Off-Policy Evaluation and Learning for Survival Outcomes Under Censoring

Kohsuke Kubota[1] , **Mitsuhiro Takahashi**[1] , **Yuta Saito**[2]
[1]NTT DOCOMO, INC.
[2]Hanjuku-kaso, Co, Ltd.
{kousuke.kubota.xt, mitsuhiro.takahashi.zp}@nttdocomo.com, saito@hanjuku-kaso.com

Abstract

Optimizing survival outcomes, such as patient survival or customer retention, is a critical objective in data-driven decision-making. Off-Policy Evaluation (OPE) provides a powerful framework for assessing such decision-making policies using logged data alone, without the need for costly or risky online experiments in high-stakes applications. However, typical estimators are not designed to handle right-censored survival outcomes, as they ignore unobserved survival times beyond the censoring time, leading to systematic underestimation of the true policy performance. To address this issue, we propose a novel framework for OPE and Off-Policy Learning (OPL) tailored for survival outcomes under censoring. Specifically, we introduce IPCW-IPS and IPCW-DR, which employ the Inverse Probability of Censoring Weighting technique to explicitly deal with censoring bias. We theoretically establish that our estimators are unbiased and that IPCW-DR achieves double robustness, ensuring consistency if either the propensity score or the outcome model is correct. Furthermore, we extend this framework to constrained OPL to optimize policy value under budget constraints. We demonstrate the effectiveness of our proposed methods through simulation studies and illustrate their practical impacts using public real-world data for both evaluation and learning tasks.

1 Introduction

Evaluating and implementing policies (treatment strategies) aimed at maximizing survival outcomes presents a critical challenge in decision-making in diverse real-world domains, ranging from extending survival in healthcare [Geng *et al.*, 2015; Murphy, 2003] to reducing customer churn in marketing [Ascarza *et al.*, 2018]. The core focus of these applications is not limited to evaluating short-term effects, but encompasses a broader understanding of a policy's impact on the long-term survival trajectory.

Off-policy evaluation (OPE) provides a powerful framework to estimate the performance of new decision-making policies using only offline logged data, which is crucial in domains where online experiments are costly or impractical [Wang *et al.*, 2017]. Conventional approaches, such as Inverse Propensity Scoring (IPS) [Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983] and Doubly Robust (DR) [Dudík *et al.*, 2011; Robins and Rotnitzky, 1995], have been widely adopted for evaluating policies using offline logged data collected under different (logging) policies.

However, despite their broad applicability, **these typical estimators are not designed to handle right-censored survival outcomes**. This is because these estimators are designed primarily for immediate rewards or outcomes defined over a fixed time horizon. They also implicitly assume that all outcomes are fully observed, which is directly violated by right-censoring, a defining characteristic of survival data. A naive application of these estimators cannot distinguish censoring from true event occurrence, and therefore wrongly treats observed times as final event times. This induces systematic underestimation bias because typical methods fail to account for potential survival beyond the censoring point.

To address this issue, we propose a novel framework for off-policy evaluation and learning under practical constraints designed for censored survival outcomes. Specifically, we develop new statistical estimators, namely **Inverse Probability of Censoring Weighted-Inverse Propensity Scoring (IPCW-IPS)** and **IPCW-Doubly Robust (IPCW-DR)**, which explicitly consider censoring induced bias. IPCW-IPS additionally re-weights logged data via estimated censoring probabilities beyond the typical importance weighting regarding policy probabilities. IPCW-DR further incorporates outcome modeling, which enjoys the doubly robust property with respect to the propensity score and the outcome model. Consequently, it ensures reliability in practical applications and yields lower variance compared to IPCW-IPS. Furthermore, we extend our framework to the problem of constrained Off-Policy Learning (OPL), enabling policy optimization for survival outcomes subject to practical budget constraints.

Finally, we empirically demonstrate the effectiveness of our proposed framework. The simulation results demonstrate that our proposed estimators, particularly IPCW-DR, achieve superior accuracy than standard OPE estimators. Moreover, the application to the Starbucks Customer Reward Program dataset highlights the practical effectiveness of our approach in complex scenarios. Specifically, we demonstrate how our framework can better optimize promotional offers to mini-

mize the customer purchase cycle under budget constraints.

Our key contributions can be summarized as follows.

- We develop censoring aware estimators for off-policy evaluation of survival outcomes leveraging the technique of inverse probability of censoring weighting and provide their theoretical advantages.
- We introduce an iterative algorithm for off-policy learning with practical constraints that targets optimization of survival outcomes.
- We demonstrate the effectiveness of the proposed methods in terms of both evaluation and learning tasks, through simulation studies and a real world application.

Due to space limitations, we provide a comprehensive review of related work in Appendix A.

2 Preliminaries

2.1 Problem Formulation

Here we consider the OPE problem for an evaluation policy π_e under right-censoring, using data collected by a logging policy π_0. Let $x \in \mathcal{X}$ be the context vector representing individual characteristics, $a \in \mathcal{A}$ be the discrete action selected according to a policy $\pi(a \mid x)$. The logged dataset $\mathcal{D} = \{(x_i, a_i, T_i, r_i)\}_{i=1}^n$ is collected under a logging policy π_0, where the underlying random variables are generated as

$$(x_i, a_i, L_i, C_i) \sim p(x)\pi_0(a_i \mid x_i)p(L_i, C_i \mid x_i, a_i). \quad (1)$$

Here, L_i and C_i denote the true (latent) survival time and the censoring time, respectively. The observed data consist only of the observed time $T_i = \min\{L_i, C_i\}$ and an event indicator $r_i = \mathbb{I}\{L_i \leq C_i\}$. Thus, for censored individuals ($r_i = 0$), we only know that their true survival time exceeds the observed time T_i. Our goal is to accurately estimate the performance (policy value) of an evaluation policy π_e regarding the true survival time using only the logged data $\mathcal{D}$.

Policy Value Definition. A key quantity of interest in our work is the expected survival probability at an arbitrary time t. Let $S(x, a, t) := P(L > t \mid x, a)$ be the survival function given context x and action a. The expected survival probability under policy π_e at time t, as a type of policy value, is defined as

$$V(\pi_e, t) := \mathbb{E}_{p(x)\pi_e(a|x)}[S(x, a, t)]. \quad (2)$$

Another important quantity of interest is the Restricted Mean Survival Time (RMST) [Uno *et al.*, 2014], which is the expected survival time up to a pre-specified finite time horizon τ. This version of the policy value is defined as follows.

$$V^\tau(\pi_e) := \int_0^\tau V(\pi_e, t)dt. \quad (3)$$

Note that τ is not a tunable hyperparameter but a domein-specific objective that defines the evaluation metric itself.

Assumptions. We begin by outlining some regularity assumptions necessary for analyzing the theoretical properties of OPE estimators under censoring. First, the following common support assumption, standard in the OPE literature [Sachdeva *et al.*, 2020], ensures that the evaluation policy does not place a positive probability on actions that the logging policy never takes. It is formally stated as follows.

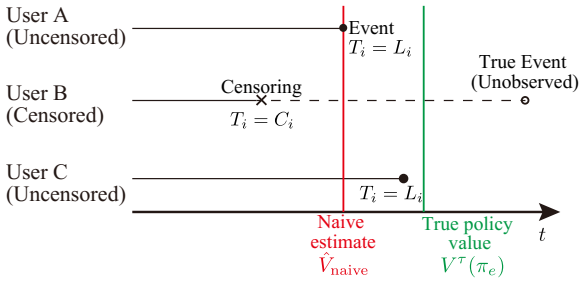


Figure 1: Illustration of the underestimation bias inherent in naive OPE. While the true event time for User B is unobserved due to censoring, naive estimators treat the censoring time as the event time, leading to a systematic underestimation of the survival time.

Assumption 2.1 (Common Support)**.** If $\pi_e(a \mid x) > 0$, then $\pi_0(a \mid x) > 0$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.

We also introduce an assumption regarding censoring.

Assumption 2.2 (Conditional Independent Censoring)**.** Given the context x and action a, the true survival time L and the censoring time C are conditionally independent

$$L \perp C \mid (x, a).$$

This assumption serves as the theoretical basis for conventional methods in survival analysis, such as the Kaplan-Meier estimator and the Cox proportional hazards model [Fleming and Harrington, 2013].

2.2 Naive OPE Under Censoring

A naive approach to addressing this estimation problem is to directly apply conventional estimators in OPE, such as IPS or DR, to the logged dataset ignoring the censoring issue [Saito and Joachims, 2021]. These standard estimators are typically designed for fully observed outcomes and, when naively applied to survival data, they treat the observed time T_i as the true event time regardless of the censoring indicator r_i. Specifically, by treating the observed indicator $\mathbb{I}\{T_i > t\}$ as the reward, the naive IPS estimator is given by

$$\hat{V}_{\text{IPS}}(\pi_e, t; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^n \frac{\pi_e(a_i \mid x_i)}{\pi_0(a_i \mid x_i)}\mathbb{I}\{T_i > t\}.$$

We can also define the naive DR estimator by combining IPS with an outcome model $\hat{S}(x, a, t)$ trained on the observed times.[1] Since these estimators are not designed to account for the censoring mechanism, they suffer from systematic underestimation bias, explicitly characterized as follows.

Proposition 2.3. Under the presence of censoring (that is, $P(L > C) > 0$) and Assumptions 2.1 and 2.2, the naive IPS and DR estimators are biased as

$$\begin{aligned}&\text{Bias}[\hat{V}_{\text{IPS}}] = \text{Bias}[\hat{V}_{\text{DR}}] \\ &:= \mathbb{E}_{p(x)\pi_e(a|x)}[S(x, a, t)(G(t \mid x, a) - 1)],\end{aligned}$$

where $G(t \mid x, a) := P(C > t \mid x, a)$ is the censoring survival function.

[1] We formally define naive DR in Eq. (D.1) in Appendix D.

The proof is provided in Appendix D.1. Since $G(t \mid x, a) \leq 1$, the bias term is always non-positive, meaning that these naive estimators systematically underestimate the policy value, as visually depicted in Figure 1. We can also see that the bias increases under severe censoring conditions, such as short follow-up periods or low event rates. In such cases, $G(t \mid x, a)$ becomes significantly smaller than one, resulting in a more severe underestimation bias.

3 Proposed Approach: Censoring-Aware Off-Policy Evaluation and Learning

To obtain unbiased estimates of expected survival probability or RMST, it is essential to incorporate the censoring mechanism directly into the OPE framework. This section proposes two new estimators that properly account for censoring and yield unbiased or at least bias-reduced policy value estimates.

3.1 Unbiased Estimation of Survival Policy Value

Our first proposed estimator, namely **Inverse Probability of Censoring Weighted-Inverse Propensity Scoring (IPCW-IPS)**, directly corrects for the censoring bias. As shown in the derivation of Proposition 2.3 (see Appendix D.1), the expectation of the naive outcome $\mathbb{I}\{T_i > t\}$ yields the confounded quantity $S(x_i, a_i, t)G(t \mid x_i, a_i)$ rather than the true survival probability $S(x_i, a_i, t)$. To deal with this bias, we propose to employ **the Inverse Probability of Censoring Weighting (IPCW)** technique [Robins *et al.*, 1995; Robins and Finkelstein, 2000]. Specifically, we define IPCW-IPS by additionally weighting each observed indicator $\mathbb{I}\{T_i > t\}$ with the inverse of the estimated probability that the sample remains uncensored at time t, denoted by $\hat{G}(t \mid x_i, a_i)$ as

$$\hat{V}_{\text{IPCW-IPS}}(\pi_e, t; \mathcal{D}) \coloneqq \frac{1}{n}\sum_{i=1}^{n} w(x_i, a_i)\frac{\mathbb{I}\{T_i > t\}}{\hat{G}(t \mid x_i, a_i)}. \quad (4)$$

where $w(x_i, a_i) \coloneqq \pi_e(a_i \mid x_i)/\pi_0(a_i \mid x_i)$.

While IPCW-IPS is designed to be unbiased, it may be susceptible to high variance particularly when censoring is heavy (i.e., when $\hat{G}(t \mid x, a)$ is small), as the inverse probability weight becomes large. To mitigate this variance inflation and improve estimation reliability further, we extend IPCW-IPS to **IPCW-Doubly Robust (IPCW-DR)** defined as

$$\begin{aligned}&\hat{V}_{\text{IPCW-DR}}(\pi_e, t; \mathcal{D}) \\ &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\Bigg(w(x_i, a_i)\left(\frac{\mathbb{I}\{T_i > t\}}{\hat{G}(t \mid x_i, a_i)} - \hat{S}(x_i, a_i, t)\right) \\ &\qquad + \mathbb{E}_{\pi_e(a|x)}[\hat{S}(x_i, a, t)]\Bigg).\end{aligned} \quad (5)$$

The IPCW-DR estimator extends the standard DR framework to the setting of censored survival outcomes. Structurally, this estimator combines a model-based estimate derived via the direct method (DM) [Beygelzimer and Langford, 2009], with a bias-correction term based on the weighted residual between the IPCW-adjusted reward and the outcome model.

We first establish the unbiasedness of our proposed estimators. The following theorem guarantees that both IPCW-IPS and IPCW-DR provide unbiased estimates of the policy value, with IPCW-DR further enjoying double robustness.

Theorem 3.1. Under Assumptions 2.1 and 2.2, provided that the censoring model $G(t \mid x, a)$ is correctly specified

1. The IPCW-IPS estimator is unbiased for the expected survival probability $V(\pi_e, t)$ if the propensity score $\pi_0(a \mid x)$ is correctly specified.
2. The IPCW-DR estimator is unbiased for $V(\pi_e, t)$ if either the propensity score $\pi_0(a \mid x)$ or the outcome model $\hat{S}(x, a, t)$ is correctly specified.

The proofs are provided in Appendix D.2.

While unbiasedness is a desirable statistical property for OPE, the variance of an estimator is an equally critical component of the MSE. The following theorem explicitly characterizes the variance of our estimators and theoretically establishes the variance reduction property of IPCW-DR.

Theorem 3.2. Under Assumptions 2.1 and 2.2, provided that the propensity score $\pi_0(a \mid x)$, the censoring model $G(t \mid x, a)$, and the outcome model $S(x, a, t)$ are correctly specified, the variance of the IPCW-IPS estimator is

$$\begin{aligned}n\mathbb{V}[\hat{V}_{\text{IPCW-IPS}}] = \mathbb{E}_{p(x)\pi_0(a|x)}\left[(w(x,a))^2\sigma^2(x,a,t)\right] \\ + \mathbb{E}_{p(x)}[\mathbb{V}_{\pi_0(a|x)}[w(x,a)S(x,a,t)]] \\ + \mathbb{V}_{p(x)}[S^{\pi_e}(x,t)],\end{aligned}$$

where $\sigma^2(x, a, t) \coloneqq S(x, a, t)/G(t|x, a) - S(x, a, t)^2$ represents the variance inflation due to censoring. Furthermore, the variance of the IPCW-DR estimator is

$$\begin{aligned}n\mathbb{V}[\hat{V}_{\text{IPCW-DR}}] = \mathbb{E}_{p(x)\pi_0(a|x)}\left[w(x,a)^2\sigma^2(x,a,t)\right] \\ + \mathbb{V}_{p(x)}[S^{\pi_e}(x,t)].\end{aligned}$$

Consequently, IPCW-DR achieves a variance reduction of $\mathbb{E}_{p(x)}[\mathbb{V}_{\pi_0(a|x)}[w(x, a)S(x, a, t)]] \geq 0$, ensuring $\mathbb{V}[\hat{V}_{\text{IPCW-DR}}] \leq \mathbb{V}[\hat{V}_{\text{IPCW-IPS}}]$. The proofs are provided in Appendix D.3.

This theorem provides a formal justification for the instability of IPCW-IPS under heavy censoring. Specifically, the term $\sigma^2(x, a, t)$ reveals that the variance is inversely proportional to the censoring survival function $G(t \mid x, a)$. Thus, when censoring is heavy (i.e., $G(t \mid x, a)$ is small), IPCW-IPS can suffer from significant variance. In contrast, IPCW-DR consistently reduces variance by leveraging the outcome model, making it a more reliable choice in practice.

While we have so far focused on estimating the pointwise survival probability $V(\pi_e, t)$, our framework naturally extends to the estimation of the RMST $V^{\tau}(\pi_e)$ defined in Eq. (3). Since the RMST is the integral of the survival probability, we can construct unbiased estimators for RMST by integrating IPCW-IPS and IPCW-DR over the time horizon $[0, \tau]$. The detailed formulations and theoretical guarantees for these RMST estimators are provided in Appendix B.1.

3.2 Off-Policy Learning for Survival Maximization

In addition to the problem of OPE, we can formulate the problem of learning an optimal policy to maximize the survival

outcome under censoring. Let Π be a predefined policy class such as neural networks. The goal of off-policy learning is to find a policy $\pi_\theta \in \Pi$ that maximizes the policy value based on the logged data $\mathcal{D}$. Due to space limitations, we will focus on the policy learning objective for the expected survival probability $V(\pi, t)$.[2] Using our proposed estimators, we can define the policy learning objectives as $\theta \coloneqq \arg\max_\theta \hat{V}(\pi_\theta, t; \mathcal{D})$.

A typical approach to solve this optimization problem is to use gradient-based approach [Huang and Jiang, 2020; Saito *et al.*, 2024], which updates the policy parameter via iterative gradient ascent as $\theta_{k+1} \leftarrow \theta_k + \nabla_\theta V(\pi_\theta, t)$. Because the true gradient

$$\nabla_\theta V(\pi_\theta, t) = \mathbb{E}_{p(x)\pi_\theta(a|x)}\left[S(x, a, t)\nabla_\theta \log \pi_\theta(a \mid x)\right]. \tag{6}$$

is unknown, we need to estimate it from the logged dataset $\mathcal{D}$ where we can apply our proposed estimators. For example, IPCW-IPS for the policy gradient can be constructed by replacing the true survival probability $S(x, a, t)$ in Eq. (6) with the IPCW-corrected outcome $\mathbb{I}\{T > t\}/\hat{G}(t \mid x, a)$ as

$$\begin{aligned}&\widehat{\nabla_\theta V}_{\text{IPCW-IPS}}(\pi_\theta, t; \mathcal{D}) \\ &\coloneqq \frac{1}{n}\sum_{i=1}^{n} w'(x, a)\frac{\mathbb{I}\{T_i > t\}}{\hat{G}(t \mid x_i, a_i)}\nabla_\theta \log \pi_\theta(a_i \mid x_i),\end{aligned} \tag{7}$$

where $w'(x, a) \coloneqq \pi_\theta(a \mid x)/\pi_0(a \mid x)$. Similarly, the IPCW-DR estimator for the policy gradient can be constructed by replacing the true survival probability $S(x, a, t)$ in Eq. (6) with the IPCW-DR outcome model as follows.

$$\begin{aligned}&\widehat{\nabla_\theta V}_{\text{IPCW-DR}}(\pi_\theta, t; \mathcal{D}) \\ &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\Bigg(w'(x_i, a_i)\hat{\delta}(x_i, a_i, t)\nabla_\theta \log \pi_\theta(a_i \mid x_i) \\ &\quad + \mathbb{E}_{\pi_\theta(a|x)}[\hat{S}(x_i, a, t)\nabla_\theta \log \pi_\theta(a_i \mid x_i)]\Bigg),\end{aligned} \tag{8}$$

where let $\hat{\delta}(x, a, t) \coloneqq \frac{\mathbb{I}\{T>t\}}{\hat{G}(t|x,a)} - \hat{S}(x, a, t)$.

These estimators are unbiased against the true policy gradient $\nabla_\theta V(\pi_\theta, t)$ under the same conditions as Theorem 3.1. IPCW-DR for the policy gradient is superior to IPCW-IPS in terms of variance under the same conditions as in Section 3.1, leading to a more efficient policy learning.

Extension to Constrained Off-Policy Learning. We can further extend our OPL framework to optimize survival outcomes *under a budget constraint*, a common requirement in many real-world applications [Cai *et al.*, 2023; Zhang *et al.*, 2021; Zhou *et al.*, 2023].

Let $c(x, a)$ be the observed cost for a given (x, a). We aim to find π_θ that maximizes $V(\pi_\theta, t)$ subject to $C(\pi_\theta) \coloneqq \mathbb{E}_{p(x)\pi_\theta(a|x)}[c(x, a)] \leq B$. This constrained optimization problem can be transformed using Lagrangian duality [Geoffrion, 1974] into a min-max problem as $\min_{\lambda\geq 0}\max_\theta L(\pi_\theta, t, \lambda)$, where $L(\pi_\theta, t, \lambda) = V(\pi_\theta, t) - \lambda(C(\pi_\theta) - B)$. To solve this problem, we can employ an iterative algorithm that alternates between updating θ (gradient ascent on L) and the Lagrangian multiplier λ. The true gradient of the Lagrangian is

$$\nabla_\theta L(\pi_\theta, t, \lambda) = \mathbb{E}_{p(x)\pi_\theta(a|x)}[R_\lambda(x, a, t)\nabla_\theta \log \pi_\theta(a \mid x)], \tag{9}$$

where $R_\lambda(x, a, t) = S(x, a, t) - \lambda c(x, a)$. Since $c(x, a)$ is fully observed and independent of censoring, we apply the IPCW correction exclusively to the survival component $S(x, a, t)$. Thus, we can naturally obtain the IPCW-IPS and IPCW-DR estimators for $\nabla_\theta L$ by simply replacing the reward term in Eqs. (7) and (8) with the Lagrangian reward $\left(\mathbb{I}\{T_i > t\}/\hat{G}(t \mid x, a) - \lambda c(x_i, a_i)\right)$ and its model-augmented counterpart, respectively. The unbiasedness of these estimators against the true Lagrangian policy gradient is proved in Appendix D.7. This approach enables stable offline policy learning that rigorously satisfies budget constraints while handling the censoring issue.

4 Simulation Studies

This section demonstrates the effectiveness of our proposed methods in controlled simulated environments, which allow us to systematically analyze their behavior under known and carefully varied experimental conditions.

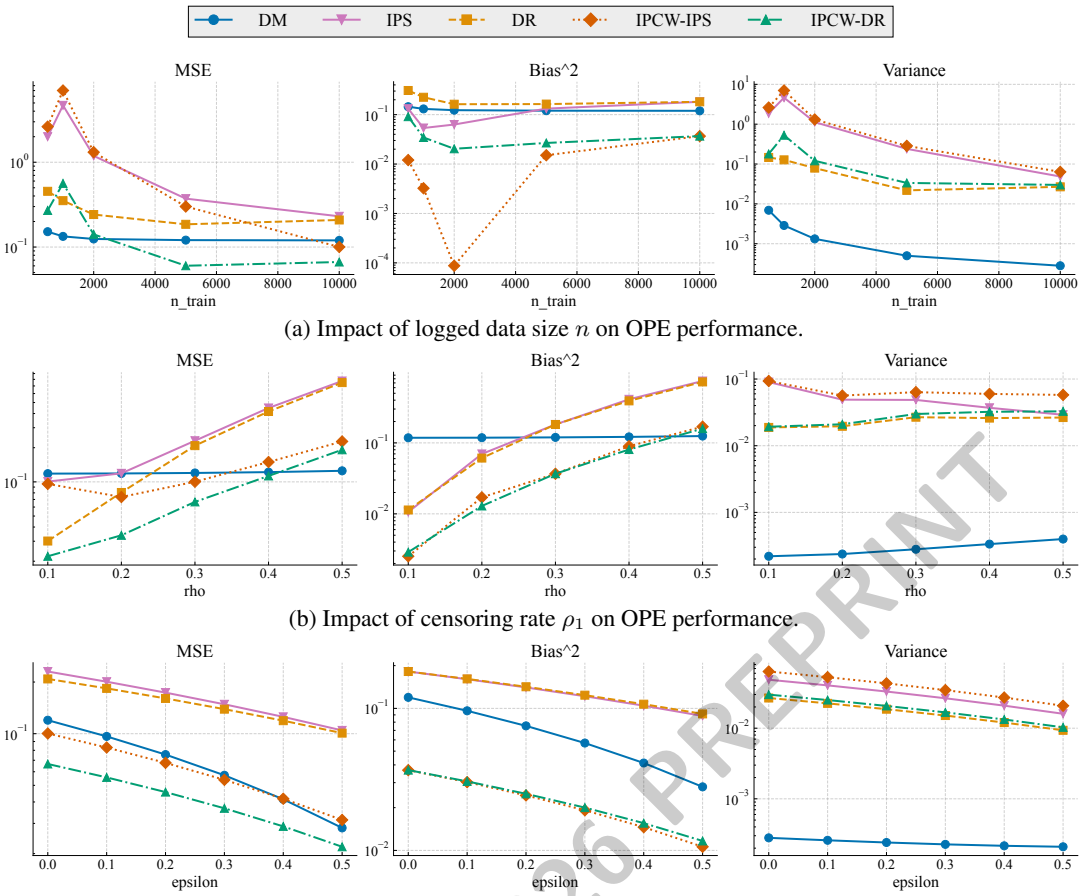
Data Generating Process. We construct the logged dataset $\mathcal{D} = \{(x_i, a_i, T_i, r_i)\}_{i=1}^n$ to simulate OPE/L for survival outcomes under right-censoring. The data-generating process samples context vectors x, selects actions a using a logging policy π_0, and generates latent survival times L along with censoring times C. This design enables direct control over the degree of censoring. Detailed descriptions of the data-generating process are provided in Appendix C.1.

OPE Performance Comparison. In the OPE experiments, we compare IPCW-IPS and IPCW-DR with the naive baselines including DM, IPS and DR, all of which ignore the censoring mechanism. We investigate the OPE accuracy under varying experimental conditions defined by three key factors: (i) training sample size n, (ii) censoring rate ρ_1, and (iii) policy divergence ϵ. First, we vary $n \in \{500, 1,000, 2,000, 5,000, 10,000\}$ to evaluate asymptotic behavior of the estimators. Second, we vary the censoring rate $\rho_1 = P(L > C)$ from $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ to test robustness against heavy censoring. We obtain each target ρ_1 by numerically adjusting the censoring distribution parameters, as described in Appendix C.1. Third, to analyze the impact of the discrepancy between the logging and evaluation policies, we define the evaluation policy π_e using the ϵ-greedy rule:

$$\pi_e(a \mid x; \epsilon) = (1 - \epsilon) \cdot \mathbb{I}\{a = \text{argmax}_{a'} V^\tau(x, a')\} + \frac{\epsilon}{|\mathcal{A}|}.$$

We configure the logging and evaluation policies in accordance with the synthetic experimental protocols in prior relevant studies [Kiyohara *et al.*, 2024; Peng *et al.*, 2023]. By varying $\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, we examine how the divergence between the logging policy π_0 and the target policy π_e affects estimation variance.

[2] Note that the formulation for maximizing RMST proceeds analogously by integrating the objective over the time horizon.



(a) Impact of logged data size n on OPE performance.

(b) Impact of censoring rate ρ_1 on OPE performance.

(c) Impact of policy divergence ϵ on OPE performance.

Figure 2: OPE performance comparison across key experimental factors. The figure illustrates how MSE changes as we vary (a) the logged data size n, (b) the censoring rate ρ_1, and (c) the divergence between the logging and evaluation policies ϵ.

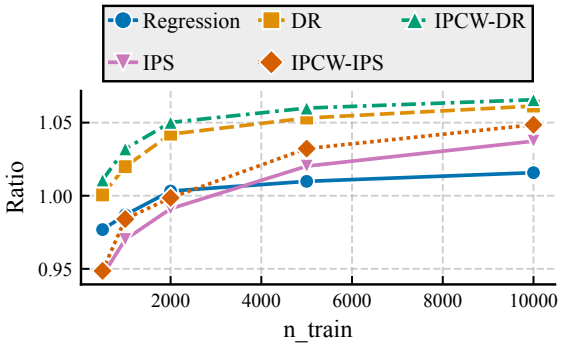
We evaluate and compare the MSE, squared bias, and variance of the estimators over 100 independent trials. In practice, constructing our estimators requires estimating nuisance components such as the censoring model $\hat{G}(t \mid x, a)$. To ensure computational stability under heavy censoring, we estimate this using Cox proportional hazards models with L_2 regularization. Additional implementation details are provided in Appendix C.2 and our code to reproduce the results is shared as a supplementary material.

OPL Performance Comparison. We evaluate the performance of OPL methods in learning a policy π_θ, parameterized by a Multi-Layer Perceptron (MLP) [LeCun *et al.*, 2015], to maximize the expected RMST. Our comparison includes two learning paradigms, including a regression-based approach [Jeunen and Goethals, 2021], which selects actions using RMST predictions from an outcome model, and policy gradient–based approaches [Huang and Jiang, 2020], which optimize differentiable OPE estimators directly as objective functions. We analyze the learning performance under varying training sample sizes n and censoring rates ρ_1. We evaluate the learned policies on an independent large-scale test set using the improvement ratio over the logging policy, defined as $V^\tau(\hat{\pi})/V^\tau(\pi_0)$. Additional implementation details are provided in Appendix C.3.

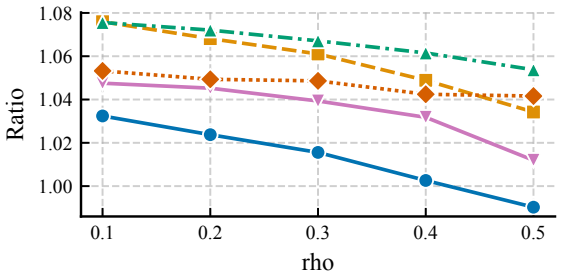
Simulation Results. Figure 2 illustrates the OPE performance comparison results across varying logged data sizes n, censoring rates ρ_1, and policy divergences ϵ.

Figure 2a interestingly shows that naive IPS and DR do not reduce squared bias even when we increase the sample size, and their MSE remains high. This observation directly reflects Proposition 2.3, which explains how censoring induces systematic underestimation for those typically unbiased estimators. In contrast, IPCW-IPS and IPCW-DR steadily lower both MSE and squared bias as n grows, even in the existence of severe censoring, and these trends align with the unbiasedness guarantees in Theorem 3.1. DM achieves the smallest variance, but it retains a non-vanishing and substantial bias because the outcome model alone cannot fully capture the underlying survival dynamics. IPCW-DR corrects this bias while maintaining a strong bias–variance balance.

Next, in Figure 2b, we observe that the performance of naive estimators deteriorates rapidly as the censoring rate increases. Our IPCW-based estimators, however, continue to produce much lower MSE even under heavy censoring.



(a) Impact of training sample size n on OPL performance.



(b) Impact of censoring rate ρ_1 on OPL performance.

Figure 3: OPL performance comparison. The improvement ratio $V^{\tau}(\hat{\pi})/V^{\tau}(\pi_0)$ quantifies how effectively each method learns a superior policy compared to the logging policy optimalities.

IPCW-DR, in particular, avoids the variance inflation that often appears in high-censoring settings, which highlights its robustness. Moreover, IPCW-DR remains highly stable under high policy divergence as shown in Figure 2c, whereas naive estimators continue to show unreliable accuracy.

Figure 3 summarizes the OPL results. In Figure 3a, policies learned based on IPCW-IPS and IPCW-DR achieve substantially higher improvement ratios compared to policy learners based on naive IPS and DR. Naive methods frequently underestimate survival times and thus fail to choose optimal actions, while IPCW-based learners appropriately deal with this bias and consistently identify better policies. Our methods continue to outperform naive learners as censoring increases as well (Figure 3b). These results demonstrate that our IPCW-based approaches achieve robust and reliable policy improvements across a range of realistic conditions.

5 Real-World Case Study: Starbucks Customer Retention

This section explores the practical effectiveness of our proposed estimators through a semi-synthetic experiment based on the Starbucks Customer Reward Program dataset[3]. Real-world datasets capture rich behavioral patterns, but they rarely offer counterfactual outcomes or complex censoring structures that allow a rigorous comparison of off-policy estimators or learners. To address these limitations and to perform yet a meaningful experiment, we first construct a data-driven "Oracle" environment directly from the Starbucks mobile app logs. The dataset contains detailed customer information such as age, income, membership tenure, and gender, together with offer attributes such as difficulty, reward amount, delivery channel, and duration. We define a discrete action space $\mathcal{A}$ containing ten offer types. In this environment, the survival time T measures how long a user takes to make the first purchase after receiving an offer. Our goal in this environment is to evaluate policies that shorten this purchase cycle, which corresponds to minimizing RMST over the validity period. We treat a purchase within the period as an event and regard all remaining observations as censored.

Furthermore, to examine estimator robustness under challenging censoring mechanisms, we introduce artificial censoring that depends on user demographics as their features. The original dataset mainly reflects simple Type I censoring due to offer expiration, which does not adequately stress-test modern OPE estimators. By generating censoring times that interact with demographic features, we create scenarios where the random censoring assumption fails in a controlled and interpretable way. This setting allows our experiment to reveal when naive estimators fail and how IPCW-based corrections address those failures. We describe the full data construction procedure and the censoring mechanism for this experiment in Appendix C.5.

Estimator	MSE	Squared bias	Variance
DM	30.7	29.2	**1.5**
IPS	216.7	213.7	3.1
DR	190.1	188.5	<u>1.6</u>
IPCW-IPS (Ours)	<u>11.8</u>	**1.4**	10.5
IPCW-DR (Ours)	**10.7**	<u>4.4</u>	6.4

Table 1: OPE results on the Starbucks dataset. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

OPE Results. Table 1 summarizes the OPE results on the Starbucks dataset. As shown in Table 1, IPCW-IPS and IPCW-DR deliver a clear and substantial reduction in MSE compared with IPS and DR. Both IPS and DR accumulate substantial squared bias, a pattern that directly reflects the theoretical insight shown in Proposition 2.3. In contrast, IPCW-IPS and IPCW-DR control squared bias effectively, demonstrating that the proposed correction mechanism performs reliably even in realistic settings that include complex and feature-dependent censoring structures.

We can also see from the table that DM records the smallest variance, but its large bias exposes its heavy reliance on the predictive accuracy of the outcome model. When the outcome model is misspecified, DM suffers from substantial bias regardless of the presence of censoring. IPCW-DR, however, counterbalances this weakness, i.e., although it shows slightly larger bias than IPCW-IPS, it significantly reduces

[3]https://www.kaggle.com/datasets/blacktile/starbucks-app-customer-reward-program-data

Learner	RMST (Mean ± Std) ↓	Cost (Mean ± Std)	Feasible Rate ↑
Regression	48.86 ± 1.21	2.54 ± 0.30	0.80
IPS	39.61 ± 1.90	2.44 ± 0.20	<u>0.95</u>
DR	<u>39.11 ± 2.06</u>	2.51 ± 0.23	0.90
IPCW-IPS (Ours)	39.41 ± 1.73	2.43 ± 0.20	**0.99**
IPCW-DR (Ours)	**38.16 ± 1.41**	2.53 ± 0.20	<u>0.95</u>

Table 2: Performance comparison of constrained OPL methods averaged over 100 trials with budget $B = 2.82$. Feasible rate denotes the proportion of trials satisfying the constraint. The symbols ↓ and ↑ indicate that lower and higher values are better, respectively. The best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

variance and achieves the lowest overall MSE among all estimators. This result highlights how the doubly robust structure stabilizes the offline estimation process. Thanks to its doubly robust structure, IPCW-DR significantly reduces variance compared to IPCW-IPS. These findings highlight the practical value of our estimators for real-world marketing and retention. By consistently outperforming baselines even under complex censoring patterns and non-linear dynamics, IPCW-IPS and IPCW-DR empower practitioners to derive effective interventional strategies from logged data that shorten purchase cycles and strengthen long-term engagement.

Constrained OPL Analysis. We further investigate the practical utility of our framework by evaluating the constrained OPL methods on the semi-synthetic Starbucks dataset. This experiment follows the same "Oracle" construction used in the OPE analysis.

To set up the constrained optimization problem, we first define the cost function $c(x, a)$ based on the reward amount attached to each offer. In line with common performance-based marketing practices, we charge this cost only when a user converts; otherwise, the cost remains zero. Next, we set the budget constraint B to 80% of the average cost incurred by the logging policy ($B = 2.82$), creating a realistic and challenging setting in which the learner must deliver faster customer engagement (lower RMST) while operating under a substantially reduced spending allowance.

We compare our Constrained IPCW-IPS and Constrained IPCW-DR with three constrained baselines: Constrained Regression, Constrained IPS, and Constrained DR. We solve the constrained optimization problem using Lagrangian duality to enable gradient-based learning. Appendix C.5 provides full details regarding the baseline implementations, neural network architecture, and optimization hyperparameters.

Table 2 reports the constrained OPL results on the Starbucks dataset. We observe that our IPCW-DR learner delivers the strongest RMST performance and clearly outperforms both naive OPE-based learners and the model-based regression approach. The naive IPS and DR learners underestimate survival outcomes as they ignore the censoring issue during policy learning, and the regression baseline performs even worse because it relies entirely on an imperfect outcome model. In contrast, IPCW-DR combines IPCW correction with a doubly robust structure, allowing the learner to optimize survival outcomes more effectively even under strict resource constraints. The IPCW-IPS and IPCW-DR learners also reduce the standard deviation of RMST substantially compared with naive approaches, indicating that our censoring-aware corrections create more stable gradient signals throughout the policy optimization process. Naive estimators unintentionally inject noise into the learning signal because they interpret censored samples as early events, leading to their unstable learning performances. Our IPCW-based learners avoid this distortion and update the policy using signals that reflect the underlying counterfactual survival outcomes more accurately. As a result, they repeatedly derive high-quality policies across different data samples and demonstrate strong reliability in practice.

We also examine the feasible rate, the proportion of experiment runs in which the learned policy respects the budget constraint. IPCW-IPS achieves the highest feasible rate of 0.99, and IPCW-DR follows closely at 0.95. Because the IPCW correction provides more accurate survival estimates, the learner can track the budget boundary more precisely and maintain consistent constraint satisfaction within the Lagrangian optimization framework. Overall, these results show that our censoring-aware OPL framework not only improves survival-based outcomes but also stabilizes the learning dynamics even with challenging constraints. This combination of predictive accuracy, robustness, and reliability positions our approach as a practical and powerful tool for real-world decision-making regarding survival outcomes.

6 Concluding Remarks

This paper proposes a novel framework for Off-Policy Evaluation (OPE) and Off-Policy Learning (OPL) with survival outcomes subject to right censoring. Specifically, we introduce two novel estimators, IPCW-IPS and IPCW-DR, which incorporate the statistical technique of Inverse Probability of Censoring Weighting to correct for systematic bias induced by censoring. Theoretically, we establish the unbiasedness of our proposed estimators and prove the doubly robust property and variance reduction effects inherent in IPCW-DR. Extensive simulation studies demonstrate that, while conventional OPE methods suffer from severe underestimation bias, our estimators maintain high estimation accuracy across a wide range of experiment conditions. Furthermore, real-world experiments using the Starbucks dataset validate that our proposed estimators are robust against complex real-world data structures and censoring mechanisms, and substantially outperform existing baselines in terms of both OPE and OPL.

References

[Ai *et al.*, 2022] Meng Ai, Biao Li, Heyang Gong, Qingwei Yu, Shengjie Xue, Yuan Zhang, Yunzhou Zhang, and Peng Jiang. LBCF: A large-scale budget-constrained causal forest algorithm. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2310–2319, New York, NY, USA, 2022. Association for Computing Machinery.

[Ascarza *et al.*, 2018] Eva Ascarza, Scott A. Neslin, Oded Netzer, et al. In pursuit of enhanced customer retention management: Review, key issues, and future directions. *Cust. Need. and Solut.*, 5:65–81, 2018.

[Beygelzimer and Langford, 2009] Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 129–138, New York, NY, USA, 2009. Association for Computing Machinery.

[Cai *et al.*, 2023] Tianchi Cai, Jiyan Jiang, Wenpeng Zhang, Shiji Zhou, Xierui Song, Li Yu, Lihong Gu, Xiaodong Zeng, Jinjie Gu, and Guannan Zhang. Marketing budget allocation with offline constrained deep reinforcement learning. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, WSDM '23, page 186–194, New York, NY, USA, 2023. Association for Computing Machinery.

[Chu *et al.*, 2023] Jianing Chu, Shu Yang, and Wenbin Lu. Multiply robust off-policy evaluation and learning under truncation by death. In *Proceedings of the International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[Cox, 1972] David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

[Du *et al.*, 2019] Shuyang Du, James Lee, and Farzin Ghaffarizadeh. Improve user retention with causal learning. In *Proceedings of Machine Learning Research*, volume 104 of *Proceedings of Machine Learning Research*, pages 34–49. PMLR, 05 Aug 2019.

[Dudík *et al.*, 2011] Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the International Conference on Machine Learning*, ICML'11, page 1097–1104, Madison, WI, USA, 2011. Omnipress.

[Fleming and Harrington, 2013] Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 2013.

[Geng *et al.*, 2015] Yuan Geng, Hao Helen Zhang, and Wenbin Lu. On optimal treatment regimes selection for mean survival time. *Statistics in medicine*, 34(7):1169–1184, 2015.

[Geoffrion, 1974] Arthur M. Geoffrion. *Lagrangean relaxation for integer programming*, pages 82–114. Springer Berlin Heidelberg, Berlin, Heidelberg, 1974.

[Glorot and Bengio, 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[Hao *et al.*, 2020] Xiaotian Hao, Zhaoqing Peng, Yi Ma, Guan Wang, Junqi Jin, Jianye Hao, Shan Chen, Rongquan Bai, Mingzhou Xie, Miao Xu, Zhenzhe Zheng, Chuan Yu, Han Li, Jian Xu, and Kun Gai. Dynamic knapsack optimization towards efficient multi-channel sequential advertising. In *Proceedings of the International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[Horvitz and Thompson, 1952] Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[Hosmer Jr *et al.*, 2013] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[Huang and Jiang, 2020] Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *Proceedings of the International Conference on Machine Learning*, pages 4434–4443. PMLR, 2020.

[Ishwaran *et al.*, 2008] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841 – 860, 2008.

[Jenkins, 2005] Stephen P Jenkins. Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, 42:54–56, 2005.

[Jeunen and Goethals, 2021] Olivier Jeunen and Bart Goethals. Pessimistic reward models for off-policy learning in recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, pages 63–74, 2021.

[Jiang *et al.*, 2017] Runchao Jiang, Wenbin Lu, Rui Song, and Marie Davidian. On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society, Series B*, 79(4):1165–1185, sep 2017.

[Kaplan and Meier, 1958] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[Katzman *et al.*, 2018] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep

neural network. *BMC Medical Research Methodology*, 18(1):24, 2018.

[Kingma and Ba, 2017] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[Kiyohara *et al.*, 2024] Haruka Kiyohara, Masahiro Nomura, and Yuta Saito. Off-policy evaluation of slate bandit policies via optimizing abstraction. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3150–3161, New York, NY, USA, 2024. Association for Computing Machinery.

[LeCun *et al.*, 2015] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[Lee *et al.*, 2018] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Murphy, 2003] Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65(2):331–355, 04 2003.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010. Omnipress.

[Peng *et al.*, 2023] Jie Peng, Hao Zou, Jiashuo Liu, Shaoming Li, Yibao Jiang, Jian Pei, and Peng Cui. Offline policy evaluation in large action spaces via outcome-oriented action grouping. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1220–1230, New York, NY, USA, 2023. Association for Computing Machinery.

[Robins and Finkelstein, 2000] James M. Robins and Dianne M. Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[Robins and Rotnitzky, 1995] James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

[Robins *et al.*, 1995] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.

[Rosenbaum and Rubin, 1983] 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, 04 1983.

[Sachdeva *et al.*, 2020] Noveen Sachdeva, Yi Su, and Thorsten Joachims. Off-policy bandits with deficient support. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 965–975, 2020.

[Saito and Joachims, 2021] Yuta Saito and Thorsten Joachims. Counterfactual learning and evaluation for recommender systems: Foundations, implementations, and recent advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 828–830, New York, NY, USA, 2021. Association for Computing Machinery.

[Saito *et al.*, 2024] Yuta Saito, Jihan Yao, and Thorsten Joachims. POTEC: Off-policy learning for large action spaces via two-stage policy decomposition. *arXiv preprint arXiv:2402.06151*, 2024.

[Uno *et al.*, 2014] Hajime Uno, Brian Claggett, Lu Tian, Eisuke Inoue, Paul Gallo, Toshio Miyata, Deborah Schrag, Masahiro Takeuchi, Yoshiaki Uyama, Lihui Zhao, Hicham Skali, Scott Solomon, Susanna Jacobus, Michael Hughes, Milton Packer, and Lee-Jen Wei. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22):2380–2385, 2014. PMID: 24982461.

[Wang *et al.*, 2017] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3589–3597. PMLR, 06–11 Aug 2017.

[Wang *et al.*, 2025] Han Wang, Yang Xu, Wenbin Lu, and Rui Song. Off-policy evaluation under nonignorable missing data. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 65020–65058. PMLR, 13–19 Jul 2025.

[Zhang *et al.*, 2021] Yang Zhang, Bo Tang, Qingyu Yang, Dou An, Hongyin Tang, Chenyang Xi, Xueying Li, and Feiyu Xiong. BCORLE(λ): an offline reinforcement learning and evaluation framework for coupons allocation in e-commerce market. In *Advances in Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc.

[Zhao *et al.*, 2019] Kui Zhao, Junhao Hua, Ling Yan, Qi Zhang, Huan Xu, and Cheng Yang. A unified framework for marketing budget allocation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, page 1820–1830, New York, NY, USA, 2019. Association for Computing Machinery.

[Zhou *et al.*, 2023] Hao Zhou, Shaoming Li, Guibin Jiang, Jiaqi Zheng, and Dong Wang. Direct heterogeneous causal learning for resource allocation problems in marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023.