Off-Policy Evaluation with Out-of-Sample Guarantees

Anonymous authors
Paper under double-blind review

Abstract

We consider the problem of evaluating the performance of a decision policy using past observational data. The outcome of a policy is measured in terms of a loss (aka. disutility or negative reward) and the main problem is making valid inferences about its out-of-sample loss when the past data was observed under a different and possibly unknown policy. Using a sample-splitting method, we show that it is possible to draw such inferences with finite-sample coverage guarantees about the entire loss distribution, rather than just its mean. Importantly, the method takes into account model misspecifications of the past policy – including unmeasured confounding. The evaluation method can be used to certify the performance of a policy using observational data under a specified range of credible model assumptions.

1 Introduction

In this work, we are interested in evaluating the performance of a decision policy, denoted π, which chooses an action from a discrete action set. Each action A is taken in a context with observable covariates X and incurs a real-valued loss L (aka. disutility or negative reward). Such policies are considered, for example, in contextual bandit problems and precision medicine (Langford & Zhang, 2007; Qian & Murphy, 2011; Lattimore & Szepesvári, 2020; Tsiatis et al., 2019). For instance, A may be one of several treatment options for a patient with observable characteristics X, and L measures the severity of the outcome.

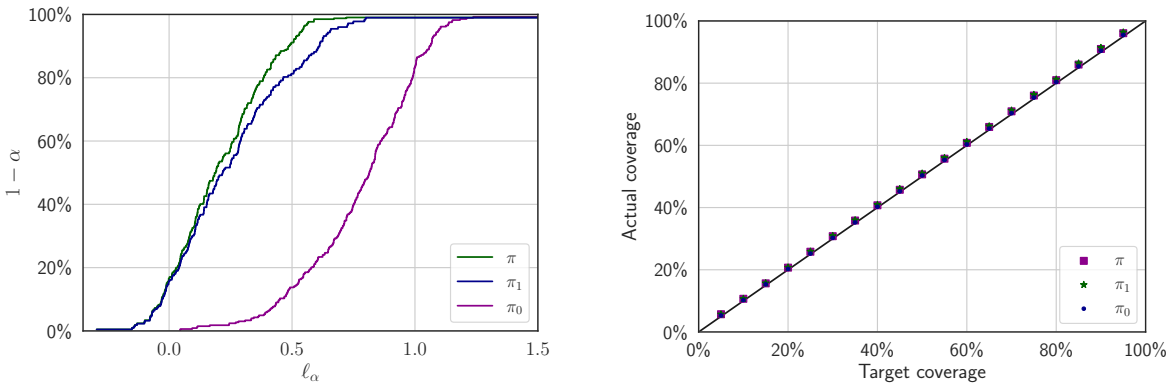
A target policy π can be evaluated using experimental data obtained from trials. Such experiments are, however, often costly to perform and may lead to rather small sample sizes in, e.g., clinical settings. Moreover, in safety-critical applications, it is often unethical to test new policies without severe restrictions on the trial population as well as the target policy. A more fundamental inferential problem, however, is the lack of 'external' validity, i.e., the limited ability to extrapolate from the trial population to the intended target population may lead to invalid inferences (Westreich, 2019; Manski, 2019). The main alternative is *off-policy* evaluation, i.e., using observational data from a past decision process to infer the performance of the target policy. For this to be valid one needs to *assume* that the past process is modeled without systematic errors, i.e, no unmeasured confounding and using well-specified models. The credibility of these assumptions, therefore, determines the 'internal' validity of inferences about π from observational data (Manski, 2003).

Inferences that lack validity are particularly serious when evaluating π in decision processes that are costly or safety-critical. In such cases, even inferences that are asymptotically valid with increasing sample size may not be adequate. Moreover, when the resulting distribution of losses is skewed or is widely dispersed, its tails are important to evaluate. Then inferring the expected loss $\mathbb{E}_\pi[L]$, as is commonly done, provides a very limited evaluation of π (Wang et al., 2018). For instance, the average loss in a population may be small but the tail losses can be unacceptably large. In such applications, we are more concerned with providing valid certifications of the overall performance (see Figure 1a), rather than inferences of a single distributional parameter.

In this paper, we propose a method for evaluating a specified target policy using observational data that

- provides finite-sample coverage guarantees for the out-of-sample loss,
- evaluates the entire loss distribution instead of, e.g., its expected value,

- and takes model misspecification, including unmeasured confounding, into account.



(a) Evaluation of out-of-sample loss L.

(b) Evaluation of coverage of limit curves.

Figure 1: Evaluating out-of-sample losses under target policy with binary decisions $A \in \{0, 1\}$. Policies π_0 and π_1 correspond to 'treat none' ($A \equiv 0$) and 'treat all' ($A \equiv 1$), respectively, while π denotes a policy that adapts to context X. (For more details, see the example in Section 5.1 with p_1.) (a) Each curve certifies that a new loss L falls below the limit ℓ_α with a probability of at least $1 - \alpha$ using $n = 1000$ data points. The certified performance of policy π dominates those of the alternative policies. (b) Evaluation of actual coverage, that is, the probability of $L \leq \ell_\alpha$, versus target coverage $1 - \alpha$.

2 Problem formulation

We consider a *target policy* π for deciding an action A in different contexts, which are described by observed and unobserved covariates X and U, respectively. The policy corresponds to a distribution $p_\pi(A|X)$, and can be either deterministic or random. Our aim is to evaluate the losses L that result from applying any given π. Each instance of contextual covariates, action, and loss, i.e., (X, U, A, L), is drawn independently from a target distribution $p_\pi(X, U, A, L)$. At our disposal is an *observational* data set

$$\mathcal{D} = \big((X_i, A_i, L_i)\big)_{i=1}^n, \tag{1}$$

and our goal is to use it to characterize the *out-of-sample loss* L_{n+1}. Note that it was collected under a possibly *different* policy than π. If we continue with the example of patients, then L_{n+1} would quantify the severity of the outcome for a new future patient that is unobserved at the time of our evaluation. Specifically, for any miscoverage level $\alpha \in (0, 1)$, we seek an informative limit $\ell_\alpha(\mathcal{D})$ on the loss such that

$$\boxed{\mathbb{P}_\pi\Big\{L_{n+1} \leq \ell_\alpha(\mathcal{D})\Big\} \geq 1 - \alpha.} \tag{2}$$

In other words, $\ell_\alpha(\mathcal{D})$ as a function of α yields a finite-sample performance certification of π as is illustrated in Figure 1a. Unlike a single point estimate, the limit curve characterizes the distribution of losses under π. For instance, consider a patient population where (X, U) denote its covariates, and A is its received treatment with an outcome loss L. Then across all coverage levels, (2) ensures that the treatment of a future patient will incur a loss no greater than $\ell_\alpha(\mathcal{D})$ under policy π with confidence $1 - \alpha$. Figure 1b shows the probability that a limit $\ell_\alpha(\mathcal{D})$, proposed below, bounds the future loss L_{n+1} across all values of $1 - \alpha$. Note that limit curves can be constructed for several alternative target policies, which enables us to compare and focus on the parts of their loss distributions that are most relevant to the specific decision problem.

The causal structure of this decision process is illustrated in Figure 2a. The target distribution admits a causal factorization

$$p_\pi(X, U, A, L) = p(X, U)\, p_\pi(A|X)\, p(L|A, X, U), \tag{3}$$

where $p(X, U)$ and $p(L|A, X, U)$ are *unknown*. The central challenge in *off-policy* evaluation of π is that (1) is not sampled from (3) but from a shifted *training* distribution which admits a causal factorization

$$p(X, U, A, L) = p(X, U)\, p(A|X, U)\, p(L|A, X, U), \tag{4}$$

where $p(A|X, U)$ characterizes a *past policy* (aka. behavioral policy) that may differ from π. The causal structure corresponding to (4) is illustrated in Figure 2b. If the past policy were known, it would be possible to adjust for the distribution shift from training to target distribution using the ratio

$$\frac{p_\pi(X, U, A, L)}{p(X, U, A, L)} \equiv \frac{p_\pi(A|X)}{p(A|X, U)}. \tag{5}$$

This is feasible in certain problems with fully automated decision-making, such as online recommendation systems, where the past policy is designed using observable covariates only, i.e., $p(A|X, U) \equiv p(A|X)$. In more general problems, however, we have only a *nominal* model of the past policy $\widehat{p}(A|X)$ (aka. propensity model), typically estimated from prior observable data. The nominal model may therefore diverge from $p(A|X, U)$ due to various modelling errors that persist even in the large-sample scenario: model misspecification and unmeasured confounding via U (Peters et al., 2017; Westreich, 2019). Here we follow the marginal sensitivity methodology of Tan (2006) and characterize the model divergence with respect to the odds of taking action A. That is, the nominal odds diverge from the unknown odds by some bounded factor $\Gamma \geq 1$ as follows:

$$\frac{1}{\Gamma} \leq \underbrace{\frac{p(A|X, U)}{1 - p(A|X, U)}}_{\text{unknown odds}} \Bigg/ \underbrace{\frac{\widehat{p}(A|X)}{1 - \widehat{p}(A|X)}}_{\text{nominal odds}} \leq \Gamma, \tag{6}$$

for all discrete actions A and almost surely all X, U. When the bound equals $\Gamma = 1$, the nominal model is perfectly specified and there is no unmeasured confounding. In general, we consider all cases where the nominal odds diverge by *at most* a factor Γ.

In summary, the problem we consider is to construct a limit $\ell_\alpha(\mathcal{D})$ for target policy π using a nominal model $\widehat{p}(A|X)$ and Γ. The resulting limit should satisfy the finite-sample guarantee (2) for all miscoverage levels α, and thereby certify the target policy performance for any specified bound Γ in (6). This enables a *robust* evaluation of target policies using observational data since it can be performed across a range of credible odds divergence bounds Γ as we will illustrate in the numerical experiments in Section 5. By increasing the divergence bound Γ, the credibility of our assumptions on $\widehat{p}(A|X)$ increases, but the strength of our conclusions about L_{n+1} decrease, cf. Manski (2003). The trade-off between credibility of assumptions and informativeness of inferences will be quantified as well.

(a) Decision process under target policy

(b) Decision process under past policy

Figure 2: Directed acyclic diagrams representing the causal structure of the decision process under (a) target policy, which yields (3), and (b) past policy, which yields (4). In (b), both contextual covariates (X, U) may joint affect actions A and outcome loss L. Thus U gives rise to unmeasured confounding.

3 Background

We place the problem considered in this paper and our proposed method within the framework of off-policy evaluation.

Expected loss: In the off-policy evaluation literature, the target quantity is often the unknown expected loss $\mathbb{E}_\pi[L]$ of policy π. A standard estimator of the mean, that dates back to Horvitz & Thompson (1952), is based on inverse propensity weighting:

$$V_{\mathrm{IPW}}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(X_i, A_i)\, L_i, \tag{7}$$

where $\widehat{w}(X, A) = \frac{p_\pi(A|X)}{\widehat{p}(A|X)}$ is a model of (5), see for instance Rosenbaum & Rubin (1983); Beygelzimer et al. (2009); Qian & Murphy (2011); Zhang et al. (2012); Zhao et al. (2012); Kallus (2018). We note that the estimator in (7) is unbiased when $\Gamma = 1$. An alternative standard estimator is based on regression modeling:

$$V_{\mathrm{RM}}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} p_\pi(a|X_i)\, \widehat{\ell}(a, X_i), \tag{8}$$

where $\widehat{\ell}(A, X)$ is a model of $\mathbb{E}[L|A, X]$.

The approaches in (7) and (8) have complementary strengths and weaknesses resulting from the challenges of modelling the past policy and the conditional mean of losses, respectively. Even when the models are well-specified, the accuracy of the estimators depends highly on the overlap of covariates X across decisions A in the training data (Oberst et al., 2020; D'Amour et al., 2021). When the overlap is weak, the variance of $V_{\mathrm{IPW}}(\mathcal{D})$ can become excessively large, even when it is unbiased. This can be mitigated by clipping the weights, which in turn causes bias (Rubin, 2001; Kang & Schafer, 2007; Schafer & Kang, 2008; Strehl et al., 2010).

When the models $\widehat{w}(X, A)$ or $\widehat{\ell}(A, X)$ exhibit systematic errors, however, the corresponding estimators in (7) and (8) are biased and may invalidate the evaluation of π. The 'doubly robust' estimator

$$V_{\mathrm{DR}}(\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{w}(X_i, A_i)\Big[L_i - \widehat{\ell}(A_i, X_i)\Big] + \sum_{a\in\mathcal{A}} p_\pi(a|X_i)\, \widehat{\ell}(a, X_i)\Big),$$

is one way of robustification in the case of *either* model $\widehat{w}(X, A)$ or $\widehat{\ell}(A, X)$ is misspecified. Moreover, it reduces the estimator variance provided $\widehat{\ell}(A, X)$ is sufficiently accurate (Bang & Robins, 2005; Dudík et al., 2011; Rotnitzky et al., 2012).

Distribution of losses: When the loss distribution is highly skewed and the tails are long, the use of expected loss to evaluate policies can be inadequate, especially in high-stakes problems. There are alternative parameters of the loss distribution, described by the cumulative distribution function $F(\ell) = \mathbb{P}_\pi\{L_{n+1} \leq \ell\}$ (cdf), that one can consider in such problems, e.g., a quantile or the conditional value at risk (Wang et al., 2018; Chandak et al., 2021; Huang et al., 2021).

Off-policy evaluation of π with respect to some alternative parameter can be achieved using cdf-estimators that are analogous to the mean estimators above, see Huang et al. (2021). In analogy to (7), the inverse propensity weighted cdf-estimator

$$\widehat{F}_{\mathrm{IPW}}(\ell; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{w}(X_i, A_i)\, \mathbb{1}(L_i \leq \ell), \tag{9}$$

is point-wise unbiased when $\Gamma = 1$. Similar to (8), the estimator

$$\widehat{F}_{\mathrm{RM}}(\ell; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} p_\pi(a|X_i)\, \widehat{c}(\ell; a, X_i),$$

requires a model $\widehat{c}(\ell; a, x)$ of the conditional distribution $\mathbb{P}\{L \leq \ell | A, X\}$. To mitigate against model misspecification that threatens the validity of the evaluation of π, one can use the 'doubly robust' estimator

$$\widehat{F}_{\mathrm{DR}}(\ell; \mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{w}(X_i, A_i)\Big[\mathbb{1}(L_i \leq \ell) - \widehat{c}(\ell; A_i, X_i)\Big] + \sum_{a\in\mathcal{A}} p_\pi(a|X_i)\, \widehat{c}(\ell; a, X_i)\Big),$$

which protects in the case that either model $\widehat{w}(X, A)$ or $\widehat{\ell}(A, X)$ is misspecified. While this estimator is consistent, it is not guaranteed to yield a proper cdf.

In this paper, we are interested in limiting the out-of-sample loss L_{n+1}. The α-quantile is the smallest ℓ_α such that $\mathbb{P}_\pi\{L_{n+1} \leq \ell_\alpha\} \geq 1 - \alpha$. It can be estimated as

$$\ell_\alpha(\mathcal{D}) = \inf\left\{\ell : \widehat{F}(\ell; \mathcal{D}) \geq 1 - \alpha\right\},$$

using one of the cdf-estimators above. We will use $\widehat{F}_{\text{IPW}}$ as a benchmark below.

Distribution-free inference: Derivations of finite-sample valid, nonparametric limits on random variables date back to the works of Wilks (1941); Wald (1943); Scheffe & Tukey (1945). More recently, the related methodology of conformal prediction has focused on developing covariate-specific prediction regions (Vovk et al., 2005; Shafer & Vovk, 2008; Vovk, 2012). See Lei & Wasserman (2014); Lei et al. (2018); Romano et al. (2019) for further developments. Tibshirani et al. (2019) extended the methodology to make it valid also under known covariate shifts. This extended methodology was used to provide context-specific prediction intervals for any given policy π, which are statistically valid under the assumption that the past policy $p(A|X, U)$ is known (Osama et al., 2020; Taufiq et al., 2022).

The marginal sensitivity methodology developed in Tan (2006) enables us to specify a more credible range of assumptions using (6). This type of methodology was used for robust policy learning in Kallus & Zhou (2021), for robust policy learning in sequential decisions in Namkoong et al. (2020) and for sensitivity analysis of treatment effects in Jin et al. (2023) in the case of unobserved confounding. The present paper considers the overall performance of π, similar to Huang et al. (2021). However, different from that paper, we focus on ensuring inferences on the out-of-sample losses that are valid even for finite training data and systematic modelling errors – including unobserved confounding – using a sample-splitting technique that leverages results derived in Jin et al. (2023).

4 Method

We show that one can limit the out-of-sample losses for π under any specified odds divergence bound $\Gamma \geq 1$ for the nominal model in (6), which we assume holds. It follows that (5) is bounded as:

$$\underline{W} \leq \frac{p_\pi(X, U, A, L)}{p(X, U, A, L)} \leq \overline{W}, \tag{10}$$

where the bounds equal

$$\underline{W} = p_\pi(A|X) \cdot \left[1 + \Gamma^{-1}\big(\widehat{p}(A|X\big)^{-1} - 1)\right] \qquad \text{and} \qquad \overline{W} = p_\pi(A|X) \cdot \left[1 + \Gamma\big(\widehat{p}(A|X)^{-1} - 1\big)\right]. \tag{11}$$

The model $\widehat{p}(A|X)$ is typically estimated from prior observable data. The above bounds are functions of X and A drawn from the training distribution (4). In order to ensure finite-sample guarantees, we randomly split the training data (1) into two separate sets,

$$\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1,$$

with samples of sizes n_0 and $n - n_0$, respectively. The set $\mathcal{D}_1$ is used to form the function

$$\widehat{F}(\ell; \mathcal{D}_1, w) = \frac{\sum_{i=n_0+1}^{n} \underline{W}_i \mathbb{1}(L_i \leq \ell)}{\sum_{i=n_0+1}^{n} \underline{W}_i \mathbb{1}(L_i \leq \ell) + \sum_{i=n_0+1}^{n} \overline{W}_i \mathbb{1}(L_i > \ell) + w}, \tag{12}$$

as a proxy for the unknown cdf of the out-of-sample loss L_{n+1} and $w > 0$ represents the ratio in (10). We use the set $\mathcal{D}_0$ to construct an upper bound on this unknown ratio. As the following result shows, this enables us to construct a valid limit ℓ_α on the future loss L_{n+1}.

Theorem 4.1. *Define the following quantile function of* (12) *as*

$$\ell_{\alpha,\beta} = \inf\left\{\ell : \widehat{F}(\ell; \mathcal{D}_1, \overline{w}_\beta(\mathcal{D}_0)) \geq \frac{1-\alpha}{1-\beta}\right\}, \tag{13}$$

where

$$\overline{w}_\beta(\mathcal{D}_0) = \begin{cases} \overline{W}_{[\lceil (n_0+1)(1-\beta)\rceil]}, & \lceil (n_0+1)(1-\beta)\rceil \leq n_0, \\ \infty, & \textit{otherwise}, \end{cases} \tag{14}$$

and $\overline{W}_{[k]}$ *denotes the* k*th order statistic of* $(\overline{W}_i)_{i=1}^{n_0}$.

For any miscoverage probability $\alpha \in (0,1)$*, construct*

$$\boxed{\ell_\alpha(\mathcal{D}) = \min_{\beta:0<\beta<\alpha} \ell_{\alpha,\beta},} \tag{15}$$

which a valid limit on the out-of-sample loss L_{n+1} *with a probability of at least* $1-\alpha$ *as specified in* (2).

The proof of Theorem 4.1 can be found in Appendix A.1.

The variable $\overline{w}_\beta(\mathcal{D}_0)$ in (14) provides a statistically valid upper bound on the unknown ratio in (10) where β only specifies its confidence level. For *any* given value of $0 < \beta < \alpha$, $\ell_{\alpha,\beta}$ in (13) bounds the loss L_{n+1} with a coverage probability of at least $1-\alpha$. Consequently, (2) chooses the tightest achievable limit.

Remark 1. *When* $\Gamma = 1$*, then* $\underline{W} = \overline{W}$ *in* (10)*. Therefore* (12) *can be thought of as an inverse propensity weighted cdf, where the weights are normalized to sum to unity.*

We now turn to the implementation of (15). We note that (12) is a piecewise constant function in ℓ and that $\overline{w}_\beta$ in (14) is a piecewise constant function in β. Therefore (13) and subsequently (15) can be solved by evaluating functions along a discrete set of points.

The computation of a limit curve is summarized in Algorithm 1, using a discrete grid of miscoverage levels α. Given a model $\widehat{p}(A|X)$ and a range of odds divergence bounds Γ, Algorithm 1 produces a set of corresponding limit curves.

Algorithm 1 Limit curve of policy π

Input: Policy $p_\pi(A|X)$, training data $\mathcal{D}$, model $\widehat{p}(A|X)$, bound $\Gamma \geq 1$ and sample split size n_0.
1: Randomly split $\mathcal{D}$ into $\mathcal{D}_0$ and $\mathcal{D}_1$.
2: **for** $\alpha \in \{0, \ldots, 1\}$ **do**
3: **for** $\beta \in \{0, \ldots, \alpha\}$ **do**
4: Compute $\overline{w}_\beta$ using (14).
5: Compute $\ell_{\alpha,\beta}$ using (13).
6: **end for**
7: Set ℓ_α to the smallest $\ell_{\alpha,\beta}$ above.
8: Store (α, ℓ_α).
9: **end for**
Output: $\{(\alpha, \ell_\alpha)\}$

5 Numerical experiments

In the experiments below, we evaluate policies using the limit curves (α, ℓ_α). In any given problem the maximum value $L_{\max}$ of the support of L provides a valid but uninformative limit for any α. We therefore quantify the *informativeness* of a valid limit curve as follows:

$$\text{Informativeness} = 1 - \alpha^*, \text{where } \alpha^* = \inf\{\alpha : \ell_\alpha(\mathcal{D}) < L_{\max}\}. \tag{16}$$

That is, the highest coverage probability at which we can informatively certify the performance of π, i.e. the right limit of the curves. We can quantify how increasing the credibility of our model assumption, i.e., by increasing Γ, affects the informativeness of the limit curve. We also consider the coverage probability of the curves:

$$\text{Miscoverage gap} = \alpha - \mathbb{P}_\pi\{L_{n+1} > \ell_\alpha(\mathcal{D})\}. \tag{17}$$

When this gap is positive, the limit is *conservative* and when the gap is negative the limit is *invalid*, respectively, at level α.

A natural benchmark for the proposed limit (15) in this problem setting is the estimated quantile

$$\ell_\alpha(\mathcal{D}) = \inf\left\{\ell : \widehat{F}_{\mathrm{IPW}}(\ell;\mathcal{D}) \geq 1-\alpha\right\}, \tag{18}$$

using the inverse propensity weighted cdf-estimator (9).

In all examples below, the limit (15) is computed using sample splits of equal size, i.e., $n_0 = \lceil n/2 \rfloor$. An additional experiment using data from the Infant Health and Development Program (IHDP) can be found in Appendix A.2.

The code used for the experiments is made available here `TODO`.

5.1 Synthetic data

In the first example, we consider synthetic data in order to evaluate the coverage of the derived limit curves. We use a simulation setting similar to Jin et al. (2023). The miscoverage gap (17) is estimated by Monte Carlo simulation using 1000 runs and for each run drawing independent 1000 new samples $(X_{n+1}, U_{n+1}, A_{n+1}, L_{n+1})$.

We consider a population of individuals with two-dimensional covariates distributed uniformly as

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{U}(0,1)^2.$$

The actions are binary $A \in \{0,1\}$ corresponding to 'not treat' and 'treat', respectively. We want to evaluate a deterministic target policy, described by

$$p_\pi(A=0|X) = \mathbb{1}(X_1 X_2 \geq \tau), \tag{19}$$

for different $\tau \in [0,1]$. That is, all individuals whose covariate product X_1X_2 falls below τ are treated. Note that $\tau = 0$ corresponds a 'treat none' policy ($A \equiv 0$ for all X) and $\tau = 1$ corresponds to a 'treat all' policy ($A \equiv 1$ for all X). Below we discuss the resulting losses under this policy using observational data with sample sizes $n \in \{250, 500, 1000\}$.

Case: Known past policy ($\Gamma = 1$). In the first scenario, we assume that the past policy is known exactly and there is no unmeasured confounding.

For the training data, the past policy has selected actions as a Bernoulli process:

$$p(A=0|X) \equiv \widehat{p}(A=0|X) = f\big(c(X_1X_2+1)\big), \quad c \in \left[\frac{1}{2}, 2\right], \tag{20}$$

where $f(\cdot)$ is the sigmoid function. The conditional loss distribution is given by

$$(L|A=0,X) \sim \mathcal{N}(1 - X_1X_2, 0.1) \quad \text{and} \quad (L|A=1,X) \sim \mathcal{N}(X_1X_2, 0.1),$$

and thus $L_{\max} = \infty$. We consider three configurations c of past policies (20), which yield inverse propensity weights in three ranges: $\frac{1}{p_1(A|X)} < 3.72$ ($c = 1/2$), $\frac{1}{p_2(A|X)} < 8.39$ ($c = 1$), and $\frac{1}{p_3(A|X)} < 55.6$ ($c = 2$). Thus we anticipate $p_3(A|X)$ to be the most challenging case.

Here we evaluate three target policies $\tau = \{0, 0.5, 1\}$ in (19) and present their resulting limit curves using data from different past policies (20), see Figure 3. The dashed line shows the corresponding past policy. The limit curves for a given target policy are quite similar across training distributions and are still informative at the 90% level using (16). The main difference is in the inferred tail losses and is notable for when $\tau = 1$ under the more challenging past policy $p_3(A|X)$.

We now turn to the evaluating miscoverage gap (17). Figure 4 presents the gaps for target policy $\tau = 0.5$ in (19). The solid lines illustrate the proposed method and the dashed lines show the benchmark (18). We see

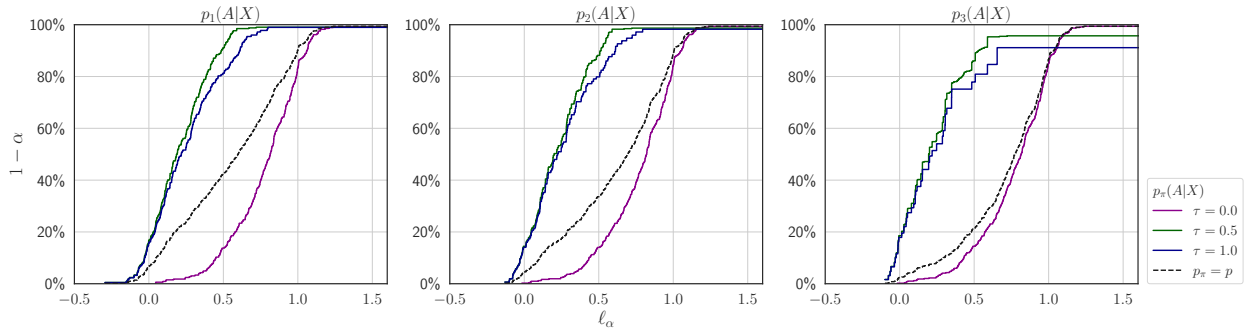


Figure 3: Limit curves when the past policy is known ($\Gamma = 1$) for three different potential target policies (i.e. $\tau = \{0.0, 0.5, 1.0\}$ in (19)). Dashed curve denotes the past policy. $n = 1000$.

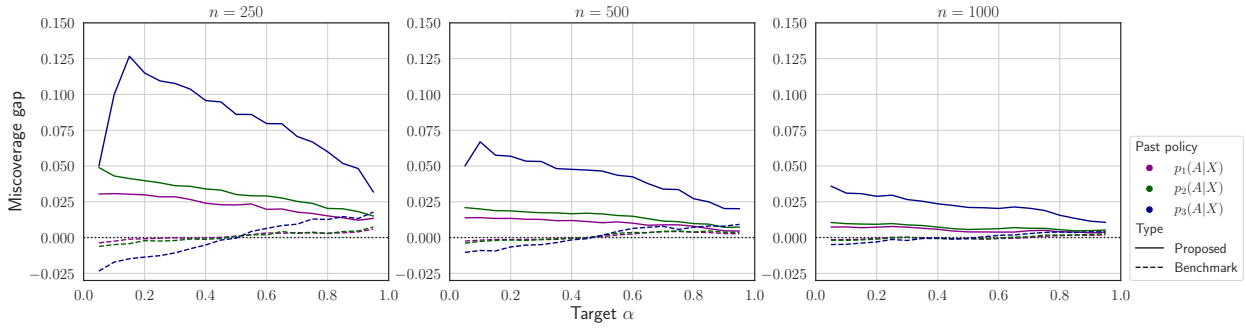


Figure 4: Miscoverage gaps (17) versus α, when the past policy is known ($\Gamma = 1$). Dashed curve denotes the benchmark (18).

that the proposed method is slightly conservative, but remains valid for all α. In contrast, the benchmark is not valid in the tail of the distribution, but is less conservative for higher α in this well-specified case.

Case: Estimated past policy ($\Gamma > 1$). In the second scenario, we assume that we only have access to an estimate $\widehat{p}(A|X)$ (given by (20)) and that there is unmeasured confounding in the training distribution. To render visually distinct curves from the previous case, we consider here a rather extreme case of confounding following Jin et al. (2023).

Specifically, we have an unobserved variable drawn as

$$(U|X) \sim \mathcal{N}(0, 0.1(X_1 + X_2)),$$

and the loss $(L|A, X, U)$ is generated as

$$L = \begin{cases} 1 - X_1 X_2 + U, & A = 0, \\ X_1 X_2 + U, & A = 1. \end{cases}$$

We define the past policy in a manner that enables us to control the divergence from the nominal model $\widehat{p}(A|X)$ in (20):

$$\begin{aligned} p(A = 0|X, U) = \mathbb{1}(U \leq t(X)) \left[1 + \Gamma_0^{-1}\big(\widehat{p}(A = 0|X)^{-1} - 1\big)\right] \\ + \mathbb{1}(U > t(X)) \left[1 + \Gamma_0\big(\widehat{p}(A = 0|X)^{-1} - 1\big)\right], \end{aligned} \tag{21}$$

where the threshold function $t(X)$ is designed empirically to ensure that the resulting median loss of the past policy for $A = 1$ is maximized. Our design of the past policy can be seen as the worst case among all unknown past policies that diverge by a factor Γ_0 in (6). We fix $\Gamma_0 = 2$ here, but treat it as unknown.

For simplicity, we consider a 'treat all' target policy ($\tau = 1$). Its limit curves, under different assumed odds divergence bounds $\Gamma = \{1, 2, 3\}$, are presented in Figure 5. Note that in the case of unmeasured confounding, the resulting curves differ notably across the training distributions unlike in Figure 3. We see that for the first and second distributions, the informativeness of all curves stays around the 90% level. However, in the more extreme third case, the informativeness drops to just above the 60% level when we increase the credibility of our model assumption to an odds bound of $\Gamma = 3$. This example illustrates an inherent trade-off between credibility and informativeness.

Figure 6 validates our guarantees using data drawn from $p_1(A|X)$. When $\Gamma \geq \Gamma_0 = 2$, the limit curves are valid and as Γ increases to 3, the limits become quite conservative. Note that the conservativeness persists even as the sample size n is increased fourfold. For $\Gamma = 1$, there is no guarantee of coverage and in this case the limit curve is invalid. The benchmark does not take the unmeasured confounding into account and is consequently invalid throughout.

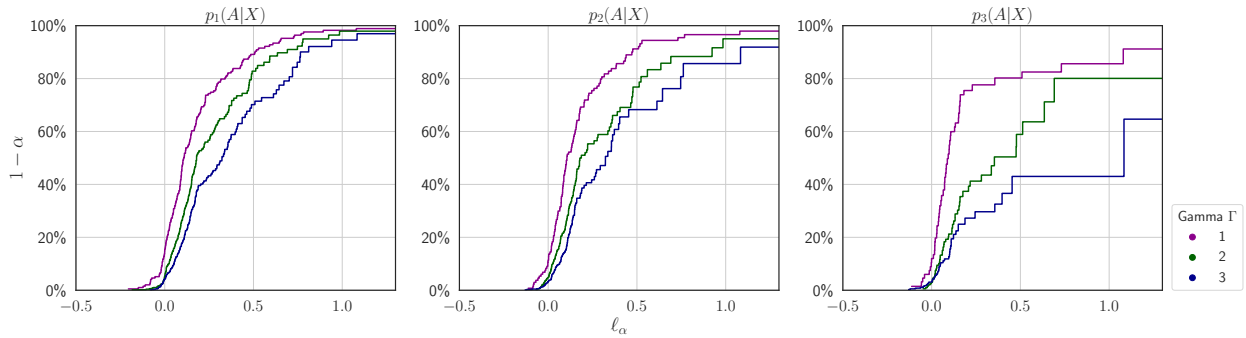


Figure 5: Limit curves for 'treat all' target policy using odds bounds $\Gamma = \{1, 2, 3\}$, when the past policy is unknown and subject to unmeasured confounding ($\Gamma_0 = 2$ in (21)). $n = 1000$.

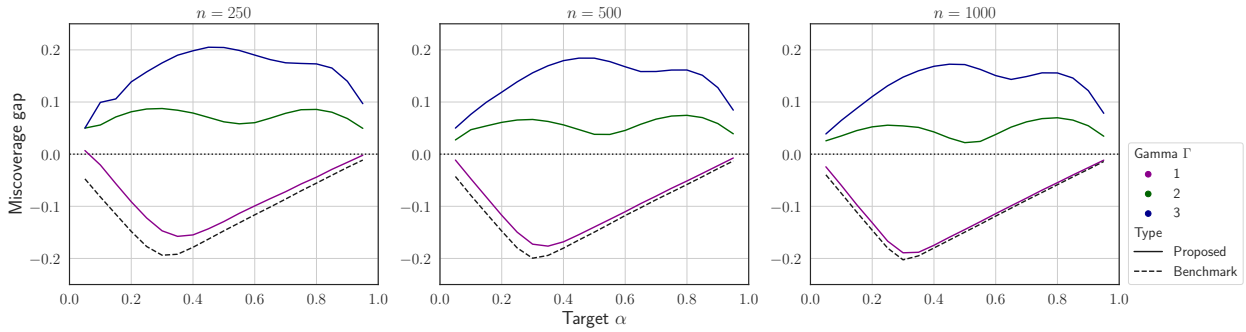


Figure 6: Miscoverage gaps (17) versus α, when the past policy is unknown and subject to unmeasured confounding. Dashed curve denotes the benchmark (18) which does not take confounding into account. Note that $\Gamma = 1$ assumes there is no odds divergence for our model, which in this case leads to an invalid limit curve.

5.2 Real data

In the second example, we use data from the National Health and Nutrition Examination Survey (NHANES) for the years 2013-2014 to illustrate the use of the proposed method. Following Zhao et al. (2019), we study the effect of seafood consumption on blood mercury levels. The action A indicates whether a person has a low or high consumption of fish or shellfish ($\leq$1 vs. >12 servings in the past month) and the loss L is the total concentration of blood mercury (in μg/L). There are 8 covariates in X for each person: gender, age, income, whether income is missing or not, race, education, ever smoked and number of cigarettes smoked

last month. 1 individual with missing education data and 7 with missing smoking data are omitted. 175 individuals have missing income data, for them we impute the median income. This preprocessing results in a data set of 1107 individuals of which 21% follow a high fish consumption diet and 79% follow a low consumption diet, see Zhao et al. (2018). We use a fitted logistic regression model for $\widehat{p}(A|X)$, where the continuous covariates in X have been standardized.

We compare two policies, π_0 and π_1, corresponding to low ($A \equiv 0$) and high ($A \equiv 1$) seafood consumption, respectively. Their corresponding limit curves are presented in Figure 7 under different odds divergences bounds Γ on the model $\widehat{p}(A|X)$. In the most extreme case, we consider the nominal odds diverging by at most a factor $\Gamma = 3$. For reference, we display the limit curve under the unknown consumption policy in the population. We see that lower mercury levels can be certified by low seafood consumption. Note that a guidance value for mercury levels is 8 μg/L for women of child-bearing age and 20 μg/L for women ≥ 50 years (Kales & Thompson, 2016). If we were to evaluate policies based on the expected loss alone, then high seafood consumption for women has an expected loss of approximately 3.3 μg/L and could be deemed safe. However, this figure masks the tail losses that substantial proportion of women may face under such a policy. For a high consumption policy, our method could only certify that approximately 80% ($\Gamma = 1$) to 50% ($\Gamma = 3$) of females would stay below the guideline level of 8 μg/L. In contrast, for the low consumption policy, the corresponding figure is approximately 95% (for $1 \leq \Gamma \leq 3$).

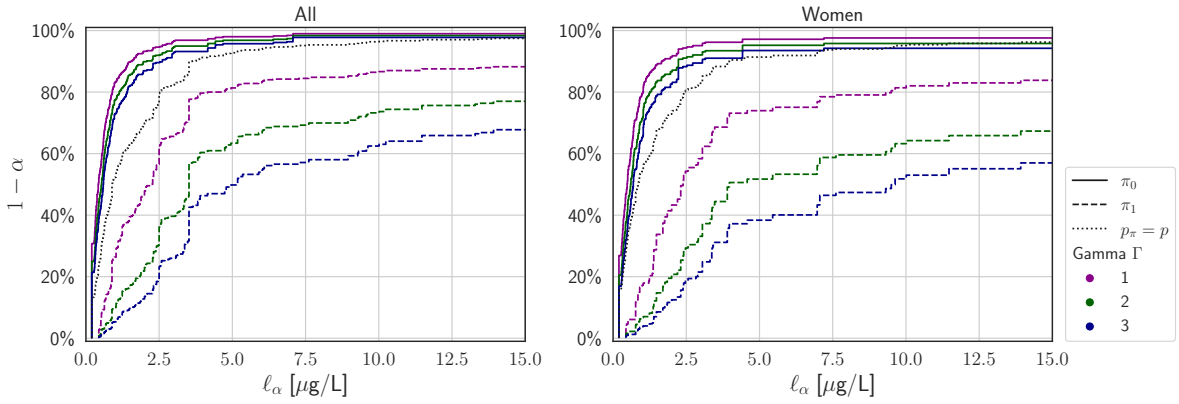


Figure 7: Limit curves for blood mercury levels in a population under different seafood consumption policies. Left: Overall population. Right: Female population. We compare two policies: low (π_0) and high (π_1) consumption. The credibility of the assumptions increases with the odds bound $\Gamma = \{1, 2, 3\}$. The past policy is indicated by dotted curves.

6 Conclusion

We have considered the problem of off-policy evaluation, i.e., making valid inferences about a target policy using past observational data obtained under a different decision process with a possibly unknown policy. Using the marginal sensitivity model, we derived a sample-splitting method that provides limit curves with finite-sample coverage guarantees even under model misspecifications, including unmeasured confounding. The validity, informativeness, and conservativeness of the resulting limit curves were demonstrated in the numerical experiments.

References

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

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 49–56, 2009.

Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.

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

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

The Infant Health and Development Program. Enhancing the outcomes of low-birth-weight, premature infants. a multisite, randomized trial. *JAMA*, 263 22:3035–42, 1990.

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

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.

Audrey Huang, Liu Leqi, Zachary Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *Advances in Neural Information Processing Systems*, 34:23714–23726, 2021.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6), 2023.

Stefanos N Kales and Aaron MS Thompson. A young woman concerned about mercury. *CMAJ*, 188(2): 133–134, 2016.

Nathan Kallus. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

Joseph DY Kang and Joseph L Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):71–96, 2014.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Charles F Manski. Identification problems in the social sciences and everyday life. *Southern Economic Journal*, 70(1):11–21, 2003.

Charles F Manski. *Patient Care Under Uncertainty*. Princeton University Press, 2019.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

Michael Oberst, Fredrik Johansson, Dennis Wei, Tian Gao, Gabriel Brat, David Sontag, and Kush Varshney. Characterization of overlap in observational studies. In *International Conference on Artificial Intelligence and Statistics*, pp. 788–798. PMLR, 2020.

Muhammad Osama, Dave Zachariah, and Peter Stoica. Learning robust decision policies from observational data. *Advances in Neural Information Processing Systems*, 33:18205–18214, 2020.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

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

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Andrea Rotnitzky, Quanhong Lei, Mariela Sued, and James M Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456, 2012.

Donald B Rubin. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2(3):169–188, 2001.

Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4):279, 2008.

Henry Scheffe and John W Tukey. Non-parametric estimation. i. validation of order statistics. *The Annals of Mathematical Statistics*, 16(2):187–192, 1945.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. Learning from logged implicit exploration data. *Advances in neural information processing systems*, 23, 2010.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Muhammad Faaiz Taufiq, Jean-Francois Ton, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal off-policy prediction in contextual bandits. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=IfgOWI5v2f`.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Anastasios A Tsiatis, Marie Davidian, Shannon T Holloway, and Eric B Laber. *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490. PMLR, 2012.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Abraham Wald. An extension of wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1):45–55, 1943.

Lan Wang, Yu Zhou, Rui Song, and Ben Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

Daniel Westreich. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press, Incorporated, 2019. ISBN 9780190665760. URL `https://books.google.se/books?id=5R2yDwAAQBAJ`.

Samuel S Wilks. Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96, 1941.

Baqun Zhang, Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

Qingyuan Zhao, Dylan S Small, and Paul R Rosenbaum. Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association*, 113(523):1070–1084, 2018.

Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):735–761, 2019.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

A Appendix

A.1 Proof of Theorem 4.1

The first part of the proof builds on techniques used to derive weighted conformal prediction intervals in Tibshirani et al. (2019).

Let us consider a sequence of $n - n_0$ samples drawn from (4) followed by a new sample drawn from (3), i.e.,

$$\mathcal{D}_+ = \big((X_{n_0+1}, U_{n_0+1}, A_{n_0+1}, L_{n_0+1}), \ldots, (X_n, U_n, A_n, L_n), (X_{n+1}, U_{n+1}, A_{n+1}, L_{n+1})\big).$$

The joint distribution of this sequence can be expressed using:

$$\prod_{i=n_+}^{n} p(x_i, u_i, a_i, \ell_i) \cdot p(x_{n+1}, u_{n+1}, a_{n+1}, \ell_{n+1}) w_{n+1} = p(\mathcal{D}_+) w_{n+1} = p(\mathcal{S}_+) w_{n+1}, \tag{22}$$

where $n_+ = n_0 + 1$ for notational simplicity, $\mathcal{S}_+$ is an unordered set of elements from $\mathcal{D}_+$, and the weight

$$w_i = \frac{p_\pi(x_i, u_i, a_i, \ell_i)}{p(x_i, u_i, a_i, \ell_i)},$$

is the (unobservable) ratio (5) that quantifies the distribution shift from training to target distribution. We shall use the expression for the joint distribution to derive the distribution function for the new loss L_{n+1}.

Suppose we are given the unordered set $\mathcal{S}_+$ alone, then the particular sequence $\mathcal{D}_+$ is unknown. Let E_i denote the event that the sample $(X_{n+1}, U_{n+1}, A_{n+1}, L_{n+1})$ equals the ith sample (x_i, u_i, a_i, ℓ_i) in the unknown sequence $\mathcal{D}_+$. We consider all possible sequences $\mathcal{D}_+$ obtained by permutations σ of elements in the set $\mathcal{S}_+$. Using the joint distribution in (22), the joint probability of the event E_i and $\mathcal{S}_+$ is then

$$\mathbb{P}\{E_i, \mathcal{S}_+\} = \sum_{\sigma:\sigma(n+1)=n+i} p(\mathcal{S}_+) w_i = p(\mathcal{S}_+) w_i n!,$$

where we have used the symbol $\mathbb{P}$ for the samples that are drawn from two different processes (4) and (3).

The conditional probability of event E_i can now be expressed as

$$p_i = \mathbb{P}\{E_i|\mathcal{S}_+\} = \frac{\mathbb{P}\{E_i, \mathcal{S}_+\}}{\sum_{j=n_+}^{n+1} \mathbb{P}\{E_j, \mathcal{S}_+\}} = \frac{w_i}{\sum_{j=n_+}^{n+1} w_j},$$

where the first equality follows from the law of total probability. The probability that the loss L_{n+1} of the new sample equals ℓ_i, when conditioning on the unordered set $\mathcal{S}_+$, is equal to

$$\mathbb{P}\{L_{n+1} = \ell_i|\mathcal{S}_+\} = \mathbb{P}\{E_i|\mathcal{S}_+\} = p_i.$$

Thus conditional on $\mathcal{S}_+$, the new loss L_{n+1} has the following cdf:

$$\mathbb{P}\{L_{n+1} \leq \ell|\mathcal{S}_+\} = \sum_{i=n_+}^{n+1} p_i \mathbb{1}(\ell_i \leq \ell) = \frac{\sum_{i=n_+}^{n+1} w_i \mathbb{1}(L_i \leq \ell)}{\sum_{i=n_+}^{n+1} w_i}. \tag{23}$$

Before marginalizing out $\mathcal{S}_+$ from (23), we consider a limit ℓ that is a function of the observable elements in $\mathcal{S}_+$. For this part, we will build on the proof technique in (Jin et al., 2023, thm. 2.2).

Specifically, using (12) we define the following limit:

$$\ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1}) = \inf\left\{\ell : \widehat{F}(\ell; \mathcal{D}_1, \overline{W}_{n+1}) \geq \frac{1-\alpha}{1-\beta}\right\}, \tag{24}$$

for any $0 < \beta < \alpha$, where $\overline{W}_{n+1} \geq W_{n+1}$ is given in (11). Now insert the limit $\ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})$ into (23) and use the law of total expectation to marginalize out $\mathcal{S}_+$:

$$\begin{aligned}\mathbb{P}\{L_{n+1} \leq \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})\} &= \mathbb{E}\big[\mathbb{P}\{L_{n+1} \leq \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})|\mathcal{S}_+\}\big] \\ &= \mathbb{E}\left[\frac{\sum_{i=n_+}^{n+1} W_i \mathbb{1}(L_i \leq \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1}))}{\sum_{i=n_+}^{n+1} W_i}\right].\end{aligned}$$

We now proceed to lower bound this probability. Note that by construction:

$$\mathbb{E}\left[\widehat{F}(\ell_\alpha; \mathcal{D}_1, \overline{W}_{n+1})\right] = \mathbb{E}\left[\frac{\sum_{i\in\mathcal{D}_1} \underline{W}_i \mathbb{1}(L_i \leq \ell_\alpha)}{\sum_{i\in\mathcal{D}_1} \underline{W}_i \mathbb{1}(L_i \leq \ell_\alpha) + \sum_{i\in\mathcal{D}_1} \overline{W}_i \mathbb{1}(L_i > \ell_\alpha) + \overline{W}_{n+1}}\right] \geq \frac{(1-\alpha)}{(1-\beta)}.$$

Using this expression, we have that

$$\begin{aligned}&\mathbb{P}\{L_{n+1} \leq \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})\} - \frac{(1-\alpha)}{(1-\beta)} \\ &\geq \mathbb{E}\left[\frac{\sum_{i=n_+}^{n+1} W_i \mathbb{1}(L_i \leq \ell_\alpha)}{\sum_{i=n_+}^{n+1} W_i}\right] - \mathbb{E}\left[\frac{\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \leq \ell_\alpha)}{\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \leq \ell_\alpha) + \sum_{i=n_+}^{n} \overline{W}_i \mathbb{1}(L_i > \ell_\alpha) + \overline{W}_{n+1}}\right] \\ &= \mathbb{E}\left[\frac{(*)}{\left[\sum_{i=n_+}^{n+1} W_i\right]\left[\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \leq \ell_\alpha) + \sum_{i=n_+}^{n} \overline{W}_i \mathbb{1}(L_i > \ell_\alpha) + \overline{W}_{n+1}\right]}\right],\end{aligned}$$

where

$$
\begin{aligned}
(*) &= \left[\sum_{i=n_+}^{n+1} W_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \le \ell_\alpha) + \sum_{i=n_+}^{n} \overline{W}_i \mathbb{1}(L_i > \ell_\alpha) + \overline{W}_{n+1}\right] - \left[\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n+1} W_i\right] \\
&\ge \left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \le \ell_\alpha) + \sum_{i=n_+}^{n} \overline{W}_i \mathbb{1}(L_i > \ell_\alpha) + \overline{W}_{n+1}\right] - \left[\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n+1} W_i\right] \\
&\ge \left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n} \overline{W}_i \mathbb{1}(L_i > \ell_\alpha) + \overline{W}_{n+1}\right] - \left[\sum_{i=n_+}^{n} \underline{W}_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i > \ell_\alpha) + W_{n+1}\right] \\
&\ge \left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i > \ell_\alpha) + W_{n+1}\right] - \left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i \le \ell_\alpha)\right]\left[\sum_{i=n_+}^{n} W_i \mathbb{1}(L_i > \ell_\alpha) + W_{n+1}\right] \\
&= 0,
\end{aligned}
$$

using the bounds in (10) to get the third inequality. Therefore we obtain a valid limit:

$$
\mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})\} \ge \frac{(1-\alpha)}{(1-\beta)}. \tag{25}
$$

However, $\overline{W}_{n+1}$ depends on a new sample drawn from the training distribution and thus the limit is unusable. In lieu of $\overline{W}_{n+1}$, we use $\overline{w}_\beta(\mathcal{D}_0)$ in (14) to define the modified limit

$$
\ell_\alpha(\mathcal{D}) = \inf\left\{\ell : \widehat{F}(\ell; \mathcal{D}_1, \overline{w}_\beta(\mathcal{D}_0)) \ge \frac{1-\alpha}{1-\beta}\right\}. \tag{26}
$$

Comparing it with (24), we see that

$$
\ell_\alpha(\mathcal{D}) \ge \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1}), \tag{27}
$$

whenever $\overline{W}_{n+1} \le \overline{w}_\beta(\mathcal{D}_0)$. By the construction in (14), the probability of this event is lower bounded by

$$
\mathbb{P}\{\overline{W}_{n+1} \le \overline{w}_\beta(\mathcal{D}_0)\} \ge 1-\beta, \tag{28}
$$

see Vovk et al. (2005); Lei et al. (2018).

We use this property to lower bound the probability of $L_{n+1} \le \ell_\alpha(\mathcal{D})$. First, note that

$$
\begin{aligned}
\mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D})\} &= \mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D}) \mid \overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\}\, \mathbb{P}\{\overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\} \\
&\quad + \mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D}) \mid \overline{W}_{n+1} > \overline{w}^\beta(\mathcal{D}_0)\}\, \mathbb{P}\{\overline{W}_{n+1} > \overline{w}^\beta(\mathcal{D}_0)\} \\
&\ge \mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D}) \mid \overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\}\, \mathbb{P}\{\overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\} + 0.
\end{aligned}
$$

The first factor can be lower bounded using (27), so that

$$
\begin{aligned}
\mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D})\} &\ge \mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1}) \mid \overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\}\, \mathbb{P}\{\overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\} \\
&= \mathbb{P}\{L_{n+1} \le \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})\}\, \mathbb{P}\{\overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\} \\
&\ge \frac{(1-\alpha)}{(1-\beta)}\, \mathbb{P}\{\overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)\} \\
&\ge 1-\alpha.
\end{aligned} \tag{29}
$$

The second line follows from using sample splitting, which ensures that $L_{n+1} \le \ell_\alpha(\mathcal{D}_1, \overline{W}_{n+1})$ and $\overline{W}_{n+1} \le \overline{w}^\beta(\mathcal{D}_0)$ are independent events. The third and fourth lines follow from (25) and (28), respectively. Since (29) holds for any $0 < \beta < \alpha$, we choose β in (26) that yields the tightest limit, cf. (15). Note that $\mathbb{P}$ here represents the probability over samples drawn from both (4) and (3). Since L_{n+1} is drawn from (3), we can write (29) as in (2) for notational clarity.

A.2 Semi-real data

The Infant Health and Development Program (IHDP) investigated the impact of early childhood interventions on the health of low birth-weight and premature infants (Health & Program, 1990). Hill (2011) used this study to assemble a data set of 25 covariates X that measured various aspects of the children and their mothers such as birth weight, weeks born preterm, head circumference and age of mother, etc. These covariates were standardized to have a zero mean and unit standard deviation. The data set also includes whether each child received special medical care or not, denoted by action A. The original study was a randomised control study, but the data were imbalanced as a non-random subset of the treated population was removed. This unbalanced data set contains 747 children, including 139 treated and 608 control units.

The outcome was the child's cognitive development score, which we reverse it to an underdevelopment score to treat it as a loss L. The outcome is generated synthetically and we use 'response surface A' from Hill (2011):

$$L|A = 0, X \sim \mathcal{N}(-\varphi^\top X, 1), \quad L|A = 1, X \sim \mathcal{N}(-(\varphi^\top X + 4), 1).$$

The vector φ is discrete valued with elements drawn from $\{0, 1, 2, 3, 4\}$ with probabilities $\{0.5, 0.2, 0.15, 0.1, 0.05\}$. We use a fitted logistic regression model for $\widehat{p}(A|X)$. The limit (15) is computed using sample splits where $n_0 = 0.1n$. Another $0.1n$ samples of the data is randomly used to evaluate the policies in Figure 9.

We compare two policies: the child did not receive special medical care π_0 ('treat none') or the child received special medical care π_1 ('treat all'). The limit curves are presented in Figure 8 under different odds divergence bounds $\Gamma = \{1, 1.5, 2\}$ on the nominal model. When $\Gamma = 1$, the 'treat all' policy is the policy resulting in the lowest limit curve. When Γ is increased to 2 the limit curves are closer to each other, but the 'treat none' policy is instead the policy with the lowest limit curve.

Figure 9 validates our guarantee in (2). The miscoverage gap (17) is estimated using 1000 simulation runs where we draw a new vector φ and a new underdevelopment score L for each run. The proposed method is valid for all values of Γ, but is more conservative for the higher values. The benchmark does not take the unmeasured confounding into account and is invalid for the 'treat all' policy (right).

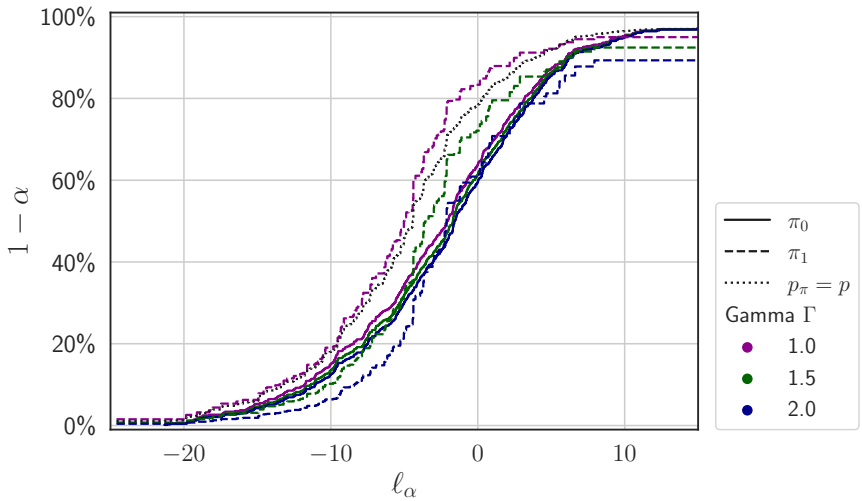


Figure 8: Limit curves for a synthetic underdevelopment score in a population under different policies. We compare two policies: the child did not receive special medical care π_0 or the child received special medical care π_1. The past policy is indicated by dotted curves ($p_\pi = p$). The credibility of the assumptions increases with the odds bound $\Gamma = \{1, 1.5, 2\}$.

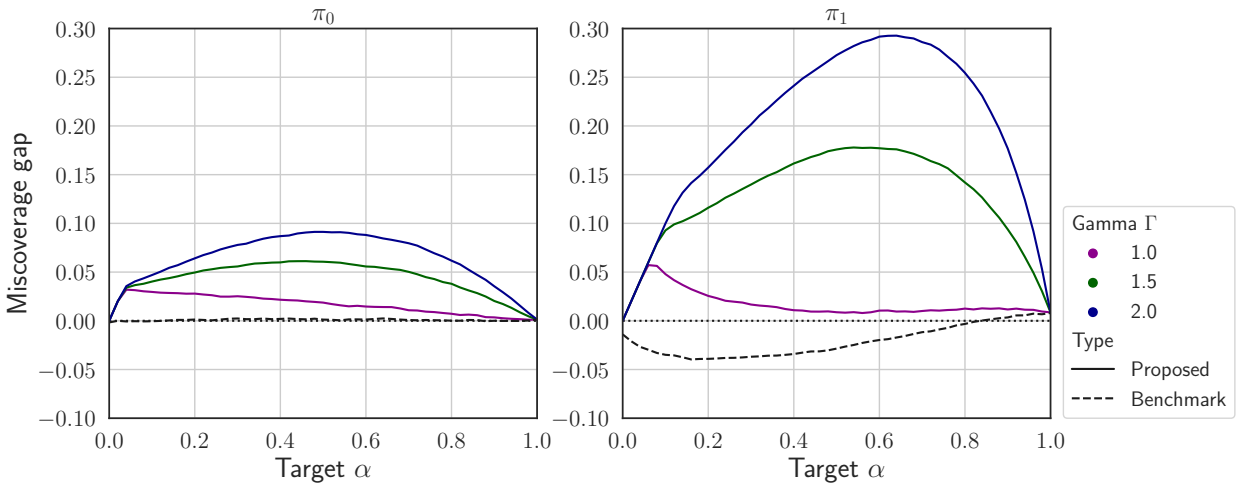


Figure 9: Miscoverage gaps (17) versus α for data from the Infant Health and Development Program (IHDP) under two different policies: the child did not receive special medical care π_0 or the child received special medical care π_1. Dashed curve denotes the benchmark (18) which does not take confounding or modelling errors into account. The credibility of the assumptions increases with the odds bound $\Gamma = \{1, 1.5, 2\}$.