficiency of $\hat{R}_{\text{SE}}(\pi)$)**.** *Under the Assumptions 1–2, if $||\hat{e}_1(x) - e_1(x)||_2 \cdot ||\hat{\mu}_a(x) - \mu_a(x)||_2 = o_{\mathbb{P}}(n^{-1/2})$ and $||\hat{s}(x) - s(x)||_2 \cdot ||\hat{\mu}_a(x) - \mu_a(x)||_2 = o_{\mathbb{P}}(n^{-1/2})$ for all $x \in \mathcal{X}$ and $a \in \{0,\ 1\}$, then $\hat{R}_{SE}(\pi)$ is a consistent estimator of $R(\pi)$, and satisfies*

$$\sqrt{n}\{\hat{R}_{SE}(\pi) - R(\pi)\} \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = Var(\varphi_R)$ is the semiparametric efficiency bound of $R(\pi)$, and $\xrightarrow{d}$ means convergence in distribution.

Theorem 2 (See Appendix A.3 for proofs) establishes the consistency and asymptotic normality of the proposed estimator $\hat{R}_{\text{SE}}(\pi)$. In addition, it shows that $\hat{R}_{\text{SE}}(\pi)$ is semiparametric efficient, i.e., it achieves the semiparametric efficiency bound. These desired properties hold under the mild condition that the nuisance parameters $\{e(x), s(x), \mu_0(x), \mu_1(x)\}$ are estimated at a rate faster than $n^{-1/4}$. These conditions are common in causal inference and can be easily satisfied using a variety of flexible machine learning methods [Chernozhukov *et al.*, 2018; Wu *et al.*, 2024a].

4.2 Learning the Optimal Policy

After estimating the reward, we now focus on learning the optimal policy. Recall that for a given hypothesis space Π, the target policy is given as $\pi^*(x) = \arg\max_{\pi\in\Pi} R(\pi)$. Through optimizing different estimators of reward $R(\pi)$, we obtain different estimator of π^*, denoted as $\hat{\pi}$, defined by

$$\hat{\pi}(x) = \arg\max_{\pi\in\Pi} \hat{R}(\pi) \tag{4}$$

where $\hat{R}(\pi)$ can be $\hat{R}_{\text{Direct}}(\pi)$, $\hat{R}_{\text{IPW}}(\pi)$, or $\hat{R}_{\text{SE}}(\pi)$. Algorithm 1 summarizes the procedures for learning π^*.

Algorithm 1 Proposed Policy Learning Approach

Input: The source dataset $\mathcal{D}_1$ and the target dataset $\mathcal{D}_0$.
Output: The learned policy $\hat{\pi}$.

1: **Stage 1:** Fit models $\hat{\mu}_1(X)$, $\hat{\mu}_0(X)$, $\hat{e}_1(X)$, $\hat{s}(X)$.
2: **Stage 2:**
3: **while** Stop condition is not reached **do**
4: Sample a batch of data from $\mathcal{D}_0 \cup \mathcal{D}_1$.
5: Minimize the loss -$\hat{R}(\pi, \hat{\mu}_0, \hat{\mu}_0, \hat{e}_1, \hat{s}, X, A, Y, G)$ to update π, using the batch sample.
6: **end while**
7: Return a learned policy.

As discussed in [Athey and Wager, 2021], when learning the policy by optimizing the estimated reward, it achieves better generalization performance if the estimated reward is more efficient. Since the estimator $\hat{R}_{\text{SE}}(\pi)$ is the most efficient one under Assumptions 1-2, as shown in Theorem 1, we then focus on exploring its finite sample properties and the learned policy obtained by optimizing it. Similar results can also be derived for the direct and IPW methods. In finite samples, we allow $\hat{\mu}_a(X), \hat{e}_1(X)$, and $\hat{s}(X)$ to be inaccurate, i.e., they may differ from $\mu_a(X)$, $e_1(X)$, and $s(X)$.

The following Proposition 2 presents the bias of $\hat{R}_{\text{SE}}(\pi)$.

Proposition 2 (Bias)**.** *Given the learned $\hat{\mu}_a(X)$ for $a = 0, 1$, $\hat{e}_1(X)$, and $\hat{s}(X)$, then for any given π, the bias of $\hat{R}_{SE}(\pi)$ is*

$$\begin{aligned}&\textit{Bias}(\hat{R}_{SE}(\pi)) = |\mathbb{E}[\hat{R}_{SE}(\pi)] - R(\pi)|\\ &= \Big|\frac{1}{n}\sum_{i=1}^{n}\Big[\frac{\pi(X_i)(\mu_1(X_i)-\hat{\mu}_1(X_i))}{1-q}\times\\ &\Big\{\frac{s(X_i)e_1(X_i)(1-\hat{s}(X_i)) - \hat{s}(X_i)\hat{e}_1(X_i)(1-s(X_i))}{\hat{e}_1(X_i)\hat{s}(X_i)}\Big\}\\ &+\frac{(1-\pi(X_i))(\mu_0(X_i)-\hat{\mu}_0(X_i))}{1-q}\times\\ &\Big\{\frac{s(X_i)e_0(X_i)(1-\hat{s}(X_i)) - \hat{s}(X_i)\hat{e}_0(X_i))(1-s(X_i))}{\hat{e}_0(X_i))\hat{s}(X_i)}\Big\}\Big]\Big|,\end{aligned}$$

where $e_0(X_i) = 1 - e_1(X_i)$ and $\hat{e}_0(X_i) = 1 - \hat{e}_1(X_i)$.

From Proposition 2 (See Appendix A.4 for proofs), the bias of $\hat{R}_{\text{SE}}(\pi)$ is the product of the estimation errors $\mu_a(X) - \hat{\mu}_a(X))$ and $s(X)e_a(X)(1-\hat{s}(X)) - \hat{s}(X)\hat{e}_a(X)(1-s(X))$ for $a = 0, 1$. Clearly, when either $\hat{\mu}_a(X)$ is close to $\mu_a(X)$, or $\hat{s}(X)$ and $\hat{e}_1(X)$ are close to $s(X)$ and $e_1(X)$, $\hat{R}_{\text{SE}}(\pi)$ will be close to $R(\pi)$. This further demonstrates the double robustness of $\hat{R}(\pi)$.

Next, we show the generalization error bound (or the regret) of the learned policy. For clarity, we define

$$\hat{\pi}_{\text{se}}(x) = \arg\max_{\pi\in\Pi} \hat{R}_{\text{SE}}(\pi),$$

which is the learned policy by optimizing $\hat{R}_{\text{SE}}(\pi)$.

Theorem 3 (Generalization Error Bound)**.** *For any finite hypothesis space* Π*, we have that*

(i) with at least probability $1-\eta$*,*

$$R(\hat{\pi}_{se}) \leq \hat{R}(\hat{\pi}_{se}) + \textit{Bias}(\hat{R}_{SE}(\hat{\pi}_{se})) + \mathcal{B}(\mathcal{D}_0, \mathcal{D}_1, \eta, \Pi),$$

where $\mathcal{B}(\mathcal{D}_0, \mathcal{D}_1, \eta, \Pi)$ *equals to*

$$\sqrt{\frac{\log(2|\Pi|/\eta)}{2n^2} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_{A_i}(X_i))^2(1-\hat{s}(X_i))^2}{(1-q)^2 \hat{e}^2_{A_i}(X_i)\hat{s}^2(X_i)}},$$

with $\hat{\mu}_{A_i}(X_i) = A_i\hat{\mu}_1(X_i) + (1-A_i)\hat{\mu}_0(X_i)$.

(ii) with at least probability $1-\eta$*,*

$$R(\hat{\pi}_{se}) \leq R(\pi^*) + \textit{Bias}(\hat{R}_{SE}(\hat{\pi}_{se})) + \textit{Bias}(\hat{R}_{SE}(\pi^*)) + 2\mathcal{B}(\mathcal{D}_0, \mathcal{D}_1, \eta, \Pi).$$

Theorem 3(i) provides the generalization error bound of the learned policy $\hat{\pi}_{se}$, and Theorem 3(ii) presents the difference between the generalization risks of the learned policy and the optimal policy π^*. Note that when $(Y_i - \hat{\mu}_A(X_i))^2$ are bounded, then $\mathcal{B}(\mathcal{D}_0, \mathcal{D}_1, \eta, \Pi)$ will converge to 0 as the sample size n goes to infinity. Thus, for a sufficiently large sample size n, if the nuisance parameters are estimated with adequate accuracy, the generalization bound of the learned policy will be approximated well by the estimated reward. Additionally, the generalization bound of the learned policy will be close to that of the optimal policy $\hat{\pi}^*$.

4.3 Generalizing Policy to the Entire Domain

The approach proposed in Sections 4.1–4.2 is designed to learn the optimal policy in the target domain. In this subsection, we extend the approach to the entire domain, aiming to learn the optimal policy that maximizes $V(\pi)$. Under Assumption 1–2, the reward $V(\pi)$ is identified as

$$V(\pi) = \mathbb{E}[\pi(X)\mu_1(X) + (1-\pi(X))\mu_0(X)]. \quad (5)$$

Similar to Theorem 1, we present the the efficient influence function and the efficient bound of $V(\pi)$.

Theorem 4 (Efficiency Bound of $V(\pi)$)**.** *Under Assumptions 1-2, the efficient influence function of* $V(\pi)$ *is*

$$\varphi_V = \pi(X)\mu_1(X) + (1-\pi(X))\mu_0(X) - V(\pi) + \frac{G}{s(X)}$$
$$\left\{\frac{\pi(X)A(Y-\mu_1(X))}{e_1(X)} + \frac{(1-\pi(X))(1-A)(Y-\mu_0(X))}{1-e_1(X)}\right\}.$$

The semiparametric efficiency bound of $V(\pi)$ *is* $\textit{Var}(\varphi_V)$*.*

From Theorem 4, we can construct the semiparametric efficient (SE) estimator of $V(\pi)$, which is given as

$$\hat{V}_{\text{SE}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{G_i}{\hat{s}(X_i)}\Bigg\{\frac{\pi(X_i)A_i\{Y_i - \hat{\mu}_1(X_i)\}}{\hat{e}_1(X_i)} + \frac{(1-\pi(X_i))(1-A_i)\{Y_i-\hat{\mu}_0(X_i)\}}{1-\hat{e}_1(X_i)}\Bigg\} + \pi(X_i)\hat{\mu}_1(X_i) + (1-\pi(X_i))\hat{\mu}_0(X_i)\Bigg].$$

Similar to Proposition 1 and Theorem 2, it can be shown that $\hat{V}_{\text{SE}}(\pi)$ possesses desirable properties, including double robustness, consistency, asymptotic normality, and semiparametric efficiency under regular conditions. Moreover, following the method described in Section 4.2, we can develop an approach to learn the optimal policy in the entire domain based on the estimated reward $\hat{V}_{\text{SE}}(\pi)$. For brevity, the detailed description is omitted here.

5 Experiments

We conduct experiments on both simulated datasets and real-world datasets to answer the following questions:

- **RQ1**: Does the proposed SE method provide a more accurate estimation of the reward?
- **RQ2**: Does the proposed method learn better policies?
- **RQ3**: How does the degree of covariate shift affect the performance of the proposed method?

5.1 Experiments on Simulated Datasets

Datasets. We generate the source and target datasets in the simulated experiment. The source dataset consists of 512 individuals. For each individual, the covariates $X = (X_1, X_2, X_3)^\top \sim \mathcal{N}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$, where $\boldsymbol{\mu_1} = [10, 3, 7]^\top$ is the mean vector and $\boldsymbol{\Sigma_1} \in \mathbb{R}^{3\times 3}$ is the covariance matrix with the (i,j)-th element being $2^{-|i-j|}$ for $i,j = 1,2,3$. The treatment A is generated from $Bern(0.5)$, and the potential outcomes are generated as follows: $Y(1) = 15 + 0.4\tilde{X}_1\tilde{X}_2 + 0.7\tilde{X}_3 + \epsilon, Y(0) = 10 + 0.1\tilde{X}_1 + 0.5\tilde{X}_2\tilde{X}_3 + \epsilon$, where $\tilde{X}_j = X_j \cdot |X|_j^{0.1} + X_j \cdot |X|_j^{0.3} + X_j \cdot |X|_j^{0.5}$ for $j = 1,2,3$ and $\epsilon \sim \mathcal{N}(0,1)$. The observed outcome is $Y = AY(1) + (1-A)Y(0)$. The variables X, A, Y are available in the source dataset. For the target dataset, we generate 2,048 individuals. The covariates $X \sim \mathcal{N}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$, where the mean vector $\boldsymbol{\mu_2} = [9,4,6]^\top$ and the covariance matrix $\boldsymbol{\Sigma_2} \in \mathbb{R}^{3\times 3}$ has (i,j)-th element given by $2^{-|i-j|+1}$ for $i,j = 1,2,3$. Only X is available in the target dataset. In addition, to assess the performance of the learned policy, we also generate $\{Y(1), Y(0)\}$ for individuals in the target dataset, using the same method as in the source dataset.

Compared Methods. The direct and IPW method. $\hat{\pi}_{\text{direct}} = \arg\max_{\pi\in\Pi} \hat{R}_{\text{Direct}}(\pi)$ and $\hat{\pi}_{\text{ipw}} = \arg\max_{\pi\in\Pi} \hat{R}_{\text{IPW}}(\pi)$.

Evaluation Metrics. The same as [Kitagawa and Tetenov, 2018; Li *et al.*, 2023b], we adopt the metrics below.

• To assess the performance of a policy learning method, we calculate the true reward in the target dataset given by

$$\hat{R}(\hat{\pi}) = n_0^{-1}\sum_{i=1}^{n}(1-G_i)[\pi(X_i)Y_i(1) + (1-\pi(X_i))Y_i(0)].$$

Also, we define the regret as $\Delta E = \hat{R}(\pi_0^*) - \hat{R}(\hat{\pi})$, representing the difference between the reward induced by the oracle policy and that induced by the learned policy $\hat{\pi}$.

• To evaluate the accuracy of the learned policy $\hat{\pi}$, we define the policy error as

$$n_0^{-1}\sum_{i=1}^{n}(1-G_i)|\pi_0^*(X_i) - \hat{\pi}(X_i)|^2,$$

Simulated Dataset	REWARDS				POLICY ERROR			WELFARE CHANGES		
METHODS	MEAN	RI	SD	ΔE	MEAN	RI	SD	ΔW	RI	SD
Direct (baseline)	455.24	-	**0.32**	40.62	0.45	-	**0.0024**	215194.26	-	**800.41**
IPW	477.17	4.82%	29.38	18.69	0.33	-63.10%	0.3165	269741.06	25.24%	41875.37
SE	**490.22***	**7.68%**	0.85	**5.49***	**0.09***	**78.51%**	0.0062	**287185.90***	**33.43%**	1940.79

Real-World Dataset	REWARDS				POLICY ERROR			WELFARE CHANGES		
METHODS	MEAN	RI	SD	ΔE	MEAN	RI	SD	ΔW	RI	SD
Direct (baseline)	47.76	-	0.3076	13.18	0.50	-	0.0083	24482.59	-	1157.53
IPW	50.00	4.68%	1.3002	10.95	0.42	15.07%	0.0391	28468.45	16.28%	1971.35
SE	**60.50***	**26.66%**	**0.0017**	**0.45***	**0.08***	**84.35%**	**0.0002**	**47175.56***	**92.69%**	**814.84**

Note: * statistically significant results (p-value ≤ 0.05) using the paired t-test compared with the baseline.

Table 1: Comparison of estimated rewards, policy errors and welfare changes on the simulated and real-world datasets. The best results are highlighted in bold. RI refers to the relative improvement over the corresponding baseline. SD indicates standard deviation.

the mean square errors between the oracle policy π_0^* and $\hat{\pi}$.

• We define welfare change as

$$\Delta W = \sum_{i=1}^{n}[(Y_i(1) - Y_i(0))\hat{\pi}(X_i)],$$

representing the difference between the total rewards induced by $\hat{\pi}$ and the null policy $\pi \equiv 0$. A policy learning method is better when it yields a larger true reward, smaller regret, lower policy error, and greater welfare change.

Performance Comparison (RQ1 and RQ2). We average 50 independent trials of policy learning and Table 1 (top panel) reports the average rewards $\hat{R}(\hat{\pi})$, regrets ΔE, policy errors, and the welfare changes for different policy learning methods. From it, we have the following observations:

• The proposed SE method achieves the highest reward, smallest regret, lowest policy error, and largest welfare change. Compared to the Direct method (baseline), these improvements are substantial—up to 7.68% in reward, 78.51% in policy error, and 33.43% in welfare change—demonstrating the superiority of the SE method.

• The standard deviations of the Direct and SE methods are significantly smaller than that of the IPW method, indicating the instability (large variance) of the IPW method.

In summary, the SE method outperforms the competing Direct and IPW methods, owing to its desirable properties, such as double robustness and semiparametric efficiency.

Effect of varying the degree of covariate shift (RQ3). We evaluate the robustness of three policy learning approaches (Direct, IPW, SE) by varying the degree of covariate shift between the source and target datasets. The degree of covariate shift is measured by the Chebyshev distance between the two mean vectors $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ in source and target datasets. The Chebyshev distance is defined as $d(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) = \max_j |\mu_{1,j} - \mu_{2,j}|$, where $\mu_{1,j}$ and $\mu_{2,j}$ are the j-th elements of $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$, respectively. Three metrics are used to measure performance: true reward induced by the learned policy, policy error, and welfare change. Figure 1 illustrates how these metrics vary with increasing covariate shift, showing that the proposed SE method remains more stable across different levels of covariate shift and consistently achieves superior performance.

Effect of varying probability of treatment. We further evaluate the robustness of the proposed methods by varying the generation mechanism of treatments in the source dataset. We generate A in the source dataset from $Bern(\sigma(-\beta\tilde{X}_2))$, where $\sigma(\cdot)$ is the sigmoid function, and $\beta = 0, 0.1, \ldots, 0.9, 1$. The probability of receiving the treatment is 0.5 when $\beta = 0$. As β increases, the probability of an individual receiving the treatment decreases (since $\tilde{X}_2$ is always greater than 0 in our setting). Figure 2 displays how the evaluation metrics change as the proportion of individuals receiving treatment decreases, demonstrating the robustness, stability, and overall superiority of the proposed SE method.

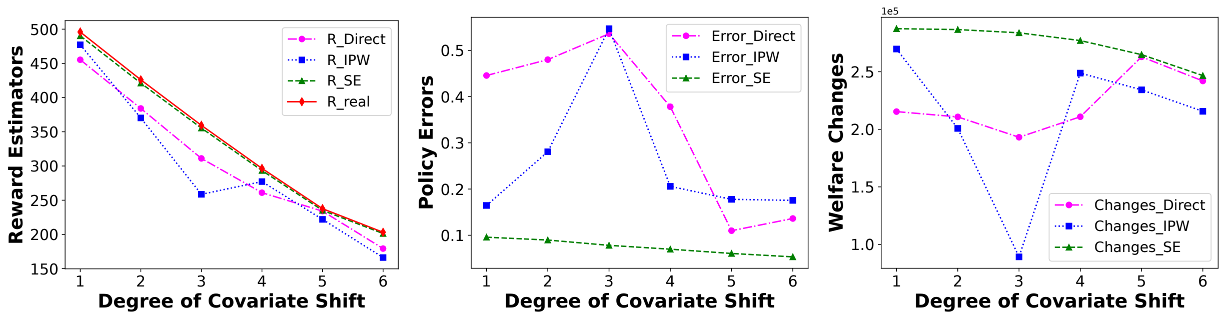
5.2 Experiments on Real-World Datasets

Datasets. The Communities and Crime dataset [Redmond, 2009] comprises 1994 records from communities in the United States, which combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. Each record includes 127 covariates, including location information (such as state and county), economic factors (such as perCapInc and HousVacant) and demographics (such as PopDens and PctBSorMore). We use records from communities in New Jersey as the source dataset and records from communities in other states as the target dataset. In addition to using the information from covariates, we simulated the treatment A and the potential outcomes $Y(1), Y(0)$. See Appendix D for the detailed data generation process.

Performance Comparison. We also average over 50 independent trials of policy learning and Table 1 (bottom panel) reports the associated results. From it, we have the following observations: (1) The proposed SE method outperforms the Direct and IPW methods across all evaluation metrics; (2) Compared to the baseline, the SE method shows substantial improvements—up to 26.66% in reward, 84.35% in policy error, and 92.69% in welfare change. These observations further demonstrate the SE method's superiority.

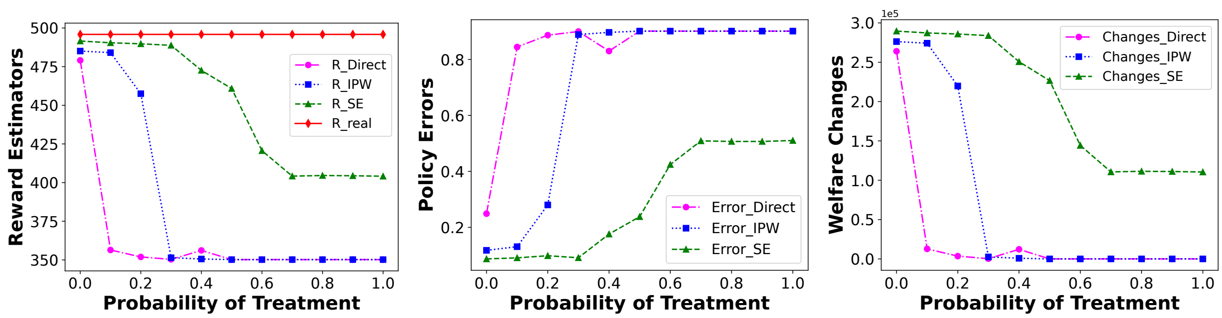
6 Related Work

Policy Learning. Policy learning seeks to identify which individuals should receive treatment to maximize the re-



(a) Reward estimators (b) Policy errors (c) Welfare changes

Figure 1: Comparison of three methods with different means of covariates in the target dataset



(a) Reward estimators (b) Policy errors (c) Welfare changes

Figure 2: Comparison of three methods with different treatments in the target dataset

ward according to their covariates [Murphy, 2003], with wide-ranging applications in fields such as precision medicine [Bertsimas *et al.*, 2017; Kitagawa and Tetenov, 2018; Kosorok and Laber, 2019], reinforcement learning [Liu *et al.*, 2021; Kwan *et al.*, 2023], and recommender systems [Ma *et al.*, 2020; Chen and Sun, 2021; Li *et al.*, 2023a]. Nevertheless, most policy learning approaches rely solely on a single labeled dataset. In scenarios where labeled data are difficult to obtain, these methods often struggle with external validity and generalizability. To address this issue, a straightforward strategy is to combine labeled and unlabeled data, utilizing techniques like transfer learning and semi-supervised learning [Kora *et al.*, 2022; Huynh *et al.*, 2022]. However, while transfer and semi-supervised learning methods are well-developed for prediction tasks, the challenge of learning optimal policies by leveraging both labeled and unlabeled data remains largely unexplored [Uehara *et al.*, 2020].

Causal Effects Generalizability. Recently, there has been growing research interest in integrating information from multiple data sources for causal inference [Hartman *et al.*, 2015; Lodi *et al.*, 2019; Colnet *et al.*, 2024; Wu *et al.*, 2025; Kallus and Mao, 2024; Wu and Mao, 2025]. However, heterogeneity in data distribution across these sources presents a significant challenge. Different heterogeneity in data distribution necessitates tailored techniques, including specific assumptions [Hotz *et al.*, 2005; Kern *et al.*, 2016b; Li *et al.*, 2023c; Hu *et al.*, 2023; Wu *et al.*, 2024c; Yang *et al.*, 2024] and structural causal models (SCM) [Pearl, 1995; Correa *et al.*, 2018; Tikka *et al.*, 2019]. While data integration has gained considerable theoretical attention, its extension to policy learning remains underexplored. In this article, we investigate how to learn the optimal policy in a target dataset by leveraging information from a source dataset.

7 Conclusion

In this article, we propose a principled policy learning method under covariate shift from a causal perspective. We detail the identifiability assumptions for the reward, derive its efficient influence function, and develop a doubly robust estimator. We establish the asymptotic properties of the proposed estimator, including consistency, normality, and semiparametric efficiency. We also provide a generalization error bound for the learned optimal policy. Extensive experiments confirm the effectiveness and reliability of our proposed method, demonstrating both theoretical and practical advantages.

Acknowledgments

This work was supported by the National Key Research and Development Program (No. 2024YFD2100700), the National Natural Science Foundation of China (No. 12301370), the BTBU Digital Business Platform Project by BMEC, the Beijing Key Laboratory of Applied Statistics and Digital Regulation, and the gift funding from ByteDance Research.

References

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

[Bertsimas *et al.*, 2017] 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.

[Buchanan *et al.*, 2018] Ashley L Buchanan, Michael G Hudgens, Stephen R Cole, and et al. Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(4):1193–1209, 2018.

[Chen and Sun, 2021] Keyu Chen and Shiliang Sun. Knowledge-based conversational recommender systems enhanced by dialogue policy learning. In *IJCKG*, pages 10–18, 2021.

[Chernozhukov *et al.*, 2018] V. Chernozhukov, D. Chetverikov, M. Demirer, and et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, 2018.

[Colnet *et al.*, 2024] Bénédicte Colnet, Imke Mayer, Guanhua Chen, and et al. Causal inference methods for combining randomized trials and observational studies: a review. *Statistical Science*, 39(1):165–191, 2024.

[Correa *et al.*, 2018] Juan Correa, Jin Tian, and Elias Bareinboim. Generalized adjustment under confounding and selection biases. In *AAAI*, pages 6355–6342, 2018.

[Dahabreh *et al.*, 2020] Issa J. Dahabreh, Sarah E. Robertson, Jon A. Steingrimsson, Elizabeth A. Stuart, and Miguel A. Hernán. Extending inferences from a randomized trial to a new target population. *Statistics in Medicine*, 39:1999–2014, 2020.

[Gama *et al.*, 2014] João Gama, Indrė Žliobaitė, Albert Bifet, and et al. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

[Hartman *et al.*, 2015] Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society Series A*, 178(3):757–778, 2015.

[Hernán and Robins, 2020] M.A. Hernán and J. M. Robins. *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC, 2020.

[Hotz *et al.*, 2005] V Joseph Hotz, Guido W Imbens, and Julie H Mortimer. Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270, 2005.

[Hu *et al.*, 2023] W. Hu, X.H. Zhou, and P. Wu. Identification and estimation of treatment effects on long-term outcomes in clinical trials with external observational data. *Statistica Sinica*, 2023.

[Huynh *et al.*, 2022] Tri Huynh, Aiden Nibali, and Zhen He. Semi-supervised learning for medical image classification using imbalanced training data. *Computer methods and programs in biomedicine*, 216:106628, 2022.

[Imbens *et al.*, 2024] Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. Long-term causal inference under persistent confounding via data combination. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024.

[Kallus and Mao, 2024] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae099, 2024.

[Kern *et al.*, 2016a] Holger L Kern, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127, 2016.

[Kern *et al.*, 2016b] Holger L Kern, Elizabeth A Stuart, Jennifer Hill, and Donald P Green. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of research on educational effectiveness*, 9(1):103–127, 2016.

[Kitagawa and Tetenov, 2018] Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

[Kora *et al.*, 2022] Padmavathi Kora, Chui Ping Ooi, Oliver Faust, and et al. Transfer learning techniques for medical image analysis: A review. *Biocybernetics and Biomedical Engineering*, 42(1):79–107, 2022.

[Kosorok and Laber, 2019] Michael R. Kosorok and Eric B. Laber. Precision medicine. *Annual Review of Statistics and Its Application*, 6:263–86, 2019.

[Kwan *et al.*, 2023] Wai-Chung Kwan, Hong-Ru Wang, Hui-Min Wang, and Kam-Fai Wong. A survey on recent advances and challenges in reinforcement learning methods for task-oriented dialogue policy learning. *Machine Intelligence Research*, 20(3):318–334, 2023.

[Lesko *et al.*, 2017] Catherine R Lesko, Ashley L Buchanan, Daniel Westreich, Jessie K Edwards, Michael G Hudgens, and Stephen R Cole. Generalizing study results: a potential outcomes perspective. *Epidemiology*, 28(4):553–561, 2017.

[Li *et al.*, 2022a] Fan Li, Ashley L. Buchanan, and Stephen R. Cole. Generalizing trial evidence to target populations in non-nested designs: Applications to aids clinical trials. *The Journal of the Royal Statistical Society, Series C*, 71(3):669–697, 2022.

[Li *et al.*, 2022b] Haoxuan Li, Yan Lyu, Chunyuan Zheng, and Peng Wu. Tdr-cl: Targeted doubly robust collaborative learning for debiased recommendations. In *ICLR*, 2022.

[Li *et al.*, 2023a] Haoxuan Li, Yanghao Xiao, Chunyuan Zheng, and Peng Wu. Balancing unobserved confounding with a few unbiased ratings in debiased recommendations. In *WWW*, 2023.

[Li *et al.*, 2023b] Haoxuan Li, Chunyuan Zheng, Yixiao Cao, Zhi Geng, Yue Liu, and Peng Wu. Trustworthy policy learning under the counterfactual no-harm criterion. In *ICML*, pages 20575–20598, 2023.

[Li *et al.*, 2023c] Xinyu Li, Wang Miao, Fang Lu, and Xiao-Hua Zhou. Improving efficiency of inference in clinical trials with external control data. *Biometrics*, 79(1):394–403, 2023.

[Liu *et al.*, 2021] Yongshuai Liu, Avishai Halev, and Xin Liu. Policy learning with constraints in model-free reinforcement learning: A survey. In *IJCAI*, 2021.

[Lodi *et al.*, 2019] Sara Lodi, Andrew Phillips, Jens Lundgren, and et al. Effect estimates in randomized trials and observational studies: comparing apples with apples. *American journal of epidemiology*, 188(8):1569–1577, 2019.

[Ma *et al.*, 2020] Jiaqi Ma, Zhe Zhao, Xinyang Yi, and et al. Off-policy learning in two-stage recommender systems. In *WWW*, 2020.

[Murphy, 2003] Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

[Newey, 1990] Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

[Pearl, 1995] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[Pesciullesi *et al.*, 2020] Giorgio Pesciullesi, Philippe Schwaller, Teodoro Laino, and Jean-Louis Reymond. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature Communications*, 11(1):4874, 2020.

[Redmond, 2009] Michael Redmond. Communities and crime. UCI Machine Learning Repository, doi.org/10.24432/C53W3X, 2009.

[Rosenbaum and Rubin, 1983] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal. *Biometric*, 70:41–55, 1983.

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

[Splawa-Neyman, 1990] Jerzy Splawa-Neyman. On the application of probability theory to agricultural experiments. essay on. *Statistical Science*, 5(4,465-480), 1990.

[Stuart *et al.*, 2011] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2):369–386, 2011.

[Sun *et al.*, 2020] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, and et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

[Tikka *et al.*, 2019] Santtu Tikka, Antti Hyttinen, and J. Karvanen. Causal effect identification from multiple incomplete data sources: A general search-based approach. *arXiv:1902.01073*, 2019.

[Tsiatis, 2006] Anastasios A Tsiatis. *Semiparametric theory and missing data*, volume 4. Springer, 2006.

[Uehara *et al.*, 2020] Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *NeurIPS*, 33:49–61, 2020.

[Wang *et al.*, 2018] Anna X Wang, Caelin Tran, Nikhil Desai, David Lobell, and Stefano Ermon. Deep transfer learning for crop yield prediction with remote sensing data. In *SIGCAS*, 2018.

[Wang *et al.*, 2020] Liansheng Wang, Yudi Jiao, Ying Qiao, Nianyin Zeng, and Rongshan Yu. A novel approach combined transfer learning and deep learning to predict tmb from histology image. *Pattern Recognition Letters*, 135:244–248, 2020.

[Wu and Mao, 2025] Peng Wu and Xiaojie Mao. The promises of multiple experiments: Identifying joint distribution of potential outcomes. *arXiv:2504.20470*, 2025.

[Wu *et al.*, 2022] Peng Wu, Haoxuan Li, Yuhao Deng, Wenjie Hu, Quanyu Dai, Zhenhua Dong, Jie Sun, Rui Zhang, and Xiao-Hua Zhou. On the opportunity of causal learning in recommendation systems: Foundation, estimation, prediction and challenges. In *IJCAI*, 2022.

[Wu *et al.*, 2024a] Peng Wu, Peng Ding, Zhi Geng, and Yue Li. Quantifying individual risk for binary outcome: Bounds and inference. *arXiv:2402.10537*, 2024.

[Wu *et al.*, 2024b] Peng Wu, Shasha Han, Xingwei Tong, and Runze Li. Propensity score regression for causal inference with treatment heterogeneity. *Statistica Sinica*, 34:747–769, 2024.

[Wu *et al.*, 2024c] Peng Wu, Ziyu Shen, Feng Xie, Zhongyao Wang, Chunchen Liu, and Yan Zeng. Policy learning for balancing short-term and long-term rewards. In *ICML*, 2024.

[Wu *et al.*, 2025] Peng Wu, Shanshan Luo, and Zhi Geng. On the comparative analysis of average treatment effects estimation via data combination. *Journal of the American Statistical Association*, 2025.

[Yang *et al.*, 2024] Qinwei Yang, Xueqing Liu, Yan Zeng, Ruocheng Guo, Yang Liu, and Peng Wu. Learning the optimal policy for balancing short-term and long-term rewards. *NeurIPS*, 2024.

[Zhuang *et al.*, 2020] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.