Uncertainty Quantification and Causal Considerations for Off-Policy Decision Making

Muhammad Faaiz Taufiq

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

Acknowledgements

Abstract

Off-policy evaluation (OPE) is a critical challenge in robust decision-making that seeks to assess the performance of a new policy using data collected under a different policy. However, the existing OPE methodologies suffer from several limitations arising from statistical uncertainty as well as causal considerations. In this thesis, we address these limitations by presenting three different works.

Firstly, we consider the problem of high variance in the importance-sampling-based OPE estimators. We propose a novel off-policy evaluation estimator, the Marginal Ratio (MR) estimator, to alleviate this problem. By focusing on the marginal distribution of outcomes rather than the policy shift directly, the MR estimator achieves significant variance reduction compared to state-of-the-art methods, while maintaining unbiasedness.

Next, we shift our attention towards uncertainty quantification in off-policy evaluation. To this end, we propose Conformal Off-Policy Prediction (COPP) as a novel approach to quantify this uncertainty with finite-sample guarantees. Unlike traditional methods focusing on point estimates of expected outcomes, COPP provides reliable predictive intervals for outcomes under a target policy. This enables robust decision-making in risk-sensitive applications and offers a more comprehensive understanding of policy performance.

Finally, we address the fundamental challenge of causal inference in off-policy evaluation. Recognizing the limitations of traditional OPE methods under unmeasured confounding, we develop novel causal bounds for sequential decision settings that remain valid under arbitrary confounding. We apply these bounds for the assessment of digital twin models without relying on strong causal assumptions. We propose a framework for causal falsification, allowing us to identify scenarios where the digital twin's predictions diverge from real-world behavior. This approach provides valuable insights into model reliability and helps ensure safe and effective decision-making.

We conclude this thesis with a discussion of our contributions and limitations of the presented work, and outline interesting avenues for future research arising from our work.

Contents

1 Introduction

Contents

The ability to make well-informed decisions is crucial across a variety of domains. Whether it is a doctor prescribing the most effective treatment for a patient or a company launching a marketing campaign that resonates with its target audience [Xu et al., 2020, Li et al., 2010, Bastani and Bayati, 2019], we constantly strive to take actions that lead to desirable outcomes. However, achieving this goal becomes increasingly challenging in the face of uncertainty. Real-world data is often noisy and incomplete, and the systems we interact with are complex and constantly evolving. As machine learning models become more integrated into critical applications, the need for robust decision-making under these challenging conditions becomes paramount.

This thesis explores the key challenges of robust decision-making in machine learning, specifically focusing on *off-policy evaluation* [Kuzborskij et al., 2021, Wang et al., 2017a, Thomas et al., 2015, Swaminathan and Joachims, 2015c,d, Dudík et al., 2014a]. Consider the example of a doctor who wants to assess a new treatment for a disease. Ideally, they would conduct a randomized controlled trial [Tsiatis et al., 2019] where patients are

randomly assigned the new treatment or a standard one. However, such trials can be expensive, time-consuming or worse, ethically problematic. Off-policy evaluation (OPE) offers a compelling alternative by allowing us to evaluate the performance of a new decision-making policy (the new treatment) using data collected under a different policy (the standard treatment). This eliminates the need for costly experimentation and allows for quicker implementation of potentially more effective strategies.

However, off-policy evaluation presents its own set of challenges. These challenges stem from two main sources of uncertainty:

- **Statistical uncertainty:** This arises from the inherent randomness in the data we have access to and the limitations of the models we use to represent the real world. For instance, the doctor might have a limited number of patients in their historical dataset, and their model might not perfectly capture a patient's response to treatment due to model misspecification. In these circumstances, the conventional OPE methods may suffer from high variance and/or bias, thereby potentially resulting in misleading conclusions [Saito et al., 2021, Su et al., 2020, Saito and Joachims, 2022].

- **Causal unidentifiability:** In many cases, it may be impossible to definitively establish the causal effects of actions even if we had access to infinite data. This arises due to factors like confounding variables, which can influence both the treatment and the outcome. Imagine the existence of some unmeasured factors, such as a patient's pre-existing conditions, that can influence both their initial treatment and their response to the treatment. This makes it challenging to isolate the true causal effect of the new treatment from the influence of these confounding variables [Tsiatis et al., 2019, Kallus and Zhou, 2018, Namkoong et al., 2020].

This thesis tackles these challenges head-on, proposing novel methods for off-policy evaluation that address both statistical and causal uncertainties. Before we go into the specifics of these challenges, we introduce the framework of contextual bandits which forms the basis of the setting considered in Chapters 2 and 3.

1.1 Contextual bandits

Contextual bandits [Lattimore and Szepesvári, 2020] provide a powerful framework for tackling decision-making problems where the effectiveness of an action depends on the specific context in which it is chosen. For instance, in medical decision-making, the optimal treatment for a patient might depend on various factors such as their age, medical history, and current symptoms. Contextual bandits allow us to model these complex decision-making scenarios by incorporating the notion of context.

In this setting, we use covariates $X \in \mathcal{X}$ to denote features which encapsulate the contextual information such as the patient's age and medical history, we use $A \in \mathcal{A}$ to represent the action chosen by some real-world agent (such as a doctor), and $Y \in \mathcal{Y}$ to denote the outcome/reward observed as a result of taking action A, for example, $Y \in \{0, 1\}$ might represent whether a patient survives ($Y = 1$) or not ($Y = 0$). The goal of a learner in contextual bandits is to choose actions A for a context X which maximises the reward Y.

1.1.1 Off-policy evaluation in contextual bandits

Off-policy evaluation (OPE) tackles a crucial challenge in decision-making: assessing the performance of a new policy using data collected under a different policy [Swaminathan and Joachims, 2015a, Wang et al., 2017b, Farajtabar et al., 2018a, Su et al., 2019b, Metelli et al., 2021, Liu et al., 2019, Sugiyama and Kawanabe, 2012, Swaminathan et al., 2017b]. This is particularly valuable when conducting controlled experiments with the new policy is impractical or unethical. Here, we formally define the OPE problem in contextual bandits which will set up the challenges tackled in Chapters 2 and 3 of this thesis.

To be more concrete, let $\mathcal{D} \coloneqq \{(x_i, a_i, y_i)\}_{i=1}^n$ be a historically logged dataset with n observations, generated by a (possibly unknown) *behaviour policy* $\pi^b(a \mid x)$, i.e. the conditional distribution of agent's actions is $A \mid X = x \sim \pi^b(\cdot \mid x)$. Next, suppose that we are given a different target policy, which we denote by $\pi^*(a \mid x)$. Our goal is to estimate what the expected outcome *would* be if actions were instead sampled from this target policy π^*.

Off-policy evaluation (OPE)

The main objective of off-policy evaluation (OPE) is to estimate the expectation of the outcome Y under a given target policy π^* using only the logged data $\mathcal{D}$.

The key challenge of OPE arises from the fact that we do not have access to samples from the target distribution which makes the estimation of off-policy value non-trivial in general. To tackle this problem, the standard OPE methods make the following assumption.

Assumption 1.1.1 (No unmeasured confounding)**.** *The agent's action in the observational data A depends only on the context X and possibly additional randomness independent of everything else. This means that when choosing the action, the agent does not rely on additional information relevant to the outcome which is not captured in the context. For instance, in a medical context, this assumption means that all of the information that clinicians use to make treatment decisions is captured in the data. This assumption is also referred to as the* strong ignorability assumption *[Tsiatis et al., 2019].*

Then, under Assumption 1.1.1, the off-policy value can be estimated using importance-sampling-based methods. However, these estimators come with their own set of limitations, which are described in the following section and form the basis of our contributions in Chapters 2 and 3.

1.2 Limitations of existing OPE methods

1.2.1 High variance in OPE estimators

The conventional off-policy value estimators use policy ratios $\rho(a, x) \coloneqq \pi^*(a \mid x)/\pi^b(a \mid x)$ as importance weights. In cases where the two policies are significantly different, the policy ratios $\rho(a, x)$ attain extreme values leading to a high variance in the OPE estimators. To alleviate this high variance, Dudík et al. [2014b] proposed a *Doubly Robust (DR)* estimator for OPE. DR uses a control variate to decrease the variance of conventional OPE estimators. However, DR still relies on policy ratios as importance weights and as a result, also suffers from high variance when the policy shift is large. This problem is further exacerbated as the sizes of the action and context spaces grow [Sachdeva et al., 2020, Saito and Joachims, 2022]. Chapter 2 of this thesis specifically focuses on this limitation of OPE.

Besides using control variates (as in DR estimator), several techniques have been proposed to address the variance issues associated with importance weights.

Trading off variance for bias Swaminathan and Joachims [2015a,b], London and Sandler [2019] attempt to bound the importance weights within a certain range to prevent them from becoming excessively large. Besides this, the *Direct Method (DM)* [Beygelzimer and Langford, 2008] avoids the use of importance-sampling by estimating the reward function from observational data. Similarly, Switch-DR [Wang et al., 2017b] aims to circumvent the high variance in conventional DR estimator by switching to the Direct Method when the importance weights are large. However, these approaches introduce a bias-variance trade-off, as clipping the weights or using the learned reward function can introduce bias into the estimates.

Marginalization-based techniques Several works explore marginalisation techniques for variance reductions. For example, Saito and Joachims [2022] propose Marginalized Inverse Propensity Score (MIPS), which considers the marginal shift in the distribution of a lower dimensional embedding of the action space, denoted by E, instead of considering the shift in the policies explicitly. While this approach reduces the variance, we show in Chapter 2 that MIPS relies on an additional assumption regarding the action embeddings E which does not hold in general.

In addition, various marginalisation ideas have also been proposed in the context of reinforcement learning (RL). For example, Liu et al. [2018], Xie et al. [2019b], Kallus and Uehara [2022] use methods which consider the shift in the marginal distribution of the states, and apply importance weighting with respect to this marginal shift rather than the trajectory distribution. Similarly, Fujimoto et al. [2021] use marginalisation for OPE in deep RL, where the goal is to consider the shift in marginal distributions of state and action. Although marginalization is a key trick of these estimators, these techniques are aimed at resolving the curse of horizon, a problem specific to RL.

1.2.2 Lack of uncertainty quantification

Most techniques for OPE in contextual bandits focus on evaluating policies based on their *expected* outcomes [Kuzborskij et al., 2021, Wang et al., 2017a, Thomas et al., 2015, Swaminathan and Joachims, 2015c,d, Dudík et al., 2014a]. However, this can be problematic as methods that are only concerned with the average outcome do not take into account any notions of variance, for example. Therefore, in risk-sensitive settings such as econometrics,

where we want to minimize the potential risks, metrics such as CVaR (Conditional Value at Risk) might be more appropriate [Keramati et al., 2020]. Additionally, when only small sample sizes of observational data are available, the average outcomes under finite data can be misleading, as they are prone to outliers and hence, metrics such as medians or quantiles are more robust in these scenarios [Altschuler et al., 2019]. Next, we outline some recent works which tackle this challenge by developing methodologies to account for the uncertainty in off-policy performance using available data.

Off-policy risk assessment in contextual bandits Instead of estimating bounds on the expected outcomes, Huang et al. [2021], Chandak et al. [2021] establish finite-sample bounds for a general class of metrics (e.g., Mean, CVaR, CDF) on the outcome. Their methods can be used to estimate quantiles of the outcomes under the target policy and are therefore robust to outliers. For example, Chandak et al. [2021] proposed a non-parametric Weighted Importance Sampling (WIS) estimator for the empirical CDF of Y under π^*, which can be used to construct predictive intervals on the outcome under target policy. This can help us quantify the range of plausible outcomes Y that are likely to occur if actions are chosen according to target policy π^*. However, the resulting bounds do not depend on the context X (i.e., are not adaptive w.r.t. X). This can lead to overly conservative intervals, which may not be very informative. In Chapter 3, we circumvent this problem by proposing a methodology of constructing predictive intervals on Y under target policy π^* which are adaptive w.r.t. context X and are therefore considerably more informative.

1.3 Sequential decision setting

Contextual bandits encapsulate the single-decision regimes where, for each observed context, we take a single action and observe the resulting outcome. This is analogous to a doctor choosing a single treatment for a patient based on their current state. However, many real-world decision-making scenarios involve multiple interventions over time, where each action not only affects the immediate outcome but also influences the context for future decisions. To capture this complexity, we introduce the sequential decision setting in this section.

This setting extends the framework of contextual bandits to handle sequential decision-making problems, allowing us to model more complex scenarios where interventions unfold over time and the context evolves dynamically.

We consider a setting with a fixed number of decisions per episode (i.e., a fixed time horizon) $T \in \{1, 2, \ldots\}$. For each $t \in \{0, \ldots, T\}$, we assume that the process gives rise to an observation at time t, denoted by X_t which takes values in some space $\mathcal{X}_t \coloneqq \mathbb{R}^{d_t}$. Moreover, at time $t \in \{1, \ldots, T\}$ a real-world agent (such as a doctor) chooses an action A_t which takes values in some space $\mathcal{A}_t$. The agent's choice of A_t may depend on the historical observations $(X_0, \ldots, X_{t-1})$ or any additional information not captured in historical observations that the agent can access. For example, in a medical context, the observations may consist of a patient's vital signs, and the actions may consist of possible treatments or interventions that the doctor chooses based on patient history. The actions taken up till time t, i.e. $(A_1, A_2, \ldots, A_t)$ can influence the future observations $(X_t, X_{t+1}, \ldots, X_T)$. This describes the sequential decision setting, of which the contextual bandits are a special case when $T = 1$.

1.3.1 Assumption of no unmeasured confounding

Most of the standard OPE methods for contextual bandits can be straightforwardly extended to sequential decision settings [Uehara et al., 2022]. However, like in contextual bandits, these estimators assume no unmeasured confounding (outlined in Assumption 1.1.1) in the available observational data. This assumption is unverifiable from the observational data alone and is violated in many real-world circumstances where some information not captured in the data influences not only the action chosen A_t, but also the future observations $(X_t, X_{t+1}, \ldots, X_T)$. This can happen when the real-world agent has access to more information than is captured in the data. In such circumstances, the causal effect of a given action sequence may be unidentifiable from the available observational data, making it impossible to accurately estimate the value of the target policy. To make this concrete, we provide an intuitive illustration of this phenomenon below using a toy example where the available observational suffers from unmeasured confounding.

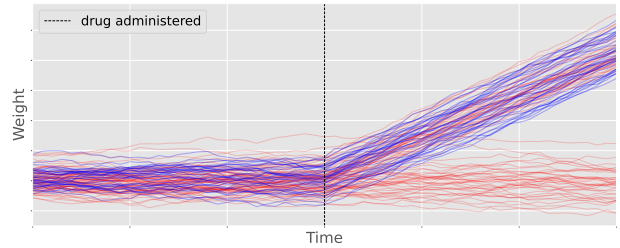


Figure 1.1: The discrepancy between observational data and interventional behaviour in the presence of unmeasured confounding: the range of outcomes observed in the data for patients who were administered the drug (blue) differs from what *would* be observed if the drug were administered to the general population (red).

Toy example: Unmeasured confounding in medical decision-making

Suppose that we are interested in estimating the effect of a drug on the weight of patients in a certain population. Moreover, assume that this drug interacts with an enzyme that is only present in part of the population. Denote by $U \in \{0, 1\}$ the presence or absence of the enzyme in a patient, and assume that when $U = 1$ the patient's weight increases after action the drug is administered, and that when $U = 0$ the drug has no effect. Additionally, suppose that, among the patients whose data we have obtained, the drug was only prescribed to those for whom $U = 1$, perhaps on the basis of some initial lab reports available to the prescriber. Finally, suppose that these lab results were *not* included in the context X captured in the observational dataset $\mathcal{D}$, so that the value of U for each patient cannot be determined from the data we have available.

In this setup, since the drug was only administered to patients with $U = 1$, it would appear from the data that the drug causes patient weight to increase. However, when the drug is administered to the general population, i.e. regardless of the value of U, we would observe that the drug has no effect on patients for whom $U = 0$. Figure 1.1 illustrates this discrepancy under a toy model for this scenario. In this example, since the data $\mathcal{D}$ contains no information about the presence or absence of the enzyme in patients, U, it is impossible to determine using the data $\mathcal{D}$ alone how the drug will affect a given population of patients.

Causal considerations under unmeasured confounding

Here, we elaborate further on the implications of unmeasured confounding on the identifiability of causal effects in sequential decision setting. This topic forms the central

focus of Chapter 4 of this thesis.

Informally, the assumption of no unmeasured confounding holds when each action A_t is chosen by the behavioural agent solely on the basis of the information available at time t that is actually recorded in the dataset, namely $X_0, A_1, X_1, \ldots, A_{t-1}, X_{t-1}$, as well as possibly some additional randomness that is independent of the real-world process, such as the outcome of a coin toss. Unobserved confounding is present whenever this does not hold, i.e. whenever some unmeasured factor simultaneously influences both the agent's choice of action and the observation produced by the real-world process.

While it may be reasonable to assume that the data are unconfounded in certain contexts. For example, in certain situations it may be possible to gather data in a way that specifically guarantees there is no confounding. Randomised controlled trials, which ensure that each A_t is chosen via a carefully designed randomisation procedure [Lavori and Dawson, 2004, Murphy, 2005], constitute a widespread example of this approach. However, for typical datasets, it is widely acknowledged that the assumption of no unmeasured confounding will rarely hold, and so OPE procedures based on this assumption may yield unreliable results in practice [Murphy, 2003, Tsiatis et al., 2019]. This result is formalised in a foundational result from the causal inference literature, often referred to as the *fundamental problem of causal inference* [Holland, 1986].

Fundamental problem of causal inference (informal statement)

The causal effect of an action is not uniquely identified by the observational data distribution in the presence of unmeasured confounding (without additional assumptions).

Partial identification Since the precise identifiability of causal effects is not possible in the presence of unmeasured confounding, a notable line of work instead explores partial identification techniques [Manski, 1990, 1989, 2003]. Instead of the point identification of causal effects which may require strong unconfounding assumptions, partial identification typically considers the range of causal effects which may occur in the presence of confounding. For example, Manski [1990] constructs sharp bounds on the causal effects which can be readily estimated using the available observational data. While these bounds do not require any strong assumptions, they can be conservative.

Sensitivity analysis Slightly stronger assumptions yield inferences that may be more powerful but less credible. To this end, Rosenbaum [2002] proposes a classical model of confounding for a single binary decision setting which posits that the unobserved confounders have a limited influence on the agent's actions in the real world. Namkoong et al. [2020] extend this model to the multi-action sequential decision-making setting, and subsequently use this to obtain bounds on the off-policy value.

The Rosenbaum model is also closely related to (albeit different from) the marginal sensitivity model introduced by Tan [2006] which also assumes bounds on the strength of unmeasured confounding on agent's actions. Subsequently, Kallus and Zhou [2020] uses the marginal sensitivity model to develop a policy learning algorithm which remains robust to unmeasured confounding. However, these models impose assumptions on the strength of unmeasured confounding which can be impossible to verify using observational data alone, and therefore the inferences obtained may be misleading in many cases.

Proxy causal learning This comprises methodologies for estimating the causal effect of actions on outcomes in the presence of unobserved confounding, using *proxy variables* which contain relevant side information about the unmeasured confounders [Xu et al., 2021, Tchetgen et al., 2020, Xu and Gretton, 2024]. This usually involves a two-stage regression. First, the relationship between action and proxies is modelled and subsequently, this model is used to learn the causal effect of actions on the outcomes. Kuroki and Pearl [2014] outline the necessary conditions on proxy variables to obtain the true causal effects. While proxy causal learning may be effective in cases where such proxy variables are available, in many real-world settings the available proxy variables may not satisfy the necessary conditions for identification of true causal effects.

Chapter 4 of this thesis considers the challenges posed by unmeasured confounding in sequential decision setting. We propose a set of novel bounds on the causal effects in this setting, which remain valid in the presence of arbitrary unmeasured confounding and rely on minimal assumptions making them highly applicable to a wide variety of real-world settings.

1.4 Contributions and thesis outline

Having outlined some of the key challenges associated with off-policy evaluation, we dedicate the rest of this thesis to addressing each of these individually. Specifically, this thesis is organised as follows:

Chapter 2: Variance reduction [Taufiq et al., 2023b] The first challenge we consider is that of high variance in existing OPE estimators based on importance sampling. As we mentioned in Section 1.2.1, this variance is exacerbated in cases where there is low overlap between behaviour and target policies, or where the action or context space is high-dimensional. To address this challenge, we propose a novel OPE estimator for contextual bandits, the Marginal Ratio (MR) estimator, which uses a marginalisation technique to focus on the shift in the marginal distribution of outcomes Y directly, instead of the policies themselves. Unlike the conventional approaches like IPW and DR estimators, intuitively our proposed estimator treats actions A and contexts X as latent variables. As a result, the resulting estimator is significantly more robust to the overlap between policies and the sizes of action and/or context spaces. This chapter also includes extensive theoretical and empirical analyses demonstrating the benefits of the MR estimator compared to the state-of-the-art OPE estimators for contextual bandits.

Chapter 3: Uncertainty quantification [Taufiq et al., 2022] As explained in Section 1.2.2, most OPE methods have focused on the expected outcome of a policy which does not capture the variability of the outcome Y. In addition, many of these methods provide only asymptotic guarantees of validity at best. In this chapter, we address these limitations by considering a novel application of conformal prediction [Vovk et al., 2005] to contextual bandits. Given data collected under a behavioral policy, we propose *conformal off-policy prediction* (COPP), which can output reliable predictive intervals for the outcome under a new target policy. We provide theoretical finite-sample guarantees without making any additional assumptions beyond the standard contextual bandit setup, and empirically demonstrate the utility of COPP compared with existing methods on synthetic and real-world data.

Chapter 4: Causal considerations [Cornish et al., 2023] In this chapter we consider the sequential decision setting, where the available observational data may suffer from unmeasured confounding. As mentioned in Section 1.3.1, fundamental results from causal inference mean that in this setting the interventional behaviour of outcomes is unidentifiable from the observational distribution. To address this challenge, we provide a novel set of longitudinal causal bounds that remain valid under arbitrary unmeasured confounding.

Chapter 4 focuses on the application of these bounds for assessing the accuracy of *digital twin models*. These models are virtual systems designed to predict how a real-world process will evolve in response to interventions. To be considered accurate, these models must correctly capture the true interventional behaviour of outcomes. Unfortunately, the causal unidentifiability results mean observational data cannot be used to certify a twin in this sense if the data are confounded. To circumvent this, we instead use our proposed causal bounds to find situations in which the twin *is not* correct, and present a general-purpose statistical procedure for doing so. Our approach yields reliable and actionable information about the twin under only the assumption of an i.i.d. dataset of observational trajectories, and remains sound even if the data are confounded.

Chapter 5: Conclusion Finally, we conclude by summarising the main findings of the works presented in this thesis. In this chapter, we also discuss some of the limitations of our proposed methodologies and mention some interesting avenues for future research arising from these works.

1.5 An overview of work conducted during the DPhil

In this section, we provide an overview of the research conducted during the doctoral studies by listing the papers which are included in this thesis, as well those which have been omitted.

1.5.1 Works included in the thesis

Each chapter of this thesis is based on a paper. These papers are listed in chronological order here for completeness.

1. **Muhammad Faaiz Taufiq***, Jean-Francois Ton*, Rob Cornish, Yee Whye Teh, and Arnaud Doucet. Conformal Off-Policy Prediction in Contextual Bandits. In *Advances in Neural Information Processing Systems, 2022*. [Taufiq et al., 2022]

2. Rob Cornish*, **Muhammad Faaiz Taufiq***, Arnaud Doucet, and Chris Holmes. Causal Falsification of Digital Twins, 2023. *Preprint*. [Cornish et al., 2023]

3. **Muhammad Faaiz Taufiq**, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal Density Ratio for Off-Policy Evaluation in Contextual Bandits. In *Advances in Neural Information Processing Systems, 2023*. [Taufiq et al., 2023b]

1.5.2 Works omitted from the thesis

For the purposes of coherence and conciseness, several works which were part of the doctoral research have been ommitted from this thesis. Here, we list these papers along with a brief description in chronological order for completeness.

1. **Muhammad Faaiz Taufiq**, Patrick Blöbaum, and Lenon Minorics. Manifold Restricted Interventional Shapley Values. In *International Conference on Artificial Intelligence and Statistics, 2023*. [Taufiq et al., 2023a]

2. **Muhammad Faaiz Taufiq**, Jean-Francois Ton, and Yang Liu. Achievable Fairness on your Data with Utility Guarantees. Under review at *NeurIPS 2024*. [Taufiq et al., 2024]

3. Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, **Muhammad Faaiz Taufiq**, and Hang Li. Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment, 2023. In *NeurIPS 2023 Workshop on Socially Responsible Language Modelling Research (SoLaR)*. [Liu et al., 2024]

In Taufiq et al. [2023a], we consider the robustness of Shapley values, which are model-agnostic methods for explaining model predictions. Many commonly used methods of computing Shapley values, known as off-manifold methods, are sensitive to model behaviour outside the data distribution. This makes Shapley explanations highly sensitive to off-manifold perturbation of models, resulting in misleading explanations. To circumvent

this problem, we propose *ManifoldShap*, which respects the model's domain of validity by restricting model evaluations to the data manifold. We show, theoretically and empirically, that ManifoldShap is robust to off-manifold perturbations of the model and leads to more accurate and intuitive explanations than existing state-of-the-art Shapley methods.

Beyond this, Taufiq et al. [2024] considers fairness within the context of machine learning models. In this setting, training models that minimize disparity across different sensitive groups often leads to diminished accuracy, a phenomenon known as the fairness-accuracy tradeoff. The severity of this trade-off inherently depends on dataset characteristics such as dataset imbalances or biases and therefore, using a uniform fairness requirement across diverse datasets remains questionable. To address this, we present a computationally efficient approach to approximate the fairness-accuracy trade-off curve tailored to individual datasets, backed by rigorous statistical guarantees. Crucially, we introduce a novel methodology for quantifying uncertainty in our estimates, thereby providing practitioners with a robust framework for auditing model fairness while avoiding false conclusions due to estimation errors.

Finally, Liu et al. [2024] presents a comprehensive survey of key dimensions that are crucial to consider when assessing the trustworthiness of Large Language Models (LLMs). The survey covers seven major categories of LLM trustworthiness: reliability, safety, fairness, resistance to misuse, explainability and reasoning, adherence to social norms, and robustness. The empirical results presented in this study indicate that, in general, more aligned models tend to perform better in terms of overall trustworthiness. However, the effectiveness of alignment varies across the different trustworthiness categories considered. This highlights the importance of conducting more fine-grained analyses, testing, and making continuous improvements on LLM alignment.

2

Marginal Density Ratio for Off-Policy Evaluation in Contextual Bandits

Contents

Abstract

Off-Policy Evaluation (OPE) in contextual bandits is crucial for assessing new policies using existing data without costly experimentation. However, current OPE methods, such as Inverse Probability Weighting (IPW) and Doubly Robust (DR) estimators, suffer from high variance, particularly in cases of low overlap between target and behavior policies or large action and context spaces. In this paper, we introduce a new OPE estimator for contextual bandits, the Marginal Ratio (MR) estimator, which focuses on the shift in the marginal distribution of outcomes Y instead of the policies themselves. Through rigorous theoretical analysis, we demonstrate the benefits of the MR estimator compared to conventional methods like IPW and DR in terms of variance reduction. Additionally, we establish a connection between the MR estimator and the state-of-the-art Marginalized Inverse Propensity Score (MIPS) estimator, proving that MR achieves lower variance among a generalized family of MIPS estimators. We further illustrate the utility of the MR estimator in causal inference settings, where it exhibits enhanced performance in estimating Average Treatment Effects (ATE). Our experiments on synthetic and real-world datasets corroborate our theoretical findings and highlight the practical advantages of the MR estimator in OPE for contextual bandits.

2.1 Introduction

In contextual bandits, the objective is to select an action A, guided by contextual information X, to maximize the resulting outcome Y. This paradigm is prevalent in many real-world applications such as healthcare, personalized recommendation systems, or online advertising [Li et al., 2010, Bastani and Bayati, 2019, Xu et al., 2020]. The objective is to perform actions, such as prescribing medication or recommending items, which lead to desired outcomes like improved patient health or higher click-through rates. Nonetheless, updating the policy presents challenges, as naïvely implementing a new, untested policy may raise ethical or financial concerns. For instance, prescribing a drug based on a new policy poses risks, as it may result in unexpected side effects. As a result, recent research [Swaminathan and Joachims, 2015a, Wang et al., 2017b, Farajtabar et al., 2018a, Su et al., 2019b, Metelli et al., 2021, Liu et al., 2019, Sugiyama and Kawanabe, 2012, Swaminathan et al., 2017b] has concentrated on evaluating the performance of new policies (target policy) using only existing data that was generated using the current policy (behaviour policy). This problem is known as Off-Policy Evaluation (OPE).

Current OPE methods in contextual bandits, such as the Inverse Probability Weighting (IPW) [Horvitz and Thompson, 1952] and Doubly Robust (DR) [Dudík et al., 2014b] estimators primarily account for the policy shift by re-weighting the data using the ratio of the target and behaviour polices to estimate the target policy value. This can be problematic as it may lead to high variance in the estimators in cases of substantial policy shifts. The issue is further exacerbated in situations with large action or context spaces [Saito and Joachims, 2022], since in these cases the estimation of policy ratios is even more difficult leading to extreme bias and variance.

In this work we show that this problem of high variance in OPE can be alleviated by using methods which directly consider the shift in the marginal distribution of the outcome Y resulting from the policy shift, instead of considering the policy shift itself (as in IPW and DR). To this end, we propose a new OPE estimator for contextual bandits called the Marginal Ratio (MR) estimator, which weights the data directly based on the shift in the marginal distribution of outcomes Y and consequently is much more robust to increasing sizes of action and context spaces than existing methods like IPW or DR. Our extensive theoretical analyses show that MR enjoys better variance properties

than the existing methods making it highly attractive for a variety of applications in addition to OPE. One such application is the estimation of Average Treatment Effect (ATE) in causal inference, for which we show that MR provides greater sample efficiency than the most commonly used methods.

Our contributions in this paper are as follows:

- Firstly, we introduce MR, an OPE estimator for contextual bandits, that focuses on the shift in the marginal distribution of Y rather than the joint distribution of (X, A, Y). We show that MR has favourable theoretical properties compared to existing methods like IPW and DR. Our analysis also encompasses theory on the approximation errors of our estimator.

- Secondly, we explicitly lay out the connection between MR and Marginalized Inverse Propensity Score (MIPS) [Saito and Joachims, 2022], a recent state-of-the-art contextual bandits OPE method, and prove that MR attains lowest variance among a generalized family of MIPS estimators.

- Thirdly, we show that the MR estimator can be applied in the setting of causal inference to estimate average treatment effects (ATE), and theoretically prove that the resulting estimator is more data-efficient with higher accuracy and lower variance than commonly used methods.

- Finally, we verify all our theoretical analyses through a variety of experiments on synthetic and real-world datasets and empirically demonstrate that the MR estimator achieves better overall performance compared to current state-of-the-art methods.

2.2 Background

2.2.1 Setup and Notation

We consider the standard contextual bandit setting. Let $X \in \mathcal{X}$ be a context vector (e.g., user features), $A \in \mathcal{A}$ denote an action (e.g., recommended website to the user), and $Y \in \mathcal{Y}$ denote a scalar reward or outcome (e.g., whether the user clicks on the website). The outcome and context are sampled from unknown probability distributions $p(y \mid x, a)$ and $p(x)$ respectively. Let $\mathcal{D} \coloneqq \{(x_i, a_i, y_i)\}_{i=1}^n$ be a historically logged dataset with n

observations, generated by a (possibly unknown) *behaviour policy* $\pi^b(a \mid x)$. Specifically, $\mathcal{D}$ consists of i.i.d. samples from the joint density under *behaviour policy*,

$$p_{\pi^b}(x, a, y) \coloneqq p(y \mid x, a)\, \pi^b(a \mid x)\, p(x). \tag{2.1}$$

We denote the joint density of (X, A, Y) under the *target policy* as

$$p_{\pi^*}(x, a, y) \coloneqq p(y \mid x, a)\, \pi^*(a \mid x)\, p(x). \tag{2.2}$$

Moreover, we use $p_{\pi^b}(y)$ to denote the marginal density of Y under the behaviour policy,

$$p_{\pi^b}(y) = \int_{\mathcal{A}\times\mathcal{X}} p_{\pi^b}(x, a, y)\, \mathrm{d}a\, \mathrm{d}x,$$

and likewise for the target policy π^*. Similarly, we use $\mathbb{E}_{\pi^b}$ and $\mathbb{E}_{\pi^*}$ to denote the expectations under the joint densities $p_{\pi^b}(x, a, y)$ and $p_{\pi^*}(x, a, y)$ respectively.

Off-policy evaluation (OPE) The main objective of OPE is to estimate the expectation of the outcome Y under a given target policy π^*, i.e., $\mathbb{E}_{\pi^*}[Y]$, using only the logged data $\mathcal{D}$.

Throughout this work, we assume that the support of the target policy π^* is included in the support of the behaviour policy π^b. This is to ensure that importance sampling yields unbiased off-policy estimators, and is satisfied for exploratory behaviour policies such as the ϵ-greedy policies.

Assumption 2.2.1 (Support)**.** *For any* $x \in \mathcal{X}, a \in \mathcal{A}$, $\pi^*(a \mid x) > 0 \implies \pi^b(a \mid x) > 0$.

2.2.2 Existing off-policy evaluation methodologies

Next, we will present some of the most commonly used OPE estimators before outlining the limitations of these methodologies. This motivates our proposal of an alternative OPE estimator.

The value of the target policy can be expressed as the expectation of outcome Y under the target data distribution $p_{\pi^*}(x, a, y)$. However in most cases, we do not have access to samples from this target distribution and hence we have to resort to importance sampling methods.

Inverse Probability Weighting (IPW) estimator One way to compute the target policy value, $\mathbb{E}_{\pi^*}[Y]$, when only given data generated from $p_{\pi^b}(x, a, y)$ is to rewrite the policy value as follows:

$$\mathbb{E}_{\pi^*}[Y] = \int y\, p_{\pi^*}(x,a,y)\,\mathrm{d}y\,\mathrm{d}a\,\mathrm{d}x = \int y\, \underbrace{\frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(x,a,y)}}_{\rho(a,x)}\, p_{\pi^b}(x,a,y)\,\mathrm{d}y\,\mathrm{d}a\,\mathrm{d}x = \mathbb{E}_{\pi^b}\left[Y\,\rho(A,X)\right],$$

where $\rho(a,x) \coloneqq \frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(x,a,y)} = \frac{\pi^*(a|x)}{\pi^b(a|x)}$, given the factorizations in Eqns. (2.1) and (2.2). This leads to the commonly used *Inverse Probability Weighting (IPW)* [Horvitz and Thompson, 1952] estimator:

$$\hat{\theta}_{\mathrm{IPW}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho(a_i, x_i)\, y_i.$$

When the behaviour policy is known, IPW is an unbiased and consistent estimator. However, it can suffer from high variance, especially as the overlap between the behaviour and target policies decreases.

Doubly Robust (DR) estimator To alleviate the high variance of IPW, Dudík et al. [2014b] proposed a *Doubly Robust (DR)* estimator for OPE. DR uses an estimate of the conditional mean $\hat{\mu}(a,x) \approx \mathbb{E}[Y \mid X = x, A = a]$ (*outcome model*), as a control variate to decrease the variance of IPW. It is also doubly robust in that it yields accurate value estimates if either the importance weights $\rho(a,x)$ or the outcome model $\hat{\mu}(a,x)$ is well estimated [Dudík et al., 2014b, Jiang and Li, 2016]. The DR estimator for $\mathbb{E}_{\pi^*}[Y]$ can be written as follows:

$$\hat{\theta}_{\mathrm{DR}} = \frac{1}{n}\sum_{i=1}^{n} \rho(a_i, x_i)\,(y_i - \hat{\mu}(a_i, x_i)) + \hat{\eta}(\pi^*),$$

where $\hat{\eta}(\pi^*) = \frac{1}{n}\sum_{i=1}^{n}\sum_{a'\in\mathcal{A}} \hat{\mu}(a', x_i)\pi^*(a' \mid x_i) \approx \mathbb{E}_{\pi^*}[\hat{\mu}(A,X)]$. Here, $\hat{\eta}(\pi^*)$ is referred to as the Direct Method (DM) as it uses $\hat{\mu}(a,x)$ directly to estimate target policy value.

2.2.3 Limitation of existing methodologies

To estimate the value of the target policy π^*, the existing methodologies consider the shift in the joint distribution of (X, A, Y) as a result of the policy shift (by weighting samples by policy ratios). As we show in Section 2.3.1, considering the joint shift can lead to inefficient policy evaluation and high variance especially as the policy shift increases [Li et al., 2018]. Since our goal is to estimate $\mathbb{E}_{\pi^*}[Y]$, we will show in the next section that considering

only the shift in the marginal distribution of the outcomes Y from $p_{\pi^b}(Y)$ to $p_{\pi^*}(Y)$, leads to a more efficient OPE methodology compared to existing approaches.

To better comprehend why only considering the shift in the marginal distribution is advantageous, let us examine an extreme example where we assume that $Y \perp\!\!\!\perp A \mid X$, i.e., the outcome Y for a user X is independent of the action A taken. In this specific instance, $\mathbb{E}_{\pi^*}[Y] = \mathbb{E}_{\pi^b}[Y] \approx 1/n \sum_{i=1}^{n} y_i$, indicating that an unweighted empirical mean serves as a suitable unbiased estimator of $\mathbb{E}_{\pi^*}[Y]$. However, IPW and DR estimators use policy ratios $\rho(a, x) = \frac{\pi^*(a|x)}{\pi^b(a|x)}$ as importance weights. In case of large policy shifts, these ratios may vary significantly, leading to high variance in IPW and DR.

In this particular example, the shift in policies is inconsequential as it does not impact the distribution of outcomes Y. Hence, IPW and DR estimators introduce additional variance due to the policy ratios when they are not actually required. This limitation is not exclusive to this special case; in general, methodologies like IPW and DR exhibit high variance when there is low overlap between target and behavior policies [Li et al., 2018] even if the resulting shift in marginals of the outcome Y is not significant.

Therefore, we propose the *Marginal Ratio (MR)* OPE estimator for contextual bandits in the subsequent section, which circumvents these issues by focusing on the shift in the marginal distribution of the outcomes Y. Additionally, we provide extensive theoretical insights on the comparison of MR to existing state-of-the-art methods, such as IPW and DR.

2.3 Marginal Ratio (MR) estimator

Our method's key insight involves weighting outcomes by the marginal density ratio of outcome Y:

$$\mathbb{E}_{\pi^*}[Y] = \int_{\mathcal{Y}} y \, p_{\pi^*}(y) \, \mathrm{d}y = \int_{\mathcal{Y}} y \, \frac{p_{\pi^*}(y)}{p_{\pi^b}(y)} \, p_{\pi^b}(y) \, \mathrm{d}y = \mathbb{E}_{\pi^b}\left[Y \, w(Y)\right],$$

where $w(y) \coloneqq \frac{p_{\pi^*}(y)}{p_{\pi^b}(y)}$. This leads to the Marginal Ratio OPE estimator:

$$\hat{\theta}_{\mathrm{MR}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} w(y_i) \, y_i.$$

In Section 2.3.1 we prove that by only considering the shift in the marginal distribution of outcomes, the MR estimator achieves a lower variance than the standard OPE methods. In fact, this estimator does not depend on the shift between target and behaviour policies directly. Instead, it depends on the shift between the marginals $p_{\pi^b}(y)$ and $p_{\pi^*}(y)$.

Estimation of $w(y)$ When the weights $w(y)$ are known exactly, the MR estimator is unbiased and consistent. However, in practice the weights $w(y)$ are often not known and must be estimated using the logged data $\mathcal{D}$. Here, we outline an efficient way to estimate $w(y)$ by first representing it as a conditional expectation, which can subsequently be expressed as the solution to a regression problem.

Lemma 2.3.1

Let $w(y) = \frac{p_{\pi^*}(y)}{p_{\pi^b}(y)}$ and $\rho(a,x) = \frac{\pi^*(a|x)}{\pi^b(a|x)}$, then $w(y) = \mathbb{E}_{\pi^b}\left[\rho(A,X) \mid Y = y\right]$, and,

$$w = \arg\min_f \mathbb{E}_{\pi^b}\left[(\rho(A,X) - f(Y))^2\right]. \tag{2.3}$$

Lemma 2.3.1 allows us to approximate $w(y)$ using a parametric family $\{f_\phi : \mathbb{R} \to \mathbb{R} \mid \phi \in \Phi\}$ (e.g. neural networks) and defining $\hat{w}(y) \coloneqq f_{\phi^*}(y)$, where ϕ^* solves the regression problem in Eq. (2.3).

Note that MR can also be estimated alternatively by directly estimating $h(y) \coloneqq w(y)\, y$ using a similar regression technique as above and computing $\hat{\theta}_{\text{MR}} = 1/n \sum_{i=1}^n h(y_i)$. We include additional details along with empirical comparisons in Appendix A.6.1.

2.3.1 Theoretical analysis

Recall that the traditional OPE estimators like IPW and DR use importance weights which account for the the shift in the joint distributions of (X, A, Y). In this section, we prove that by considering only the shift in the marginal distribution of Y instead, MR achieves better variance properties than these estimators. Our analysis in this subsection assumes that the ratios $\rho(a,x)$ and $w(y)$ are known exactly. Since the OPE estimators considered are unbiased in this case, our analysis of variance is analogous to that of the mean squared error (MSE) here. We address the case where the weights are not known exactly in Section 2.3.1. First, we make precise our intuition that the shift in the joint distribution of (X, A, Y) is 'greater' than the shift in the marginal distribution of outcomes Y. We formalise this using the notion of f-divergences.

Proposition 2.3.1

Let $f : [0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$, and $\mathrm{D}_f(P||Q)$ denotes the f-divergence between distributions P and Q. Then,

$$\mathrm{D}_f\left(p_{\pi^*}(x, a, y) \,||\, p_{\pi^b}(x, a, y)\right) \geq \mathrm{D}_f\left(p_{\pi^*}(y) \,||\, p_{\pi^b}(y)\right).$$

Intuition Proposition 2.3.1 shows that the shift in the joint distributions is at least as 'large' as the shift in the marginals of the outcome Y. Traditional OPE estimators, therefore take into consideration more of a distribution shift than needed, and consequently lead to inefficient estimators. In contrast, the MR estimator mitigates this problem by only considering the shift in the marginal distributions of outcomes resulting from the policy shift. This provides further intuition on why the MR estimator has lower variance compared to existing methods.

Proposition 2.3.2 (Variance comparison with IPW estimator)

When the weights $\rho(a, x)$ and $w(y)$ are known exactly, we have that $\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] \leq \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}]$. In particular,

$$\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] = \frac{1}{n}\mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\rho(A, X) \mid Y\right] Y^2\right] \geq 0.$$

Intuition Proposition 2.3.2 shows that the variance of MR estimator is smaller than that of the IPW estimator when the weights are known exactly. Moreover, the proposition also shows that the difference between the two variances will increases as the variance $\mathrm{Var}_{\pi^b}\left[\rho(A, X) \mid Y\right]$ increases. This variance is likely to be large when the policy shift between π^b and π^* is large, or when the dimensions of contexts X and/or the actions A is large, and therefore in these cases the MR estimator will perform increasingly better than the IPW estimator. A similar phenomenon occurs for DR as we show next, even though in this case the variance of MR is not in general smaller than that of DR.

Proposition 2.3.3 (Variance comparison with DR estimator)

When the weights $\rho(a, x)$ and $w(y)$ are known exactly and $\mu(A, X) := \mathbb{E}[Y \mid X, A]$, we have that,

$$\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{DR}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] \geq \frac{1}{n}\mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, Y \mid Y\right] - \mathrm{Var}_{\pi^b}\left[\rho(A, X)\, \mu(A, X) \mid X\right]\right].$$

Intuition Proposition 2.3.3 shows that if the conditional variance $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, Y \mid Y\right]$ is greater than $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, \mu(A, X) \mid X\right]$ on average, the variance of the MR estimator will be less than that of the DR estimator. Intuitively, this will occur when the dimension of context space $\mathcal{X}$ is high because in this case the conditional variance over X and A, $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, Y \mid Y\right]$ is likely to be greater than the conditional variance over A, $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, \mu(A, X) \mid X\right]$. Our empirical results in Appendix A.6.2 are consistent with this intuition. Additionally, we also provide theoretical comparisons with other extensions of DR, such as Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020] in Appendix A.2, and show that a similar intuition applies for these results. We emphasise that the well known results in Wang et al. [2017b] which show that IPW and DR estimators achieve the optimal *worst case* variance (where the worst case is taken over a class of possible outcome distributions $Y \mid X, A$) are not at odds with our results presented here (as the distribution of $Y \mid X, A$ is fixed in our setting).

Comparison with Marginalised Inverse Propensity Score (MIPS) [Saito and Joachims, 2022]

In this section, we compare MR against the recently proposed Marginalised Inverse Propensity Score (MIPS) estimator [Saito and Joachims, 2022], which uses a marginalisation technique to reduce variance and provides a robust OPE estimate specifically in contextual bandits with large action spaces. We prove that the MR estimator achieves lower variance than the MIPS estimator and doesn't require new assumptions.

MIPS estimator As we mentioned earlier, the variance of the IPW estimator may be high when the action A is high dimensional. To mitigate this, the MIPS estimator assumes the existence of a (potentially lower dimensional) action embedding E, which summarises all 'relevant' information about the action A. Formally, this assumption can be written as follows:

Assumption 2.3.1. *The action A has no direct effect on the outcome Y, i.e.,*

$$Y \perp\!\!\!\perp A \mid X, E.$$

For example, in the setting of a recommendation system where A corresponds to the items recommended, E may correspond to the item categories. Assumption 2.3.1 then intuitively means that item category E encodes all relevant information about the item A which determines the outcome Y. Assuming that such action embedding E exists, Saito and Joachims [2022] prove that the MIPS estimator $\hat{\theta}_{\text{MIPS}}$, defined as

$$\hat{\theta}_{\text{MIPS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{p_{\pi^*}(e_i, x_i)}{p_{\pi^b}(e_i, x_i)}\, y_i = \frac{1}{n}\sum_{i=1}^{n} \frac{p_{\pi^*}(e_i \mid x_i)}{p_{\pi^b}(e_i \mid x_i)}\, y_i,$$

provides an unbiased estimator of target policy value $\mathbb{E}_{\pi^*}[Y]$. Moreover, $\text{Var}_{\pi^b}[\hat{\theta}_{\text{MIPS}}] \leq \text{Var}_{\pi^b}[\hat{\theta}_{\text{IPW}}]$.

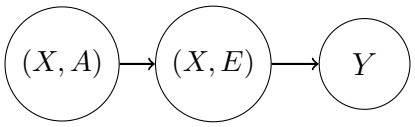


Figure 2.1: Bayesian network corresponding to Assumption 2.3.1.

Intuition The context-embedding pair (X, E) can be seen as a representation of the context-action pair (X, A) which contains less 'redundant information' regarding the outcome Y. Intuitively, the MIPS estimator, which only considers the shift in the distribution of (X, E) is therefore more efficient than the IPW estimator (which considers the shift in the distribution of (X, A) instead).

MR achieves lower variance than MIPS Given the intuition above, we should achieve greater variance reduction as the amount of redundant information in the representation (X, E) decreases. We formalise this in Appendix A.4 and show that the variance of MIPS estimator decreases as the representation gets closer to Y in terms of information content. As a result, we achieve the greatest variance reduction by considering the marginal shift in the outcome Y itself (as in MR) rather than the shift in the representation (X, E) (as in MIPS). The following result formalizes this finding.

Theorem 2.3.2

When the weights $w(y)$, $\frac{p_{\pi^*}(e,x)}{p_{\pi^b}(e,x)}$ and $\rho(a,x)$ are known exactly, then under Assumption 2.3.1,

$$\mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] = \mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{MIPS}}] = \mathbb{E}_{\pi^*}[Y], \quad \text{and} \quad \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] \leq \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MIPS}}] \leq \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}].$$

This analysis provides a link between the MR and MIPS estimators in the framework of contextual bandits, and shows that the MR estimator achieves lower variance than MIPS estimator while not requiring any additional assumptions (e.g. Assumption 2.3.1 as in MIPS). We also verify this empirically in Section 2.5.1 by reproducing the experimental setup in Saito and Joachims [2022] along with the MR baseline.

Weight estimation error

Our analysis so far assumes prior knowledge of the behavior policy π^b and the marginal ratios $w(y)$. However, in practice, both quantities are often unknown and must be estimated from data. To this end, we assume access to an additional training dataset $\mathcal{D}_{\mathrm{tr}} = \{(x_i^{\mathrm{tr}}, a_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^m$ (for weight estimation), in addition to the evaluation dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$ (for computing the OPE estimate). The estimation of $\hat{w}(y)$ involves a two-step process that exclusively utilizes data from $\mathcal{D}_{\mathrm{tr}}$:

(i) First, we estimate the policy ratio $\hat{\rho}(a,x) \approx \frac{\pi^*(a|x)}{\pi^b(a|x)}$. This can be achieved by estimating the behaviour policy $\widehat{\pi}^b$, and defining $\hat{\rho}(a,x) \coloneqq \frac{\pi^*(a|x)}{\widehat{\pi}^b(a|x)}$. Alternatively, $\hat{\rho}(a,x)$ can also be estimated directly by using density ratio estimation techniques as in Sondhi et al. [2020].

(ii) Secondly, we estimate the weights $\hat{w}(y)$ using Eq. (2.3) with $\hat{\rho}$ instead of ρ.

In practice, one may consider splitting $\mathcal{D}_{\mathrm{tr}}$ for each estimation step outlined above. Moreover, each approximation step may introduce bias and therefore, the MR estimator may have two sources of bias. While classical OPE methods like IPW and DR also suffer from bias because of $\hat{\rho}$ estimation, the estimation of $\hat{w}(y)$ is specific to MR. However, we show below that given any policy ratio estimate $\hat{\rho}$, if $\hat{w}(y)$ approximates $\mathbb{E}_{\pi^b}[\hat{\rho}(A,X) \mid Y = y]$ 'well enough' (i.e., the estimation step (ii) shown above is 'accurate enough'), then MR achieves a lower variance than IPW and incurs little extra bias.

Proposition 2.3.4

Suppose that the IPW and MR estimators are defined as,

$$\tilde{\theta}_{\text{IPW}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}(a_i, x_i)\, y_i, \quad \text{and} \quad \tilde{\theta}_{\text{MR}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{w}(y_i)\, y_i,$$

and define the approximation error as $\epsilon \coloneqq \hat{w}(Y) - \tilde{w}(Y)$, where $\tilde{w}(Y) \coloneqq \mathbb{E}_{\pi^b}[\hat{\rho}(A, X) \mid Y]$. Then we have that, $\text{Bias}(\tilde{\theta}_{\text{MR}}) - \text{Bias}(\tilde{\theta}_{\text{IPW}}) = \mathbb{E}_{\pi^b}[\epsilon\, Y]$. Moreover,

$$\text{Var}_{\pi^b}[\tilde{\theta}_{\text{IPW}}] - \text{Var}_{\pi^b}[\tilde{\theta}_{\text{MR}}] = \frac{1}{n}(\underbrace{\mathbb{E}_{\pi^b}[\text{Var}_{\pi^b}[\hat{\rho}(A, X)\, Y \mid Y]]}_{\geq 0} - \text{Var}_{\pi^b}[\epsilon\, Y] - 2\,\text{Cov}(\tilde{w}(Y)\, Y, \epsilon\, Y)). \tag{2.4}$$

Intuition The ϵ term defined in Proposition 2.3.4 denotes the error of the second approximation step outlined above. As a direct consequence of this result, we show in Appendix A.3 that as the error ϵ becomes small (specifically as $\mathbb{E}_{\pi^b}[\epsilon^2] \to 0$), the difference between biases of MR and IPW estimator becomes negligible. Likewise, the terms $\text{Var}_{\pi^b}[\epsilon\, Y]$ and $\text{Cov}(\tilde{w}(Y)\, Y, \epsilon\, Y)$ in Eq. (2.4) will also be small and as a result the variance of MR will be lower than that of IPW (as the first term is positive).

In fact, using recent results regarding the generalisation error of neural networks [Lai et al., 2023], we show that when using 2-layer wide neural networks to approximate the weights $\hat{w}(y)$, the estimation error ϵ declines with increasing training data size m. Specifically, under certain regularity assumptions we obtain $\mathbb{E}_{\pi^b}[\epsilon^2] = O(m^{-2/3})$. Using this we show that as the training data size m increases, the biases of MR and IPW estimators become roughly equal with a high probability, and

$$\text{Var}_{\pi^b}[\tilde{\theta}_{\text{IPW}}] - \text{Var}_{\pi^b}[\tilde{\theta}_{\text{MR}}] = \frac{1}{n}\,\mathbb{E}_{\pi^b}[\text{Var}_{\pi^b}[\hat{\rho}(A, X)\, Y \mid Y]] + O(m^{-1/3}).$$

Therefore the variance of MR estimator falls below that of IPW for large enough m. The empirical results shown in Appendix A.6.2 are consistent with this result. Due to space constraints, the main technical result has been included in Appendix A.3.

2.3.2 Application to causal inference

Beyond contextual bandits, the variance reduction properties of the MR estimator make it highly useful in a wide variety of other applications. Here, we show one such application in the field of causal inference, where MR can be used for the estimation of average treatment

effect (ATE) [Pearl, 2009] and leads to some desirable properties in comparison to the conventional ATE estimation approaches. Specifically, we illustrate that the MR estimator for ATE utilizes the evaluation data $\mathcal{D}$ more efficiently and achieves lower variance than state-of-the-art ATE estimators and consequently provides more accurate ATE estimates. To be concrete, the goal in this setting is to estimate ATE, defined as follows:

$$\text{ATE} \coloneqq \mathbb{E}[Y(1) - Y(0)].$$

Here $Y(a)$ corresponds to the outcome under a deterministic policy $\pi_a(a' \mid x) \coloneqq \mathbb{1}(a' = a)$. Hence any OPE estimator can be used to estimate $\mathbb{E}[Y(a)]$ (and therefore ATE) by considering target policy $\pi^* = \pi_a$. An important distinction between MR and existing approaches (like IPW or DR) is that, when estimating $\mathbb{E}[Y(a)]$, the existing approaches only use datapoints in $\mathcal{D}$ with $A = a$. To see why this is the case, we note that the policy ratios $\frac{\pi^*(A|X)}{\pi^b(A|X)} = \frac{\mathbb{1}(A=a)}{\pi^b(A|X)}$ are zero when $A \neq a$. In contrast, the MR weights $\frac{p_{\pi^*}(Y)}{p_{\pi^b}(Y)}$ are not necessarily zero for datapoints with $A \neq a$, and therefore the MR estimator uses all evaluation datapoints when estimating $\mathbb{E}[Y(a)]$.

As such we show that MR applied to ATE estimation leads to a smaller variance than the existing approaches. Moreover, because MR is able to use all datapoints when estimating $\mathbb{E}[Y(a)]$, MR will generally be more accurate than the existing methods especially in the setting where the data is imbalanced, i.e., the number of datapoints with $A = a$ is small for a specific action a. In Appendix A.5, we formalise this variance reduction of the MR ATE estimator compared to IPW and DR estimators, by deriving analogous results to Propositions 2.3.2 and 2.3.3. In addition, we also show empirically in Section 2.5.3 that the MR ATE estimator outperforms the most commonly used ATE estimators.

2.4 Related work

Off-policy evaluation is a central problem both in contextual bandits [Dudík et al., 2014b, Wang et al., 2017b, Liu et al., 2018, Farajtabar et al., 2018a, Su et al., 2019b, 2020, Kallus et al., 2021, Metelli et al., 2021, Saito et al., 2020] and in RL [Thomas and Brunskill, 2016, Xie et al., 2019b, Kallus and Uehara, 2022, Liu et al., 2020]. Existing OPE methodologies can be broadly categorised into Direct Method (DM), Inverse Probability Weighting (IPW), and Doubly Robust (DR). While DM typically has a low variance, it suffers from high

bias when the reward model is misspecified [Voloshin et al., 2021]. On the other hand, IPW [Horvitz and Thompson, 1952] and DR [Dudík et al., 2014b, Wang et al., 2017b, Su et al., 2020] use policy ratios as importance weights when estimating policy value and suffer from high variance as overlap between behaviour and target policies increases or as the action/context space gets larger [Sachdeva et al., 2020, Saito and Joachims, 2022]. To circumvent this problem, techniques like weight clipping or normalisation [Swaminathan and Joachims, 2015a,b, London and Sandler, 2019] are often employed, however, these can often increase bias.

In contrast to these approaches, Saito and Joachims [2022] propose MIPS, which considers the marginal shift in the distribution of a lower dimensional embedding of the action space. While this approach reduces the variance associated with IPW, we show in Section 2.3.1 that the MR estimator achieves a lower variance than MIPS while not requiring any additional assumptions (like Assumption 2.3.1).

In the context of Reinforcement Learning (RL), various marginalisation techniques of importance weights have been used to propose OPE methodologies. Liu et al. [2018], Xie et al. [2019b], Kallus and Uehara [2022] use methods which considers the shift in the marginal distribution of the states, and applies importance weighting with respect to this marginal shift rather than the trajectory distribution. Similarly, Fujimoto et al. [2021] use marginalisation for OPE in deep RL, where the goal is to consider the shift in marginal distributions of state and action. Although marginalization is a key trick of these estimators, these techniques do not consider the marginal shift in reward as in MR and are aimed at resolving the curse of horizon, a problem specific to RL. Apart from this, Rowland et al. [2020] propose a general framework of OPE based on conditional expectations of importance ratios for variance reduction. While their proposed framework includes reward conditioned importance ratios, this is not the main focus and there is little theoretical and empirical comparison of their proposed methodology with existing state-of-the-art methods like DR.

Finally we note that the idea of approximating the ratio of intractable marginal densities via leveraging the fact that this ratio can be reformulated as the conditional expectation of a ratio of tractable densities is a standard idea in computational statistics [Meng and Wong, 1996] and has been exploited more recently to perform likelihood-free inference [Brehmer et al., 2020]. In particular, while Meng and Wong [1996] typically

approximates this expectation through Markov chain Monte Carlo, Brehmer et al. [2020] uses regression instead, however without any theory.

2.5 Empirical evaluation

In this section, we provide empirical evidence to support our theoretical results by investigating the performance of our MR estimator against the current state-of-the-art OPE methods. The code to reproduce our experiments has been made available at: github.com/faaizT/MR-OPE.

2.5.1 Experiments on synthetic data

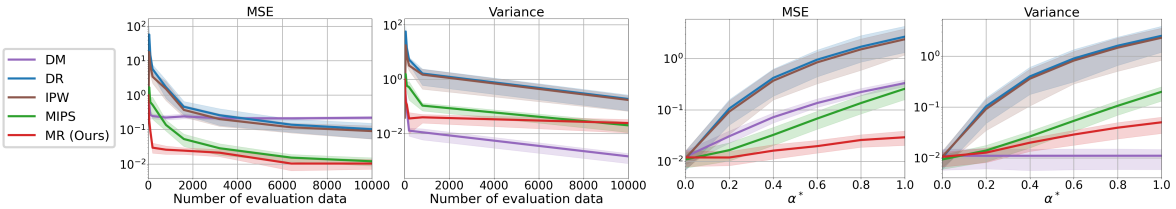
For our synthetic data experiment, we reproduce the experimental setup for the synthetic data experiment in Saito and Joachims [2022] by reusing their code with minor modifications. Specifically, $\mathcal{X} \subseteq \mathbb{R}^d$, for various values of d as described below. Likewise, the action space $\mathcal{A} = \{0, \dots, n_a - 1\}$, with n_a taking a range of different values. Additional details regarding the reward function, behaviour policy π^b, and the estimation of weights $\hat{w}(y)$ have been included in Appendix A.6.2 for completeness.

Target policies To investigate the effect of increasing policy shift, we define a class of policies,

$$\pi^{\alpha^*}(a|x) = \alpha^* \, \mathbb{1}(a = \arg\max_{a' \in \mathcal{A}} q(x, a')) + \frac{1 - \alpha^*}{|\mathcal{A}|} \quad \text{where} \quad q(x, a) \coloneqq \mathbb{E}[Y \mid X = x, A = a],$$

where $\alpha^* \in [0, 1]$ allows us to control the shift between π^b and π^*. In particular, as we show later, the shift between π^b and π^* increases as $\alpha^* \to 1$. Using the ground truth behaviour policy π^b, we generate a dataset which is split into training and evaluation datasets of sizes m and n respectively.

Baselines We compare our estimator with DM, IPW, DR and MIPS estimators. Our setup includes action embeddings E satisfying Assumption 2.3.1, and so MIPS is unbiased. Additional baselines have been considered in Appendix A.6.2. For MR, we split the training data to estimate $\widehat{\pi}^b$ and $\hat{w}(y)$, whereas for all other baselines we use the entire training data to estimate $\widehat{\pi}^b$ for a fair comparison.



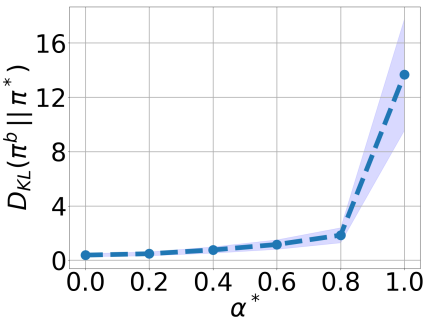
(a) Results with varying evaluation data size n. **(b)** Results with varying α^*.

Figure 2.2: Results for synthetic data experiment. In 2.2a we have $\alpha^* = 0.8$ and in 2.2b we have $n = 800$.

Results We compute the target policy value using the n evaluation datapoints. Here, the MSE of the estimators is computed over 10 different sets of logged data replicated with different seeds. The results presented have context dimension $d = 1000$, number of actions $n_a = 100$ and training data size $m = 5000$. More experiments for a variety of parameter values are included in Appendix A.6.2.

Varying number of evaluation data n In Figure 2.2a we plot the results with increasing size of evaluation data n increases. MR achieves the smallest MSE among all the baselines considered when n is small, with the MSE of MR being at least an order of magnitude smaller than every baseline for $n \leq 500$. This shows that MR is significantly more accurate than the baselines when the size of the evaluation data is small. As $n \to \infty$, the difference between the results for MR and MIPS decreases. However, MR attains smaller variance and MSE than MIPS generally, verifying our analysis in Section 2.3.1. Moreover, Figure 2.2a shows that while the variance of MR is greater than that of DM, it still achieves the lowest MSE overall, owing to the high bias of DM.

Varying α^* As α^* parameter of the target policy increases, so does the shift between the policies π^b and π^{α^*} as illustrated by the figure on the right, which plots the KL-divergence $D_{\mathrm{KL}}(\pi^b \,||\, \pi^{\alpha^*})$ as a function of α. Figure 2.2b plots the results for increasing policy shift. Overall, the MSE of MR estimator is lowest among all the baselines. Moreover, while the MSE and variance of all estimators increase with increasing α^* the increase in these quantities is lower for the MR estimator than for the other baselines. Therefore,



the relative performance of MR estimator improves with increasing policy shift and MR remains robust to increase in policy shift.

Additional ablation studies In Appendix A.6.2, we investigate the effect of varying context dimensions d, number of actions n_a and number of training data m. In every case, we observe that the MR estimator has a smaller MSE than all other baselines considered. In particular, MR remains robust to increasing n_a whereas the MSE and variance of IPW and DR estimators degrade substantially when $n_a \geq 2000$. Likewise, MR outperforms the baselines even when the training data size m is small.

2.5.2 Experiments on classification datasets

Following previous works on OPE in contextual bandits [Dudík et al., 2014b, Kallus et al., 2021, Farajtabar et al., 2018b, Wang et al., 2017b], we transform classification datasets into contextual bandit feedback data in this experiment. We consider five UCI classification datasets [Dua and Graff, 2017] as well as Mnist [Deng, 2012] and CIFAR-100 [Krizhevsky, 2009] datasets, each of which comprises $\{(x_i, a_i^{\mathrm{gt}})\}_i$, where $x_i \in \mathcal{X}$ are feature vectors and $a_i^{\mathrm{gt}} \in \mathcal{A}$ are the ground-truth labels. In the contextual bandits setup, the feature vectors x_i are considered to be the contexts, whereas the actions correspond to the possible class of labels. For the context vector x_i and the action a_i, the reward y_i is defined as $y_i \coloneqq \mathbb{1}(a_i = a_i^{\mathrm{gt}})$, i.e., the reward is 1 when the action is the same as the ground truth label and 0 otherwise. Here, the baselines considered include the DM, IPW and DR estimators as well as Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020]. We do not consider a MIPS baseline here as there is no natural embedding E of A. Additional details are provided in Appendix A.6.3.

In Table 2.1, we present the results with number of evaluation data $n = 1000$ and number of training data $m = 500$. The table shows that across all datasets, MR achieves the lowest MSE among all methods. Moreover, for the Letter and CIFAR-100 datasets the IPW and DR yield large bias and variance arising from poor policy estimates $\widehat{\pi}^b$. Despite this, the MR estimator which utilizes the *same* $\widehat{\pi}^b$ for the estimation of $\hat{w}(y)$ leads to much more accurate results. We also verify that MR outperforms the baselines for increasing policy shift and evaluation data n in Appendix A.6.3.

Table 2.1: Mean squared error of target policy value with standard errors over 10 different seeds for different classification datasets. Here, number of evaluation data $n = 1000$, and $\alpha^* = 0.6$.

Dataset	Digits	Letter	OptDigits	PenDigits	SatImage	Mnist	CIFAR-100
DM	0.1508±0.0015	0.0886±0.0026	0.0485±0.0016	0.0520±0.0016	0.0208±0.0009	0.1109±0.0014	0.0020±0.0001
DR	0.1334±0.0400	35.085±17.768	0.0464±0.0061	0.2343±0.1404	0.0560±0.0395	0.2617±0.0139	3823.9±2023.2
DRos	0.0847±0.0025	0.2363±0.0586	0.0384±0.0025	0.0138±0.0029	0.0078±0.0008	0.2151±0.0061	0.2628±0.1087
IPW	0.1632±0.0462	45.253±22.057	0.0844±0.0056	0.1342±0.0531	0.0900±0.0676	0.3359±0.0118	4116.9±2097.9
SwitchDR	0.0982±0.0032	0.2387±0.0507	0.0557±0.0047	0.0342±0.0090	0.0136±0.0012	0.2750±0.0102	1.1644±0.8227
MR (Ours)	**0.0034±0.0001**	**0.0018±0.0004**	**0.0006±0.0002**	**0.0008±0.0002**	**0.0016±0.0003**	**0.0121±0.0009**	**0.0007±0.0002**

Table 2.2: Mean absolute ATE estimation error ϵ_{ATE} with standard errors over 10 different seeds, for increasing number of evaluation data n.

n	50	200	1600	3200
DM	0.092±0.003	0.092±0.003	0.092±0.004	0.092±0.004
DR	0.101±0.024	**0.065±0.009**	0.071±0.005	0.069±0.004
DRos	0.100±0.017	0.089±0.006	0.093±0.004	0.087±0.004
IPW	0.092±0.024	0.088±0.014	0.067±0.007	0.067±0.007
SwitchDR	0.101±0.024	**0.065±0.009**	0.071±0.005	0.069±0.004
MR (Ours)	**0.062±0.007**	**0.065±0.007**	**0.061±0.005**	**0.061±0.006**

2.5.3 Application to ATE estimation

In this experiment, we investigate the empirical performance of the MR estimator for ATE estimation.

Twins dataset We use the Twins dataset studied in Louizos et al. [2017], which comprises data from twin births in the USA between 1989-1991. The treatment $a = 1$ corresponds to being born the heavier twin and the outcome Y corresponds to the mortality of each of the twins in their first year of life. Specifically, $Y(1)$ corresponds to the mortality of the heavier twin (and likewise for $Y(0)$). To simulate the observational study, we follow a similar strategy as in Louizos et al. [2017] to selectively hide one of the two twins as explained in Appendix A.6.4. We obtain a total of 11,984 datapoints, of which 5000 datapoints are used to train the behaviour policy $\widehat{\pi}^b$ and outcome model $\hat{q}(x, a)$.

Here, we consider the same baselines as the classification data experiments in previous section. For our evaluation, we consider the absolute error in ATE estimation, ϵ_{ATE}, defined as: $\epsilon_{\text{ATE}} \coloneqq |\hat{\theta}^{(n)}_{\text{ATE}} - \theta_{\text{ATE}}|$. Here, $\hat{\theta}^{(n)}_{\text{ATE}}$ denotes the value of the ATE estimated using n evaluation datapoints. We compute the ATE value using the n evaluation datapoints, over 10 different sets of observational data (using different seeds). Table 2.2 shows that MR achieves the lowest estimation error ϵ_{ATE} for all values of n considered here. While the performance of other baselines improves with increasing n, MR outperforms them all.

2.6 Discussion

In this paper, we proposed an OPE method for contextual bandits called marginal ratio (MR) estimator, which considers only the shift in the marginal distribution of the outcomes

resulting from the policy shift. Our theoretical and empirical analysis showed that MR achieves better variance and MSE compared to the current state-of-the-art methods and is more data efficient overall. Additionally, we demonstrated that MR applied to ATE estimation provides more accurate results than most commonly used methods. Next, we discuss limitations of our methodology and possible avenues for future work.

Limitations The MR estimator requires the additional step of estimating $\hat{w}(y)$ which may introduce an additional source of bias in the value estimation. However, $\hat{w}(y)$ can be estimated by solving a simple 1d regression problem, and as we show empirically in Appendix A.6, MR achieves the smallest bias among all baselines considered in most cases. Most notably, our ablation study in Appendix A.6.2 shows that even when the training data is reasonably small, MR outperforms the baselines considered.

Future work The MR estimator can also be applied to policy optimisation problems, where the data collected using an 'old' policy is used to learn a new policy. This approach has been used in Proximal Policy Optimisation (PPO) [Schulman et al., 2017] for example, which has gained immense popularity and has been applied to reinforcement learning with human feedback (RLHF) [Lambert et al., 2022]. We believe that the MR estimator applied to these methodologies could lead to improvements in the stability and convergence of these optimisation schemes, given its favourable variance properties.

Acknowledgements

We would like to thank Jake Fawkes, Siu Lun Chau, Shahine Bouabid and Robert Hu for their useful feedback. We also appreciate the insights and constructive criticisms provided by the anonymous reviewers. MFT acknowledges his PhD funding from Google DeepMind.

3

Conformal Off-Policy Prediction in Contextual Bandits

Contents

Abstract

Most off-policy evaluation methods for contextual bandits have focused on the expected outcome of a policy, which is estimated via methods that at best provide only asymptotic guarantees. However, in many applications, the expectation may not be the best measure of performance as it does not capture the variability of the outcome. In addition, particularly in safety-critical settings, stronger guarantees than asymptotic correctness may be required. To address these limitations, we consider a novel application of conformal prediction to contextual bandits. Given data collected under a behavioral policy, we propose *conformal off-policy prediction* (COPP), which can output reliable predictive intervals for the outcome under a new target policy. We provide theoretical finite-sample guarantees without making any additional assumptions beyond the standard contextual bandit setup, and empirically demonstrate the utility of COPP compared with existing methods on synthetic and real-world data.

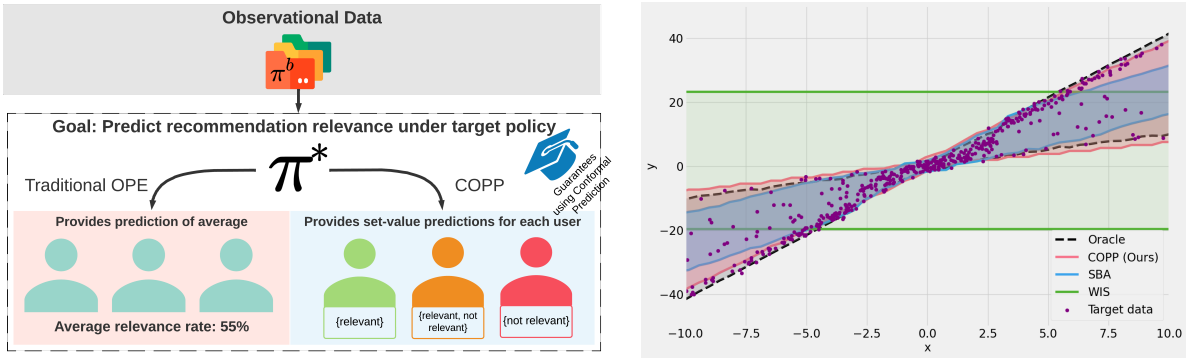


Figure 3.1: Left (a): Conformal Off-Policy Prediction against standard off-policy evaluation methods. **Right (b):** 90% predictive intervals for Y against X for COPP, competing methods and the oracle.

3.1 Introduction

Before deploying a decision-making policy to production, it is usually important to understand the plausible range of outcomes that it may produce. However, due to resource or ethical constraints, it is often not possible to obtain this understanding by testing the policy directly in the real-world. In such cases we have to rely on observational data collected under a different policy than the target. Using this observational data to evaluate the target policy is known as off-policy evaluation (OPE).

Traditionally, most techniques for OPE in contextual bandits focus on evaluating policies based on their **expected** outcomes; see e.g., Kuzborskij et al. [2021], Wang et al. [2017a], Thomas et al. [2015], Swaminathan and Joachims [2015c,d], Dudík et al. [2014a]. However, this can be problematic as methods that are only concerned with the average outcome do not take into account any notions of variance, for example. Therefore, in risk-sensitive settings such as econometrics, where we want to minimize the potential risks, metrics such as CVaR (Conditional Value at Risk) might be more appropriate [Keramati et al., 2020]. Additionally, when only small sample sizes of observational data are available, the average outcomes under finite data can be misleading, as they are prone to outliers and hence, metrics such as medians or quantiles are more robust in these scenarios [Altschuler et al., 2019].

Notable exceptions in the OPE literature are Huang et al. [2021], Chandak et al. [2021]. Instead of estimating bounds on the expected outcomes, Huang et al. [2021], Chandak et al. [2021] establish finite-sample bounds for a general class of metrics (e.g., Mean, CVaR, CDF) on the outcome. Their methods can be used to estimate quantiles of the outcomes

under the target policy and are therefore robust to outliers. However, the resulting bounds do not depend on the covariates X (not adaptive w.r.t. X). This can lead to overly conservative intervals, as we will show in our experiments and can become uninformative when the data are heteroscedastic (see Fig. 3.1b).

In this paper, we propose Conformal Off-Policy Prediction (COPP), a novel algorithm that uses Conformal Prediction (CP) [Vovk et al., 2005] to construct predictive interval/sets for outcomes in contextual bandits (see Fig.3.1a) using an observational dataset. COPP enjoys both finite-sample theoretical guarantees and adaptivity w.r.t. the covariates X, and, to the best of our knowledge, is the first such method based on CP that can be applied to stochastic policies and continuous action spaces.

In summary, our contributions are:

(i) We propose an application of CP to construct predictive intervals for bandit outcomes that is more general (applies to stochastic policies and continuous actions) than previous work.

(ii) We provide theoretical guarantees for COPP, including finite-sample guarantees on marginal coverage and asymptotic guarantees on conditional coverage.

(iii) We show empirically that COPP outperforms standard methods in terms of coverage and predictive interval width when assessing new policies.

3.1.1 Problem setup

Let $\mathcal{X}$ be the covariate space (e.g., user data), $\mathcal{A}$ the action space (e.g., recommended items) and $\mathcal{Y}$ the outcome space (e.g., relevance to the user). We allow both $\mathcal{A}$ and $\mathcal{Y}$ to be either discrete or continuous. In our setting, we are given logged observational data $\mathcal{D}_{obs} = \{x_i, a_i, y_i\}_{i=1}^{n_{obs}}$ where actions are sampled from a behavioural policy π^b, i.e. $A \mid x \sim \pi^b(\cdot \mid x)$ and $Y \mid x, a \sim P(\cdot \mid x, a)$. We assume that we do not suffer from unmeasured confounding. At test time, we are given a state x^{test} and a new policy π^*. While π^b may be unknown, we assume the target policy π^* to be known.

We consider the task of rigorously quantifying the performance of π^* without any distributional assumptions on X or Y. Many existing approaches estimate $\mathbb{E}_{\pi^*}[Y]$, which is useful for comparing two policies directly as they return a single number. However, the

expectation does not convey fine-grained information about how the policy performs for a specific value of X, nor does it account for the uncertainty in the outcome Y.

Here, we aim to construct intervals/sets on the outcome Y which are (i) adaptive w.r.t. X, (ii) capture the variability in the outcome Y and (iii) provide finite-sample guarantees. Current methods lack at least one of these properties (see Sec. 3.5). One way to achieve these properties is to construct a set-valued function of x, $\hat{C}(x)$ which outputs a *subset* of $\mathcal{Y}$. Given any finite dataset $\mathcal{D}_{obs}$, this subset is guaranteed to contain the true value of Y with any pre-specified probability, i.e.

$$1-\alpha \leq \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) \leq 1-\alpha + o_{n_{obs}}(1) \tag{3.1}$$

where n_{obs} is the size of available observational data and $P^{\pi^*}_{X,Y}$ is the joint distribution of (X,Y) under target policy π^*. In practice, $\hat{C}(x)$ can be used as a diagnostic tool downstream for a granular assessment of likely outcomes under a target policy. The probability in (3.1) is taken over the joint distribution of (X,Y), meaning that (3.1) holds marginally in X (marginal coverage) and not for a given $X=x$ (conditional coverage). In Sec. 3.4.2, we provide additional regularity conditions under which not only marginal but also conditional coverage holds. Next, we introduce the Conformal Prediction framework, which allows us to construct intervals $\hat{C}(x)$ that satisfy (3.1) along with properties (i)-(iii).

3.2 Background

Conformal prediction [Vovk et al., 2005, Shafer and Vovk, 2008] is a methodology that was originally used to compute distribution-free prediction sets for regression and classification tasks. Before introducing COPP, which applies CP to contextual bandits, we first illustrate how CP can be used in standard regression.

3.2.1 Standard conformal prediction

Consider the problem of regressing $\mathrm{Y} \in \mathcal{Y}$ against $X \in \mathcal{X}$. Let $\hat{f}$ be a model trained on the *training* data $\mathcal{D}_{tr} = \{X_i^0, Y_i^0\}_{i=1}^m \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ and let the *calibration* data $\mathcal{D}_{cal} = \{X_i, Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{X,Y}$ be independent of $\mathcal{D}_{tr}$. Given a desired coverage rate $1-\alpha \in (0,1)$,

we construct a band $\hat{C}_n : \mathcal{X} \to \{\text{subsets of } \mathcal{Y}\}$, based on the calibration data such that, for a new i.i.d. test data $(X, Y) \sim P_{X,Y}$,

$$1 - \alpha \leq \mathbb{P}_{(X,Y)\sim P_{X,Y}}(Y \in \hat{C}_n(X)) \leq 1 - \alpha + \frac{1}{n+1}, \tag{3.2}$$

where the probability is taken over X, Y and $\mathcal{D}_{cal} = \{X_i, Y_i\}_{i=1}^n$ and is conditional upon $\mathcal{D}_{tr}$.

In order to obtain $\hat{C}_n$ satisfying (3.2), we introduce a non-conformity score function $V_i = s(X_i, Y_i)$, e.g., $(\hat{f}(X_i) - Y_i)^2$. We assume here $\{V_i\}_{i=1}^n$ have no ties almost surely. Intuitively, the non-conformity score V_i uses the outputs of the predictive model $\hat{f}$ on the calibration data, to measure how far off these predictions are from the ground truth response. Higher scores correspond to worse fit between x and y according to $\hat{f}$. We define the empirical distribution of the scores $\{V_i\}_{i=1}^n \cup \{\infty\}$

$$\hat{F}_n \coloneqq \frac{1}{n+1}\sum_{i=1}^{n} \delta_{V_i} + \frac{1}{n+1}\delta_{\infty} \tag{3.3}$$

with which we can subsequently construct the conformal interval $\hat{C}_n$ that satisfies (3.2) as follows:

$$\hat{C}_n(x) \coloneqq \{y : s(x, y) \leq \eta\} \tag{3.4}$$

where η is an empirical quantile of $\{V_i\}_{i=1}^n$, i.e. $\eta = \text{Quantile}_{1-\alpha}(\hat{F}_n)$ is the $1 - \alpha$ quantile.

Intuitively, for roughly $100 \cdot (1-\alpha)\%$ of the calibration data, the score values will be below η. Therefore, if the new datapoint (X, Y) and $\mathcal{D}_{cal}$ are i.i.d., the probability $\mathbb{P}(s(X, Y) \leq \eta)$ (which is equal to $\mathbb{P}(Y \in \hat{C}_n(X))$ by (3.4)) will be roughly $1 - \alpha$. Exchangeability of the data is crucial for the above to hold. In the next section we will explain how Tibshirani et al. [2019] relax the exchangeability assumption.

3.2.2 Conformal prediction under covariate shift

Tibshirani et al. [2019] extend the CP framework beyond the setting of exchangeable data, by constructing valid intervals even when the calibration data and test data are not drawn from the same distribution. The authors focus on the *covariate shift* scenario i.e. the distribution of the covariates changes at test time:

$$(X_i, Y_i) \overset{\text{i.i.d}}{\sim} P_{X,Y} = P_X \times P_{Y|X}, \quad i = 1, \ldots, n$$
$$(X, Y) \sim \tilde{P}_{X,Y} = \tilde{P}_X \times P_{Y|X}, \text{ independently}$$

where the ratio $w(x) \coloneqq \mathrm{d}\tilde{P}_X/\mathrm{d}P_X(x)$ is known. The key realization in Tibshirani et al. [2019] is that the requirement of *exchangeability* in CP can be relaxed to a more general property, namely *weighted exchangeability* (see Def. B.1.1). They propose a weighted version of conformal prediction, which shifts the empirical distribution of non-conformity scores, $\hat{F}_n$, at a point x, using weights $w(x)$. This adjusts $\hat{F}_n$ for the covariate shift, before picking the quantile η:

$$\hat{F}_n^x \coloneqq \sum_{i=1}^{n} p_i^w(x)\delta_{V_i} + p_{n+1}^w(x)\delta_\infty \quad \text{where,}$$

$$p_i^w(x) = \frac{w(X_i)}{\sum_{j=1}^n w(X_j) + w(x)}, \quad p_{n+1}^w(x) = \frac{w(x)}{\sum_{j=1}^n w(X_j) + w(x)}.$$

In standard CP (without covariate shift), the weight function satisfies $w(x) = 1$ for all x, and we recover (3.3). Next, we construct the conformal prediction intervals $\hat{C}_n$ as in standard CP using (3.4) where η now depends on x due to $p_i^w(x)$. The resulting intervals, $\hat{C}_n$, satisfy:

$$\mathbb{P}_{(X,Y)\sim\tilde{P}_{X,Y}}(Y \in \hat{C}_n(X)) \geq 1 - \alpha$$

As mentioned previously in Sec. 3.1.1, the above demonstrates marginal coverage guarantees over test point X and calibration dataset $\mathcal{D}_{cal}$, not conditional on a given $X = x$ or a fixed $\mathcal{D}_{cal}$. We will discuss this nuance later on in Sec. 3.4.2. In addition, previous work by Vovk shows that conditioned on a single calibration dataset, standard CP can achieve coverage that is 'close' to the required coverage with high probability. However, this has not been extended to the case where the distribution shifts. This is out of the scope of this paper and an interesting future direction.

Thus Tibshirani et al. [2019] show that the CP algorithm can be extended to the setting of covariate shift with the resulting predictive intervals satisfying the coverage guarantees when the weights are known. The extension of these results to approximate weights was proposed in Lei and Candès [2021] and is generalized to our setting in Sec. 3.4.

3.3 Conformal Off-Policy Prediction (COPP)

In the contextual bandits introduced in Sec. 3.1.1, we assume that the observational data $\mathcal{D}_{obs} = \{x_i, a_i, y_i\}_{i=1}^{n_{obs}}$ is generated from a behavioural policy π^b. At inference time

Algorithm 1: Conformal Off-Policy Prediction (COPP)

Inputs: Observational data $\mathcal{D}_{obs} = \{X_i, A_i, Y_i\}_{i=1}^{n_{obs}}$, conf. level α, a score function $s(x, y) \in \mathbb{R}$, new data point x^{test}, target policy π^* ;

Output: Predictive interval $\hat{C}_n(x^{test})$;

Split $\mathcal{D}_{obs}$ into training data ($\mathcal{D}_{tr}$) and calibration data ($\mathcal{D}_{cal}$) of sizes m and n respectively;

Use $\mathcal{D}_{tr}$ to estimate weights $\hat{w}(\cdot, \cdot)$ using (3.7);

Compute $V_i := s(X_i, Y_i)$ for $(X_i, A_i, Y_i) \in \mathcal{D}_{cal}$;

Let $\hat{F}_n^{x,y}$ be the weighted distribution of scores

$$\hat{F}_n^{x,y} := \sum_{i=1}^{n} p_i^{\hat{w}}(x, y)\delta_{V_i} + p_{n+1}^{\hat{w}}(x, y)\delta_{\infty}$$

where $p_i^{\hat{w}}(x, y) = \frac{\hat{w}(X_i,Y_i)}{\sum_{j=1}^{n} \hat{w}(X_j,Y_j)+\hat{w}(x,y)}$ and $p_{n+1}^{\hat{w}}(x, y) = \frac{\hat{w}(x,y)}{\sum_{j=1}^{n} \hat{w}(X_j,Y_j)+\hat{w}(x,y)}$;

For x^{test} construct: $\hat{C}_n(x^{test}) := \{y : s(x^{test}, y) \leq \text{Quantile}_{1-\alpha}(\hat{F}_n^{x^{test},y})\}$

Return $\hat{C}_n(x^{test})$

we are given a new target policy π^* and want to provide intervals on the outcomes Y for covariates X that satisfy (3.1).

The key insight of our approach is to consider the following joint distribution of (X, Y):

$$P^{\pi^b}(x, y) = P(x) \int P(y|x, a)\pi^b(a|x)\mathrm{d}a = P(x)P^{\pi^b}(y|x)$$
$$P^{\pi^*}(x, y) = P(x) \int P(y|x, a)\pi^*(a|x)\mathrm{d}a = P(x)P^{\pi^*}(y|x)$$

Therefore, the change of policies from π^b to π^* causes a shift in the joint distributions of (X, Y) from $P_{X,Y}^{\pi^b}$ to $P_{X,Y}^{\pi^*}$. More precisely, a shift in the conditional distribution of $Y|X$. As a result, our problem boils down to using CP in the setting where the conditional distribution $P_{Y|X}^{\pi^b}$ changes to $P_{Y|X}^{\pi^*}$ due to the different policies, while the covariate distribution P_X remains the same.

Hence our problem is not concerned about covariate shift as addressed in Tibshirani et al. [2019], but instead uses the idea of *weighted exchangeability* to extend CP to the setting of policy shift. To account for this distributional mismatch, our method shifts the empirical distribution of non-conformity scores at a point (x, y) using the weights $w(x, y) = \mathrm{d}P_{X,Y}^{\pi^*}/\mathrm{d}P_{X,Y}^{\pi^b}(x, y) = \mathrm{d}P_{Y|X}^{\pi^*}/\mathrm{d}P_{Y|X}^{\pi^b}(x, y)$ as follows:

$$\hat{F}_n^{x,y} := \sum_{i=1}^{n} p_i^{w}(x, y)\delta_{V_i} + p_{n+1}^{w}(x, y)\delta_{\infty}, \tag{3.5}$$

where,

$$p_i^w(x,y) = \frac{w(X_i,Y_i)}{\sum_{j=1}^n w(X_j,Y_j) + w(x,y)} \quad \text{and,}$$
$$p_{n+1}^w(x,y) = \frac{w(x,y)}{\sum_{j=1}^n w(X_j,Y_j) + w(x,y)}.$$

The intervals are then constructed as below which we call Conformal Off-Policy Prediction (see Algorithm 1).

$$\hat{C}_n(x^{test}) := \{y : s(x^{test},y) \leq \eta(x^{test},y)\} \text{ where, } \eta(x,y) := \text{Quantile}_{1-\alpha}(\hat{F}_n^{x,y}). \tag{3.6}$$

Remark The weights $w(x,y)$ in (3.5) depend on x and y, as opposed to only x. In particular, finding the set of y's satisfying (3.6) becomes more complicated than for the standard covariate shifted CP which only requires a single computation of $\eta(x)$ for a given x as shown in (3.4). In our case however, we have to create a k sized grid of potential values of y for every x to find $\hat{C}_n(x)$. This operation is embarrassingly parallel and hence does not add much computational overhead compared to the standard CP, especially because CP mainly focuses on scalar predictions.

3.3.1 Estimation of weights $w(x,y)$

So far we have been assuming that we know the weights $w(x,y)$ exactly. However, in most real-world settings, this will not be the case. Therefore, we must resort to estimating $w(x,y)$ using observational data. In order to do so, we first split the observational data into training ($\mathcal{D}_{tr}$) and calibration ($\mathcal{D}_{cal}$) data. Next, using $\mathcal{D}_{tr}$, we estimate $\hat{\pi}^b(a \mid x) \approx \pi^b(a \mid x)$ and $\hat{P}(y \mid x,a) \approx P(y \mid x,a)$ (which is independent of the policy). We then compute a Monte Carlo estimate of weights using the following:

$$\hat{w}(x,y) = \frac{\frac{1}{h}\sum_{k=1}^h \hat{P}(y|x,A_k^*)}{\frac{1}{h}\sum_{k=1}^h \hat{P}(y|x,A_k)} \approx \frac{\int P(y|x,a)\pi^*(a|x)\mathrm{d}a}{\int P(y|x,a)\pi^b(a|x)\mathrm{d}a}, \tag{3.7}$$

where $A_k \sim \hat{\pi}^b(\cdot \mid x)$, $A_k^* \sim \pi^*(\cdot \mid x)$ and h is the number of Monte Carlo samples.

Why not construct intervals using $\hat{P}(y|x,a)$ directly?

We could directly construct predictive intervals $\hat{C}_n(x)$ over outcomes by sampling

$$Y_j \overset{\text{i.i.d.}}{\sim} \hat{P}^{\pi^*}(y|x) = \int \hat{P}(y|x,a)\pi^*(a|x)\mathrm{d}a.$$

However, the coverage of these intervals directly depends on the estimation error of $\hat{P}(y|x,a)$. This is not the case in COPP, as the coverage does not depend on $\hat{P}(y|x,a)$ directly but rather on the estimation of $\hat{w}(x,y)$ (see Prop. 3.4.2). We hypothesize that this indirect dependence of COPP on $\hat{P}(y|x,a)$ makes it less sensitive to the estimation error. In Sec. 3.6, our empirical results support this hypothesis as COPP provides more accurate coverage than directly using $\hat{P}(y|x,a)$ to construct intervals. Lastly, in Appendix B.2.5 we show how we can avoid estimating $\hat{P}(y|x,a)$ by proposing an alternative method for estimating the weights directly. We leave this for future work.

3.4 Theoretical guarantees

3.4.1 Marginal coverage

In this section we provide theoretical guarantees on marginal coverage $\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}_n(X))$ for the cases where the weights $w(x,y)$ are known exactly as well as when they are estimated. Using the idea of *weighted exchangeability*, we extend [Tibshirani et al., 2019, Theorem 2] to our setting.

Proposition 3.4.1

Let $\{X_i, Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P^{\pi^b}_{X,Y}$ be the calibration data. For any score function s, and any $\alpha \in (0,1)$, define the conformal predictive interval at a point $x \in \mathbb{R}^d$ as

$$\hat{C}_n(x) \coloneqq \{y \in \mathbb{R} : s(x,y) \leq \eta(x,y)\}$$

where $\eta(x,y) \coloneqq \text{Quantile}_{1-\alpha}(\hat{F}_n^{x,y})$, and $\hat{F}_n^{x,y}$ is as defined in (3.5) with exact weights $w(x,y)$. If $P^{\pi^*}(y|x)$ is absolutely continuous w.r.t. $P^{\pi^b}(y|x)$, then $\hat{C}_n$ satisfies

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}_n(X)) \geq 1-\alpha.$$

Proposition 3.4.1 assumes exact weights $w(x,y)$, which is usually not the case. For

CP under covariate shift, Lei and Candès [2021] showed that even when the weights are approximated, i.e., $\hat{w}(x, y) \neq w(x, y)$, we can still provide finite-sample upper and lower bounds on the coverage, albeit with an error term Δ_w (defined in (3.8)). Next, we extend this result to our setting when the weight function $w(x, y)$ is approximated as in Section 3.3.1.

Proposition 3.4.2

Let $\hat{C}_n$ be the conformal predictive intervals obtained as in Proposition 3.4.1, with weights $w(x, y)$ replaced by approximate weights $\hat{w}(x, y) = \hat{w}(x, y; \mathcal{D}_{tr})$, where the training data, $\mathcal{D}_{tr}$, is fixed. Assume that $\hat{w}(x, y)$ satisfies $(\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X, Y)^r])^{1/r} \leq M_r < \infty$ for some $r \geq 2$. Define Δ_w as,

$$\Delta_w \coloneqq \tfrac{1}{2}\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}} \mid \hat{w}(X, Y) - w(X, Y) \mid . \tag{3.8}$$

$$\text{Then,} \quad \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}_n(X)) \geq 1 - \alpha - \Delta_w.$$

If, in addition, non-conformity scores $\{V_i\}_{i=1}^n$ have no ties almost surely, then we also have

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}_n(X)) \leq 1 - \alpha + \Delta_w + cn^{1/r-1},$$

for some positive constant c depending only on M_r and r.

Proposition 3.4.2 provides finite-sample guarantees with approximate weights $\hat{w}(\cdot, \cdot)$. Note that if the weights are known exactly then the above proposition can be simplified by setting $\Delta_w = 0$. In the case where the weight function is estimated *consistently*, we recover the exact coverage asymptotically. A natural question to ask is whether the consistency of $\hat{w}(x, y)$ implies the consistency of $\hat{P}(y|x, a)$; in which case one could use $\hat{P}(y|x, a)$ directly to construct the intervals. We prove that this is not the case in general and provide detailed discussion in Appendix B.2.5.

3.4.2 Conditional coverage

So far we only considered marginal coverage (3.1), where the probability is over both X and Y. Here, we provide results on conditional coverage $\mathbb{P}_{Y\sim P^{\pi^*}_{Y|X}}(Y \in \hat{C}_n(X) \mid X)$ which is a strictly stronger notion of coverage than marginal coverage [Foygel Barber et al., 2021]. Vovk [2012], Lei and Wasserman [2014] prove that exact conditional coverage cannot be achieved without making additional assumptions. However, we show that, in the case where

Y is a continuous random variable and we can estimate the quantiles of $P^{\pi^*}_{Y|X}$ consistently, we get an approximate conditional coverage guarantee using the below proposition.

Proposition 3.4.3 (Asymptotic conditional coverage)

Let m, n be the number of training and calibration data respectively, $\hat{q}_{\beta,m}(x) = \hat{q}_{\beta,m}(x; \mathcal{D}_{tr})$ be an estimate of the β-th conditional quantile $q_\beta(x)$ of $P^{\pi^*}_{Y|X=x}$, $\hat{w}_m(x, y) = \hat{w}_m(x, y; \mathcal{D}_{tr})$ be an estimate of $w(x, y)$ and $\hat{C}_{m,n}(x)$ be the conformal interval resulting from algorithm 1 with score function $s(x, y) = \max\{y - \hat{q}_{\alpha_{hi}}(x), \hat{q}_{\alpha_{lo}}(x) - y\}$ where $\alpha_{hi} - \alpha_{lo} = 1 - \alpha$. Assume that the following hold:

1. $\lim_{m\to\infty} \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}} |\hat{w}_m(X, Y) - w(X, Y)| = 0.$
2. there exists $r, b_1, b_2 > 0$ such that $P^{\pi^*}(y \mid x) \in [b_1, b_2]$ uniformly over all (x, y) with $y \in [q_{\alpha_{lo}}(x) - r, q_{\alpha_{lo}}(x) + r] \cup [q_{\alpha_{hi}}(x) - r, q_{\alpha_{hi}}(x) + r]$,
3. $\exists k > 0$ s.t. $\lim_{m\to\infty} \mathbb{E}_{X\sim P_X}[H^k_m(X)] = 0$ where

$$H_m(x) = \max\{|\hat{q}_{\alpha_{lo},m}(x) - q_{\alpha_{lo}}(x)|, |\hat{q}_{\alpha_{hi},m}(x) - q_{\alpha_{hi}}(x)|\}$$

Then for any $t > 0$, we have that $\lim_{m,n\to\infty} \mathbb{P}(\mathbb{P}_{Y\sim P^{\pi^*}_{Y|X}}(Y \in \hat{C}_{m,n}(X) \mid X) \leq 1-\alpha-t) = 0.$

One caveat of Prop. 3.4.3 is that Assumption 3 is rather strong. In general, consistently estimating the quantiles under the target policy π^* is not straightforward given that we only have access to observational data from π^b. While one can use a weighted pinball loss to estimate quantiles under π^*, consistent estimation of these quantiles would require a consistent estimate of the weights (see Appendix B.3). Hence, unlike [Lei and Candès, 2021, Theorem 1], our Prop. 3.4.3 is not a "*doubly robust*" result.

Towards group balanced coverage

As pointed out by Angelopoulos and Bates [2021], we may want predictive intervals that have the same coverage across different groups, e.g., across male and female users [Romano et al., 2020]. Standard CP will not necessarily achieve this, as the coverage guarantee (3.1) is over the entire population of users. However, we can use COPP on each subgroup separately to obtain group balanced coverage. A more detailed discussion on how to construct such intervals has been included in Appendix B.2.4.

3.5 Related work

Conformal prediction A number of works have explored the use of CP under distribution shift. The works of Tibshirani et al. [2019] and Lei and Candès [2021] are particularly notable as they extend CP to the general setting of *weighted exchangeability*. In particular, Lei and Candès [2021] use CP for counterfactual inference where the goal is to obtain predictive intervals on the outcomes of treatment and control groups. The authors formulate the counterfactual setting into that of covariate shift in the input space $\mathcal{X}$ and show that under certain assumptions, finite-sample coverage can be guaranteed.

Fundamentally, our work differs from Lei and Candès [2021] by framing the problem as a shift in the conditional $P_{Y|X}$ rather than as a shift in the marginal P_X. The resulting methodology we obtain from this then differs from Lei and Candès [2021] in a variety of ways. For example, while Lei and Candès [2021] assume a deterministic target policy, COPP can also be applied to stochastic target policies, which have been used in a variety of applications, such as recommendation systems or RL applications [Swaminathan et al., 2017a, Su et al., 2020, Farajtabar et al., 2018a]. Likewise, unlike Lei and Candès [2021], COPP is applicable to continuous action spaces, e.g., doses of medication administered.

In addition, when the target policy is deterministic, there is an important methodological difference between COPP and Lei and Candès [2021]. In particular, Lei and Candès [2021] construct the intervals on outcomes by splitting calibration data w.r.t. actions. In contrast, it can be shown that COPP uses the entire calibration data when constructing intervals on outcomes. This is a consequence of integrating out the actions in the weights $w(x, y)$ (3.7), and empirically leads to smaller variance in coverage compared to Lei and Candès [2021]. See B.2.1 for the experimental results comparing COPP to Lei and Candès [2021] for deterministic policies.

Osama et al. [2020] propose using CP to *construct* robust policies in contextual bandits with discrete actions. Their methodology uses CP to choose actions and does not involve evaluating target policies. Hence, the problem being considered is orthogonal to ours. There has also been concurrent work adapting CP to individual treatment effect (ITE) sensitivity analysis model [Jin et al., 2021, Yin et al., 2021]. Similar to our approach, these works formulate the sensitivity analysis problem as one of CP under the joint distribution

shift $P_{X,Y}$. While our methodologies are related, the application of CP explored in these works, i.e. ITE estimation under unobserved confounding, is fundamentally different.

Uncertainty in contextual bandits Recall from the introduction, that most works in this area have focused on quantifying uncertainty in expected outcome (policy value) [Dudík et al., 2014a, Kuzborskij et al., 2021]. Despite providing finite sample-guarantees on the expectation, these methods do not account for the variability in the outcome itself and in general are not adaptive w.r.t. X, i.e. they do not satisfy properties (i), (ii) from Sec. 3.1.1. Huang et al. [2021], Chandak et al. [2021] on the other hand, propose off-policy assessment algorithms for contextual bandits w.r.t. a more general class of risk objectives such as Mean, CVaR etc. Their methodologies can be applied to our problem, to construct predictive intervals for off-policy outcomes. However, unlike COPP, these intervals are not adaptive w.r.t. X, i.e. do not satisfy property (i) in Sec. 3.1.1. Moreover, they do not provide upper bounds on coverage probability, which often leads to overly conservative intervals, as shown in our experiments. Lastly, while distributional perspective has been explored in reinforcement learning [Bellemare et al., 2017], no finite sample-guarantees are available to the best of our knowledge.

3.6 Experiments

Baselines for comparison Given our problem setup, there are no established baselines. Instead, we compare our proposed method COPP to the following competing methods, which were constructed to capture the uncertainty in the outcome distribution and take into account the policy shift.

Weighted Importance Sampling (WIS) CDF estimator Given observational dataset $\mathcal{D}_{obs} = \{x_i, a_i, y_i\}_{i=1}^{n_{obs}}$, Huang et al. [2021] proposed a non-parametric WIS-based estimator for the empirical CDF of Y under π^*, $\hat{F}_{WIS}(t) \coloneqq \frac{\sum_{i=1}^{n_{obs}} \hat{\rho}(a_i,x_i)\mathbb{1}(y_i \leq t)}{\sum_{i=1}^{n_{obs}} \hat{\rho}(a_i,x_i)}$ where $\hat{\rho}(a,x) \coloneqq \frac{\pi^*(a|x)}{\hat{\pi}^b(a|x)}$ are the importance weights. We can use $\hat{F}_{WIS}$ to get predictive intervals $[y_{\alpha/2}, y_{1-\alpha/2}]$ where $y_\beta \coloneqq \text{Quantile}_\beta(\hat{F}_{WIS})$. The intervals $[y_{\alpha/2}, y_{1-\alpha/2}]$ do not depend on x.

Table 3.1: Toy experiment results with required coverage 90%. While WIS intervals provide required coverage, the mean interval length is huge compared to COPP (see table 3.1b).

(a) Mean coverage as a function of policy shift with 2 standard errors over 10 runs.

Coverage	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$
COPP (Ours)	**0.90 ± 0.01**	**0.90 ± 0.01**	**0.91 ± 0.01**
WIS	**0.89 ± 0.01**	**0.91 ± 0.02**	0.94 ± 0.02
SBA	**0.90 ± 0.01**	0.88 ± 0.01	0.87 ± 0.01
COPP (GT weights Ours)	**0.90 ± 0.01**	**0.90 ± 0.01**	**0.90 ± 0.01**
CP (no policy shift)	**0.90 ± 0.01**	0.87 ± 0.01	0.85 ± 0.01
CP (union)	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01

(b) Mean interval length as a function of policy shift with 2 standard errors over 10 runs.

Interval Lengths	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$
COPP (Ours)	9.08 ± 0.10	9.48 ± 0.22	9.97 ± 0.38
WIS	24.14 ± 0.30	32.96 ± 1.80	43.12 ± 3.49
SBA	8.78 ± 0.12	8.94 ± 0.10	8.33 ± 0.09
COPP (GT weights Ours)	8.91 ± 0.09	9.25 ± 0.12	9.59 ± 0.20
CP (no policy shift)	9.00 ± 0.10	9.00 ± 0.10	9.00 ± 0.10
CP (union)	10.66 ± 0.18	11.04 ± 0.2	11.4 ± 0.26

Sampling Based Approach (SBA) As mentioned in Sec. 3.3.1, we can directly use the estimated $\hat{P}(y \mid x, a)$ to construct the predictive intervals as follows. For a given x^{test}, we generate $A_i \overset{\text{i.i.d.}}{\sim} \pi^*(\cdot \mid x^{test})$, and $Y_i \sim \hat{P}(\cdot \mid x^{test}, A_i)$ for $i \leq \ell$. We then define the predictive intervals for x^{test} using the $\alpha/2$ and $1-\alpha/2$ quantiles of $\{Y_i\}_{i \leq \ell}$. While SBA is not a standard baseline, it is a natural comparison to make to answer the question of "why not construct the intervals using $\hat{P}(y|x,a)$ directly"?

3.6.1 Toy experiment

We start with synthetic experiments and an ablation study, in order to dissect and understand our proposed methodology in more detail. We assume that our policies are stationary and there is overlap between the behaviour and target policy, both of which are standard assumptions [Huang et al., 2021, Su et al., 2019a, Xie et al., 2019a].

Synthetic data experiments setup

In order to understand how COPP works, we construct a simple experimental setup where we can control the amount of *"policy shift"* and know the ground truth. In this experiment, $X \in \mathbb{R}$, $A \in \{1, 2, 3, 4\}$ and $Y \in \mathbb{R}$, where X and $Y \mid x, a$ are normal random variables. Further details and additional experiments on continuous action spaces are given in Appendix B.4.1.

Behaviour and target policies We define a family of policies $\pi_\epsilon(a \mid x)$, where we use the parameter $\epsilon \in (0, 1/3)$ to control the policy shift between target and behaviour policies. Exact form of $\pi_\epsilon(a \mid x)$ is given in B.4.1. For the behaviour policy π^b, we use $\epsilon^b = 0.3$ (i.e. $\pi^b(a \mid x) \equiv \pi_{0.3}(a \mid x)$), and for target policies π^*, we use $\epsilon^* \in \{0.1, 0.2, 0.3\}$. Using the true behaviour policy, π^b, we generate observational data $\mathcal{D}_{obs} = \{x_i, a_i, y_i\}_{i=1}^{n_{obs}}$ which is then split into training ($\mathcal{D}_{tr}$) and calibration ($\mathcal{D}_{cal}$) datasets, of sizes m and n respectively.

Estimation of ratios, $\hat{w}(x, y)$ Using the training dataset $\mathcal{D}_{tr}$, we estimate $P(y|x, a)$ as $\hat{P}(y|x, a) = \mathcal{N}(\mu(x, a), \sigma(x, a))$, where $\mu(x, a), \sigma(x, a)$ are both neural networks (NNs). Similarly, we use NNs to estimate the behaviour policy $\hat{\pi}^b$ from $\mathcal{D}_{tr}$. Next, to estimate $\hat{w}(x, y)$, we use (3.7) with $h = 500$.

Score For the score function, we use the same formulation as in Romano et al. [2019], i.e. $s(x, y) = \max\{\hat{q}_{\alpha_{lo}}(x) - y, y - \hat{q}_{\alpha_{hi}}(x)\}$, where $\hat{q}_\beta(x)$ denotes the β quantile estimate of $P^{\pi^b}_{Y|X=x}$ trained using pinball loss.

Lastly, our weights $w(x, y)$ depend on x **and** y and hence we use a grid of 100 equally spaced out y's in our experiments to determine the predictive interval which satisfies $\hat{C}_n(x) := \{y : s(x, y) \leq \text{Quantile}_{1-\alpha}(\hat{F}^{x,y}_n)\}$. This is parallelizable and hence does not add much computational overhead.

Results Table 3.1a shows the coverages of different methods as the policy shift $\Delta_\epsilon = \epsilon^b - \epsilon^*$ increases. The behaviour policy $\pi^b = \pi_{0.3}$ is fixed and we use $n = 5000$ calibration datapoints, across 10 runs. Table 3.1a shows, how COPP stays very close to the required coverage of 90% across all target policies compared to WIS and SBA. WIS intervals are overly conservative i.e. above the required coverage, while the SBA intervals suffer from under-coverage i.e. below the required coverage. These results supports our hypothesis from Sec. 3.3.1, which stated that COPP is less sensitive to estimation errors of $\hat{P}(y|x, a)$ compared to directly using $\hat{P}(y|x, a)$ for the intervals, i.e. SBA.

Next, Table 3.1b shows the mean interval lengths and even though WIS has reasonable coverage for $\Delta_\epsilon = 0.0$ and 0.1, the average interval length is huge compared to COPP. Fig. 3.1b shows the predictive intervals for one such experiment with $\pi^* = \pi_{0.1}$ and $\pi^b = \pi_{0.3}$. We can see that SBA intervals are overly optimistic, while WIS intervals are too wide and are not adaptive w.r.t. X. COPP produces intervals which are much closer to the oracle intervals.

Ablation study

To isolate the effect of weight estimation error and policy shift, we conduct an ablation study, comparing COPP with estimated weights to COPP with Ground Truth (GT) weights and standard CP (assuming no policy shift). Table 3.1a shows that at $\Delta_\epsilon = 0$, i.e. no policy shift, standard CP achieves the required coverage as expected. However the coverage of

Table 3.2: Mean coverage as a function of policy shift Δ_ϵ and 2 standard errors over 10 runs. COPP attains the required coverage of 90%, whereas the competing methods, WIS and SBA, are over-conservative i.e. coverage above 90%. In addition, when we do not account for the policy shift, standard CP becomes progressively worse with increasing policy shift.

	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.3$	$\Delta_\epsilon = 0.4$
COPP (Ours)	**0.90 ± 0.00**	**0.90 ± 0.02**	**0.90 ± 0.01**	**0.89 ± 0.01**	**0.91 ± 0.01**
WIS	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.00	0.94 ± 0.00	0.91 ± 0.00
SBA	0.99 ± 0.00	0.99 ± 0.00	0.98 ± 0.00	0.97 ± 0.00	0.96 ± 0.00
CP (no policy shift)	**0.91 ± 0.02**	**0.92 ± 0.02**	0.93 ± 0.01	0.94 ± 0.01	0.96 ± 0.01

standard CP intervals decreases as the policy shift Δ_ϵ increases. COPP, on the other hand, attains the required coverage of 90%, by adapting the predictive intervals with increasing policy shift. Table 3.1b shows that the average interval length of COPP increases with increasing policy shift Δ_ϵ. Furthermore, Table 3.1a illustrates that while COPP achieves the required coverage for different target policies, on average it is slightly more conservative than using COPP with GT weights. This can be explained by the estimation error in $\hat{w}(x, y)$. Additionally, to investigate the effect of integrating out the actions in (3.7), we also perform CP for each action a separately (as in Lei and Candès [2021]) and then take the union of the intervals across these actions. In the union method, the probability of an action being chosen is not taken into account, (i.e., intervals are independent of π^*) and hence the coverage is overly conservative as expected.

Lastly, we investigate how increasing the number of calibration data n affects the coverage for all the methodologies. We observe that coverage of COPP is closer to the required coverage of 90% compared to the competing methodologies. Additionally, the coverage of COPP converges to the required coverage as n increases; see Appendix B.4.1 for detailed experimental results.

3.6.2 Experiments on Microsoft Ranking Dataset

We now apply COPP onto a real dataset i.e. the Microsoft Ranking dataset 30k [Qin and Liu, 2013, Swaminathan et al., 2017a, Bietti et al., 2018]. Due to space constraints, we have added additional extensive experiments on UCI datasets in Appendix B.4.3.

Dataset The dataset contains relevance scores for websites recommended to different users, and comprises of $30,000$ user-website pairs. For each user-website pair, the data contains a 136-dimensional feature vector, which consists of user's attributes corresponding

to the website, such as length of stay or number of clicks on the website. Furthermore, for each user-website pair, the dataset also contains a relevance score, i.e. how relevant the website was to the user.

First, given a user, we sample (with replacement) 5 websites from the data corresponding to that user. Next, we reformulate this into a contextual bandit where $a_i \in \{1, 2, 3, 4, 5\}$ corresponds to the action of recommending the a_i'th website to the user i. x_i is obtained by combining the 5 user-website feature vectors corresponding to the user i i.e. $x_i \in \mathbb{R}^{5\times 136}$. $y_i \in \{0, 1, 2, 3, 4\}$ corresponds to the relevance score for the a_i'th website, i.e. the recommended website. The goal is to construct prediction sets that are guaranteed to contain the true relevance score with a probability of 90%.

Behaviour and target policies We first train a NN classifier model, $\hat{f}_\theta$, mapping each 136-dimensional user-website feature vector to the softmax scores for each relevance score class. We use this trained model $\hat{f}_\theta$ to define a family of policies which pick the most relevant website as predicted by $\hat{f}_\theta$ with probability ϵ and the rest uniformly with probability $(1-\epsilon)/4$ (see Appendix B.4.2 for more details). Like the previous experiment, we use ϵ to control the shift between behaviour and target policies. For π^b, we use $\epsilon^b = 0.5$ and for π^*, $\epsilon^* \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

Estimation of ratios $\hat{w}(X, Y)$ To estimate the $\hat{P}(y \mid x, a)$ we use the trained model $\hat{f}_\theta$ as detailed in Appendix B.4.2. To estimate the behaviour policy $\hat{\pi}^b$, we train a neural network classifier model $\mathcal{X} \to \mathcal{A}$, and we use (3.7) to estimate the weights $\hat{w}(x, y)$.

Score The space of outcomes $\mathcal{Y}$ in this experiment is discrete. We define $\hat{P}^{\pi^b}(y \mid x) = \sum_{i=1}^{5} \hat{\pi}^b(A = i|x)\hat{P}(y|x, A = i)$. Using similar formulation as in Angelopoulos and Bates [2021], we define the score:

$$s(x, y) = \sum_{y'=0}^{4} \hat{P}^{\pi^b}(y' \mid x)\mathbb{1}(\hat{P}^{\pi^b}(y' \mid x) \geq \hat{P}^{\pi^b}(y \mid x)).$$

Since $\mathcal{Y}$ is discrete, we no longer need to construct a grid of y values on which to compute $\text{Quantile}_{1-\alpha}(\hat{F}_n^{x,y})$. Instead, we will simply compute this quantity on each $y \in \mathcal{Y}$, when constructing the predictive sets $\hat{C}_n(x^{test})$.

Results Table 3.2 shows the coverages of different methodologies across varying target policies π_{ϵ^*}. The behaviour policy $\pi^b = \pi_{0.5}$ is fixed and we use $n = 5000$ calibration datapoints, across 10 runs. Table 3.2 also shows that the coverage of WIS and SBA sets is dependent upon the policy shift, with both being overly conservative across the different target policies as compared to COPP. Recall that the WIS sets do not depend on x^{test} and as a result we get the same set for each test data point. This becomes even more problematic when Y is discrete – if, for each label y, $\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y = y) > 10\%$, then WIS sets (with the required coverage of 90%) are likely to contain every label $y \in \mathcal{Y}$. In comparison, COPP is able to stay much closer to the required coverage of 90% across all target policies. We have also added standard CP without policy shift as a sanity check, and observed that the sets get increasingly conservative as the policy shift increases.

Finally, we also plotted how the coverage changes as the number of calibration data n increases. We observe again that the coverage of COPP is closer to the required coverage of 90% compared to the competing methodologies. Due to space constraints, we have added the plots in Appendix B.4.2.

Class-balanced conformal prediction Using the methodology described in Sec. 3.4.2, we construct predictive sets, $\hat{C}^{\mathcal{Y}}_n(x)$, which offer label conditioned coverage guarantees (see B.2.4), i.e. for all $y \in \mathcal{Y}$,

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}^{\mathcal{Y}}_n(X) \mid Y = y) \geq 1 - \alpha.$$

We empirically demonstrate that $\hat{C}^{\mathcal{Y}}_n$ provides label conditional coverage, while $\hat{C}_n$ obtained using alg. 1 may not. Due to space constraints, details on construction of $\hat{C}^{\mathcal{Y}}_n$ as well as experimental results have been included in Appendix B.4.2.

3.7 Conclusion and limitations

In this paper, we propose COPP, an algorithm for constructing predictive intervals on off-policy outcomes, which are adaptive w.r.t. covariates X. We theoretically prove that COPP can guarantee finite-sample coverage by adapting the framework of conformal prediction to our setup. Our experiments show that conventional methods cannot guarantee any user pre-specified coverage, whereas COPP can. For future work, it would be interesting

to apply COPP to policy training. This could be a step towards robust policy learning by optimising the worst case outcome [Stutz et al., 2022].

We conclude by mentioning several limitations of COPP. Firstly, we do not guarantee conditional coverage in general. We outline conditions under which conditional coverage holds asymptotically (Prop. 3.4.3), however, this relies on somewhat strong assumptions. Secondly, our current method estimates the weights $w(x, y)$ through $P(y \mid x, a)$, which can be challenging. We address this limitation in Appendix B.2.5, where we propose an alternative method to estimate the weights directly, without having to model $P(y \mid x, a)$. Lastly, reliable estimation of our weights $\hat{w}(x, y)$ requires sufficient overlap between behaviour and target policies. The results from COPP may suffer in cases where this assumption is violated, which we illustrate empirically in Appendix B.4.1. We believe these limitations suggest interesting research questions that we leave to future work.

Acknowledgements

We would like to thank Andrew Jesson, Sahra Ghalebikesabi, Robert Hu, Siu Lun Chau and Tim Rudner for useful feedback. JFT is supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). MFT acknowledges his PhD funding from Google DeepMind. RC and AD are supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme [Grant number EP/R018561/1].

4

Causal Falsification of Digital Twins

Contents

Abstract

Digital twins are simulation-based models designed to predict how a real-world process will evolve in response to interventions. This modelling paradigm holds substantial promise in many applications, but rigorous procedures for assessing their accuracy are essential for safety-critical settings. We consider how to assess the accuracy of a digital twin using real-world data. We formulate this as causal inference problem, which leads to a precise definition of what it means for a twin to be "correct". Unfortunately, fundamental results from causal inference mean observational data cannot be used to certify a twin in this sense unless potentially tenuous assumptions are made, such as that the data are unconfounded. To avoid these assumptions, we propose instead to find situations in which the twin *is not* correct, and present a general-purpose statistical procedure for doing so. Our approach yields reliable and actionable information about the twin under only the assumption of an i.i.d. dataset of observational trajectories, and remains sound even if the data are confounded. We apply our methodology to a large-scale, real-world case study involving sepsis modelling within the Pulse Physiology Engine, which we assess using the MIMIC-III dataset of ICU patients.

4.1 Introduction

4.1.1 Motivation

There is increasing interest in the use of simulation-based models for obtaining *causal* insights. Such models aim to describe what *would* occur when different actions or interventions are applied to some real-world process of interest, thereby allowing planning and decision-making to be done with a fuller understanding of the different outcomes that may result. In many applications, models of this kind are referred to as *digital twins* [Barricelli et al., 2019, Jones et al., 2020, Niederer et al., 2021]. These have been considered for a wide range of use-cases including aviation [Bellinger et al., 2011], manufacturing [Lu et al., 2020], healthcare [Corral-Acero et al., 2020, Coorey et al., 2022], civil engineering [Sacks et al., 2020], and agriculture [Jans-Singh et al., 2020].

Many applications of digital twins are considered safety-critical, which means the cost of deploying an inaccurate twin to production is potentially very high. As such, methodology for assessing the performance of a twin before its deployment is essential for the safe, widespread adoption of digital twins in practice [Niederer et al., 2021]. In this work, we consider the problem of assessing twin accuracy and propose a concrete, theoretically grounded, and general-purpose methodology to this end. We focus specifically on the use of statistical methods that leverage data obtained from the real-world process that the twin is designed to model. Such strategies are increasingly viable for many applications as datasets grow larger, and offer the promise of lower overheads compared with alternative strategies that rely for instance on domain expertise.

We formulate twin assessment as a problem of *causal inference* [Rubin, 1974, 2005, Pearl, 2009, Hernán and Robins, 2020]. In particular, we consider a twin to be accurate if it correctly captures the behaviour of a real-world process of interest in response to certain interventions, rather than the behaviour of the process as it evolves on its own. This seems in keeping with the overall (if often implicit) design objectives that underlie many applications, including those cited above. Our causal assessment approach also highlights certain pitfalls associated with conventional methods that do not account for causal factors, and that can give rise to misleading inferences about the twin as a result.

In most cases, it is desirable for an assessment procedure to be reliable and robust, and for its conclusions about the twin to be highly trustworthy. As such, our goal in this

paper is to obtain a methodology that is always *sound*, even possibly at the expense of being conservative: we prefer not to draw any conclusion about the accuracy of the twin at all than to draw some conclusion that is potentially misleading. To this end, we rely on minimal assumptions about the twin and the real-world process of interest. In addition to improving robustness, this also means our resulting methodology is very general, and may be applied to a wide variety of twins across application domains.

4.1.2 Contribution

We begin by providing a causal model for a general-purpose twin and the data we have available for assessment. We use this to show precisely it is not possible to use observational data to *certify* that the twin is causally accurate unless strong and often tenuous assumptions are made about the data-generating process, such as that the data are free of unmeasured confounding. To avoid these assumptions, we propose an assessment paradigm instead based on *falsification*: we search for specific cases when the twin is *not* accurate, rather than trying to quantify its accuracy in a more holistic sense.

To obtain a practical methodology suitable for real twins, we provide a novel set of longitudinal causal bounds that hold without additional causal assumptions. These bounds generalize the classical bounds of Manski [1990], and can be considerably more informative in comparison. We use this result as the basis for a general-purpose statistical testing procedure for falsifying a twin. Overall, our method relies on only the assumption of an independent and identically distributed (i.i.d.) dataset of observational trajectories: it does not require modelling the dynamics of the real-world process or any internal implementation details of the twin, and remains sound in the presence of arbitrary unmeasured confounding. We demonstrate the effectiveness of our procedure through a large-scale, real-world case study in which we use the MIMIC-III ICU dataset [Johnson et al., 2016] to assess the Pulse Physiology Engine [Bray et al., 2019], an open-source model for human physiology simulation.

4.1.3 Related work

Various high-level guidelines and workflows have been proposed for the assessment of digital twins in the literature to-date [Roy and Oberkampf, 2011, Grieves and Vickers, 2017, Khan et al., 2018, Corral-Acero et al., 2020, Kochunas and Huan, 2021, Niederer et al., 2021,

Dahmen et al., 2022]. In some cases, these guidelines have been codified as standards: for example, the ASME V&V40 Standard [AMSE, 2018] provides a risk-based framework for assessing the credibility of a model from a variety of factors that include source code quality and the mathematical form of the model [Galappaththige et al., 2022]. However, a significant gap still exists between these guidelines and a practical implementation that could be deployed for real twins, and the need for a rigorous lower-level framework to enable the systematic assessment of twins has been noted in this literature [Corral-Acero et al., 2020, Niederer et al., 2021, Kapteyn et al., 2021, Masison et al., 2021]. We contribute towards this effort by describing a precise statistical methodology for twin assessment that can be readily implemented in practice, and which is accompanied by theoretical guarantees of robustness that hold under minimal assumptions.

In addition, a variety of concrete assessment procedures have been applied to certain specific digital twin models in the literature. For example, the Pulse Physiology Engine [Bray et al., 2019], which we consider in our empirical case study, as well as the related BioGears [McDaniel et al., 2019, McDaniel and Baird, 2019] were both assessed by comparing their outputs with ad hoc values based either on available medical literature or the opinions of subject matter experts. Other twins have been assessed by comparing their outputs with real-world data through a variety of bespoke numerical schemes [Larrabide et al., 2012, Hemmler et al., 2019, Lal et al., 2021, Jans-Singh et al., 2020, Galappaththige et al., 2022]. In contrast, our paper proposes a general-purpose statistical procedure for assessing twins that may be applied generically across many applications and architectures. To the best of our knowledge, our paper is also the first to identify the need for a causal approach to twin assessment and the pitfalls that arise when causal considerations are not properly accounted for.

4.2 Causal formulation

4.2.1 The real-world process

We begin by providing a causal model the real-world process that the twin is designed to simulate. We do so in the language of *potential outcomes* [Rubin, 1974, 2005], although we note that we could have used the alternative framework of directed acyclic graphs and structural causal models [Pearl, 2009] (see also Imbens [2020] for a comparison of the

two). We assume the real-world process operates over a fixed time horizon $T \in \{1, 2, \ldots\}$. This simplifies our presentation in what follows, and it is straightforward to generalize our methodology to variable length time horizons if needed. For each $t \in \{0, \ldots, T\}$, we assume the process gives rise to a *observation* at time t, which takes values in some real-valued space $\mathcal{X}_t \coloneqq \mathbb{R}^{d_t}$. We also assume that the process can be influenced by some *action* taken at each time $t \in \{1, \ldots, T\}$. We denote the space of actions available at time t by $\mathcal{A}_t$, which in this work we assume is always finite. For example, in a robotics context, the observations may consist of all the readings of all the sensors of the robot, and the actions may consist of commands that can be input by an external user. In a medical context, the observations may consist of the vital signs of a patient, and the actions may consist of possible treatments or interventions. To streamline notation, we will index these spaces using vector notation, so that e.g. $\mathcal{A}_{1:t}$ denotes the cartesian product $\mathcal{A}_1 \times \cdots \times \mathcal{A}_t$, and $a_{1:t} \in \mathcal{A}_{1:t}$ is a choice of $a_1 \in \mathcal{A}_1, \ldots, a_t \in \mathcal{A}_t$.

We model the dynamics of the real-world process via the longitudinal potential outcomes framework proposed by Robins [1986], which imposes only a weak temporal structure on the underlying phenomena of interest and so may be applied across a wide range of applications in practice. In particular, for each $a_{1:T} \in \mathcal{A}_{1:T}$, we posit the existence of random variables or *potential outcomes* $X_0, X_1(a_1), \ldots, X_T(a_{1:T})$, where $X_t(a_{1:t})$ takes values in $\mathcal{X}_t$. We will denote this sequence more concisely as $X_{0:T}(a_{1:T})$. Intuitively, X_0 represents data available before the first action, while $X_{1:T}(a_{1:T})$ represents the sequence of real-world outcomes that *would* occur if actions $a_{1:T}$ were taken successively. These quantities are therefore of fundamental interest for planning a course of actions to achieve some desired result.

As random variables, each $X_t(a_{1:t})$ may depend on additional randomness that is not explicitly modelled, and so in particular may be influenced by all the previous potential outcomes $X_{0:t-1}(a_{1:t-1})$, and possibly other random quantities. This models a process whose initial state is determined by external factors, such as when a patient from some population first presents at a hospital, and where the process then evolves according both to specific actions chosen from $\mathcal{A}_{1:T}$ as well as additional external factors. It is clear that this structure applies to a wide range of phenomena occurring in practice.

4.2.2 The digital twin

We think of the twin as a computational device that, when executed, outputs a sequence of values intended to simulate a possible future trajectory of the real-world process when certain actions in $\mathcal{A}_{1:T}$ are chosen, conditional on some initial data in $\mathcal{X}_0$. We allow the twin to make use of an internal random number generator to produce outputs that vary stochastically even under fixed inputs (although our framework encompasses twins that evolve deterministically also). By executing the twin repeatedly, a user may therefore estimate the range of behaviours that the real-world process may exhibit under different action sequences, which can then inform planning and decision-making downstream.

Precisely, we model the output the twin would produce at timestep $t \in \{1, \dots, T\}$ after receiving initialisation $x_0 \in \mathcal{X}_0$ and successive inputs $a_{1:t} \in \mathcal{A}_{1:t}$ as the quantity $h_t(x_0, a_{1:t}, U_{1:t})$, where h_t is a measurable function taking values in $\mathcal{X}_t$, and each U_s is some (possibly vector-valued) random variable. We will denote $\widehat{X}_t(x_0, a_{1:t}) \coloneqq h_t(x_0, a_{1:t}, U_{1:t})$, which we also refer to as a potential outcome. A full twin trajectory therefore consists of $\widehat{X}_1(x_0, a_1), \dots, \widehat{X}_T(x_0, a_{1:T})$, which we write more compactly by $\widehat{X}_{1:T}(x_0, a_{1:T})$. Conceptually, $h_1, \dots, h_T$ constitute the program that executes inside the twin, and $U_{1:T}$ may be thought of as the collection of all outputs of the internal random number generator that the twin uses. We assume these random numbers $U_{1:T}$ and the real-world outcomes $(X_{0:T}(a_{1:T}) : a_{1:T} \in \mathcal{A}_{1:T})$ are independent, which is mild in practice. We also assume that repeated executions of the twin give rise to i.i.d. copies of $U_{1:T}$. This means that, given fixed inputs x_0 and $a_{1:T}$, repeated executions of the twin produce i.i.d. copies of $\widehat{X}_{1:T}(x_0, a_{1:T})$. Otherwise, we make no assumptions about the precise form of either the h_t or the U_t, which allows our model to encompass a wide variety of possible twin implementations.

4.2.3 Correctness

Before we can consider how to assess the twin, we must first define how we want the twin ideally to behave. The following condition seems appropriate for many applications.

Definition 4.2.1 (Correctness)

The twin is *interventionally correct* if, for $\mathrm{Law}[X_0]$-almost all x_0 and $a_{1:T} \in \mathcal{A}_{1:T}$, distribution of $\widehat{X}_{1:T}(x_0, a_{1:T})$ is equal to the conditional distribution of $X_{1:T}(a_{1:T})$ given $X_0 = x_0$.

Operationally, if a twin is interventionally correct, then by repeatedly executing the twin and applying Monte Carlo techniques, it is possible to approximate arbitrarily well the conditional distribution of the future of the real-world process under each possible choice of action sequence. The same can also be shown to hold when each action at each time t is chosen dynamically on the basis of previous observations in $\mathcal{X}_{0:t}$. As a result, an interventionally correct twin may be used for *planning*, or in other words may be used to select a policy for choosing actions that will yield a desirable distribution over observations at each step. We emphasise that interventional correctness does not mean the twin will accurately predict the behaviour of any *specific* trajectory of the real-world process in an almost sure sense (unless the real-world process is deterministic), but only the distribution of outcomes that will be observed over repeated independent trajectories. However, this is sufficient for many applications, and appears to be the strongest guarantee possible when dealing with real-world phenomena whose underlying behaviour is stochastic.

Definition 4.2.1 introduces some technical difficulties that arise in the general case when conditioning on events with probability zero (e.g. $\{X_0 = x_0\}$ if X_0 is continuous). In what follows, it is more convenient to consider an unconditional formulation of interventional correctness. This is supplied by the following result, which considers the behaviour of the twin when it is initialised with the (random) value of X_0 taken from the real-world process, rather than with a fixed choice of x_0. See Section C.2 of the Appendix for a proof.

Proposition 4.2.1

The twin is interventionally correct if and only if, for all choices of $a_{1:T} \in \mathcal{A}_{1:T}$, the distribution of $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T}))$ is equal to the distribution of $X_{0:T}(a_{1:T})$.

4.2.4 Online prediction

Our model here represents a twin at time $t = 0$ making predictions about all future timesteps $t \in \{1, \ldots, T\}$ under different choices of inputs $a_{1:T}$. In practice, many twins are designed to receive new information at each timestep in an online fashion and update their predictions for subsequent timesteps accordingly [Grieves and Vickers, 2017, Niederer et al., 2021]. Various notions of correctness can be devised for this online setting. We describe two possibilities in Section C.3 of the Appendix, and show that these notions of correctness essentially reduce to Definition 4.2.1, which motivates our focus on that notion in what follows.

4.3 Data-driven twin assessment

4.3.1 Overall setup

There are many conceivable methods for assessing the accuracy of a twin, including static analysis of the twin's source code and the solicitation of domain expertise, and in practice it seems most robust to use a combination of different techniques rather than relying on any single one [AMSE, 2018, Niederer et al., 2021]. However, in this paper, we focus on what we will call *data-driven assessment*, which we see as an important component of a larger assessment pipeline. That is, we consider the use of statistical methods that rely solely on a dataset of trajectories obtained from the real-world process and the twin. We show in this section that without further assumptions, it is not possible to obtain a data-driven assessment procedure that can *certify* that a twin is interventionally correct. We instead propose a strategy based on *falsifying* the twin, which we develop into a concrete statistical testing procedure in later sections.

We will assume access to a dataset of trajectories obtained by observing the interaction of some behavioural agents with the real-world process. We model each trajectory as follows. First, we represent the action chosen by the agent at time $t \in \{1, \ldots, T\}$ as an $\mathcal{A}_t$-valued random variable A_t. We then obtain a trajectory in our dataset by recording at each step the action A_t chosen and the observation $X_t(A_{1:t})$ corresponding to this choice of action. As a result, each observed trajectory has the following form:

$$X_0, A_1, X_1(A_1), \ldots, A_T, X_T(A_{1:T}). \tag{4.1}$$

This corresponds to the standard *consistency* assumption in causal inference [Hernán and Robins, 2020], and intuitively means that the potential outcome $X_t(a_{1:t})$ is observed in the data when the agent actually chose $A_{1:t} = a_{1:t}$. We model our full dataset as a set of i.i.d. copies of (4.1).

4.3.2 Certification is unsound in general

A natural high-level strategy for twin assessment has the following structure. First, some hypothesis $\mathcal{H}$ is chosen with the following property:

$$\text{If } \mathcal{H} \text{ is true, then the twin is interventionally correct.} \tag{4.2}$$

Data is then used to try to show $\mathcal{H}$ is true, perhaps up to some level of confidence. If successful, it follows by construction that the twin is interventionally correct. Assessment procedures designed to *certify* the twin in this way are appealing because they promise a strong guarantee of accuracy for certified twins. Unfortunately, the following foundational result from the causal inference literature (often referred to as the *fundamental problem of causal inference* [Holland, 1986]) means that data-driven certification procedures of this kind are in general unsound, as we explain next. For completeness, Section C.4 of the Appendix includes a self-contained proof of this result in our notation.

Theorem 4.3.1

If $\mathbb{P}(A_{1:T} \neq a_{1:T}) > 0$, then the distribution of $X_{0:T}(a_{1:T})$ is not uniquely identified by the distribution of the data in (4.1) without further assumptions.

Since the distribution of the data encodes the information that would be contained in an infinitely large dataset of trajectories, Theorem 4.3.1 imposes a fundamental limit on what can be learned about the distribution of $X_{0:T}(a_{1:T})$ from the data we have assumed. It follows that if $\mathcal{H}$ is any hypothesis satisfying (4.2), then $\mathcal{H}$ cannot be determined to be true from even an infinitely large dataset. This is because, if we could do so, then we could also determine the distribution of $X_{0:T}(a_{1:T})$, since by Proposition 4.2.1 this would be equal to the distribution of $(X_0, \widehat{X}_T(X_0, a_{1:T}))$. In other words, we cannot use the data alone to certify that the twin is interventionally correct.

4.3.3 The assumption of no unmeasured confounding

Theorem 4.3.1 is true in the general case, when no additional assumptions about the data-generating process are made. One way forward is therefore to introduce assumptions under which the distribution of $X_{0:T}(a_{1:T})$ *can* be identified. This would mean it is possible to certify that the twin is interventionally correct, since, at least in principle, we could simply check whether this matches the distribution of $(X_0, \widehat{X}_{1:T}(X_0, a_{1:T}))$ produced by the twin.

The most common such assumption in the causal inference literature is that the data are free of *unmeasured confounding*. Informally, this holds when each action A_t is chosen by the behavioural agent solely on the basis of the information available at time t that is actually recorded in the dataset, namely $X_0, A_1, X_1(A_1), \ldots, A_{t-1}, X_{t-1}(A_{1:t-1})$, as well as possibly some additional randomness that is independent of the real-world process, such as

the outcome of a coin toss. (This can be made precise via the *sequential randomisation assumption* introduced by Robins [1986].) Unobserved confounding is present whenever this does not hold, i.e. whenever some unmeasured factor simultaneously influences both the agent's choice of action and the observation produced by the real-world process.

It is reasonable to assume that the data are unconfounded in certain contexts. For example, in certain situations it may be possible to gather data in a way that specifically guarantees there is no confounding. Randomised controlled trials, which ensure that each A_t is chosen via a carefully designed randomisation procedure [Lavori and Dawson, 2004, Murphy, 2005], constitute a widespread example of this approach. Likewise, it is possible to show that the data are unconfounded if each $X_t(a_{1:t})$ is a deterministic function of $X_{0:t-1}(a_{1:t-1})$ and a_t, which may be reasonable to assume for example in certain low-level physics or engineering contexts. (See Section C.5 of the Appendix for a proof.) However, for stochastic phenomena and for typical datasets, it is widely acknowledged that the assumption of no unmeasured confounding will rarely hold, and so assessment procedures based on this assumption may yield unreliable results in practice [Murphy, 2003, Tsiatis et al., 2019]. Section C.6 of the Appendix illustrates this concretely with a toy scenario.

4.3.4 General-purpose assessment via falsification

Our goal is to obtain an assessment methodology that is general-purpose, and as such we would like to avoid introducing assumptions such as unconfoundedness that do not hold in general. To achieve this, borrowing philosophically from Popper [2005], we propose a strategy that replaces the goal of verifying the interventional correctness of the twin with that of *falsifying* it. Specifically, we consider hypotheses $\mathcal{H}$ with the dual property to (4.2), namely:

$$\text{If the twin is interventionally correct, then } \mathcal{H} \text{ is true.} \tag{4.3}$$

We will then try to show that each such $\mathcal{H}$ is *false*. Whenever we are successful, we will thereby have gained some knowledge about a failure mode of the twin, since by construction the twin can only be correct if $\mathcal{H}$ is true. In effect, each $\mathcal{H}$ we falsify will constitute a *reason* that the twin is not correct, and may suggest concrete improvements to its design, or may identify cases where its output should not be trusted.

Importantly, unlike for (4.2), Theorem 4.3.1 does not preclude the possibility of data-driven assessment procedures based on (4.3). As we show below, there do exist hypotheses $\mathcal{H}$ satisfying (4.3) that can in principle be determined to be false from the data alone without additional assumptions. In this sense, falsification provides a means for *sound* data-driven twin assessment, whose results can be relied upon across a wide range of circumstances. On the other hand, falsification approaches cannot provide a *complete* guarantee about the accuracy of a twin: even if we fail to falsify many $\mathcal{H}$ satisfying (4.3), we cannot then infer that the twin is correct. As such, in situations where (for example) it is reasonable to believe that the data are in fact unconfounded, it may be desirable to use this assumption to obtain additional information about the twin than is possible from falsification alone.

4.4 Longitudinal causal bounds

4.4.1 Statement of result

One possible means for obtaining interventional information about the twin is via the classical bounds proposed by Manski [1990]. These bounds hold without further assumptions, and so could in principle give rise to a sound falsification procedure of the kind we are seeking. However, although they have been successfully applied in various cases, Manski's bounds are often very conservative, and so would not lead to very informative results if used directly. To address this, we propose a novel generalisation of these bounds that explicitly accounts for the temporal structure of our setting. As we explain below, our bounds can become considerably more informative than those of Manski, while also not requiring the addition of untestable causal assumptions. We provide these bounds next, along with several theoretical results about their behaviour and optimality. In Section 4.5, we use these bounds to define a class of $\mathcal{H}$ with the desired property (4.3), which then yields a procedure for falsifying twins through hypothesis testing techniques.

Theorem 4.4.1

Suppose $(Y(a_{1:t}) : a_{1:t} \in \mathcal{A}_{1:t})$ are real-valued potential outcomes defined jointly with $(X_{0:T}(a_{1:T}) : a_{1:T} \in \mathcal{A}_{1:T})$ and $A_{1:T}$, and that for some $t \in \{1, \dots, T\}$, $a_{1:t} \in \mathcal{A}_{1:t}$, measurable $B_{0:t} \subseteq \mathcal{X}_{0:t}$, and $y_{\text{lo}}, y_{\text{up}} \in \mathbb{R}$ we have

$$\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}) > 0 \tag{4.4}$$

$$\mathbb{P}(y_{\text{lo}} \leq Y(a_{1:t}) \leq y_{\text{up}} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) = 1. \tag{4.5}$$

Then it holds that

$$\mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}] \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \leq \mathbb{E}[Y_{\text{up}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}]. \tag{4.6}$$

where we define $N \coloneqq \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\}$, and similarly

$$Y_{\text{lo}} \coloneqq \mathbb{1}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{lo}}$$
$$Y_{\text{up}} \coloneqq \mathbb{1}(A_{1:t} = a_{1:t})\, Y(A_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{up}}.$$

(See Section C.7 of the Appendix for a proof.)

For brevity, in what follows we will write the terms in (4.6) as Q_{lo}, Q, and Q_{up} respectively, so the conclusion of this result becomes $Q_{\text{lo}} \leq Q \leq Q_{\text{up}}$.

Intuitively, $Y(a_{1:t})$ here may be thought of as some quantitative outcome of interest. For example, in a medical context, $Y(a_{1:t})$ might represent the heart rate of a patient at time t after receiving some treatments $a_{1:t}$. When defining our hypotheses below, we consider the specific form $Y(a_{1:t}) \coloneqq f(X_{0:t}(a_{1:t}))$, where $f : \mathcal{X}_{0:t} \to \mathbb{R}$ is some scalar function. The value Q is then simply the (conditional) average behaviour of this outcome. By Theorem 4.3.1, Q is in general not identified by the data since it depends on $X_{0:t}(a_{1:t})$. On the other hand, both Q_{lo} and Q_{up} *are* identified, since the relevant random variables Y_{lo}, Y_{up}, N, and $X_{0:N}(A_{1:N})$ can all be expressed as functions of the observed data $X_{0:t}(A_{1:t})$ and $A_{1:t}$. In this way, Theorem 4.4.1 bounds the behaviour of a non-identifiable quantity in terms of identifiable ones. At a high level, this is achieved by replacing $Y(a_{1:t})$, whose value is only observed on the event $\{A_{1:t} = a_{1:t}\}$, with Y_{lo} and Y_{up}, which are equal to $Y(a_{1:t})$ when its value is observed (i.e. when $A_{1:t} = a_{1:t}$), and which fall back to the worst-case values of y_{lo}

and y_{up} otherwise. We emphasise that Theorem 4.4.1 does not require any additional causal assumptions, and in particular remains true under arbitrary unmeasured confounding.

In the structural causal modelling framework [Pearl, 2009], a related result to Theorem 4.4.1 was given as Corollary 1 by Zhang and Bareinboim [2019]. However, their result involves a complicated ratio of unknown quantities that makes estimation of their bounds difficult, since it is not obvious how to obtain an unbiased estimator for their ratio term. In contrast, our proposed causal bounds are considerably simpler, since both Q_{lo} and Q_{up} here are expressed as (conditional) expectations. This makes their unbiased estimation straightforward, which we use to obtain exact confidence intervals for both terms in Section 4.5.2.

4.4.2 Informativeness

For Theorem 4.4.1 to be useful in practice, we would like the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ to be relatively narrower than the worst-case bounds $[y_{\text{lo}}, y_{\text{up}}]$ that are trivially implied by (4.5). We can quantify the extent to which this occurs by the ratio

$$\frac{Q_{\text{up}} - Q_{\text{lo}}}{y_{\text{up}} - y_{\text{lo}}} = 1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}), \tag{4.7}$$

where the equality here follows from the definitions of Q_{lo} and Q_{up} together with some straightforward manipulations. In other words, the (relative) tightness of our bounds is determined by the value of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, which is itself closely related to the classical *propensity score* in the causal inference literature [Rosenbaum and Rubin, 1983]. Intuitively, as this probability grows larger, so too does $\mathbb{P}(Y(a_{1:t}) = Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, which means the effect of unmeasured confounding on the value of Q is reduced, leading to tighter bounds.

Theorem 4.4.1 is a generalisation of the bounds proposed by Manski [1990], which can be recovered as the case where $B_{0:t} = \mathcal{X}_{0:t}$, so that (4.6) becomes $\mathbb{E}[Y_{\text{lo}}] \leq \mathbb{E}[Y(a_{1:t})] \leq \mathbb{E}[Y_{\text{up}}]$. In practice, Manski's result is often regarded as quite uninformative. From (4.7), this is true whenever $\mathbb{P}(A_{1:t} = a_{1:t})$ is small, which often occurs in many applications, particularly for longer action sequences. On the other hand, in many contexts it seems reasonable to anticipate that certain longer action sequences will be fairly likely to occur when conditioned on some intermediate observations. In other words, $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$

may be large, even if $\mathbb{P}(A_{1:t} = a_{1:t})$ is not. By choosing $B_{0:t}$ carefully, we can therefore obtain tighter bounds than would be possible by using Manski's original result. The following straightforward result provides a sufficient condition for this to hold. In Section 4.6 below, we also show empirically that Theorem 4.4.1 yields more informative results in our case study compared with Manski's original bounds.

Proposition 4.4.1

Consider the same setup as Theorem 4.4.1, where also $y_{\text{lo}} \leq Y(a_{1:t}) \leq y_{\text{up}}$ almost surely. If $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}) > \mathbb{P}(A_{1:t} = a_{1:t})$, then the width of Manski's bounds exceeds that of Theorem 4.4.1, i.e. $Q_{\text{up}} - Q_{\text{lo}} < \mathbb{E}[Y_{\text{up}}] - \mathbb{E}[Y_{\text{lo}}]$.

Beyond allowing us to obtain tighter bounds, the conditional nature of Theorem 4.4.1 also appears of interest simply for its own sake. In particular, Theorem 4.4.1 describes the interventional behaviour of $Y(a_{1:t})$ conditional on the behaviour of the trajectory $X_{0:t}(a_{1:t})$, thereby providing more granular information than can be obtained from unconditional bounds of Manski [1990] alone.

4.4.3 Optimality

The following result shows that Theorem 4.4.1 cannot be improved without further assumptions. Intuitively speaking, there always exists *some* family of potential outcomes that produces the same observational data as our model, but that attains the worst-case bounds Q_{lo} or Q_{up}. Therefore, we cannot rule out the possibility that the true potential outcomes achieve Q_{lo} or Q_{up} from the observational data alone.

Proposition 4.4.2

Under the same setup as in Theorem 4.4.1, there always exists potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}), \tilde{Y}(a'_{1:t}) : a'_{1:T} \in \mathcal{A}_{1:T})$ also satisfying (4.5) (mutatis mutandis) with

$$(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T})$$

but for which

$$\mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}] = Q_{\text{lo}}.$$

The corresponding statement is also true for Q_{up}.

Apart from attempting to tighten our bounds on Q, in some cases we may wish to consider bounding the alternative quantity $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ that conditions on the *value* of $X_{0:t}(a_{1:t})$ rather than on the event $\{X_{0:t}(a_{1:t}) \in B_{0:t}\}$. To achieve this, it is natural to generalize our assumption (4.5) by supposing we now have measurable functions $y_{\text{lo}}, y_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ such that

$$y_{\text{lo}}(X_{0:t}(a_{1:t})) \leq Y(a_{1:t}) \leq y_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely,} \tag{4.8}$$

and our goal is to obtain measurable functions $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ such that

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})] \leq g_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely.} \tag{4.9}$$

As we describe in Section C.7.4 of the Appendix, bounds of this kind can be obtained directly from Theorem 4.4.1 if $X_{0:t}(a_{1:t})$ is discrete, or by a simple modification of the proof of Theorem 4.4.1 if $X_{1:t}(a_{1:t})$ is discrete (but X_0 is possibly continuous). However, somewhat surprisingly, in general we cannot obtain nontrivial bounds of this kind without further assumptions beyond the discrete case. To make this precise, we will say that a given g_{lo} and g_{up} are *permissible* if (4.9) holds when $(X_{0:T}(a_{1:t}), Y(a_{1:t}), A_{1:T} : a'_{1:T} \in \mathcal{A}_{1:T})$ are replaced by any potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}), \tilde{Y}(a'_{1:t}), \tilde{A}_{1:T} : a'_{1:T} \in \mathcal{A}_{1:T})$ for which $\text{Law}[\tilde{X}_{0:T}(\tilde{A}_{1:T}), \tilde{A}_{1:T}, \tilde{Y}(\tilde{A}_{1:t})] = \text{Law}[X_{0:T}(A_{1:T}), A_{1:T}, Y(A_{1:t})]$, and which also satisfy (4.8) (mutatis mutandis). Intuitively, this means that g_{lo} and g_{up} depend only on the information we have available, i.e. the observational distribution and our assumed worst-case values. We then have the following:

Theorem 4.4.2

Suppose X_0 is almost surely constant, $\mathbb{P}(A_1 \neq a_1) > 0$, and for some $s \in \{1, \ldots, t\}$ we have $\mathbb{P}(X_s(A_{1:s}) = x_s) = 0$ for all $x_s \in \mathcal{X}_s$. Then $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ are permissible bounds only if they are trivial, i.e.

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq y_{\text{lo}}(X_{0:t}(a_{1:t})) \quad \text{and} \quad g_{\text{up}}(X_{0:t}(a_{1:t})) \geq y_{\text{up}}(X_{0:t}(a_{1:t})) \quad \text{almost surely.}$$

Here the assumption that X_0 is constant essentially means we consider a special case of our model where there are no covariates available before the first action is taken, and serves mainly to simplify the proof. We conjecture that Theorem 4.4.2 holds more generally, provided the other assumptions are accordingly made to be conditional on X_0 also. In

any case, this result shows that general purpose bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ are not forthcoming, and Theorem 4.4.1 is the best we can hope for in general. We show below that this result is nevertheless powerful enough to obtain useful information in practice.

4.5 Falsification methodology

4.5.1 Hypotheses derived from causal bounds

We now use Theorem 4.4.1 to obtain a hypothesis testing procedure that can be used to falsify the twin, and that does not rely on any further assumptions than we have already provided. To this end, we first define the hypotheses $\mathcal{H}$ satisfying (4.3) that we will consider. Each of these will depend on a specific choice of the following parameters:

- A timestep $t \in \{1, \dots, T\}$
- A sequence of actions $a_{1:t} \in \mathcal{A}_{1:t}$
- A measurable function $f : \mathcal{X}_{0:t} \to \mathbb{R}$
- A sequence of subsets $B_{0:t} \subseteq \mathcal{X}_{0:t}$.

To streamline notation, in this section, we will consider these parameters to be fixed. However, we emphasize that our construction can be instantiated for many different choices of these parameters, and indeed we will do so in our case study below. We think of f as expressing a specific outcome of interest at time t in terms of the data we have assumed. Accordingly, for each $a'_{1:t} \in \mathcal{A}_{1:t}$, we define new potential outcomes $Y(a'_{1:t}) \coloneqq f(X_{0:t}(a'_{1:t}))$. For example, in a medical context, if $X_{0:t}(a'_{1:t})$ represents a full patient history at time t after treatments $a'_{1:t}$, then $Y(a'_{1:t})$ might represent the patient's heart rate after these treatments. Likewise, $B_{0:t}$ selects a subgroup of patients of interest, e.g. elderly patients whose blood pressure values were above some threshold at some timesteps before t.

The hypotheses we consider are based on the corresponding outcome produced by the twin when initialised at X_0, which we define for $a'_{1:t} \in \mathcal{A}_{1:t}$ as $\widehat{Y}(a'_{1:t}) \coloneqq f(X_0, \widehat{X}_{1:t}(X_0, a'_{1:t}))$. Supposing it holds that

$$\mathbb{P}(\widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}) > 0, \tag{4.10}$$

we may then define $\widehat{Q} \coloneqq \mathbb{E}[\widehat{Y}(a_{1:t}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]$, i.e. the analogue of Q for the twin. By Proposition 4.2.1, if the twin is interventionally correct, then $\widehat{Q} = Q$. Theorem

4.4.1 therefore implies that the following hypotheses have our desired property (4.3):

$$\mathcal{H}_{\mathrm{lo}}\text{: If (4.4), (4.5), and (4.10) hold, then } \widehat{Q} \geq Q_{\mathrm{lo}}$$
$$\mathcal{H}_{\mathrm{up}}\text{: If (4.4), (4.5), and (4.10) hold, then } \widehat{Q} \leq Q_{\mathrm{up}}.$$

Moreover, $\mathcal{H}_{\mathrm{lo}}$ and $\mathcal{H}_{\mathrm{up}}$ can in principle be determined to be true or false from the information we have assumed available, since Q_{lo} and Q_{up} depend only on the observational data, and $\widehat{Q}$ can be estimated by generating trajectories from the twin.

When either $\mathcal{H}_{\mathrm{lo}}$ or $\mathcal{H}_{\mathrm{up}}$ is falsified, it immediately follows that the twin is not interventionally correct. However, even more than this, a falsification describes a concrete failure mode with various potential implications downstream, which is considerably more useful information about the twin in practice. For example, if $\mathcal{H}_{\mathrm{lo}}$ is false (i.e. if $\widehat{Q} < Q_{\mathrm{lo}}$), it follows that, among those trajectories for which $(X_0, \widehat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$, the mean of $\widehat{Y}(a_{1:t})$ is too small. (Higher moments could also be considered by choosing f appropriately.) In light of this, a user might choose not to rely on outputs of the twin produced under these circumstances, while a developer seeking to improve the twin could focus their attention on the specific parts of its implementation that give rise to this behaviour. We illustrate this concretely through our case study in Section 4.6.

4.5.2 Exact testing procedure

We now describe a procedure for testing $\mathcal{H}_{\mathrm{lo}}$ and $\mathcal{H}_{\mathrm{up}}$ using a finite dataset that obtains exact control over type I error without relying on additional assumptions or asymptotic approximations. We show in our case study below that this procedure is nevertheless powerful enough to obtain useful information about a twin in practice.

We focus here on obtaining a p-value for $\mathcal{H}_{\mathrm{lo}}$ given a fixed choice of parameters $(t, f, a_{1:t}, B_{0:t})$. Our procedure for $\mathcal{H}_{\mathrm{up}}$ is symmetrical, or may be regarded as a special case of testing $\mathcal{H}_{\mathrm{lo}}$ by replacing f with $-f$. Multiple hypotheses may then be handled via standard techniques such as the method of Holm [1979] (which we use in our case study) or of Benjamini and Yekutieli [2001], both of which apply without additional assumptions.

As above, we assume access to a finite dataset $\mathcal{D}$ of i.i.d. copies of (4.1). We also assume access to a dataset $\widehat{\mathcal{D}}(a_{1:t})$ of i.i.d. copies of

$$X_0, \widehat{X}_1(X_0, a_1), \ldots, \widehat{X}_t(X_0, a_{1:t}). \tag{4.11}$$

In practice, these copies can be obtained by initialising the twin with some value X_0 taken from $\mathcal{D}$ without replacement and supplying inputs $a_{1:t}$. If each X_0 in $\mathcal{D}$ is used to initialize the twin at most once, then the resulting trajectories in $\widehat{\mathcal{D}}(a_{1:t})$ are guaranteed to be i.i.d., since we assumed in Section 4.2 that the potential outcomes $\widehat{X}_t(x_0, a_{1:t})$ produced by the twin are independent across runs. We adopt this approach in our case study.

Observe that $\mathcal{H}_{\text{lo}}$ is false only if (4.4), (4.5), and (4.10) all hold. We account for this in our testing procedure as follows. First, (4.5) immediately follows if

$$y_{\text{lo}} \leq f(x_{0:t}) \leq y_{\text{up}} \qquad \text{for all } x_{0:t} \in B_{0:t}. \tag{4.12}$$

This holds automatically in certain cases, such as for binary outcomes (e.g. patient survival), or otherwise can be enforced simply by clipping the value of f to live within $[y_{\text{lo}}, y_{\text{up}}]$. We describe a practical means for choosing f in this way in Section 4.6.

To account for (4.4) and (4.10), we simply check whether there exists some trajectory in $\mathcal{D}$ with $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and some trajectory in $\widehat{\mathcal{D}}(a_{1:t})$ with $(X_0, \widehat{X}_{1:t}(X_0, a_{1:t})) \in B_{0:t}$. If there are not, then we refuse to reject $\mathcal{H}_{\text{lo}}$ at any significance level; otherwise, we proceed to test $\widehat{Q} \geq Q_{\text{lo}}$ as described next. It easily follows that there is zero probability we will reject $\mathcal{H}_{\text{lo}}$ if (4.4) and (4.10) do not in fact hold, and so our overall type I error is controlled at the desired level.

To test $\widehat{Q} \geq Q_{\text{lo}}$, we begin by constructing a one-sided lower confidence interval for Q_{lo}, and a one-sided upper confidence interval for $\widehat{Q}$. In detail, for each significance level $\alpha \in (0, 1)$, we obtain R^{α}_{lo} and $\widehat{R}^{\alpha}$ as functions of $\mathcal{D}$ and $\widehat{\mathcal{D}}(a_{1:t})$ such that

$$\mathbb{P}(Q_{\text{lo}} \geq R^{\alpha}_{\text{lo}}) \geq 1 - \frac{\alpha}{2} \qquad \mathbb{P}(\widehat{Q} \leq \widehat{R}^{\alpha}) \geq 1 - \frac{\alpha}{2}. \tag{4.13}$$

We will also ensure that these are nested, i.e. $R^{\alpha}_{\text{lo}} \leq R^{\alpha'}_{\text{lo}}$ and $\widehat{R}^{\alpha'} \leq \widehat{R}^{\alpha}$ if $\alpha \leq \alpha'$. We describe two methods for obtaining R^{α}_{lo} and $\widehat{R}^{\alpha}$ satisfying these conditions below.

From these confidence intervals, we obtain a test for the hypothesis $\widehat{Q} \geq Q_{\text{lo}}$ that rejects when $\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}$. A straightforward argument given in Section C.8.1 of the Appendix shows that this controls the type I error at the desired level α. Nestedness also yields a p-value obtained as the smallest value of α for which this test rejects, i.e. $p_{\text{lo}} \coloneqq \inf\{\alpha \in (0, 1) \mid \widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}\}$, or 1 if the test does not reject at any level.

We now consider how to obtain confidence intervals for Q_{lo} and $\widehat{Q}$ satisfying the desired conditions above. To this end, observe that both quantities are (conditional) expectations

involving random variables that can be computed from $\mathcal{D}$ or $\widehat{\mathcal{D}}(a_{1:t})$. This allows both to be estimated unbiasedly, which in turn can be used to derive confidence intervals via standard techniques. For example, consider the subset of trajectories in $\widehat{\mathcal{D}}(a_{1:t})$ with $(X_0, \widehat{X}_t(X_0, a_{1:t})) \in B_{0:t}$. For each such trajectory, we obtain a corresponding value of $\widehat{Y}(a_{1:t})$ that is i.i.d. and has expectation $\widehat{Q}$. Similarly, for Q_{lo}, we extract the subset of trajectories in $\mathcal{D}$ for which $X_{0:N}(A_{1:N}) \in B_{0:N}$ holds. The values of Y_{lo} obtained from each such trajectory are then i.i.d. and have expectation Q_{lo}.

At this point, our problem now reduces to that of constructing a confidence interval for the expectation of a random variable using i.i.d. copies of it. Various techniques exist for this, and we consider two possibilities in our case study. The first leverages the fact that $\widehat{Y}(a_{1:t})$ and Y_{lo} are bounded in $[y_{\text{lo}}, y_{\text{up}}]$, which gives rise to R^{α}_{lo} and $\widehat{R}^{\alpha}$ via an application of Hoeffding's inequality. This approach has the appealing property that (4.13) holds exactly, although often at the expense of conservativeness. In practice, this could be mitigated by instead obtaining confidence intervals via (for example) the bootstrap [Efron, 1979], although at the expense of requiring (often mild) asymptotic assumptions. Section C.8.2 of the Appendix describes both methods in greater detail. Our empirical results reported in the next section all use Hoeffding's inequality are hence exact, but we also provide additional results using bootstrapping in Section C.9.7 of the Appendix.

4.6 Case Study: Pulse Physiology Engine

4.6.1 Experimental setup

We applied our assessment methodology to the Pulse Physiology Engine [Bray et al., 2019], an open-source model for human physiology simulation. Pulse simulates trajectories of various physiological metrics for patients with conditions like sepsis, COPD, and ARDS. We describe the main steps of our experimental procedure and results below, with full details given in the Section C.9 of the Appendix.

We utilized the MIMIC-III dataset [Johnson et al., 2016], a comprehensive collection of longitudinal health data from critical care patients at the Beth Israel Deaconess Medical Center (2001-2012). We focused on patients meeting the sepsis-3 criteria [Singer et al., 2016], following the methodology of Komorowski et al. [2018] for selecting these. This yielded 11,677 sepsis patient trajectories. We randomly selected 5% of these (583 trajectories,

denoted as $\mathcal{D}_0$) to use for choosing the parameters of our hypotheses via a sample splitting approach [Cox, 1975], with the remaining 95% (11,094 trajectories, denoted as $\mathcal{D}$) reserved for the actual testing.

We considered hourly observations of each patient over the the first four hours of their ICU stay, i.e. $T = 4$. We defined the observation spaces $\mathcal{X}_{0:T}$ using a total of 17 features included in our extracted MIMIC trajectories for this time period, including static demographic quantities and patient vitals. Following Komorowski et al. [2018], the actions we considered involved the administration of intravenous fluids and vasopressors, which both play a primary role in the treatment of sepsis in clinical practice. Since these are recorded in MIMIC as continuous doses, we discretised their values via the same procedure as Komorowski et al. [2018], obtaining finite action spaces $\mathcal{A}_1 = \cdots = \mathcal{A}_4$, each with 25 distinct actions.

We defined a collection of hypothesis parameters $(t, f, a_{1:t}, B_{0:t})$, each of which we then used to define an $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ to test. For this, we chose 14 different physiological quantities of interest to assess, including heart rate, skin temperature, and respiration rate (see Table C.3 in the Appendix for a complete list). For each of these, we selected combinations of t, $a_{1:t}$, and $B_{0:t}$ observed for at least one patient trajectory in $\mathcal{D}_0$. We took y_{lo} and y_{up} to be the .2 and .8 quantiles of the same physiological quantity as was recorded in $\mathcal{D}_0$, and defined f as the function that extracts this quantity from $\mathcal{X}_t$ and clips its value between y_{lo} and y_{up}, so that (4.12) holds. We describe this procedure in full in Section C.9.5 of the Appendix. We also investigated the sensitivity of our procedure to the choice of y_{lo} and y_{up} and found it to be relatively stable: see Section C.9.9 of the Appendix. We obtained 721 unique parameter choices, leading to 1,442 total hypotheses.

We generated data from Pulse to test the chosen hypotheses. For each $a_{1:t}$ occurring in any of our hypotheses, we obtained the dataset $\widehat{\mathcal{D}}(a_{1:t})$ as described in Section 4.5.2. Specifically, we sampled X_0 without replacement from $\mathcal{D}$, and used this to initialize a twin trajectory. Then, at each hour $t' \in \{1, \ldots, 4\}$ in the simulation, we administered a dose of intravenous fluids and vasopressors corresponding to $a_{t'}$ and recorded the resulting patient features generated by Pulse. We describe this procedure in full in Section C.9.6 in the Appendix. This produced a total of 26,115 simulated trajectories.

Physiological quantity	Ours		Manski	
	Rejs.	Hyps.	Rejs.	Hyps.
Chloride Blood Concentration (Chloride)	24	94	1	46
Sodium Blood Concentration (Sodium)	21	94	9	46
Potassium Blood Concentration (Potassium)	13	94	0	46
Skin Temperature (Temp)	10	86	9	46
Calcium Blood Concentration (Calcium)	5	88	0	46
Glucose Blood Concentration (Glucose)	5	96	1	46
Arterial CO_2 Pressure ($paCO_2$)	3	70	0	46
Bicarbonate Blood Concentration (HCO_3)	2	90	1	46
Systolic Arterial Pressure (SysBP)	2	154	0	46

Table 4.1: Total hypotheses (Hyps.) and rejections (Rejs.) per physiological quantity using our causal bounds, as well as those of Manski [1990]

4.6.2 Hypothesis rejections

We tested the hypotheses just described using our methodology from Section 4.5.2. Here we report the results when using Hoeffding's inequality to obtain confidence intervals for Q_{lo}, Q_{up}, and $\widehat{Q}$. We also tried confidence intervals obtained via bootstrapping, and obtained similar if less conservative results (see Section C.9 of the Appendix). We used the Holm-Bonferroni method to adjust for multiple tests, with family-wise error rate of 0.05.

We obtained rejections for hypotheses corresponding to 10 different physiological quantities shown in Table 4.1. (Table C.3 in the Appendix shows all hypotheses we tested, including those not rejected.) We may therefore infer that, at a high level, Pulse does not simulate these quantities accurately for the population of sepsis patients we consider. This appears of interest in a variety of downstream settings: for example, a developer could use this information when considering how to improve the accuracy of Pulse, while a practitioner using Pulse may wish to rely less on these outputs as a result.

To assess the relative performance of our bounds from Theorem 4.4.1 compared with the unconditional bounds of Manski [1990], we also reran this analysis with each t, f, and $a_{1:t}$ chosen as before, but with each $B_{0:t}$ now set to the whole space $\mathcal{X}_{0:t}$. This in turn led to fewer hypotheses, namely 690 in total, which were evenly divided between hypotheses of the form $\mathbb{E}[\widehat{Y}(a_{1:t})] \geq \mathbb{E}[Y_{\text{lo}}]$ and those of the form $\mathbb{E}[\widehat{Y}(a_{1:t})] \leq \mathbb{E}[Y_{\text{up}}]$. We kept all other aspects of our procedure the same as just described, including our method for obtaining confidence intervals and controlling for multiple testing. The number of rejections that we obtained in this case is also shown in Table 4.1. As anticipated by our discussion in Section 4.4.2, the use of Manski's bounds led to considerably fewer

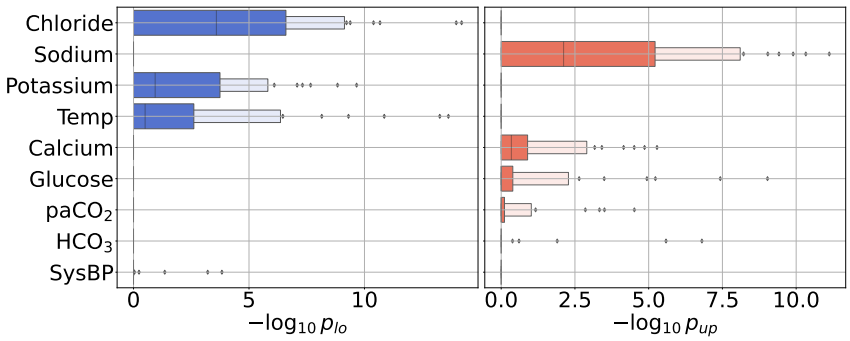


Figure 4.1: Distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ across hypotheses, grouped by physiological quantity. Higher values indicate greater evidence in favour of rejection.

rejections than our more general result given in Theorem 4.4.1, even when considered as a proportion of the total hypotheses we tested.

4.6.3 p-value plots

To obtain more granular information about the failure modes of the twin just identified, we examined the p-values obtained for each hypothesis $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ tested using our causal bounds, which we denote here by p_{lo} and p_{up}. Figure 4.1 shows the distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ that we obtained for all physiological quantities for which some hypothesis was rejected. (The remaining p-values are shown in Figure C.3 in the Appendix.) Notably, in each row, one distribution is always tightly concentrated at $-\log_{10} p = 0$ (i.e. $p = 1$). This means that, for all physiological outcomes of interest, there was either very little evidence in favour of rejecting any $\mathcal{H}_{\text{lo}}$, or very little in favour of rejecting any $\mathcal{H}_{\text{up}}$. In other words, across configurations of $(t, f, a_{1:t}, B_{0:t})$ that were rejected, the twin consistently either underestimated or overestimated each quantity on average. For example, Pulse consistently underestimated chloride blood concentration and skin temperature, while it consistently overestimated sodium and glucose blood concentration levels. Like Table 4.1, this information appears of interest and actionable in a variety of downstream tasks.

4.6.4 Pitfalls of naive assessment

A naive approach to twin assessment involves simply comparing the output of the twin with the observational data directly, without accounting for causal considerations. We now show that, unlike our methodology, the results produced in this way can be potentially

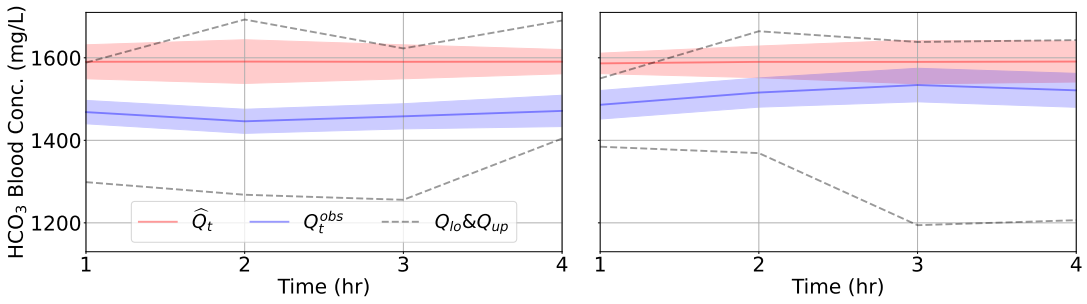


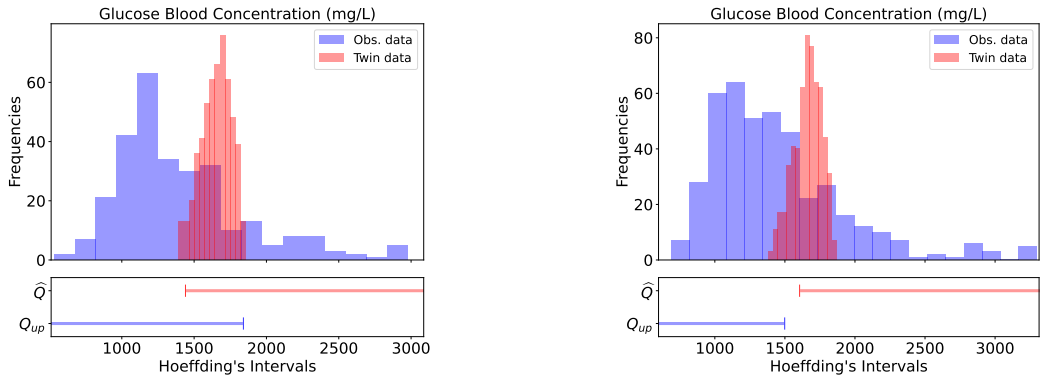
Figure 4.2: Estimates and 95% confidence intervals for $\widehat{Q}_t$ and Q_t^{obs} at each $1 \leq t \leq 4$ for two choices of $(B_{0:4}, a_{1:4})$, where $\widehat{Y}(a_{1:t})$ and $Y(a_{1:t})$ correspond to HCO_3 concentration. The dashed lines indicate lower and upper 95% confidence intervals for $Q_{\text{lo}}, Q_{\text{up}}$ respectively.

misleading. In Figure 4.2, for two different choices of $(a_{1:4}, B_{1:4})$, we plot estimates of $\widehat{Q}_t$ and Q_t^{obs} for $t \in \{1, \ldots, 4\}$, where

$$\widehat{Q}_t := \mathbb{E}[\widehat{Y}(a_{1:t}) \mid X_0 \in B_0, \widehat{X}_{1:t}(X_0, a_{1:t}) \in B_{1:t}]$$
$$Q_t^{\text{obs}} := \mathbb{E}[Y(A_{1:t}) \mid X_{0:t}(A_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}].$$

Here $\widehat{Q}_t$ is just $\widehat{Q}$ as defined above with its dependence on t made explicit. Each plot also shows one-sided 95% confidence intervals on Q_{lo} and Q_{up} at each $t \in \{1, \ldots, 4\}$ obtained from Hoeffding's inequality. Directly comparing the estimates of $\widehat{Q}_t$ and Q_t^{obs} would suggest that the twin is comparatively more accurate for the right-hand plot, as these estimates are closer to one another in that case. However, the output of the twin in the right-hand plot is falsified at $t = 1$, as can be seen from the fact that confidence interval for $\widehat{Q}_1$ lies entirely above the one-sided confidence interval for Q_{up} at that timestep. On the other hand, the output of the twin in the left-hand plot is not falsified at any of the timesteps shown, so that the twin may in fact be accurate for these $(a_{1:4}, B_{1:4})$, contrary to what a naive assessment strategy would suggest. Our methodology provides a principled means for twin assessment that avoids drawing potentially misleading inferences like this.

A similar phenomenon appears in Figure 4.3, which for two choices of $B_{0:t}$ and $a_{1:t}$ shows histograms of raw glucose values obtained from the observational data conditional on $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and from the twin conditional on $\widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}$. Below each histogram we also show 95% confidence intervals for Q_{up} and $\widehat{Q}$ obtained from Hoeffding's inequality. While Figures 4.3a and 4.3b appear visually very similar, the inferences produced by our testing procedure are different: the hypothesis corresponding to the right-hand plot is rejected, since there is no overlap between the confidence intervals



(a) Not rejected **(b)** Rejected

Figure 4.3: Raw glucose values from the observational data and twin for two choices of $(B_{0:t}, a_{1:t})$, with confidence intervals for $\widehat{Q}$ and Q_{up} shown below. The horizontal axes are truncated to the .025 and .975 quantiles of the observational data for clarity. Untruncated plots are shown in Figure C.4 of the Appendix.

underneath, while the hypothesis corresponding to the left-hand plot is not. This was not an isolated case and several other examples of this phenomenon are shown in Figure C.4 in the Appendix. This demonstrates that the inferences obtained from our procedure do not depend only on the distribution of observed outcomes (which is essentially the same for both cases). Instead, as discussed in Section 4.4.1, these also account for the worst-case effects of unmeasured confounding that may exist in the observational data.

4.7 Discussion

We have advocated for a causal approach to digital twin assessment, and have presented a statistical procedure for doing so that obtains rigorous theoretical guarantees under minimal assumptions. We now highlight the key limitations of our approach. Importantly, our methodology implicitly assumes that there is no distribution shift between testing and deployment time. If the conditional distribution of $X_{1:T}(a_{1:T})$ given X_0 changes at deployment time, then so too does the set of twins that are interventionally correct, and if this change is significant enough, our assessment procedure may yield misleading results. Distribution shift in this sense is a separate issue to unobserved confounding, and arises in a wide variety of statistical problems beyond ours.

Additionally, the procedure we used in our case study to choose the hypothesis parameters $B_{0:t}$ was ad hoc. For scalability, it would likely be necessary to obtain $B_{0:t}$ via a more automated procedure. It may also be desirable to choose $B_{0:t}$ dynamically in

light of previous hypotheses tested, zooming in to regions containing possible failure modes to obtain increasingly granular information about the twin. We see opportunities here for using machine learning techniques, but leave this to future work.

Various other extensions and improvements appear possible. For example, it is possible to replace our one-sided confidence intervals for Q_{lo}, Q_{up}, and $\widehat{Q}$ with two-sided ones, and thereby to obtain a procedure that may yield more precise information about the twin than we obtain by rejecting one of $\mathcal{H}_{\text{lo}}$ or $\mathcal{H}_{\text{up}}$. One can also leverage ideas from the literature on partial identification [Manski, 2003] to obtain greater statistical efficiency, for example by building on the line of work initiated by Imbens and Manski [2004] for obtaining more informative confidence intervals. Beyond this, it may sometimes be useful to consider additional assumptions that lead to less conservative assessment results. For example, various methods for *sensitivity analysis* have been proposed that model the *degree* to which the actions of the behavioural agent are confounded [Rosenbaum, 2002, Tan, 2006, Yadlowsky et al., 2022]. This can yield tighter bounds on Q than are implied by Theorem 4.4.1, albeit at the expense of less robustness if these assumptions are violated.

Acknowledgements

The authors are highly appreciative of the troubleshooting and development assistance provided by the Pulse team. RC, AD, and CH were supported by the Engineering and Physical Sciences Research Council (EPSRC) through the Bayes4Health programme [Grant number EP/R018561/1]. MFT was funded by Google DeepMind. The authors declare there are no competing interests.

5 Conclusion and Future Work

Contents

5.1 Discussion

Before deploying a decision-making policy to production, it is usually important to understand the plausible range of outcomes that it may produce. However, due to resource or ethical constraints, it is often not possible to obtain this understanding by testing the policy directly in the real-world. In such cases we have to rely on off-policy evaluation (OPE), which uses observational data collected under a different behavioural policy to evaluate the target policy in some way. In this thesis, we have considered the different challenges posed by the current OPE methodologies and proposed novel solutions to each of these individually. We have also demonstrated the practical utility of these solutions by applying them to a range of real-world problems involving large scale datasets. To be more specific, below we provide a brief summary of the challenges tackled in this thesis:

- In Chapter 2 we consider the problem of variance reduction in OPE estimators. To address this, we propose the Marginal Ratio (MR) estimator, which uses a marginalization technique to provide a more efficient and robust estimator for contextual bandits, and may also be of interest in other domains such as causal inference.

- Next, in Chapter 3 we provide a novel methodology of uncertainty quantification in off-policy outcomes based on conformal prediction [Vovk et al., 2005]. Our proposed technique can help practitioners quantify the plausible range of outcomes that are likely to occur under the target policy, and comes with sound finite-sample probabilistic guarantees.

- Finally, in Chapter 4 we explore the case when the assumption of no unmeasured confounding (needed for the existing OPE methodologies) is violated. We provide a set of novel causal bounds which remain valid in this case, and subsequently use these bounds to develop a procedure for robust assessment of digital twin models using observational data which remains valid under only the assumption of an i.i.d. dataset of observational trajectories.

5.2 Limitations

Here, we outline some of the limitations of the methodologies described in this thesis.

Distributional shift in data generating mechanism Our methodologies highlighted in this thesis implicitly assume that the data generating process remains unchanged between testing and deployment times. Technically, for contextual bandits this assumption means that the conditional distribution of Y given (X, A) does not shift when the target policy is deployed. If this distribution changes at deployment time, then so too does the distribution of outcomes that *would* be observed under the target policy. If this shift is significant enough our methodologies may yield misleading results.

Estimation errors The techniques mentioned in Chapters 2 and 3 involve an additional estimation of marginal density ratios for importance sampling. While we outline straightforward regression methodologies for estimating these importance ratios directly, this additional step may still introduce an additional source of bias in the value estimation.

Scalability limitations Increasing data dimensionality may pose additional challenges for our solutions, especially those presented in Chapter 3 and 4. For example, in Chapter 3, the estimation of importance ratios for our conformal off-policy prediction (COPP) algorithm may become more challenging when (X, A) is high dimensional, thereby yielding biased results. Likewise in Chapter 4, the procedure we used in our case study to choose the hypothesis parameters was ad hoc, which may not scale to high dimensional datasets. For scalability, it would likely be necessary to obtain these parameters via a more automated procedure.

5.3 Directions for future work

Our work in this thesis opens up several interesting avenues for future research. We highlight some of these below.

Off-policy learning This thesis has largely focused on robust off-policy assessment methodologies. However, our findings are highly applicable to robust policy optimisation problems, where the data collected using an 'old' policy is used to learn a new policy. Proximal Policy Optimisation (PPO) [Schulman et al., 2017] is one such approach which has gained immense popularity and has been applied to reinforcement learning with human feedback (RLHF) [Lambert et al., 2022]. We believe that our MR estimator proposed in Chapter 2 applied to these methodologies could lead to improvements in the stability and convergence of these optimisation schemes, given its favourable variance properties. Similarly, our conformal off-policy prediction (COPP) algorithm when applied to off-policy learning could be a step towards robust policy learning by optimising the worst case outcome [Stutz et al., 2022].

Addressing the curse of horizon in sequential decision-making Chapters 2 and 3 of this thesis specifically consider OPE in contextual bandits. This setting offers a strong foundational framework for conducting rigorous theoretical and empirical analyses, however, it would be interesting to extend the application of these methodologies to sequential decision frameworks. While some follow-up works have attempted to apply our COPP algorithm to Markov Decision Processes [Foffano et al., 2023, Zhang et al., 2023, Kuipers

et al., 2024], the obtained confidence sets become increasingly conservative with increasing time horizon. It is worth exploring methodologies for obtaining intervals which remain valid and informative even in sequential decision settings with large time horizons.

Application to transfer learning Finally, our solutions in Chapters 2 and 3 involves learning importance ratios which may also be of interest in other domains beyond OPE. One such area is *transfer learning* which considers cases where the testing data distribution is different from the training data distribution. Classical transfer learning methods rely on importance weighting to handle the distribution mismatch [Shimodaira, 2000, Sugiyama et al., 2007, Huang et al., 2007, Sugiyama et al., 2008, Lu et al., 2021]. Our proposed regression techniques may be of interest for obtaining these weights efficiently in high-dimensional datasets in this setting.

Appendices

A

Marginal Density Ratio for Off-Policy Evaluation in Contextual Bandits

Contents

A.1 Proofs

Proof of Lemma 2.3.1. First, we express the weights $w(y)$ as the conditional expectation as follows:

$$\begin{aligned}
w(y) &= \frac{p_{\pi^*}(y)}{p_{\pi^b}(y)} \\
&= \int_{\mathcal{X},\mathcal{A}} \frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(y)} \,\mathrm{d}a\,\mathrm{d}x \\
&= \int_{\mathcal{X},\mathcal{A}} \frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(y)} \frac{p_{\pi^b}(x,a \mid y)}{p_{\pi^b}(x,a \mid y)} \,\mathrm{d}a\,\mathrm{d}x \\
&= \int_{\mathcal{X},\mathcal{A}} \frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(x,a,y)} p_{\pi^b}(x,a \mid y) \,\mathrm{d}a\,\mathrm{d}x \\
&= \int_{\mathcal{X},\mathcal{A}} \rho(a,x)\, p_{\pi^b}(x,a \mid y) \,\mathrm{d}a\,\mathrm{d}x \\
&= \mathbb{E}_{\pi^b}[\rho(A,X) \mid Y = y],
\end{aligned}$$

where $\rho(a,x) = \frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(x,a,y)} = \frac{\pi^*(a|x)}{\pi^b(a|x)}$. Since conditional expectations can be defined as the solution of regression problem, the result follows. □

Proof of Proposition 2.3.1. We have

$$\begin{aligned}
\mathrm{D}_f\left(p_{\pi^*}(x,a,y) \,||\, p_{\pi^b}(x,a,y)\right) &= \mathbb{E}_{\pi^b}\left[f\left(\frac{p_{\pi^*}(X,A,Y)}{p_{\pi^b}(X,A,Y)}\right)\right] \\
&= \mathbb{E}_{\pi^b}\left[f\left(\frac{\pi^*(A \mid X)}{\pi^b(A \mid X)}\right)\right] \\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[f\left(\frac{\pi^*(A \mid X)}{\pi^b(A \mid X)}\right)\middle| Y\right]\right] \\
&\geq \mathbb{E}_{\pi^b}\left[f\left(\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A \mid X)}{\pi^b(A \mid X)}\middle| Y\right]\right)\right] \quad \text{(Jensen's inequality)} \\
&= \mathbb{E}_{\pi^b}\left[f\left(\frac{p_{\pi^*}(Y)}{p_{\pi^b}(Y)}\right)\right] \\
&= \mathrm{D}_f\left(p_{\pi^*}(y) \,||\, p_{\pi^b}(y)\right).
\end{aligned}$$

□

Proof of Proposition 2.3.2. Since $\mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] = \mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] = \mathbb{E}_{\pi^*}[Y]$, we have that,

$$\begin{aligned}
\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] &= \mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}]^2 - \mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}]^2 \\
&= \frac{1}{n}\left(\mathbb{E}_{\pi^b}\left[\rho(A,X)^2\, Y^2\right] - \mathbb{E}_{\pi^b}\left[w(Y)^2\, Y^2\right]\right) \\
&= \frac{1}{n}\left(\mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}[\rho(A,X)^2 \mid Y]\, Y^2\right] - \mathbb{E}_{\pi^b}\left[w(Y)^2\, Y^2\right]\right) \\
&= \frac{1}{n}\left(\mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}[\rho(A,X)^2 \mid Y]\, Y^2\right] - \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}[\rho(A,X) \mid Y]^2\, Y^2\right]\right) \\
&= \frac{1}{n}\mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\rho(A,X) \mid Y\right]\, Y^2\right].
\end{aligned}$$

In the second last step above, we use the fact that $w(y) = \mathbb{E}_{\pi^b}[\rho(A,X) \mid Y = y]$. □

Proof of Proposition 2.3.3. Let $\hat{\mu}(a,x) \approx \mathbb{E}[Y \mid X = x, A = a]$ denote the outcome model in DR estimator. Then, using multiple applications of the law of total variance we get that

$$\begin{aligned}
n\,\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{DR}}] &= \mathrm{Var}_{\pi^b}\left[\rho(A,X)\,(Y - \hat{\mu}(A,X)) + \sum_{a' \in \mathcal{A}} \hat{\mu}(a',X)\,\pi^*(a' \mid X)\right] \\
&= \mathrm{Var}_{\pi^b}\left[\rho(A,X)\,(Y - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^*}[\hat{\mu}(A,X) \mid X]\right] \\
&= \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\rho(A,X)\,(Y - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^*}[\hat{\mu}(A,X) \mid X] \mid X, A]] \\
&\quad + \mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,(Y - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^*}[\hat{\mu}(A,X) \mid X] \mid X, A]] \\
&= \mathbb{E}_{\pi^b}[\rho(A,X)^2 \mathrm{Var}[Y \mid X, A]] \\
&\quad + \mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,(Y - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^b}[\rho(A,X)\,\hat{\mu}(A,X) \mid X] \mid X, A]] \\
&= \mathbb{E}_{\pi^b}[\rho(A,X)^2 \mathrm{Var}[Y \mid X, A]] \\
&\quad + \mathrm{Var}_{\pi^b}[\rho(A,X)\,(\mu(A,X) - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^b}[\rho(A,X)\,\hat{\mu}(A,X) \mid X]] \\
&= \mathbb{E}_{\pi^b}[\rho(A,X)^2 \mathrm{Var}[Y \mid X, A]] \\
&\quad + \mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,(\mu(A,X) - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^b}[\rho(A,X)\,\hat{\mu}(A,X) \mid X] \mid X]] \\
&\quad + \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\rho(A,X)\,(\mu(A,X) - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^b}[\rho(A,X)\,\hat{\mu}(A,X) \mid X] \mid X]] \\
&= \mathbb{E}_{\pi^b}[\rho(A,X)^2 \mathrm{Var}[Y \mid X, A]] + \mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]] \\
&\quad + \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\rho(A,X)\,(\mu(A,X) - \hat{\mu}(A,X)) \mid X]] \\
&\geq \mathbb{E}_{\pi^b}[\rho(A,X)^2 \mathrm{Var}[Y \mid X, A]] + \mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]].
\end{aligned}$$

Using this, we get that

$$\begin{aligned}
&n(\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{DR}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}]) \\
&\quad \geq \mathbb{E}_{\pi^b}[\rho(A,X)^2 \mathrm{Var}[Y \mid X, A]] + \mathrm{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]\right] - \mathrm{Var}_{\pi^b}[w(Y)\,Y].
\end{aligned}$$

Again, using the law of total variance,

$$\begin{aligned}\mathrm{Var}_{\pi^b}[\rho(A,X)\,Y] &= \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\rho(A,X)\,Y \mid X,A]] + \mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,Y \mid X,A]] \\ &= \mathbb{E}_{\pi^b}[\rho(A,X)^2\mathrm{Var}[Y \mid X,A]] + \mathrm{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X)] \\ &= \mathbb{E}_{\pi^b}[\rho(A,X)^2\mathrm{Var}[Y \mid X,A]] + \mathrm{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]\right] \\ &\quad + \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]\right].\end{aligned}$$

Rearranging and substituting back into the expression earlier, we get that

$$\begin{aligned}&n(\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{DR}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}]) \\ &\quad \geq \mathrm{Var}_{\pi^b}[\rho(A,X)\,Y] - \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]\right] - \mathrm{Var}_{\pi^b}[w(Y)\,Y].\end{aligned}$$

Now, from Proposition 2.3.2 we know that

$$n(\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}]) = \mathrm{Var}_{\pi^b}[\rho(A,X)\,Y] - \mathrm{Var}_{\pi^b}[w(Y)\,Y] = \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\rho(A,X) \mid Y\right]\,Y^2\right].$$

Therefore,

$$\begin{aligned}&n(\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{DR}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}]) \\ &\quad \geq \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\rho(A,X) \mid Y\right]\,Y^2\right] - \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]\right] \\ &\quad = \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\rho(A,X)\,Y \mid Y\right] - \mathrm{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]\right].\end{aligned}$$

□

Proof of Theorem 2.3.2. This result follows straightforwardly from Proposition A.4.2 in Appendix A.4. □

Proof of Proposition 2.3.4.

$$\begin{aligned}\mathrm{Bias}(\hat{\theta}_{\mathrm{IPW}}) &= \mathbb{E}_{\pi^b}[\hat{\rho}(A,X)\,Y] - \mathbb{E}_{\pi^*}[Y] \\ &= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}[\hat{\rho}(A,X) \mid Y]\,Y\right] - \mathbb{E}_{\pi^*}[Y] \\ &= \mathbb{E}_{\pi^b}[\hat{w}(Y)\,Y] - \mathbb{E}_{\pi^b}[\epsilon\,Y] - \mathbb{E}_{\pi^*}[Y] \\ &= \mathrm{Bias}(\hat{\theta}_{\mathrm{MR}}) - \mathbb{E}_{\pi^b}[\epsilon\,Y].\end{aligned}$$

Next, to prove the variance result, we first use the law of total variance to obtain

$$\begin{aligned}\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] &= \frac{1}{n}\mathrm{Var}_{\pi^b}[\hat{\rho}(A,X)\,Y] \\ &= \frac{1}{n}\left(\mathrm{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\hat{\rho}(A,X)\,Y \mid Y]] + \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\hat{\rho}(A,X)\,Y \mid Y]]\right) \\ &= \frac{1}{n}\left(\mathrm{Var}_{\pi^b}[\tilde{w}(Y)\,Y] + \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\hat{\rho}(A,X)\,Y \mid Y]]\right).\end{aligned}$$

Moreover, using the fact that $\hat{w}(Y) = \tilde{w}(Y) + \epsilon$ we get that,

$$\begin{aligned}\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] &= \frac{1}{n}\mathrm{Var}_{\pi^b}[\hat{w}(Y)\,Y] \\ &= \frac{1}{n}\mathrm{Var}_{\pi^b}[(\tilde{w}(Y)+\epsilon)\;Y] \\ &= \frac{1}{n}\left(\mathrm{Var}_{\pi^b}[\tilde{w}(Y)\,Y] + \mathrm{Var}_{\pi^b}[\epsilon\,Y] + 2\,\mathrm{Cov}(\tilde{w}(Y)\,Y, \epsilon\,Y)\right).\end{aligned}$$

Putting together the two variance expressions derived above, we get that

$$\begin{aligned}&\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] \\ &\quad= \frac{1}{n}\Big(\mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\hat{\rho}(A,X) \mid Y]\,Y^2] - \mathrm{Var}_{\pi^b}[\epsilon\,Y] - 2\,\mathrm{Cov}(\tilde{w}(Y)\,Y, \epsilon\,Y)\Big).\end{aligned}$$

□

A.2 Comparison with extensions of the doubly robust estimator

In this section, we theoretically investigate the variance of MR against the commonly used extensions of the DR estimator, namely Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020]. At a high level, these estimators seek to reduce the variance of the vanilla DR estimator by considering modified importance weights, thereby trading off the variance for additional bias. Below, we provide the explicit definitions of these estimators for completeness.

Switch-DR estimator The original DR estimator can still have a high variance when the importance weights are large due to a large policy shift. Switch-DR [Wang et al., 2017b] aims to circumvent this problem by switching to DM when the importance weights are large:

$$\hat{\theta}_{\mathrm{SwitchDR}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho(a_i,x_i)\,(y_i - \hat{\mu}(a_i,x_i))\mathbb{1}(\rho(a_i,x_i) \le \tau) + \hat{\eta}(\pi^*),$$

where $\tau \geq 0$ is a hyperparameter, $\hat{\mu}(a, x) \approx \mathbb{E}[Y \mid X = x, A = a]$ is the outcome model, and

$$\hat{\eta}(\pi^*) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a' \in \mathcal{A}} \hat{\mu}(a', x_i) \pi^*(a' \mid x_i) \approx \mathbb{E}_{\pi^*}[\hat{\mu}(A, X)]$$

where $a_i^* \sim \pi^*(\cdot \mid x_i)$.

Doubly Robust with Optimal Shrinkage (DRos) DRos proposed by [Su et al., 2020] uses new weights $\hat{\rho}_\lambda(a_i, x_i)$ which directly minimises sharp bounds on the MSE of the resulting estimator,

$$\hat{\theta}_{\text{DRos}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_\lambda(a_i, x_i) \left(y_i - \hat{\mu}(a_i, x_i)\right) + \hat{\eta}(\pi^*),$$

where $\lambda \geq 0$ is a pre-defined hyperparameter and $\hat{\rho}_\lambda$ is defined as

$$\hat{\rho}_\lambda(a, x) \coloneqq \frac{\lambda}{\rho^2(a, x) + \lambda} \rho(a, x).$$

When $\lambda = 0$, $\hat{\rho}_\lambda(a, x) = 0$ leads to DM, whereas as $\lambda \to \infty$, $\hat{\rho}_\lambda(a, x) \to \rho(a, x)$ leading to DR.

More generally, both of these estimators can be written as follows:

$$\hat{\theta}_{\text{DR}}^{\tilde{\rho}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \tilde{\rho}(a_i, x_i) \left(y_i - \hat{\mu}(a_i, x_i)\right) + \hat{\eta}(\pi^*).$$

Here, when $\tilde{\rho}(a, x) = \rho(a, x) \mathbb{1}(\rho(a_i, x_i) \leq \tau)$, we recover the Switch-DR estimator and likewise when $\tilde{\rho}(a, x) = \hat{\rho}_\lambda(a, x)$, we recover DRos.

A.2.1 Variance comparison with the DR extensions

Next, we provide a theoretical result comparing the variance of the MR estimator with these DR extension methods.

Proposition A.2.1

When the weights $w(y)$ are known exactly and the outcome model is exact, i.e., $\hat{\mu}(a, x) = \mu(a, x) = \mathbb{E}[Y \mid X = x, A = a]$ in the DR estimator $\hat{\theta}_{\text{DR}}^{\tilde{\rho}}$ defined above,

$$\begin{aligned} &\text{Var}_{\pi^b}[\hat{\theta}_{\text{DR}}^{\tilde{\rho}}] - \text{Var}_{\pi^b}[\hat{\theta}_{\text{MR}}] \\ &\qquad \geq \frac{1}{n} \mathbb{E}_{\pi^b} \left[\text{Var}_{\pi^b} \left[\rho(A, X) \mid Y \right] Y^2 - \text{Var}_{\pi^b} \left[\rho(A, X) \mu(A, X) \mid X \right] \right] - \Delta, \end{aligned}$$

where $\Delta \coloneqq \frac{1}{n} \mathbb{E}_{\pi^b} \left[(\rho^2(A, X) - \tilde{\rho}^2(A, X)) \, \text{Var}[Y \mid X, A] \right]$.

Proof of Proposition A.2.1. Using the fact that $\hat{\mu}(a,x) = \mu(a,x)$ and the law of total variance, we get that

$$\begin{aligned}
n \operatorname{Var}_{\pi^b}[\hat{\theta}^{\tilde{\rho}}_{\mathrm{DR}}] &= \operatorname{Var}_{\pi^b}[\tilde{\rho}(A,X)\,(Y - \hat{\mu}(A,X)) + \sum_{a' \in \mathcal{A}} \hat{\mu}(a',X)\pi^*(a' \mid X)] \\
&= \operatorname{Var}_{\pi^b}[\tilde{\rho}(A,X)\,(Y - \hat{\mu}(A,X)) + \mathbb{E}_{\pi^*}[\hat{\mu}(A,X) \mid X]] \\
&= \operatorname{Var}_{\pi^b}[\tilde{\rho}(A,X)\,(Y - \mu(A,X)) + \mathbb{E}_{\pi^*}[\mu(A,X) \mid X]] \\
&= \operatorname{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\tilde{\rho}(A,X)\,(Y - \mu(A,X)) + \mathbb{E}_{\pi^*}[\mu(A,X) \mid X] \mid X, A]] \\
&\quad + \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\tilde{\rho}(A,X)\,(Y - \mu(A,X)) + \mathbb{E}_{\pi^*}[\mu(A,X) \mid X] \mid X, A]] \\
&= \operatorname{Var}_{\pi^b}[\mathbb{E}_{\pi^*}[\mu(A,X) \mid X]] + \mathbb{E}_{\pi^b}[\tilde{\rho}^2(A,X)\operatorname{Var}[Y \mid X, A]] \\
&= \operatorname{Var}_{\pi^b}[\mathbb{E}_{\pi^*}[\mu(A,X) \mid X]] + \mathbb{E}_{\pi^b}[\rho^2(A,X)\operatorname{Var}[Y \mid X, A]] \\
&\quad + \underbrace{\mathbb{E}_{\pi^b}[(\tilde{\rho}^2(A,X) - \rho^2(A,X))\operatorname{Var}[Y \mid X, A]]}_{-n\,\Delta} \\
&= \operatorname{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]] + \mathbb{E}_{\pi^b}[\rho^2(A,X)\operatorname{Var}[Y \mid X, A]] - n\,\Delta.
\end{aligned}$$

Again, using the law of total variance we can rewrite the second term on the RHS above as,

$$\begin{aligned}
&\mathbb{E}_{\pi^b}[\rho^2(A,X)\operatorname{Var}[Y \mid X, A]] \\
&\quad = \operatorname{Var}_{\pi^b}[\rho(A,X)\,Y] - \operatorname{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X)] \\
&\quad = \operatorname{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X) \mid Y]\,Y] + \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\rho(A,X) \mid Y]\,Y^2] \\
&\qquad - \operatorname{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X)] \\
&\quad = \operatorname{Var}_{\pi^b}[w(Y)\,Y] + \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\rho(A,X) \mid Y]\,Y^2] - \operatorname{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X)] \\
&\quad = n \operatorname{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] + \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\rho(A,X) \mid Y]\,Y^2] - \operatorname{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X)].
\end{aligned}$$

Putting this together, we get that

$$\begin{aligned}
&n \operatorname{Var}_{\pi^b}[\hat{\theta}^{\tilde{\rho}}_{\mathrm{DR}}] \\
&\quad = n \operatorname{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] + \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\rho(A,X) \mid Y]\,Y^2] - \operatorname{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X)] \\
&\qquad + \operatorname{Var}_{\pi^b}[\mathbb{E}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]] - n\,\Delta \\
&\quad = n \operatorname{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MR}}] + \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\rho(A,X) \mid Y]\,Y^2] - \mathbb{E}_{\pi^b}[\operatorname{Var}_{\pi^b}[\rho(A,X)\,\mu(A,X) \mid X]] - n\,\Delta,
\end{aligned}$$

where in the last step above, we again use the law of total variance. Rearranging the above leads us to the result. □

Intuition Note that for both of the DR extensions under consideration, the modified ratios $\tilde{\rho}(a, x)$ satisfy $0 \leq \tilde{\rho}(a, x) \leq \rho(a, x)$ and hence $\Delta \geq 0$ (using the definition of Δ in Proposition A.2.1). When the modified ratios $\tilde{\rho}(a, x)$ are 'close' to the true policy ratios $\rho(a, x)$, then using the definition of Δ, we have that $\Delta \approx 0$. In this case, the result above provides a similar intuition to Proposition 2.3.3 in the main text. Specifically, in this case we have that if $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, Y \mid Y\right]$ is greater than $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, \mu(A, X) \mid X\right]$ on average, the variance of the MR estimator will be less than that of the DR extension under consideration. Intuitively, this will occur when the dimension of context space $\mathcal{X}$ is high because in this case the conditional variance over X and A, $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, Y \mid Y\right]$ is likely to be greater than the conditional variance over A, $\mathrm{Var}_{\pi^b}\left[\rho(A, X)\, \mu(A, X) \mid X\right]$.

In contrast if the modified ratios $\tilde{\rho}(a, x)$ differ substantially from $\rho(a, x)$, then Δ will be large and the variance of MR may be higher than that of the resulting DR extension. However, this comes at the cost of significantly higher bias in the DR extension and consequently MSE of the DR extension will be high in this case.

A.3 Weight estimation error

In this section, we theoretically investigate the effects of using the estimated importance weights $\hat{w}(y)$ rather than $\hat{\rho}(a, x)$ on the bias and variance of the resulting OPE estimator. Further to our discussion in Section 2.3.1, we focus in this section on the approximation error when using a wide neural network to estimate the weights $\hat{w}(y)$. To this end, we use recent results regarding the generalization of wide neural networks [Lai et al., 2023] to show that the estimation error of the approximation step (ii) in the Section 2.3.1 declines with increasing number of training data when $\hat{w}(y)$ is estimated using wide neural networks. Before providing the main result, we explicitly lay out the assumptions needed.

A.3.1 Using wide neural networks to approximate the weights $\hat{w}(y)$

Assumption A.3.1. *Let $\tilde{w}(y) \coloneqq \mathbb{E}_{\pi^b}[\hat{\rho}(A, X) \mid Y = y]$. Suppose $\tilde{w} \in \mathcal{H}_1$ and $||\tilde{w}||_{\mathcal{H}_1} \leq R$ for some constant R, where $\mathcal{H}_1$ is the reproducing kernel Hilbert space (RKHS) associated with the Neural Tangent Kernel K_1 associated with 2 layer neural network defined on $\mathbb{R}$.*

Assumption A.3.2. *There exists an $M \in [0, \infty)$ such that $\mathbb{P}_{\pi^b}(|Y| \leq M) = 1$.*

Assumption A.3.3. *$\hat{\rho}(a_i, x_i)$ satisfies*

$$\hat{\rho}(a_i, x_i) = \tilde{w}(y_i) + \eta_i,$$

where $\eta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

Theorem A.3.4

Suppose that the IPW and MR estimators are defined as,

$$\tilde{\theta}_{\text{IPW}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{\rho}(a_i, x_i)\, y_i, \quad \text{and} \quad \tilde{\theta}_{\text{MR}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \hat{w}_m(y_i)\, y_i,$$

where $\hat{w}_m(y)$ is obtained by regressing to the estimated policy ratios $\hat{\rho}(a, x)$ using m i.i.d. training samples $\mathcal{D}_{\text{tr}} \coloneqq \{(x_i^{\text{tr}}, a_i^{\text{tr}}, y_i^{\text{tr}})\}_{i=1}^{m}$, i.e., by minimising the loss

$$\mathcal{L}(\phi) = \mathbb{E}_{(X,A,Y)\sim\mathcal{D}_{\text{tr}}}\left[\left(\hat{\rho}(A, X) - f_\phi(Y)\right)^2\right].$$

Suppose Assumptions A.3.1-A.3.3 hold, then for any given $\delta \in (0, 1)$, if f_ϕ is a two-layer neural network with width k that is sufficiently large and stops the gradient flow at time $t_* \propto m^{2/3}$, then for sufficiently large m, there exists a constant C_1 independent of δ and m, such that

$$|\text{Bias}(\tilde{\theta}_{\text{MR}}) - \text{Bias}(\tilde{\theta}_{\text{IPW}})| \leq C_1\, m^{-1/3} \log\frac{6}{\delta}$$

holds with probability at least $(1 - \delta)(1 - o_k(1))$. Moreover, there exist constants C_2, C_3 independent of δ and m such that

$$\begin{aligned} &n(\text{Var}_{\pi^b}[\tilde{\theta}_{\text{IPW}}] - \text{Var}_{\pi^b}[\tilde{\theta}_{\text{MR}}]) \\ &\qquad \geq \underbrace{\mathbb{E}_{\pi^b}[\text{Var}_{\pi^b}[\hat{\rho}(A, X)\, Y \mid Y]]}_{\geq 0} - C_2\, m^{-2/3} \log^2\frac{6}{\delta} - C_3\, m^{-1/3} \log\frac{6}{\delta} \end{aligned}$$

holds with probability at least $(1 - \delta)(1 - o_k(1))$. Here, the randomness comes from the joint distribution of training samples and random initialization of parameters in the neural network f_ϕ.

Proof of Theorem A.3.4. The proof of this theorem relies on [Lai et al., 2023, Theorem 4.1]. Recall the definition $\tilde{w}(Y) \coloneqq \mathbb{E}_{\pi^b}[\hat{\rho}(A, X) \mid Y]$. We can rewrite our setup in the setting of [Lai et al., 2023, Theorem 4.1], by relabelling $\hat{\rho}(a, x)$ in our setup as y in their setup and relabelling y in our setup as x in their setup. Then, given $\delta \in (0, 1)$, from [Lai

et al., 2023, Theorem 4.1], it follows that under Assumptions A.3.1-A.3.3 that there exists a constant C independent of δ and m, such that

$$\mathbb{E}_{\pi^b}[\epsilon^2] \leq C\, m^{-2/3}\, \log^2 \frac{6}{\delta} \tag{A.1}$$

holds with probability at least $(1-\delta)(1-o_k(1))$, where $\epsilon \coloneqq \hat{w}_m(Y) - \tilde{w}(Y)$. Recall from Proposition 2.3.4 that

$$|\mathrm{Bias}(\tilde{\theta}_{\mathrm{MR}}) - \mathrm{Bias}(\tilde{\theta}_{\mathrm{IPW}})| = |\mathbb{E}_{\pi^b}[\epsilon\, Y]|.$$

From this it follows using Cauchy-Schwarz inequality that,

$$|\mathrm{Bias}(\tilde{\theta}_{\mathrm{MR}}) - \mathrm{Bias}(\tilde{\theta}_{\mathrm{IPW}})| = |\mathbb{E}_{\pi^b}[\epsilon\, Y]| \leq \left(\mathbb{E}_{\pi^b}[\epsilon^2]\mathbb{E}_{\pi^b}[Y^2]\right)^{1/2}.$$

Combining the above with Eqn. (A.1), it follows that,

$$|\mathrm{Bias}(\tilde{\theta}_{\mathrm{MR}}) - \mathrm{Bias}(\tilde{\theta}_{\mathrm{IPW}})| \leq C^{1/2}\, m^{-1/3}\, \log \frac{6}{\delta} (\mathbb{E}_{\pi^b}[Y^2])^{1/2} = C_1\, m^{-1/3}\, \log \frac{6}{\delta}$$

holds with probability at least $(1-\delta)(1-o_k(1))$, where $C_1 = C^{1/2}\, (\mathbb{E}_{\pi^b}[Y^2])^{1/2}$.

Next, to prove the variance result, we recall from Proposition 2.3.4 that

$$n(\mathrm{Var}_{\pi^b}[\tilde{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\tilde{\theta}_{\mathrm{MR}}]) = \mathbb{E}_{\pi^b}[\mathrm{Var}_{\pi^b}[\hat{\rho}(A,X) \mid Y]\, Y^2] - \mathrm{Var}_{\pi^b}[\epsilon\, Y] - 2\,\mathrm{Cov}(\epsilon\, Y, \tilde{w}(Y)\, Y)$$

Now note that, under Assumption A.3.2,

$$\mathrm{Var}_{\pi^b}[\epsilon\, Y] \leq \mathbb{E}_{\pi^b}[(\epsilon\, Y)^2] \leq M^2 \mathbb{E}_{\pi^b}[\epsilon^2] \leq C\, M^2\, m^{-2/3}\, \log^2 \frac{6}{\delta} = C_2\, m^{-2/3}\, \log^2 \frac{6}{\delta},$$

holds with probability at least $(1-\delta)(1-o_k(1))$, where $C_2 = C\, M^2$. Similarly, we have that with probability at least $(1-\delta)(1-o_k(1))$,

$$\begin{aligned}
|\mathrm{Cov}(\epsilon\, Y, \tilde{w}(Y)\, Y)| &= |\mathbb{E}_{\pi^b}[\epsilon\, \tilde{w}(Y)\, Y^2] - \mathbb{E}_{\pi^b}[\epsilon\, Y]\mathbb{E}_{\pi^b}[\tilde{w}(Y)\, Y]| \\
&\leq |\mathbb{E}_{\pi^b}[\epsilon\, \tilde{w}(Y)\, Y^2]| + |\mathbb{E}_{\pi^b}[\epsilon\, Y]\mathbb{E}_{\pi^b}[\tilde{w}(Y)\, Y]| \\
&\leq \left(\mathbb{E}_{\pi^b}[\epsilon^2]\mathbb{E}_{\pi^b}[\tilde{w}(Y)^2\, Y^4]\right)^{1/2} + (\mathbb{E}_{\pi^b}[\epsilon^2]\mathbb{E}_{\pi^b}[Y^2])^{1/2} |\mathbb{E}_{\pi^b}[\tilde{w}(Y)\, Y]| \\
&= (\mathbb{E}_{\pi^b}[\epsilon^2])^{1/2} \left((\mathbb{E}_{\pi^b}[\tilde{w}(Y)^2\, Y^4])^{1/2} + (\mathbb{E}_{\pi^b}[Y^2])^{1/2}\, |\mathbb{E}_{\pi^b}[\tilde{w}(Y)\, Y]|\right) \\
&\leq C_3\, m^{-1/3}\, \log \frac{6}{\delta},
\end{aligned}$$

where $C_3 = C\, (\mathbb{E}_{\pi^b}[\tilde{w}(Y)^2\, Y^4])^{1/2} + (\mathbb{E}_{\pi^b}[Y^2])^{1/2}\, |\mathbb{E}_{\pi^b}[\tilde{w}(Y)\, Y]|$, and we use Cauchy-Schwarz inequality in the third step above. Putting this together, we obtain the required result. □

Intuition This theorem shows that as the number of training samples m increases, the biases of MR and IPW estimators become roughly equal, whereas the variance of MR estimator falls below that of the IPW estimator. The empirical results shown in Appendix A.6.2 are consistent with this result. Moreover, in Theorem A.3.4, the estimated policy ratio $\hat{\rho}(a, x)$ is fixed for increasing m, i.e., we do not update $\hat{\rho}(a, x)$ as more training data becomes available. While this may seem as a disadvantage for the IPW estimator, we point out that the result also holds when the policy ratio is exact (i.e., $\hat{\rho}(a, x) = \rho(a, x)$) and hence the IPW estimator is unbiased.

Relaxing Assumption A.3.3 Lai et al. [2023][Theorem 4.1] suppose that the data has the relationship shown in Assumption A.3.3. However, the theorem relies on Corollary 4.4 in Lin et al. [2020], which requires a strictly weaker assumption (Assumption 1 in Lin et al. [2020]). Therefore, we can relax Assumption A.3.3 to the following assumption.

Assumption A.3.5. *There exists positive constants Q and M such that for all $l \geq 2$ with $l \in \mathbb{N}$*

$$\mathbb{E}_{\pi^b}[\hat{\rho}(A, X)^l \mid Y] \leq \frac{1}{2}\, l!\, M^{l-2}\, Q^2$$

p_{π^b}-almost surely.

It is easy to check that Assumption A.3.5 is strictly weaker than Assumption A.3.3, and is also satisfied if the policy ratio $\hat{\rho}(A, X)$ is almost surely bounded. For simplicity, we use the stronger assumption in our Proposition A.3.4.

A.4 Generalised formulation of the MIPS estimator [Saito and Joachims, 2022]

As described in Section 2.3.1, the MIPS estimator proposed by Saito and Joachims [2022] assumes the existence of *action embeddings* E which summarise all relevant information about the action A, and achieves a lower variance than the IPW estimator. To achieve this, the MIPS estimator only considers the shift in the distribution of (X, E) as a result of policy shift, instead of considering the shift in (X, A) (as in IPW estimator). In this section, we show that this idea can be generalised to instead consider general representations R of the context-action pair (X, A), which encapsulate all relevant information about

the outcome Y. The MIPS estimator is a special case of this generalised setting where the representation R is of the form (X, E).

Generalised MIPS (G-MIPS) estimator Suppose that there exists an embedding R of the context-action pair (X, A), with the Bayesian network shown in Figure A.1. Here, R may be a lower-dimensional representation of the (X, A) pair which contains all the information necessary to predict the outcome Y. This corresponds to the following conditional independence assumption:

Assumption A.4.1. *The context-action pair (X, A) has no direct effect on the outcome Y given R, i.e., $Y \perp\!\!\!\perp (X, A) \mid R$.*

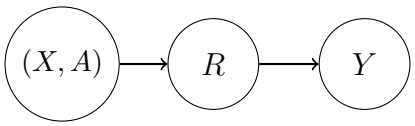


Figure A.1: Bayesian network corresponding to Assumption A.4.1.

As illustrated in Figure A.1, Assumption A.4.1 means that the embedding R fully mediates every possible effect of (X, A) on Y. The generalised MIPS estimator $\hat{\theta}_{\text{G-MIPS}}$ of target policy value, $\mathbb{E}_{\pi^*}[Y]$, is defined as

$$\hat{\theta}_{\text{G-MIPS}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \frac{p_{\pi^*}(r_i)}{p_{\pi^b}(r_i)} y_i,$$

where $p_{\pi^b}(r)$ denote the density of R under the behaviour policy (likewise for $p_{\pi^*}(r)$). Under assumption A.4.1, $\hat{\theta}_{\text{G-MIPS}}$ provides an unbiased estimator of target policy value. Similar to Lemma 2.3.1, the density ratio $\frac{p_{\pi^*}(r)}{p_{\pi^b}(r)}$ can be estimated by solving the regression problem

$$\arg\min_{f} \mathbb{E}_{\pi^b} \left(\frac{\pi^*(A \mid X)}{\pi^b(A \mid X)} - f\left(R\right) \right)^2. \tag{A.2}$$

A.4.1 Variance reduction of G-MIPS estimator

By only considering the shift in the embedding R, the G-MIPS estimator achieves a lower variance relative to the vanilla IPW estimator. The following result, which is a straightforward extension of [Saito and Joachims, 2022, Theorem 3.6], formalises this.

Proposition A.4.1 (Variance reduction of G-MIPS)

When the ratios $\rho(a,x)$ and $\frac{p_{\pi^*}(r)}{p_{\pi^b}(r)}$ are known exactly then under Assumption A.4.1, we have that $\mathbb{E}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] = \mathbb{E}_{\pi^b}[\hat{\theta}_{\text{G-MIPS}}] = \mathbb{E}_{\pi^*}[Y]$. Moreover,

$$\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\text{G-MIPS}}] \geq \frac{1}{n}\mathbb{E}_{\pi^b}\left[\mathbb{E}[Y^2 \mid R]\mathrm{Var}_{\pi^b}[\rho(A,X) \mid R]\right] \geq 0.$$

Proof of Proposition A.4.1. The following proof, which is included for completeness, is a straightforward extension of [Saito and Joachims, 2022, Theorem 3.6].

$$\begin{aligned}
&n(\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MIPS}}]) \\
&\quad= \mathrm{Var}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\right] - \mathrm{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,Y\right] \\
&\quad= \mathrm{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\middle|R\right]\right] + \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\middle|R\right]\right] - \mathrm{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,Y\middle|R\right]\right] \\
&\qquad- \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,Y\middle|R\right]\right]
\end{aligned}$$

Now using the conditional independence Assumption A.4.1, the first term on the RHS above becomes,

$$\begin{aligned}
\mathrm{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\middle|R\right]\right] &= \mathrm{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\middle|R\right]\,\mathbb{E}_{\pi^b}\left[Y|R\right]\right] \\
&= \mathrm{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,\mathbb{E}_{\pi^b}\left[Y|R\right]\right],
\end{aligned}$$

where in the last step above we use the fact that

$$\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\middle|R\right] = \frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}.$$

Putting this together, we get that

$$\begin{aligned}
&n(\mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{IPW}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}_{\mathrm{MIPS}}]) \\
&\quad= \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\middle|R\right]\right] - \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,Y\middle|R\right]\right].
\end{aligned} \tag{A.3}$$

Since we have that

$$\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\middle|R\right] = \mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\middle|R\right]\,\mathbb{E}_{\pi^b}\left[Y|R\right] = \frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,\mathbb{E}_{\pi^b}\left[Y|R\right],$$

Eq. (A.3) becomes,

$$
\begin{aligned}
&\mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\middle|R\right]\right] - \mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,Y\middle|R\right]\right] \\
&\quad= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\left(\frac{\pi^*(A|X)}{\pi^b(A|X)}\,Y\right)^2\middle|R\right] - \mathbb{E}_{\pi^b}\left[\left(\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\,Y\right)^2\middle|R\right]\right] \\
&\quad= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\left(\frac{\pi^*(A|X)}{\pi^b(A|X)}\right)^2\middle|R\right]\,\mathbb{E}_{\pi^b}\left[Y^2|R\right] - \left(\frac{p_{\pi^*}(R)}{p_{\pi^b}(R)}\right)^2\,\mathbb{E}_{\pi^b}\left[Y^2|R\right]\right] \\
&\quad= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[Y^2|R\right]\left(\mathbb{E}_{\pi^b}\left[\left(\frac{\pi^*(A|X)}{\pi^b(A|X)}\right)^2\middle|R\right] - \left(\mathbb{E}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\middle|R\right]\right)^2\right)\right] \\
&\quad= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[Y^2|R\right]\,\mathrm{Var}_{\pi^b}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\middle|R\right]\right].
\end{aligned}
$$

□

Intuition Here, R contains all relevant information regarding the outcome Y. Moreover, intuitively R can be thought of as the state obtained by 'filtering out' relevant information about Y from (X, A). Therefore, R contains less 'redundant' information regarding the outcome Y as compared to the covariate-action pair (X, A). As a result, the G-MIPS estimator which only considers the shift in the marginal distribution of R due to the policy shift is more efficient than the IPW estimator, which considers the shift in the joint distribution of (X, A) instead. In fact, as the amount of 'redundant' information regarding Y decreases in the embedding R, the G-MIPS estimator becomes increasingly efficient with decreasing variance. We formalise this as follows:

Assumption A.4.2. *Assume there exist embeddings $R^{(1)}, R^{(2)}$ of the covariate-action pair (X, A), with Bayesian network shown in Figure A.2. This corresponds to the following conditional independence assumptions:*

$$R^{(2)} \perp\!\!\!\perp (X, A) \mid R^{(1)}, \qquad \text{and} \qquad Y \perp\!\!\!\perp (R^{(1)}, X, A) \mid R^{(2)}.$$

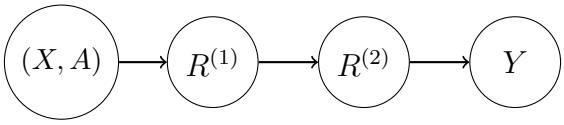


Figure A.2: Bayesian network corresponding to Assumption A.4.2.

We can define G-MIPS estimators for these embeddings to obtain unbiased OPE estimators under Assumption A.4.2 as follows:

$$\hat{\theta}^{(j)}_{\text{G-MIPS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{p_{\pi^*}(r_i^{(j)})}{p_{\pi^b}(r_i^{(j)})}\, y_i,$$

for $j \in \{1, 2\}$. Here, $\frac{p_{\pi^*}(r^{(j)})}{p_{\pi^b}(r^{(j)})}$ is the ratio of marginal densities of $R^{(j)}$ under target and behaviour policies. We next show that the variance of $\hat{\theta}^{(j)}_{\text{G-MIPS}}$ decreases with increasing j.

Proposition A.4.2

When the ratios $\rho(a, x)$, $w(y)$ and $\frac{p_{\pi^*}(r^{(j)})}{p_{\pi^b}(r^{(j)})}$ are known exactly for $j \in \{1, 2\}$, then under Assumption A.4.2 we get that

$$\mathbb{E}_{\pi^b}[\hat{\theta}_{\text{IPW}}] = \mathbb{E}_{\pi^b}[\hat{\theta}^{(1)}_{\text{G-MIPS}}] = \mathbb{E}_{\pi^b}[\hat{\theta}^{(2)}_{\text{G-MIPS}}] = \mathbb{E}_{\pi^b}[\hat{\theta}_{\text{MR}}] = \mathbb{E}_{\pi^*}[Y].$$

Moreover,

$$\text{Var}_{\pi^b}[\hat{\theta}_{\text{IPW}}] \geq \text{Var}_{\pi^b}[\hat{\theta}^{(1)}_{\text{G-MIPS}}] \geq \text{Var}_{\pi^b}[\hat{\theta}^{(2)}_{\text{G-MIPS}}] \geq \text{Var}_{\pi^b}[\hat{\theta}_{\text{MR}}].$$

Proof of Proposition A.4.2. First, we prove that the G-MIPS estimators are unbiased using induction on j. We define $R^{(0)} \coloneqq (X, A)$ and $\hat{\theta}^{(0)}_{\text{G-MIPS}}$ defined as

$$\hat{\theta}^{(0)}_{\text{G-MIPS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \frac{p_{\pi^*}(r_i^{(0)})}{p_{\pi^b}(r_i^{(0)})}\, y_i,$$

recovers the IPW estimator $\hat{\theta}_{\text{IPW}}$. When $j = 0$, we know that $\hat{\theta}^{(0)}_{\text{G-MIPS}} = \hat{\theta}_{\text{IPW}}$ is unbiased. Now, assume that $\mathbb{E}_{\pi^b}[\hat{\theta}^{(j)}_{\text{G-MIPS}}] = \mathbb{E}_{\pi^*}[Y]$.

Conditional on $R^{(j)}$, $R^{(j+1)}$ does not depend on the policy. Therefore,

$$\frac{p_{\pi^*}(r^{(j)})}{p_{\pi^b}(r^{(j)})} = \frac{p_{\pi^*}(r^{(j)})\, p(r^{(j+1)} \mid r^{(j)})}{p_{\pi^b}(r^{(j)})\, p(r^{(j+1)} \mid r^{(j)})} = \frac{p_{\pi^*}(r^{(j)}, r^{(j+1)})}{p_{\pi^b}(r^{(j)}, r^{(j+1)})}.$$

And therefore,

$$\begin{aligned}
\frac{p_{\pi^*}(r^{(j+1)})}{p_{\pi^b}(r^{(j+1)})} &= \int_{r^{(j)}} \frac{p_{\pi^*}(r^{(j)}, r^{(j+1)})}{p_{\pi^b}(r^{(j)}, r^{(j+1)})}\, p_{\pi^b}(r^{(j)} \mid r^{(j+1)})\, \mathrm{d}r^{(j)} \\
&= \int_{r^{(j)}} \frac{p_{\pi^*}(r^{(j)})}{p_{\pi^b}(r^{(j)})}\, p_{\pi^b}(r^{(j)} \mid r^{(j+1)})\, \mathrm{d}r^{(j)} \\
&= \mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})} \middle| R^{(j+1)} = r^{(j+1)} \right].
\end{aligned}$$

Using this and the fact that $R^{(j)} \perp\!\!\!\perp Y \mid R^{(j+1)}$, we get that

$$\begin{aligned}
\mathbb{E}_{\pi^b}\left[\hat{\theta}^{(j+1)}_{\text{G-MIPS}}\right] &= \mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\,Y\right] \\
&= \mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\,\mathbb{E}_{\pi^b}[Y|R^{(j+1)}]\right] \\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\middle| R^{(j+1)}\right]\,\mathbb{E}_{\pi^b}[Y|R^{(j+1)}]\right] \\
&= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\middle| R^{(j+1)}\right]\right] \\
&= \mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\right] \\
&= \mathbb{E}_{\pi^b}\left[\hat{\theta}^{(j)}_{\text{G-MIPS}}\right] = \mathbb{E}_{\pi^*}[Y].
\end{aligned}$$

Next, to prove the variance result we consider the difference

$$\begin{aligned}
&\text{Var}_{\pi^b}[\hat{\theta}^{(j)}_{\text{G-MIPS}}] - \text{Var}_{\pi^b}[\hat{\theta}^{(j+1)}_{\text{G-MIPS}}] \\
&= \frac{1}{n}\left(\text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\right] - \text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\,Y\right]\right) \\
&= \frac{1}{n}\left(\text{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\middle| R^{(j+1)}\right]\right] + \mathbb{E}_{\pi^b}\left[\text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\middle| R^{(j+1)}\right]\right]\right. \\
&\qquad \left. - \text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\,\mathbb{E}_{\pi^b}[Y \mid R^{(j+1)}]\right] - \mathbb{E}_{\pi^b}\left[\left(\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\right)^2 \text{Var}_{\pi^b}[Y \mid R^{(j+1)}]\right]\right)
\end{aligned}$$

where in the last step we use the law of total variance. Now, using the fact that $R^{(j)} \perp\!\!\!\perp Y \mid R^{(j+1)}$, we can rewrite the expression above as

$$\begin{aligned}
&= \frac{1}{n}\left(\text{Var}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\middle| R^{(j+1)}\right]\,\mathbb{E}_{\pi^b}[Y|R^{(j+1)}]\right] + \mathbb{E}_{\pi^b}\left[\text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\middle| R^{(j+1)}\right]\right]\right. \\
&\qquad \left. - \text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\,\mathbb{E}_{\pi^b}[Y \mid R^{(j+1)}]\right] - \mathbb{E}_{\pi^b}\left[\left(\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\right)^2 \text{Var}_{\pi^b}[Y \mid R^{(j+1)}]\right]\right) \\
&= \frac{1}{n}\left(\mathbb{E}_{\pi^b}\left[\text{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\middle| R^{(j+1)}\right]\right] - \mathbb{E}_{\pi^b}\left[\left(\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\right)^2 \text{Var}_{\pi^b}[Y \mid R^{(j+1)}]\right]\right).
\end{aligned}$$

Moreover, again using the conditional independence $R^{(j)} \perp\!\!\!\perp Y \mid R^{(j+1)}$, we can expand the

first term in the expression above as follows:

$$\begin{aligned}
\mathbb{E}_{\pi^b}\left[\mathrm{Var}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\,Y\middle|R^{(j+1)}\right]\right] &= \mathbb{E}_{\pi^b}\left[\mathbb{E}_{\pi^b}\left[\frac{p^2_{\pi^*}(R^{(j)})}{p^2_{\pi^b}(R^{(j)})}\middle|R^{(j+1)}\right]\,\mathbb{E}_{\pi^b}[Y^2|R^{(j+1)}]\right. \\
&\qquad \left. - \left(\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\middle|R^{(j+1)}\right]\mathbb{E}_{\pi^b}[Y|R^{(j+1)}]\right)^2\right] \\
&\geq \mathbb{E}_{\pi^b}\left[\left(\mathbb{E}_{\pi^b}\left[\frac{p_{\pi^*}(R^{(j)})}{p_{\pi^b}(R^{(j)})}\middle|R^{(j+1)}\right]\right)^2\,\mathbb{E}_{\pi^b}[Y^2|R^{(j+1)}]\right. \\
&\qquad \left. - \left(\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\mathbb{E}_{\pi^b}[Y|R^{(j+1)}]\right)^2\right] \\
&= \mathbb{E}_{\pi^b}\left[\left(\frac{p_{\pi^*}(R^{(j+1)})}{p_{\pi^b}(R^{(j+1)})}\right)^2\,\mathrm{Var}_{\pi^b}[Y \mid R^{(j+1)}]\right].
\end{aligned}$$

Here, to get the inequality above, we use the fact that $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$. Putting this together, we get that $\mathrm{Var}_{\pi^b}[\hat{\theta}^{(j)}_{\text{G-MIPS}}] - \mathrm{Var}_{\pi^b}[\hat{\theta}^{(j+1)}_{\text{G-MIPS}}] \geq 0$.

Moreover, the result $\mathrm{Var}_{\pi^b}[\hat{\theta}^{(2)}_{\text{G-MIPS}}] \geq \mathrm{Var}_{\pi^b}[\hat{\theta}_{\text{MR}}]$ follows straightforwardly from above by defining $R^{(3)} \coloneqq Y$. Then, the embeddings satisfy the causal structure

$$R^{(0)} \to R^{(1)} \to R^{(2)} \to R^{(3)} \to Y.$$

Using the result above, we know that $\mathrm{Var}_{\pi^b}[\hat{\theta}^{(2)}_{\text{G-MIPS}}] \geq \mathrm{Var}_{\pi^b}[\hat{\theta}^{(3)}_{\text{G-MIPS}}]$. But now it is straightforward to see that $\hat{\theta}^{(3)}_{\text{G-MIPS}} = \hat{\theta}_{\text{MR}}$, and the result follows. □

Intuition Here, $R^{(j+1)}$ can be thought of as the embedding obtained by 'filtering out' relevant information about Y from $R^{(j)}$. As such, the amount of 'redundant' information regarding the outcome Y decreases successively along the sequence $R^{(0)}(\coloneqq (X, A)), R^{(1)}, R^{(2)}$. As a result, the G-MIPS estimators which only consider the shift in the marginal distributions of $R^{(j)}$ due to policy shift become increasingly efficient with decreasing variance as j increases. Define the representation $R^{(3)} \coloneqq Y$, then the corresponding G-MIPS estimator reduces to the MR estimator, i.e., $\hat{\theta}^{(3)}_{\text{G-MIPS}} = \hat{\theta}_{\text{MR}}$. Moreover, this estimator has minimum variance among all the G-MIPS estimators $\{\hat{\theta}^{(j)}_{\text{G-MIPS}}\}_{0\leq j\leq k}$, as the representation $R^{(3)}$ contains precisely the least amount of information necessary to obtain the outcome Y. In other words, Y itself serves as the 'best embedding' of covariate-action pair $R^{(0)}$ which contains all relevant information regarding Y. We verify this empirically in Appendix A.6.2 by

reproducing the experimental setup in Saito and Joachims [2022] along with the MR baseline. Additionally, the MR estimator does not rely on assumptions like A.4.1 for unbiasedness.

In addition to this, solving the regression problem in Eq. (A.2) will typically be more difficult when R is higher dimensional (as is likely to be the case for many choices of embeddings R), leading to high bias. In contrast, for MR the embedding $R = Y$ is one dimensional and therefore the regression problem is significantly easier to solve and yields lower bias. Our empirical results in Appendix A.6 confirm this.

A.4.2 Doubly robust G-MIPS estimators

Consider the setup for the G-MIPS estimator shown in Figure A.1. In this case, we can derive a doubly robust extension of the G-MIPS estimator, denoted as GM-DR, which uses an estimate of the conditional mean $\tilde{\mu}(r) \approx \mathbb{E}[Y \mid R = r]$ as a control variate to decrease the variance of G-MIPS estimator. This can be explicitly written as follows:

$$\tilde{\theta}_{\text{DM-DR}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} \frac{p_{\pi^*}(r_i)}{p_{\pi^b}(r_i)} \left(y_i - \tilde{\mu}(r_i)\right) + \tilde{\eta}(\pi^*). \tag{A.4}$$

where $\tilde{\eta}(\pi^*) = \frac{1}{n} \sum_{i=1}^{n} \sum_{r' \in \mathcal{R}} \tilde{\mu}(r')\, p_{\pi^*}(r' \mid x_i)$ is the analogue of the direct method. Here, $\mathcal{R}$ denotes the space of the possible of the representations R[1]. Moreover, given the density $p(r \mid x, a)$, we can compute $p_{\pi^*}(r \mid x)$ using

$$p_{\pi^*}(r \mid x) = \sum_{a' \in \mathcal{A}} p(r \mid x, a')\, \pi^*(a' \mid x).$$

It is straightforward to extend ideas from Dudík et al. [2014b] to show that estimator $\tilde{\theta}_{\text{DM-DR}}$ is doubly robust in that it will yield accurate value estimates if either the importance weights $\frac{p_{\pi^*}(r)}{p_{\pi^b}(r)}$ or the outcome model $\tilde{\mu}(r)$ is well estimated.

There is no analogous DR extension of the MR estimator A consequence of considering the embedding $R = Y$ (as in MR) is that in this case we do not have an analogous doubly robust extension as above. To see why this is the case, note that when $R = Y$, we get that $\tilde{\mu}(r) = \mathbb{E}[Y \mid R = r] = \mathbb{E}[Y \mid Y = y] = y$. If we substitute this $\tilde{\mu}(r)$ in (A.4), we are simply left with $\tilde{\eta}(\pi^*)$ on the right hand side (as the first term cancels out). This means that the resulting estimator does not retain the doubly robust nature as we no longer obtain an accurate estimate if either the outcome model or the importance ratios are well estimated.

[1] the $\sum_{r' \in \mathcal{R}}$ can be replaced with $\int_{r' \in \mathcal{R}} \mathrm{d}r'$ when $\mathcal{R}$ is continuous

A.5 Application to causal inference

In this section, we investigate the application of the MR estimator for the estimation of average treatment effect (ATE). In this setting, we suppose that $\mathcal{A} = \{0, 1\}$, and the goal is to estimate ATE defined as follows:

$$\text{ATE} \coloneqq \mathbb{E}[Y(1) - Y(0)]$$

Here, we use the potential outcomes notation [Robins, 1986] to denote the outcome under a deterministic policy $\pi^*(a' \mid x) = \mathbb{1}(a' = a)$ as $Y(a)$.

Specifically, the IPW estimator applied to ATE estimation yields:

$$\widehat{\text{ATE}}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n} \rho_{\text{ATE}}(a_i, x_i) \times y_i,$$

where

$$\rho_{\text{ATE}}(a, x) \coloneqq \frac{\mathbb{1}(a = 1) - \mathbb{1}(a = 0)}{\pi^b(a|x)}.$$

Similarly, the MR estimator can be written as

$$\widehat{\text{ATE}}_{\text{MR}} = \frac{1}{n}\sum_{i=1}^{n} w_{\text{ATE}}(y_i) \times y_i,$$

where

$$w_{\text{ATE}}(y) = \frac{p_{\pi^{(1)}}(y) - p_{\pi^{(0)}}(y)}{p_{\pi^b}(y)},$$

and $\pi^{(a)}(a' \mid x) \coloneqq \mathbb{1}(a' = a)$ for $a \in \{0, 1\}$.

Again, using the fact that $w_{\text{ATE}}(Y) \overset{\text{a.s.}}{=} \mathbb{E}[\rho_{\text{ATE}}(A, X) \mid Y]$, we can obtain w_{ATE} by minimising a simple mean-squared loss:

$$w_{\text{ATE}} = \arg\min_f \mathbb{E}_{\pi^b}\left[\frac{\mathbb{1}(A = 1) - \mathbb{1}(A = 0)}{\pi^b(A|X)} - f(Y)\right]^2.$$

Proposition A.5.1 (Variance comparison with IPW ATE estimator)

When the weights $\rho_{\text{ATE}}(a, x)$ and $w_{\text{ATE}}(y)$ are known exactly, we have that $\text{Var}[\widehat{\text{ATE}}_{\text{MR}}] \leq \text{Var}[\widehat{\text{ATE}}_{\text{IPW}}]$. Specifically,

$$\text{Var}[\widehat{\text{ATE}}_{\text{IPW}}] - \text{Var}[\widehat{\text{ATE}}_{\text{MR}}] = \frac{1}{n}\mathbb{E}\left[\text{Var}\left[\rho_{\text{ATE}}(A, X)|Y\right]\, Y^2\right] \geq 0.$$

Proof of Proposition A.5.1. We have

$$\mathrm{Var}[\widehat{\mathrm{ATE}}_{\mathrm{IPW}}] - \mathrm{Var}[\widehat{\mathrm{ATE}}_{\mathrm{MR}}] = \frac{1}{n}\left(\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,Y] - \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y]\right). \tag{A.5}$$

Using the tower law of variance, we get that

$$\begin{aligned}
\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,Y] &= \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid Y]] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid Y]] \\
&= \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2] \\
&= \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2].
\end{aligned}$$

Putting this together with (A.5) we obtain,

$$\mathrm{Var}[\widehat{\mathrm{ATE}}_{\mathrm{IPW}}] - \mathrm{Var}[\widehat{\mathrm{ATE}}_{\mathrm{MR}}] = \frac{1}{n}\mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2],$$

which straightforwardly leads to the result. □

Given the above definitions, the IPW estimator for $\mathbb{E}[Y(a)]$ would only consider datapoints with $A = a$, as it weights the samples using the policy ratios $\mathbb{1}(A = a)/\pi^b(A|X)$ which are only non-zero when $A = a$. This is however not the case with the MR estimator, as it uses the weights $p_{\pi^*}(Y)/p_{\pi^b}(Y)$ which are not necessarily zero for $A \neq a$. Therefore, MR uses all evaluation datapoints $\mathcal{D}$ when estimating $\mathbb{E}[Y(a)]$. The MR estimator therefore leads to a more efficient use of evaluation data in this example.

Likewise, the doubly robust (DR) estimator applied to ATE estimation yields,

$$\widehat{\mathrm{ATE}}_{\mathrm{DR}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} \rho_{\mathrm{ATE}}(a_i, x_i)\,\left(y_i - \hat{\mu}(a_i, x_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}(1, x_i) - \hat{\mu}(0, x_i)\right),$$

where $\hat{\mu}(a,x) \approx \mathbb{E}[Y \mid X = x, A = a]$. Like in classical off-policy evaluation, DR yields an accurate estimator of ATE when either the weights $\rho_{\mathrm{ATE}}(a,x)$ or the outcome model i.e., $\hat{\mu}(a,x) = \mathbb{E}[Y \mid X = x, A = a]$, are well estimated. However, despite this doubly robust nature of the estimator, we can show that the variance of the DR estimator may be higher than that of the MR estimator in many cases. The following result formalises this variance comparison between the DR and MR estimators, and is analogous to the result in Proposition 2.3.3 derived for classical off-policy evaluation.

Proposition A.5.2 (Variance comparison with DR ATE estimator)

When the weights $\rho_{\text{ATE}}(a, x)$ and $w_{\text{ATE}}(y)$ are known exactly,

$$\begin{aligned}\text{Var}[\widehat{\text{ATE}}_{\text{DR}}] - \text{Var}[\widehat{\text{ATE}}_{\text{MR}}] \\ \geq \frac{1}{n}\mathbb{E}\left[\text{Var}\left[\rho_{\text{ATE}}(A, X)\, Y \mid Y\right] - \text{Var}\left[\rho_{\text{ATE}}(A, X)\mu(A, X) \mid X\right]\right],\end{aligned}$$

where $\mu(A, X) \coloneqq \mathbb{E}[Y \mid X, A]$.

Proof of Proposition A.5.2. Using the law of total variance, we get that

$$\begin{aligned}n\,\text{Var}[\widehat{\text{ATE}}_{\text{DR}}] &= \text{Var}[\rho_{\text{ATE}}(A, X)\,(Y - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X))] \\ &= \text{Var}[\mathbb{E}[\rho_{\text{ATE}}(A, X)\,(Y - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X)) \mid X, A]] \\ &\quad + \mathbb{E}[\text{Var}[\rho_{\text{ATE}}(A, X)\,(Y - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X)) \mid X, A]] \\ &= \text{Var}[\rho_{\text{ATE}}(A, X)\,(\mu(A, X) - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X))] \\ &\quad + \mathbb{E}[\rho^2_{\text{ATE}}(A, X)\text{Var}[Y \mid X, A]].\end{aligned}$$

Again, using the law of total variance we can rewrite the first term on the RHS above as,

$$\begin{aligned}&\text{Var}[\rho_{\text{ATE}}(A, X)\,(\mu(A, X) - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X))] \\ &= \text{Var}[\mathbb{E}[\rho_{\text{ATE}}(A, X)\,(\mu(A, X) - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X)) \mid X]] \\ &\quad + \mathbb{E}[\text{Var}[\rho_{\text{ATE}}(A, X)\,(\mu(A, X) - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X)) \mid X]] \\ &\geq \text{Var}[\mathbb{E}[\rho_{\text{ATE}}(A, X)\,(\mu(A, X) - \hat{\mu}(A, X)) + (\hat{\mu}(1, X) - \hat{\mu}(0, X)) \mid X]] \\ &= \text{Var}[\mathbb{E}[\rho_{\text{ATE}}(A, X)\,(\mu(A, X) - \hat{\mu}(A, X)) + \rho_{\text{ATE}}(A, X)\,\hat{\mu}(A, X) \mid X]] \\ &= \text{Var}[\mathbb{E}[\rho_{\text{ATE}}(A, X)\,\mu(A, X) \mid X]],\end{aligned}$$

where, in the second last step above we use the fact that

$$\mathbb{E}[\rho_{\text{ATE}}(A, X)\,\hat{\mu}(A, X) \mid X] = \hat{\mu}(1, X) - \hat{\mu}(0, X).$$

Putting this together, we get that

$$n\,\text{Var}[\widehat{\text{ATE}}_{\text{DR}}] \geq \text{Var}[\mathbb{E}[\rho_{\text{ATE}}(A, X)\,\mu(A, X) \mid X]] + \mathbb{E}[\rho^2_{\text{ATE}}(A, X)\text{Var}[Y \mid X, A]].$$

Therefore,

$$
\begin{aligned}
&n\left(\mathrm{Var}[\widehat{\mathrm{ATE}}_{\mathrm{DR}}] - \mathrm{Var}[\widehat{\mathrm{ATE}}_{\mathrm{MR}}]\right) \\
&\quad \geq \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] + \mathbb{E}[\rho^2_{\mathrm{ATE}}(A,X)\mathrm{Var}[Y \mid X,A]] - \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] \\
&\quad = \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid X,A]] - \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] \\
&\quad = \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] + \mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,Y] - \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid X,A]] \\
&\qquad - \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] \\
&\quad = \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] + \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2] \\
&\qquad - \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid X,A]] - \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] \\
&\quad = \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] + \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2] \\
&\qquad - \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid X,A]] - \mathrm{Var}[w_{\mathrm{ATE}}(Y)\,Y] \\
&\quad = \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] - \mathrm{Var}[\mathbb{E}[\rho_{\mathrm{ATE}}(A,X)\,Y \mid X,A]] + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2] \\
&\quad = \mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X)] - \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X) \mid X]] - \mathrm{Var}[\rho_{\mathrm{ATE}}(A,X)\,\mu(A,X)] \\
&\qquad + \mathbb{E}[\mathrm{Var}[\rho_{\mathrm{ATE}}(A,X) \mid Y]\,Y^2] \\
&\quad = \mathbb{E}\left[\mathrm{Var}\left[\rho_{\mathrm{ATE}}(A,X) \mid Y\right]\,Y^2 - \mathrm{Var}\left[\rho_{\mathrm{ATE}}(A,X)\mu(A,X) \mid X\right]\right].
\end{aligned}
$$

□

Proposition A.5.2 shows that if $\mathrm{Var}\left[Y\,\rho_{\mathrm{ATE}}(A,X) \mid Y\right]$ is greater than the conditional variance $\mathrm{Var}\left[\rho_{\mathrm{ATE}}(A,X)\mu(A,X) \mid X\right]$ on average, the variance of the MR estimator will be less than that of the DR estimator. Intuitively, this is likely to happen when the dimension of context space $\mathcal{X}$ is high because in this case, the conditional variance over X and A, $\mathrm{Var}\left[Y\,\rho_{\mathrm{ATE}}(A,X) \mid Y\right]$ is likely to be greater than the conditional variance over A, $\mathrm{Var}\left[\rho_{\mathrm{ATE}}(A,X)\mu(A,X) \mid X\right]$.

A.6 Experimental Results

In this section, we provide additional experimental details for the results presented in the main text. We also include extensive experimental results to provide further empirical evidence in favour of the MR estimator.

Computational details We ran our experiments on Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz with 8GB RAM per core. We were able to use 150 CPUs in parallel to iterate over different configurations and seeds. However, we would like to note that for each run our algorithms only requires 1 CPU and at most 30 minutes to run as our neural networks are relatively small. Throughout our experiments, whenever the outcome Y was continuous, we used a fully connected neural network with three hidden layers with 512, 256 and 32 nodes respectively (and ReLU activation function) to estimate the weights $\hat{w}(y)$. On the other hand, when the outcome is discrete we can directly estimate $\hat{w}(y) \approx \mathbb{E}[\hat{\rho}(A, X) \mid Y = y]$ by calculating the sample mean of $\hat{\rho}(A, X)$ on samples with $Y = y$. Additionally, for each configuration of parameters in our experiments, we ran experiments for 10 different seeds (in $\{0, 1, \ldots, 9\}$).

A.6.1 Alternative methodology of estimating MR

In addition to the OPE baselines like IPW, DM and DR estimators considered in the main text, we also include empirically investigate an alternative methodology of estimating MR. Below we describe this methodology, denoted as 'MR (alt)', in greater detail:

MR (alt)

Recall our definition of MR estimator:

$$\hat{\theta}_{\mathrm{MR}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} w(y_i)\, y_i.$$

In the main text, we propose estimating the weights $w(y)$ first and using this to estimate $\hat{\theta}_{\mathrm{MR}}$ using the above expression. Alternatively, we can estimate $h(y) \coloneqq y\, w(y)$ using

$$h = \arg\min_{f} \mathbb{E}_{\pi^b} \left[\left(Y \frac{\pi^*(A|X)}{\pi^b(A|X)} - f(Y) \right)^2 \right].$$

Subsequently, the MR estimator can be written as:

$$\hat{\theta}_{\mathrm{MR}} = \frac{1}{n} \sum_{i=1}^{n} h(y_i).$$

We refer to this alternative methodology as 'MR-alt' and compare it empirically against the original methodology (which we simply refer to as 'MR'). In general, it is difficult to say which of the two methods will perform better. Intuitively speaking, in cases

where the behaviour of the quantity $Y \frac{\pi^*(A|X)}{\pi^b(A|X)}$ with varying Y is 'smoother' than that of $\frac{\pi^*(A|X)}{\pi^b(A|X)}$, the alternative method is expected to perform better. Our empirical results in the next sections show that the relative performance of the two methods varies for different data generating mechanisms.

A.6.2 Synthetic data experiments

Here, we include additional experimental details for the synthetic data experiments presented in Section 2.5.1 for completeness. For this experiment, we use the same setup as the synthetic data experiment in Saito and Joachims [2022], reproduced by reusing their code with minor modifications.

Setup Here, we sample the d-dimensional context vectors x from a standard normal distribution. The setup used also includes 3-dimensional categorical action embeddings $E \in \mathcal{E}$, which are sampled from the following conditional distribution given action $A = a$,

$$p(e \mid a) = \prod_{k=1}^{3} \frac{\exp\left(\alpha_{a,e_k}\right)}{\sum_{e' \in \mathcal{E}_k} \exp\left(\alpha_{a,e'}\right)},$$

which is independent of the context X. $\{\alpha_{a,e_k}\}$ is a set of parameters sampled independently from the standard normal distribution. Each dimension of $\mathcal{E}$ has a cardinality of 10, i.e., $\mathcal{E}_k = \{1, 2, \ldots, 10\}$.

Reward function The expected reward is then defined as:

$$q(x, e) = \sum_{k=1}^{3} \eta_k \cdot (x^T M x_{e_k} + \theta_x^T x + \theta_e^T x_{e_k}),$$

where M, θ_x and θ_e are parameter matrices or vectors sampled from a uniform distribution with range $[-1, 1]$. x_{e_k} is a context vector corresponding to the k-th dimension of the action embedding, which is unobserved to the estimators. η_k specifies the importance of the k-th dimension of the action embedding, sampled from Dirichlet distribution so that $\sum_{k=1}^{3} \eta_k = 1$.

Behaviour and target policies The behaviour policy π^b is defined by applying the softmax function to $q(x, a) = \mathbb{E}[q(X, E) \mid A = a, X = x]$ as

$$\pi^b(a \mid x) = \frac{\exp\left(-q(x, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(-q(x, a')\right)}.$$

For the target policy, we define the class of parametric policies,

$$\pi^{\alpha^*}(a|x) = \alpha^* \, \mathbb{1}(a = \arg\max_{a' \in \mathcal{A}} q(x, a')) + \frac{1 - \alpha^*}{|\mathcal{A}|},$$

where $\alpha^* \in [0, 1]$ controls the shift between the behaviour and target policies. As shown in the main text, as $\alpha^* \to 1$, the shift between behaviour and target policies increases.

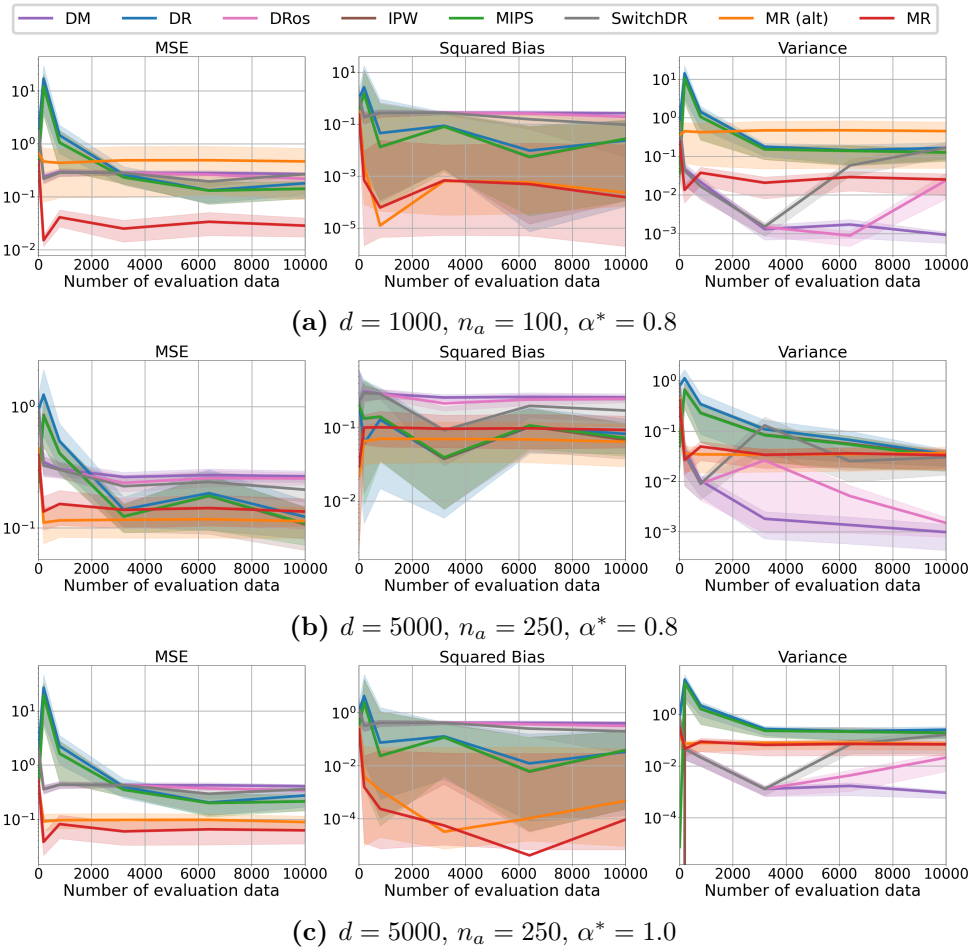
Baselines In the main text, we compare MR with DM, IPW, DR and MIPS estimators. In addition to these baselines, here we also consider Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020]. Following Saito and Joachims [2022], we use the random forest [Breiman, 2001] along with 2-fold cross-fitting [Newey and Robins, 2018] to obtain $\hat{q}(x, e)$ for DR and DM methods. To estimate $p_{\pi^b}(a \mid x, e)$ for MIPS estimator, we use logistic regression. We also include the results for MR estimated using the alternative methodology described in Section A.6.1. We refer to this as 'MR (alt)'.

Estimation of behaviour policy $\widehat{\pi}^b$ and marginal ratio $\hat{w}(y)$ We do not assume that the true behaviour policy π^b is known, and therefore estimate $\widehat{\pi}^b$ using the available training data. For the MR estimator, we estimate the behaviour policy using a random forest classifier trained on 50% of the training data and use the rest of the training data to estimate the marginal ratios $\hat{w}(y)$ using multi-layer perceptrons (MLP). Moreover, for a fair comparison we use a different behaviour policy estimate $\widehat{\pi}^b$ for all other baselines which is trained on the entire training data.

Results

For this experiment, the results are computed over 10 different sets of logged data replicated with different seeds, and in Figures A.3 - A.6 we use a total of $m = 5000$ training data.

Varying size of evaluation data n Figure A.3 shows that MR outperforms the other baselines, in terms of MSE and squared bias, when the number of evaluation data $n \leq 1000$. Additionally, we observe that in this experiment, MR estimated using our original methods ('MR'), yields better results than the alternative method of estimating MR ('MR (alt)'). Moreover, while the variance of DM is lower than that of MR, the DM method has a high bias and consequently a high MSE. We note that while the difference between MSE



(a) $d = 1000$, $n_a = 100$, $\alpha^* = 0.8$

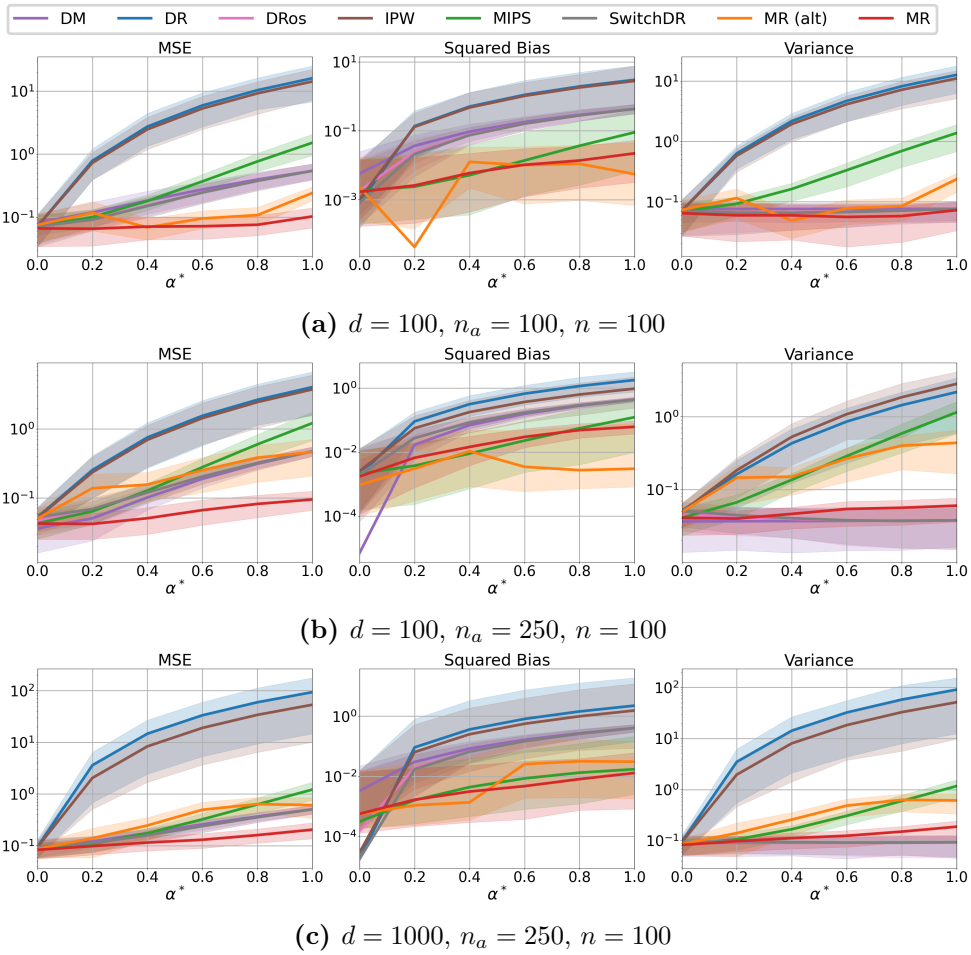
(b) $d = 5000$, $n_a = 250$, $\alpha^* = 0.8$

(c) $d = 5000$, $n_a = 250$, $\alpha^* = 1.0$

Figure A.3: MSE with varying size of evaluation dataset n for different choices of parameters.

and variance of MIPS and MR estimators decreases with increasing evaluation data size, MR still outperforms MIPS in terms of both MSE and variance.

Varying α^* Figure A.4 shows the results with increasing policy shift. It can be seen that overall MR methods achieve the smallest MSE with increasing policy shift. Moreover, the difference between MSE and variance of MR and IPW/DR methods increases with increasing policy shift, showing that MR performs especially better than these baselines when the difference between behaviour and target policies is large. Similarly, we observe in Figure A.4 that as the shift between the behaviour and target policy increases with increasing α^*, so does the difference between the MSE and variance of MR and the MIPS estimators. This shows that generally MR outperforms MIPS estimator in terms of variance and MSE, and that MR performs especially better than MIPS as the difference between behaviour and target policies increases.



(a) $d = 100$, $n_a = 100$, $n = 100$

(b) $d = 100$, $n_a = 250$, $n = 100$

(c) $d = 1000$, $n_a = 250$, $n = 100$

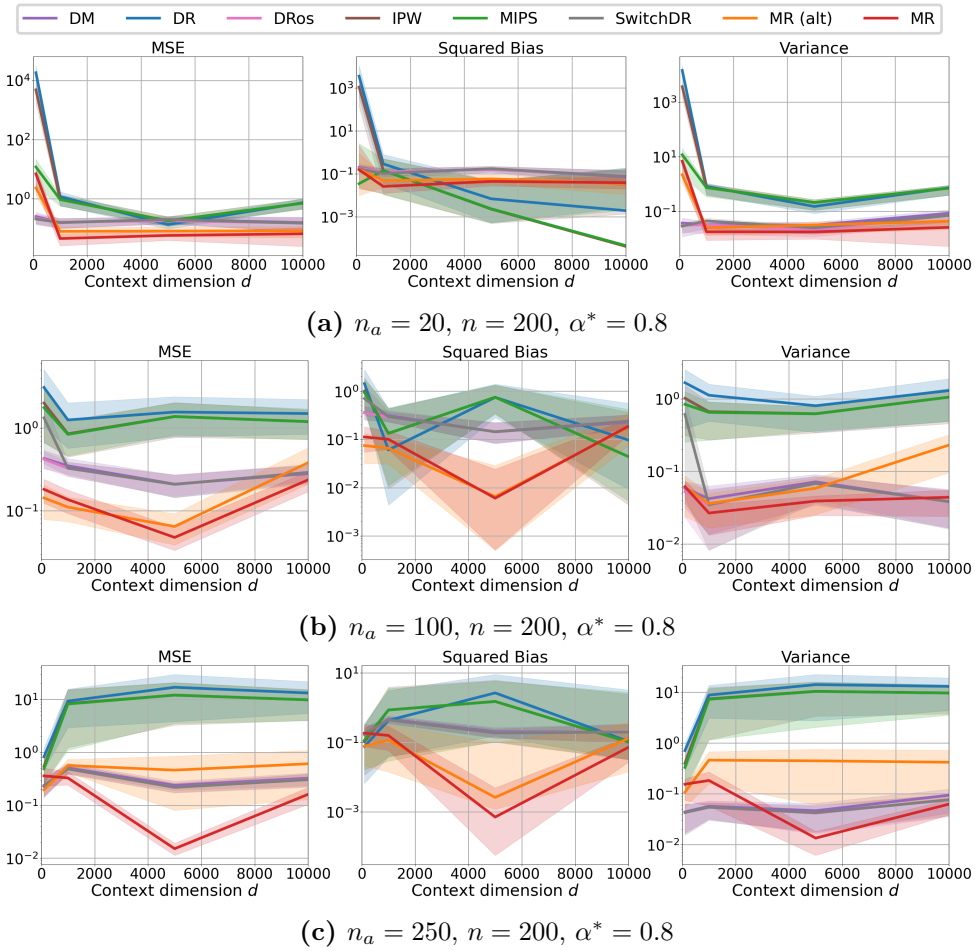
Figure A.4: MSE with varying α^* for different choices of parameters.

Varying d and n_a Figures A.5 and A.6 show that MR outperforms the other baselines as the context dimensions and/or number of actions increase. In fact, these figures show that MR is significantly robust to increasing dimensions of action and context spaces, whereas baselines like IPW and DR perform poorly in large action spaces.

Varying m Figure A.8 shows the results with increasing number of training data m. We again observe that the MR methods 'MR' and 'MR (alt)' outperforms the other baselines in terms of the MSE and squared bias even when the number of training data is low. Moreover, the variance of both the MR estimators continues to improve with increasing number of training data.

In this experiment, we observe that overall 'MR (alt)' performs worse than the original MR estimator ('MR' in the figures). However, as we observe in Appendix A.6.5, this does not happen consistently across all experiments, which suggests that the comparative

performance of the two MR methods depends on the data generating mechanism.



(a) $n_a = 20$, $n = 200$, $\alpha^* = 0.8$

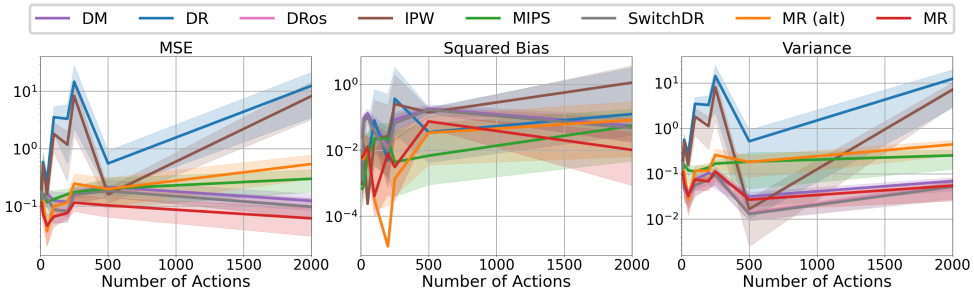
(b) $n_a = 100$, $n = 200$, $\alpha^* = 0.8$

(c) $n_a = 250$, $n = 200$, $\alpha^* = 0.8$

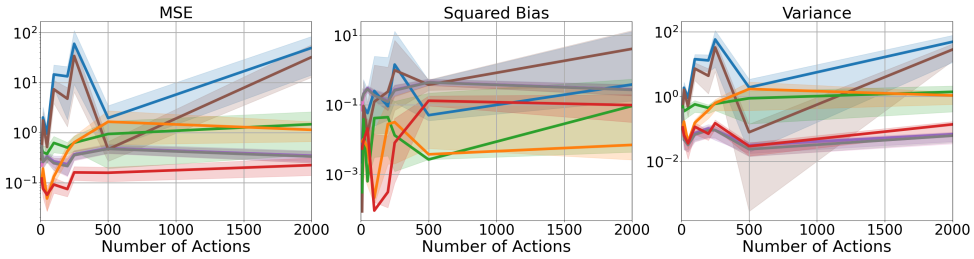
Figure A.5: MSE with varying context dimensions d for different choices of parameters.

Known policy ratios $\rho(a, x)$

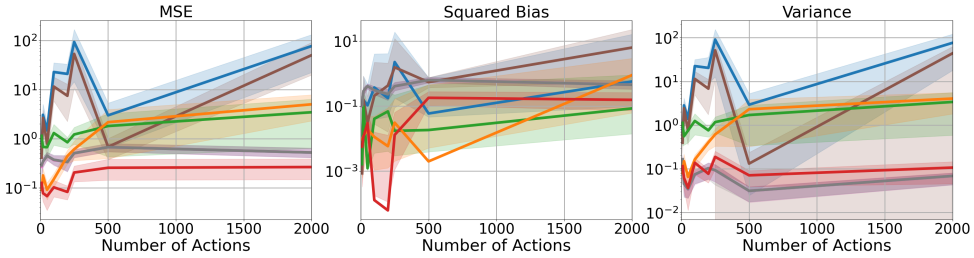
Our previous setting of unknown importance policy ratios $\rho(a, x)$ captures a wide variety of real-world applications, ranging from health care to autonomous driving. In addition, to demonstrate the utility of MR in settings with known $\rho(a, x), p(e \mid a, x)$ and unknown $w(y)$ (for our proposed method, MR), we have conducted additional experiments. Here, we use a fixed budget of datapoints (denoted by N) for each baseline and for MR we allocate $m = 2000$ of the available datapoints to estimate $\hat{w}(y)$ and use the remaining for evaluating the MR estimator (i.e., $n = N - 2000$ for MR). In contrast, for IPW and MIPS (since the importance ratios are already known), we use all of the N datapoints to evaluate the off-policy value (i.e. $n = N$ for IPW and MIPS).



(a) $d = 1000$, $n = 100$, $\alpha^* = 0.4$



(b) $d = 1000$, $n = 100$, $\alpha^* = 0.8$



(c) $d = 1000$, $n = 100$, $\alpha^* = 1.0$

Figure A.6: MSE with varying number of actions n_a for different choices of parameters.

The results included in Table A.1 show that MR achieves the smallest MSE among the baselines for $N \leq 6400$. However, we observe that the MSE of IPW, DR and MIPS (with true importance weights) falls below that of MR (with estimated weights $\hat{w}$) when the data size N is large enough (i.e., $N \geq 10,000$). This is to be expected since IPW, DR and MIPS are unbiased (i.e., use ground truth importance ratios $\rho(a, x)$) whereas MR uses estimated weights $\hat{w}(y)$ (and hence may be biased). MR still performs the best when $N \leq 6400$.

A.6.3 Experiments on classification datasets

Here, we conduct experiments on four classification datasets, OptDigits, PenDigits, SatImage and Letter datasets from the UCI repository [Dua and Graff, 2017], the Digits dataset from scikit-learn library, as well as the Mnist [Deng, 2012] and CIFAR-100 datasets [Krizhevsky, 2009].

Setup Following previous works [Dudík et al., 2014b, Kallus et al., 2021, Farajtabar et al., 2018b, Wang et al., 2017b], the classification datasets are transformed to contextual bandit feedback data. The classification dataset comprises $\{x_i, a_i^{\mathrm{gt}}\}_{i=1}^{n_0}$, where $x_i \in \mathcal{X}$ are feature vectors and $a_i^{\mathrm{gt}} \in \mathcal{A}$ are the ground-truth labels. In the contextual bandits setup, the feature vectors x_i are considered to be the contexts, whereas the actions correspond to the possible class of labels. We split the dataset into training and testing datasets of sizes m and n respectively. We present the results for a range of different values of m and n.

Reward function Let X be a context with ground truth label A^{gt}, we define the reward for action A as:

$$Y \coloneqq \mathbb{1}(A = A^{\mathrm{gt}}).$$

Behaviour and target policies Using the m training datapoints, we first train a classifier $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{A}|}$ which takes as input the feature vectors x_i and outputs a vector of softmax probabilities over labels, i.e. the a-th component of the vector $f(x)$, denoted as $(f(x))_a$ corresponds to the estimated probability $\mathbb{P}(A^{\mathrm{gt}} = a \mid X = x)$.

Next, we use f to define the ground truth behaviour policy,

$$\pi^b(a \mid x) = (f(x))_a.$$

For the target policies, we use f to define a parametric class of target policies using a trained classifier $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{A}|}$.

$$\pi^{\alpha^*}(a \mid x) = \alpha^* \cdot \mathbb{1}(a = \arg\max_{a' \in \mathcal{A}} (f(x))_{a'}) + \frac{1 - \alpha^*}{|\mathcal{A}|},$$

where $\alpha^* \in [0, 1]$. A value of α^* close to 1 leads to a near-deterministic and well-performing policy. As α^* decreases, the policy gets increasingly worse and 'noisy'. In this experiment, we consider target policies $\pi^* = \pi^{\alpha^*}$ for $\alpha^* \in \{0.0, 0.2, 0.4, \ldots, 1.0\}$.

Using the behaviour policy defined above, we generate the contextual bandits data described with training and evaluation datasets of sizes m and n respectively.

Estimation of behaviour policy $\widehat{\pi}^b$ and marginal ratio $\hat{w}(y)$ We do not assume that the behaviour policy π^b is known, and therefore estimate it using training data. To estimate the behaviour policy $\widehat{\pi}^b$, we train a random forest classifier using the training data. This estimate of behaviour policy is used for all the baselines in our experiment. Since the reward is binary, we can estimate the marginal ratios $\hat{w}(y) = \mathbb{E}_{\pi^b}[\hat{\rho}(A, X) \mid Y = y]$ by directly estimating the sample mean of $\hat{\rho}(A, X)$ for datapoints with $Y = y$. We re-use the m training datapoints to estimate this sample mean.

Baselines We compare our estimator with Direct Method (DM), IPW and DR estimators. In addition, we also consider Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020]. To estimate $\hat{q}(x, a)$ for DM and DR estimators, we use random forest classifiers (since reward Y is binary). Moreover, because of the binary nature of Y, the alternative method of estimating MR yields the same estimator as the original method, therefore we do not consider the two separately here. Additionally, in this experiment, we do not include MIPS (or G-MIPS) baseline, as there is no natural informative embedding E of the action A.

Results

For this experiment, we compute the results over 10 different sets of logged data replicated with different seeds. Figures A.9 - A.15 show the results corresponding to each baseline for the different datasets. It can be seen that across all datasets, the MR achieves the smallest MSE with increasing evaluation data size n. Moreover, across all datasets, MR attains the minimum MSE with relatively small number of evaluation data ($n \leq 100$).

Unlike the experiments in Section 2.5.1, we observe that the KL-divergence between target and behaviour policy decreases as α^* increases (see Figure A.7). Therefore, as α^* increases the shift between target and behaviour policies decreases. Figures A.9 - A.14 show that as α^* increases, the difference between the MSE, squared bias and variance of MR and the other baselines decreases. This confirms our findings from earlier experiments that MR performs especially better than the other baselines when the difference between behaviour and target policies is large.

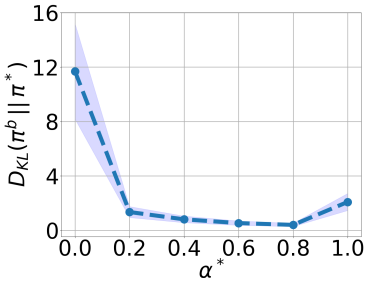
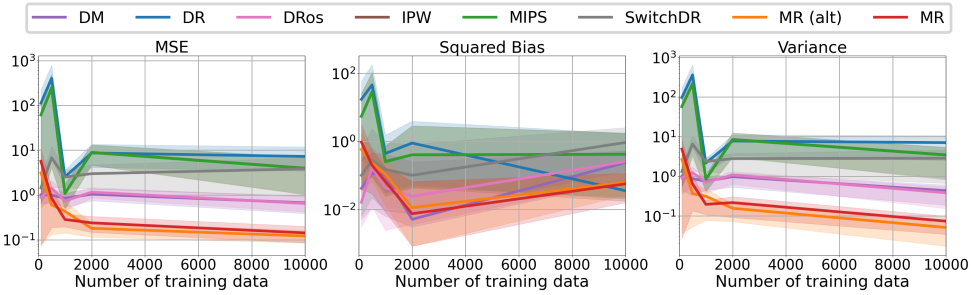
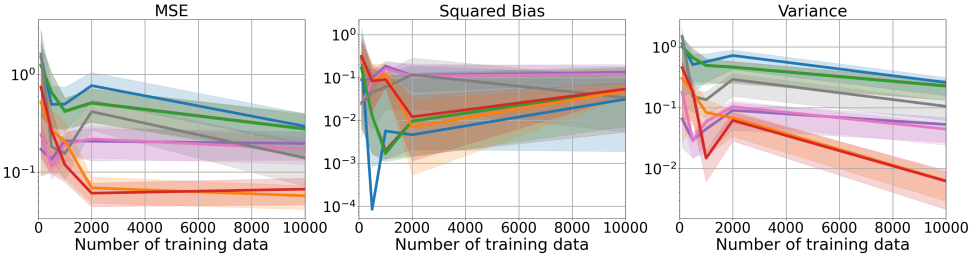


Figure A.7: KL divergence $D_{\mathrm{KL}}(\pi^b \,||\, \pi^*)$ with increasing α^* for the classification data experiments. Here, we only include the results for a specific choice of parameters for the Letter dataset. We observe similar results for other datasets and parameter choices.

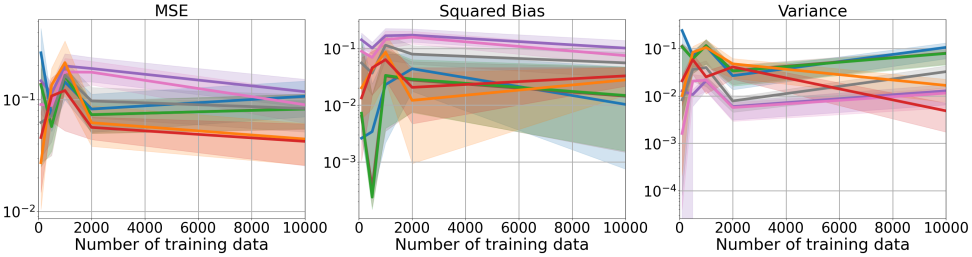
Moreover, the figures also include results with increasing number of training data m. It can be seen that MR out-performs the baselines even when the number of training data m is small ($m = 100$). Moreover, the relative advantage of MR improves with increasing m.



(a) $d = 1000$, $n = 10$, $\alpha^* = 0.8$

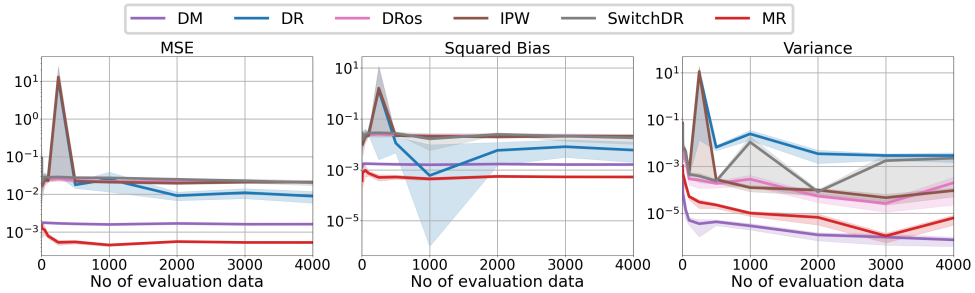


(b) $d = 1000$, $n = 200$, $\alpha^* = 0.8$

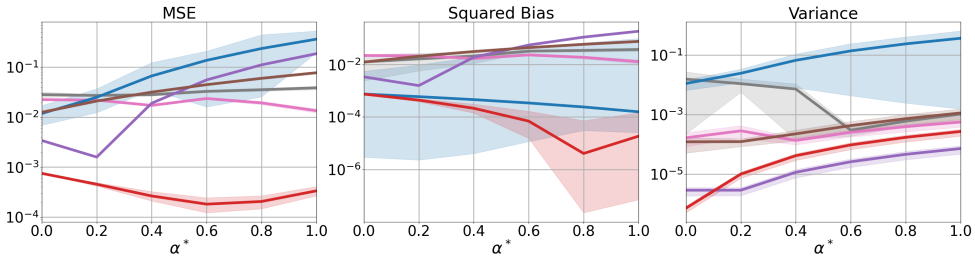


(c) $d = 1000$, $n = 800$, $\alpha^* = 0.8$

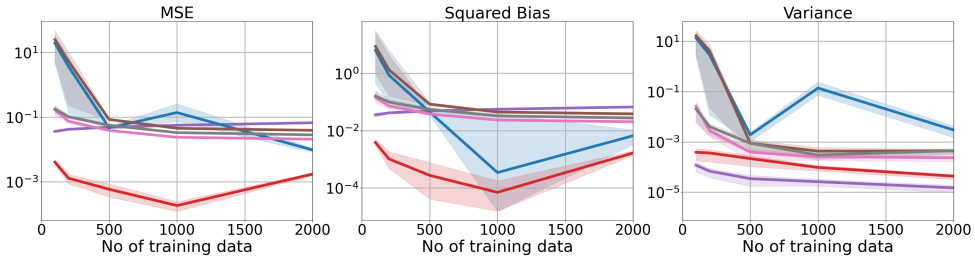
Figure A.8: MSE with varying number of training data m for different choices of parameters.



(a) Results with varying n for $\alpha^* = 0.2$ and $m = 1000$



(b) Results with varying α^* for $m = n = 1000$



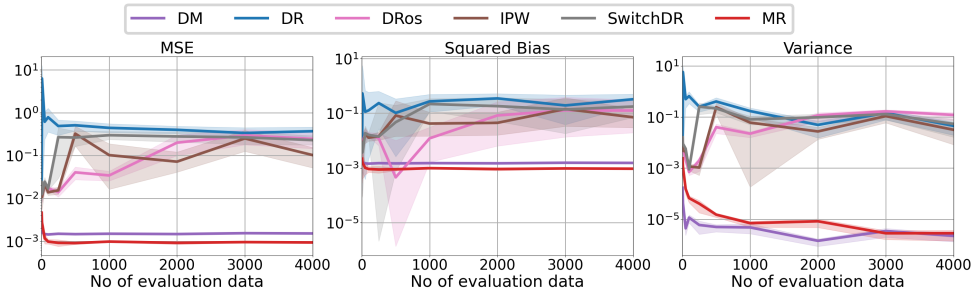
(c) Results with varying m for $n = 1000$ and $\alpha^* = 0.6$

Figure A.9: Results for OptDigits dataset

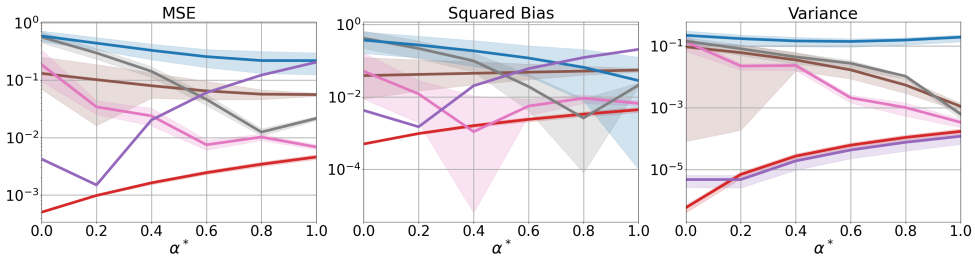
A.6.4 Application to Average Treatment Effect (ATE) estimation

In this subsection, we provide additional details for our experiment applying MR to the problem of ATE estimation presented in the main text. We begin by describing the dataset being used in this experiment.

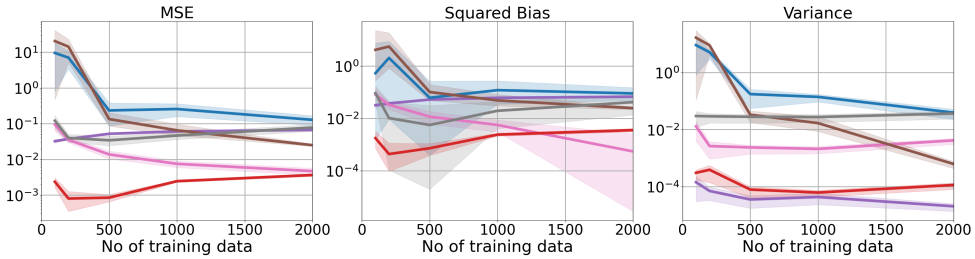
Twins dataset We use the Twins dataset as studied by Louizos et al. [2017], which comprises data from twin births in the USA between 1989-1991. The treatment $a = 1$ corresponds to being born the heavier twin and the outcome Y corresponds to the mortality of each of the twins in their first year of life. Since the data includes records for both twins, their outcomes would be considered as the two potential outcomes. Specifically, $Y(1)$ corresponds to the mortality of the heavier twin (and likewise for $Y(0)$). Closely following the methodology of Louizos et al. [2017], we only chose twins which are the same sex and weigh less than 2kgs. This provides us with a dataset of 11984 pairs of twins.



(a) Results with varying n for $\alpha^* = 0.2$ and $m = 1000$



(b) Results with varying α^* for $m = n = 1000$



(c) Results with varying m for $\alpha^* = 0.6$ and $n = 1000$

Figure A.10: Results for PenDigits dataset

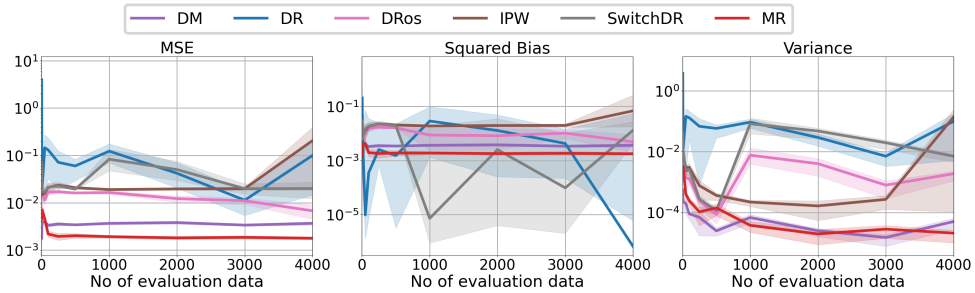
The mortality rate for the lighter twin is 18.9% and for the heavier twin is 16.4%, leading to the ATE value being $\theta_{\text{ATE}} = -2.5\%$. For each twin-pair we obtained 46 covariates relating to the parents, the pregnancy and birth.

Treatment assignment To simulate an observational study, we selectively hide one of the two twins by defining the treatment variable A which depends on the feature *GESTAT10*. This feature, which takes integer values from 0 to 9, is obtained by grouping the number of gestation weeks prior to birth into 10 groups. Then we sample actions A as follows,

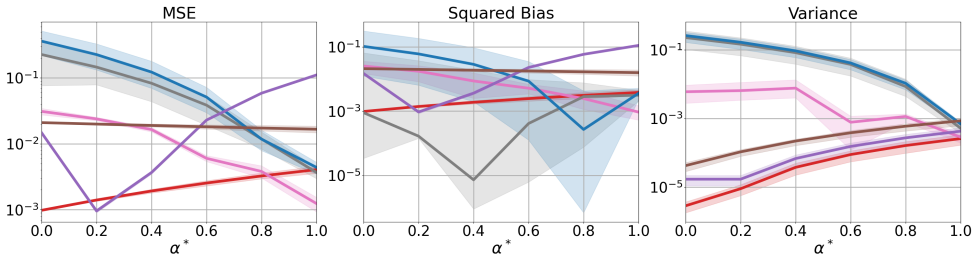
$$A \mid X \sim \text{Bern}(Z/10),$$

where Z is *GESTAT10*, and X are all the 46 features corresponding to a twin pair (including *GESTAT10*).

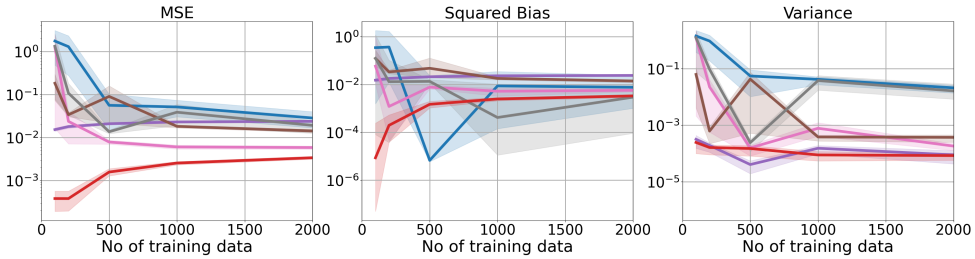
Using the treatment assignments defined above, we generate the observational data by selectively hiding one of the two twins from each pair. Next, we randomly split this dataset



(a) Results with varying n for $\alpha^* = 0.2$ and $m = 1000$



(b) Results with varying α^* for $n = 1000$



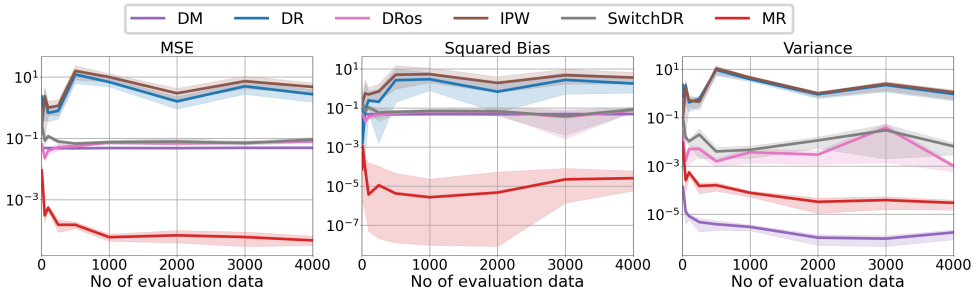
(c) Results with varying m for $\alpha^* = 0.6$ and $n = 1000$

Figure A.11: Results for SatImage dataset

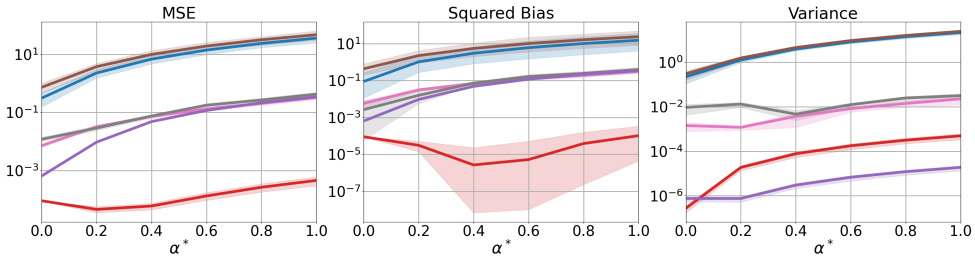
into training and evaluation datasets of sizes m and n respectively. In this experiment, we consider $m = 5000$ training datapoints.

Baselines Recall that ATE estimation can be formulated as the difference between off-policy values of deterministic policies $\pi^{(1)} := \mathbb{1}(A = 1)$ and $\pi^{(0)} := \mathbb{1}(A = 0)$. Therefore, any OPE estimator can be applied to ATE estimation. In this experiment, we compare our estimator against the baselines considered in our OPE experiments in Section A.6.3. This includes the Direct Method (DM), IPW and DR estimators as well as Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020]. To estimate $\hat{q}(x, a)$ for DM and DR estimators, we use multi-layer perceptrons (MLP) trained on the m training datapoints. Additionally, we estimate the behaviour policy $\widehat{\pi}^b$ using random forest classifier trained on the full training dataset.

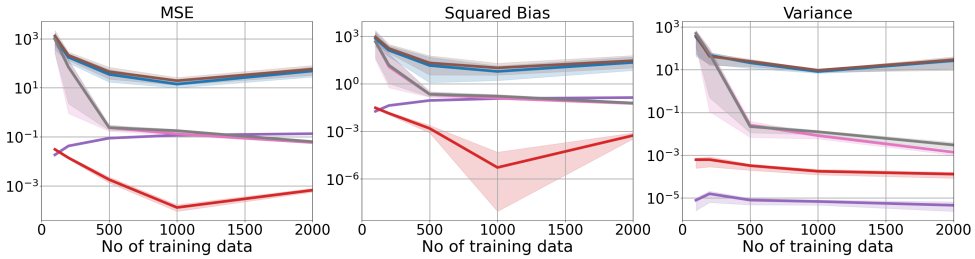
Since the outcome in this experiment is binary, we estimate the weights $w(y) =$



(a) Results with varying n for $\alpha^* = 0.2$ and $m = 1000$



(b) Results with varying α^* for $m = n = 1000$



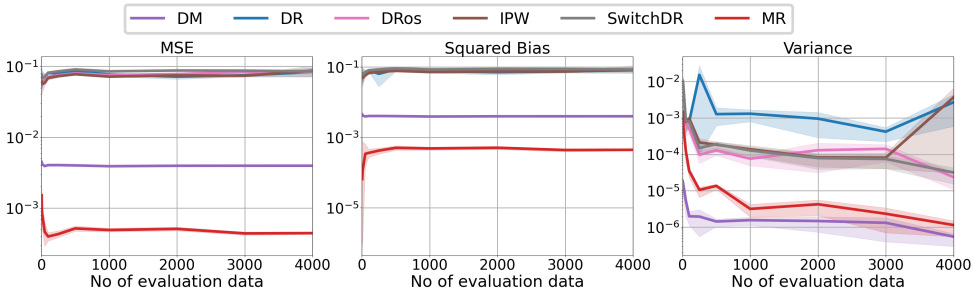
(c) Results with varying m for $\alpha^* = 0.6$ and $n = 1000$

Figure A.12: Results for Letter dataset

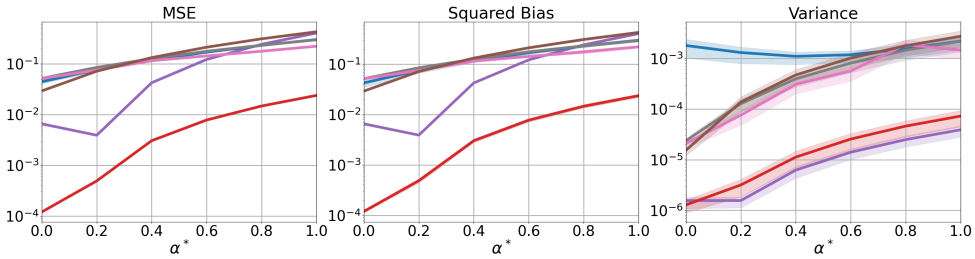
$\mathbb{E}_{\pi^b}[\hat{\rho}(A, X) \mid Y = y]$ directly by estimating the sample mean of $\hat{\rho}(A, X)$ for datapoints with $Y = y$. This means that the alternative method of estimating MR yields the same value as the default method. We therefore do not consider these estimators separately. Additionally, since there is no natural embedding R of the covariate-action space which satisfies the conditional dependence Assumption A.4.1, we do not consider the G-MIPS (or MIPS) estimator either.

Performance metric For our evaluation, we consider the absolute error in ATE estimation, ϵ_{ATE}, defined as:

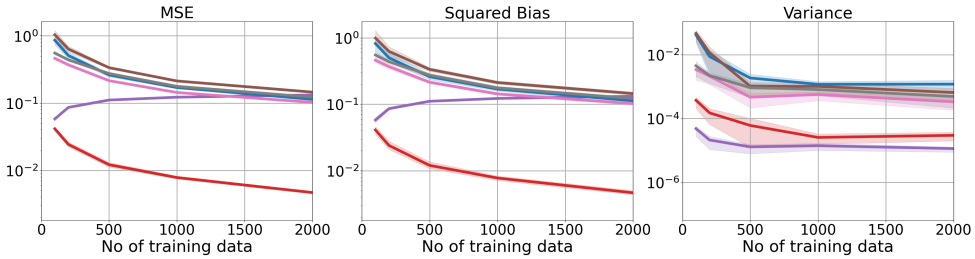
$$\epsilon_{\text{ATE}} := |\hat{\theta}^{(n)}_{\text{ATE}} - \theta_{\text{ATE}}|.$$



(a) Results with varying n for $\alpha^* = 0.2$ and $m = 1000$



(b) Results with varying α^* for $m = n = 1000$



(c) Results with varying m for $\alpha^* = 0.6$ and $n = 1000$

Figure A.13: Results for Mnist dataset

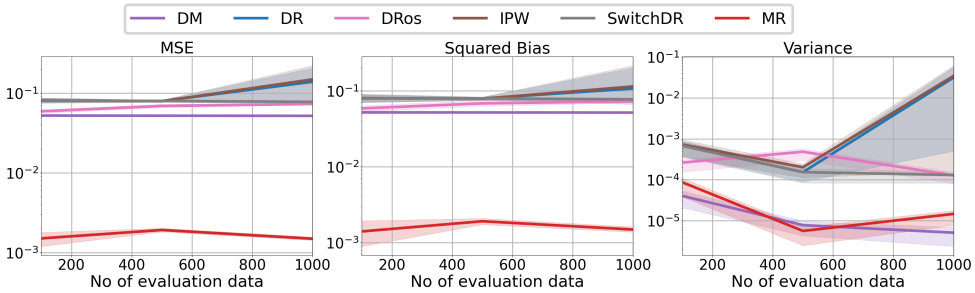
Here, $\hat{\theta}^{(n)}_{\mathrm{ATE}}$ denotes the value of the ATE estimated using n evaluation datapoints. For example, for the IPW estimator, the $\hat{\theta}^{(n)}_{\mathrm{ATE}}$ can be written as:

$$\hat{\theta}^{(n)}_{\mathrm{ATE}} = \widehat{\mathrm{ATE}}_{\mathrm{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}(a_i = 1) - \mathbb{1}(a_i = 0)}{\widehat{\pi}^b(a_i \mid x_i)}\right) y_i.$$

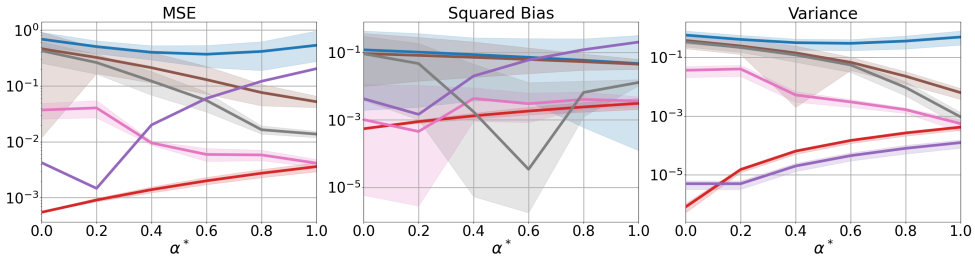
All results for this experiment are provided in the main text.

A.6.5 Additional synthetic data experiments

In addition to the synthetic data experiments provided in Section 2.5.1, we also consider an additional synthetic data setup to obtain further empirical evidence in favour of the MR estimator, and also compare it against the generalised version of the MIPS estimator (described as G-MIPS in Appendix A.4). Here, we use a similar setup to Saito and Joachims [2022] (albeit without action embeddings E) where the d-dimensional context vectors x are sampled from a standard normal distribution. Likewise, the action space



(a) Results with varying n for $\alpha^* = 0.2$ and $m = 500$



(b) Results with varying α^* for $n = 500$ and $m = 1000$

(c) Results with varying m for $\alpha^* = 0.6$ and $n = 500$

Figure A.14: Results for Digits dataset. Note that compared to other datasets we consider smaller maximum dataset sizes m, n here as the total number of available datapoints was 1797.

is finite and comprises of n_a actions, i.e. $\mathcal{A} = \{0, \ldots, n_a - 1\}$, with n_a taking a range of different values. The reward function is defined as follows:

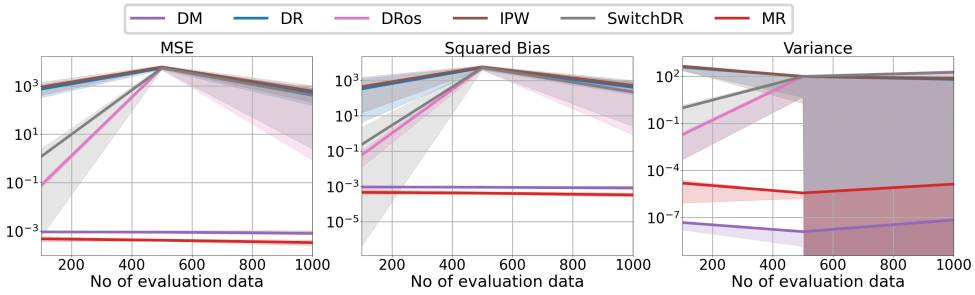
Reward function The expected reward $q(x, a) := \mathbb{E}[Y \mid x, a]$ for these experiments is defined as follows:

$$q(x, a) = \sin\left(a \cdot ||x||_2\right).$$

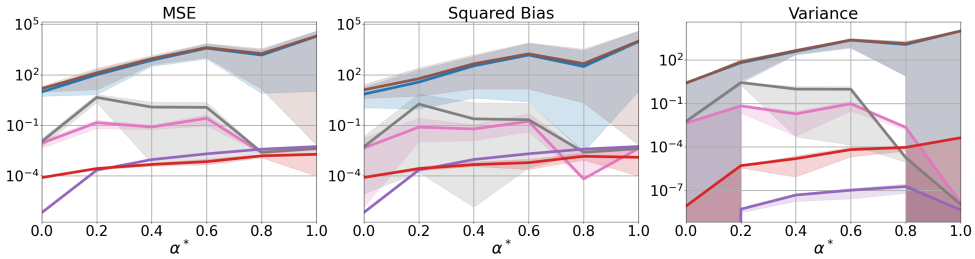
The reward Y is obtained by adding a normal noise random variable to $q(x, a)$

$$Y = q(X, A) + \epsilon,$$

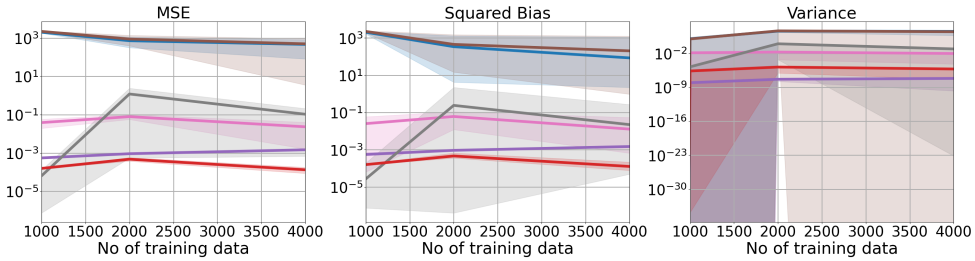
where $\epsilon \sim \mathcal{N}(0, 0.01)$. Here, it can be seen that conditional on $R = (||X||_2, A)$, the reward Y does not depend on (X, A), i.e., the embedding R satisfies the conditional independence assumption $Y \perp\!\!\!\perp (X, A) \mid R$.



(a) Results with varying n for $\alpha^* = 0.4$ and $m = 2000$



(b) Results with varying α^* for $n = 100$ and $m = 2000$



(c) Results with varying m for $\alpha^* = 0.4$ and $n = 100$

Figure A.15: Results for CIFAR-100 dataset.

Behaviour and target policies We first define a behaviour policy by applying softmax function to $q(x, a)$ as

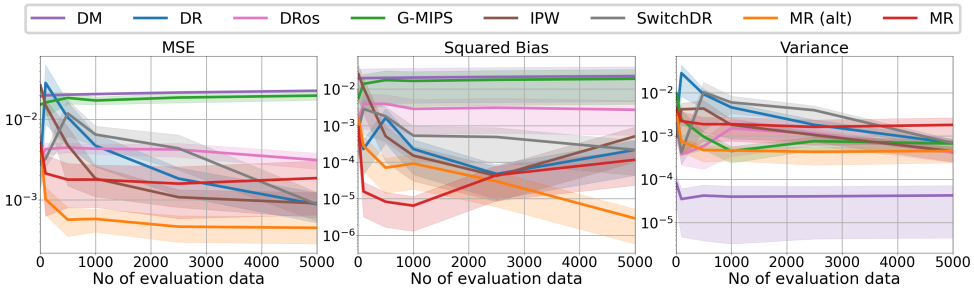
$$\pi^b(a \mid x) = \frac{\exp\left(q(x,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(q(x,a')\right)}.$$

Just like in Section 2.5.1, to investigate the effect of increasing policy shift, we define a class of policies,

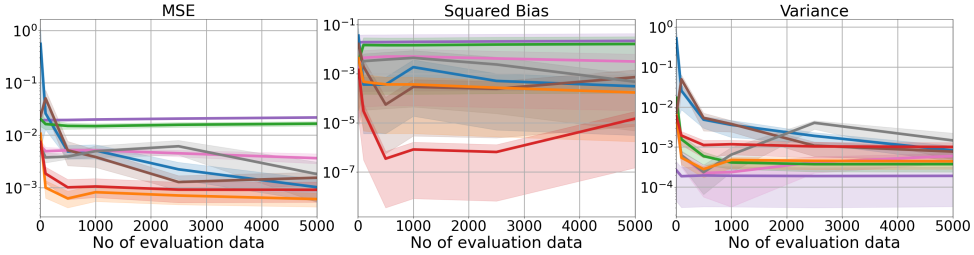
$$\pi^{\alpha^*}(a|x) = \alpha^* \, \mathbb{1}(a = \arg\max_{a' \in \mathcal{A}} q(x,a')) + \frac{1-\alpha^*}{|\mathcal{A}|} \quad \text{where} \quad q(x,a) := \mathbb{E}[Y \mid X = x, A = a],$$

where $\alpha^* \in [0, 1]$ allows us to control the shift between π^b and π^*. Again, the shift between π^b and π^* increases as $\alpha^* \to 1$. Using the ground truth behaviour policy π^b, we generate a dataset which is split into training and evaluation datasets of sizes m and n respectively.

In Figures A.16 - A.19, we present the results for this experimental setup for different choices of paramater configurations.

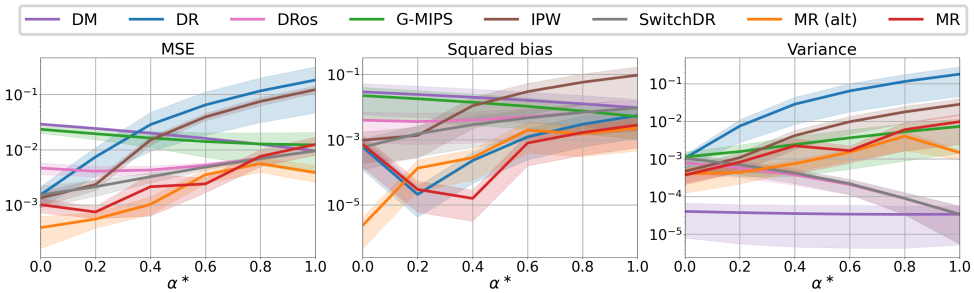


(a) $d = 1000$, $n_a = 100$, $\alpha^* = 0.4$.

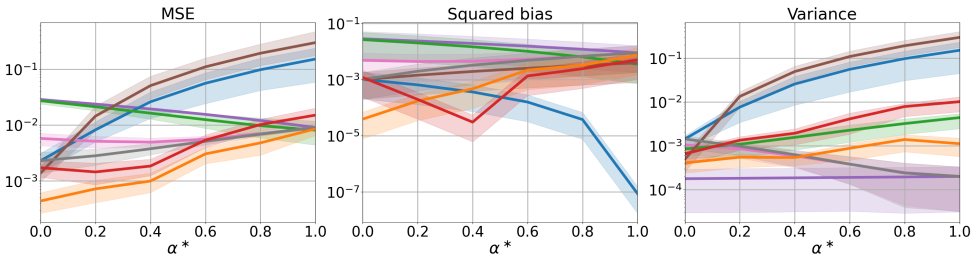


(b) $d = 10000$, $n_a = 100$, $\alpha^* = 0.4$.

Figure A.16: Results with varying size of evaluation dataset n.



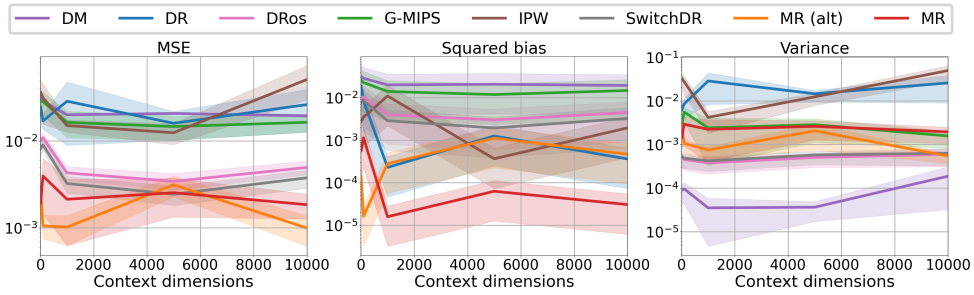
(a) $d = 1000$, $n_a = 100$, $n = 100$.



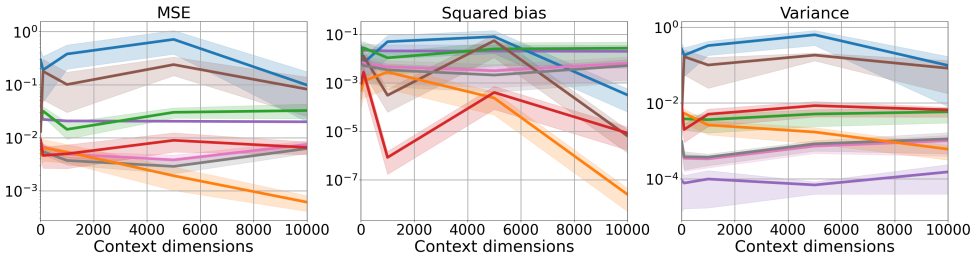
(b) $d = 10000$, $n_a = 100$, $n = 100$.

Figure A.17: Results with varying α^*.

Estimation of behaviour policy $\widehat{\pi}^b$ and marginal ratio $\hat{w}(y)$ For the MR estimator, we estimate the behaviour policy using a random forest classifier trained on 50% of the training data and use the rest of the training data to estimate the marginal ratios $\hat{w}(y)$ using multi-layer perceptrons (MLP). Moreover, for a fair comparison we use a different behaviour policy estimate $\widehat{\pi}^b$ for all other baselines which is trained on the entire training data.

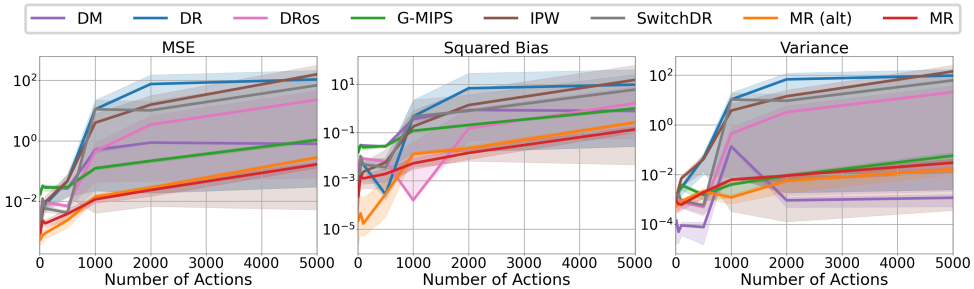


(a) $n_a = 100$, $n = 100$, $\alpha^* = 0.4$.

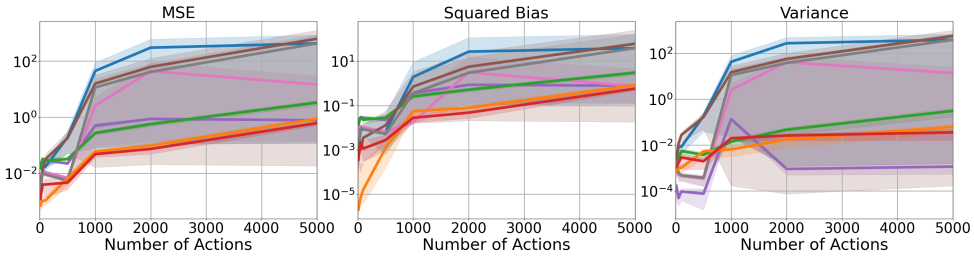


(b) $n_a = 500$, $n = 100$, $\alpha^* = 0.4$.

Figure A.18: Results with varying context dimensions d.



(a) $d = 100$, $n = 100$, $\alpha^* = 0.2$.



(b) $d = 100$, $n = 100$, $\alpha^* = 0.4$.

Figure A.19: Results with varying number of actions n_a.

Additional Baselines In addition to the baselines considered in the main text (Section 2.5.1), we also consider Switch-DR [Wang et al., 2017b] and DR with Optimistic Shrinkage (DRos) [Su et al., 2020]. In addition, we also include the results for MR estimated using the alternative method ('MR (alt)') outlined in Section A.6.1. For the G-MIPS estimator

(defined in Appendix A.4) considered here, we use $R = (a, ||x||_2)^2$. To estimate $\hat{q}(x, a)$ for DM and DR estimators, we use multi-layer perceptrons (MLPs).

Results

For this experiment, the results are computed over 10 different sets of logged data replicated with different seeds, and in Figures A.16 - A.19 we use a total of $m = 5000$ training data.

Varying n Figure A.16 shows that MR outperforms the other baselines, in terms of MSE and squared bias, when the number of evaluation data $n \leq 1000$. Additionally, we observe that in this experiment, MR esitmated using alternative methods, MR (alt), yields better results than the original method of estimating MR. Moreover, while the variance of DM is lower than that of MR, the DM method has a high bias and consequently a high MSE.

Varying α^* Figure A.17 shows the results with increasing policy shift. It can be seen that overall MR methods achieve the smallest MSE with increasing policy shift. Moreover, the difference between MSE and variance of MR and IPW/DR methods increases with increasing policy shift, showing that MR performs especially better than these baselines when the difference between behaviour and target policies is large.

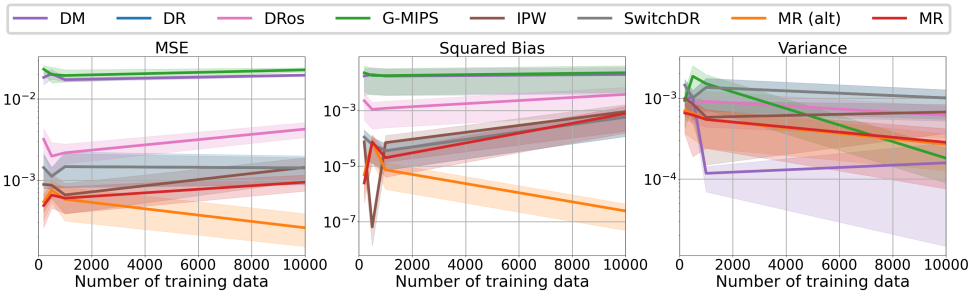
Varying d **and** n_a Figures A.18 and A.19 show that MR outperforms the other baselines as the context dimensions and/or number of actions increase. In fact, Figure A.19 shows that MR is significantly robust to increasing action space, whereas baselines like IPW and DR perform poorly in large action spaces.

Varying m Figure A.20 shows the results with increasing number of training data m. We again observe that the MR methods 'MR' and 'MR (alt)' outperforms the other baselines in terms of the MSE and squared bias even when the number of training data is low. Moreover, the variance of both the MR estimators continues to improve with increasing number of training data.

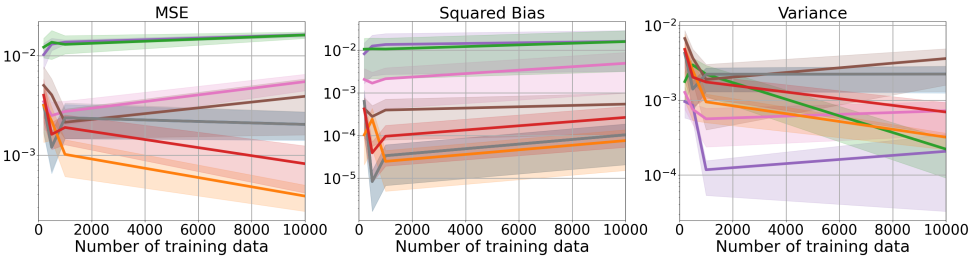
Unlike our experimental results in Section A.6.2, 'MR (alt)' performs better than the original MR estimator overall. This shows that one of these two methods is not

[2]It is easy to see that in our setup, the embedding $R = (a, ||x||_2)$ satisfies the conditional independence assumption $Y \perp\!\!\!\perp (X, A) \mid R$ needed for G-MIPS estimator to be unbiased

better than the other consistently in all cases, and their relative performance depends on the dataset under consideration.



(a) $d = 5000$, $n = 100$, $n_a = 10$, $\alpha^* = 0.2$.



(b) $d = 5000$, $n = 100$, $n_a = 10$, $\alpha^* = 0.4$.

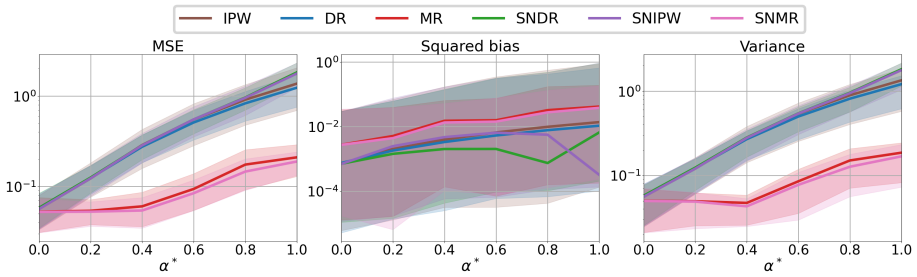
Figure A.20: Results with varying number of training data m.

A.6.6 Self-normalised MR estimator

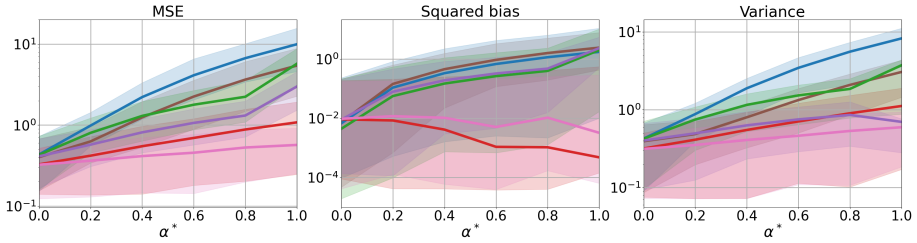
Self-normalization trick has been used in practice to reduce the variance in off-policy estimators [Swaminathan and Joachims, 2015b]. This technique is also applicable to the MR estimator, and leads to the self-normalized MR estimator (denoted as θ_{SNMR}) defined as follows:

$$\theta_{\text{SNMR}} := \sum_{i=1}^{n} \frac{w(Y_i)}{\sum_{j=1}^{n} w(Y_j)} Y_i.$$

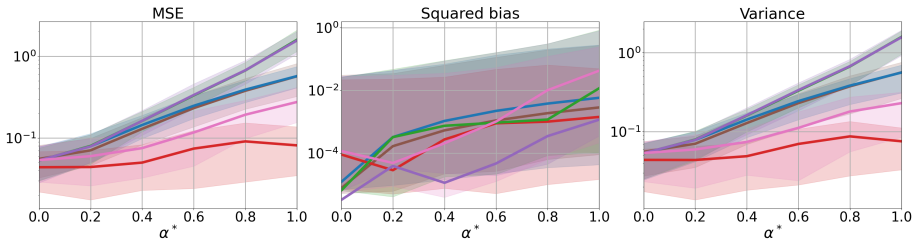
We conducted experiments to investigate the effect of self-normalisation on the performance of the IPW, DR and MR estimators. Figure A.21 shows results for three different choices of parameter configurations. Overall, we observe that in all settings, the MR and self-normalised MR (SNMR) estimator outperform all other baselines including the self-normalised IPW and DR estimators (denoted as SNIPW and SNDR respectively). Moreover, in some settings, where the importance ratios achieve very high values, self-normalisation can reduce the variance and MSE of the corresponding estimator (for example, Figure A.21b). However, we also observe cases in which self-normalization does not significantly change the results (Figure A.21a), or may even slightly worsen the MSE of the estimators (Figure A.21c).



(a) $d = 10000$, $n = 200$, $n_a = 20$, $m = 5000$.



(b) $d = 5000$, $n = 200$, $n_a = 20$, $m = 1000$.



(c) $d = 10000$, $n = 200$, $n_a = 20$, $m = 5000$.

Figure A.21: Results for self-normalised estimators with varying target policy shift α^* for synthetic data setup considered in Section 2.5.1. Here, "SN" denotes self-normalised estimators.

Table A.1: Mean-squared error results with 2 standard errors for synthetic data setup considered in Section 2.5.1 with $d = 5000$, $n_a = 50$, $\alpha^* = 0.8$. We use a fixed budget of datapoints (denoted by N) for each baseline and in the case of MR we use $m = 2000$ of the available datapoints to estimate $\hat{w}(y)$ and the rest of data to evaluate the MR estimator (i.e. $n = N - 2000$ for MR). In contrast, for IPW and MIPS since the importance ratios are already known, we use all of the N datapoints for evaluation of the off-policy value (i.e. $n = N$ for IPW and MIPS).

	N	2800	3200	6400	10000	12000
GT weights $\rho(a,x)$ and estimated reward model $\hat{\mu}(a,x)$ ($m = 2000$ used for training $\hat{\mu}(a,x)$ and $n = N - 2000$ used for evaluation)	DM	0.137±0.028	0.099±0.012	0.103±0.012	0.093±0.010	0.089±0.010
	DR	0.227±0.065	0.068±0.035	0.068±0.022	**0.024±0.011**	0.045±0.015
	DRos	0.128±0.027	0.072±0.011	0.049±0.014	0.063±0.014	0.051±0.016
	SwitchDR	0.128±0.027	0.059±0.014	0.052±0.013	0.061±0.015	0.056±0.016
GT weights (all of N datapoints are used for evaluation)	IPW	0.237±0.062	0.066±0.036	0.067±0.021	0.025±0.011	**0.044±0.014**
	MIPS	0.236±0.062	0.065±0.035	0.067±0.021	0.025±0.011	**0.044±0.014**
Estimated weights $\hat{w}(y)$ ($m = 2000$ used for training and $n = N - 2000$ used for evaluation)	MR (Ours)	**0.045±0.015**	**0.042±0.014**	**0.048±0.020**	0.049±0.020	0.047±0.016

B

Conformal Off-Policy Prediction in Contextual Bandits

Contents

B.1 Proofs

B.1.1 Proof of Proposition 3.4.1

This proof is a direct adaptation of [Tibshirani et al., 2019, Lemma 3], and has only been included for the sake of completeness.

In this proof, we use the notion of *weighted exchangeability* as defined in Section 3.2 of Tibshirani et al. [2019].

Definition B.1.1 (Weighted exchangeability)

Random variables $V_1, \ldots, V_n$ are said to be *weighted exchangeable* with weight functions $w_1, \ldots, w_n$, if the density f of their joint distribution can be factorized as

$$f(v_1, \ldots, v_n) = \prod_{i=1}^{n} w_i(v_i) g(v_1, \ldots, v_n) \tag{B.1}$$

where g is any function that does not depend on the ordering of its inputs, i.e. $g(v_{\sigma(1)}, \ldots, v_{\sigma(n)}) = g(v_1, \ldots, v_n)$ for any permutation σ of $1, \ldots, n$.

Lemma B.1.1

Let $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, ..., n+1$, be such that $\{(X_i, Y_i)\}_{i=1}^{n} \overset{\text{i.i.d.}}{\sim} P^{\pi^b}_{X,Y}$ and $(X_{n+1}, Y_{n+1}) \sim P^{\pi^*}_{X,Y}$. Then $Z_1, \ldots, Z_{n+1}$ are weighted exchangeable with weights $w_i \equiv 1$, $i \leq n$ and $w_{n+1}(X, Y) = \mathrm{d}P^{\pi^*}_{X,Y} / \mathrm{d}P^{\pi^b}_{X,Y}(X, Y)$.

Proof. The proof below is merely a verification that our proposed weights still retain the coverage guarantees and is mainly taken from Tibshirani et al. [2019]. Hence, we follow the same strategy as in Tibshirani et al. [2019], with the exception that we have the weights as in Lemma B.1.1, hence inducing a lot of simplifications. As in Tibshirani et al. [2019], we assume for simplicity that $V_1, \ldots, V_{n+1}$ are distinct almost surely, however the result holds in general case as well. We define f as the joint distribution of the random variables $\{X_i, Y_i\}_{i=1}^{n+1}$. We also denote E_z as the event of $\{Z_1, \ldots, Z_{n+1}\} = \{z_1, \ldots, z_{n+1}\}$ and let $v_i = s(z_i) = s(x_i, y_i)$, then for each i:

$$\mathbb{P}\{V_{n+1} = v_i | E_z\} = \mathbb{P}\{Z_{n+1} = z_i | E_z\} = \frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_{\sigma} f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})} \tag{B.2}$$

Now using the fact that $Z_1, \ldots, Z_{n+1}$ are weighted exchangeable:

$$\begin{aligned} \frac{\sum_{\sigma:\sigma(n+1)=i} f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_{\sigma} f(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})} &= \frac{\sum_{\sigma:\sigma(n+1)=i} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})}{\sum_{\sigma} \prod_{j=1}^{n+1} w_j(z_{\sigma(j)}) g(z_{\sigma(1)}, \ldots, z_{\sigma(n+1)})} \\ &= \frac{w_{n+1}(z_i) g(z_1, \ldots, z_{n+1})}{\sum_{j=1}^{n+1} w_{n+1}(z_j) g(z_1, \ldots, z_{n+1})} \\ &= p_i^w(z_{n+1}) \end{aligned} \tag{B.3}$$

where we recall that

$$p_i^w(x, y) := \frac{w(X_i, Y_i)}{\sum_{j=1}^{n} w(X_j, Y_j) + w(x, y)}.$$

We get simplifications in (B.3) due to the weights defined in Lemma B.1.1, i.e. $w_i \equiv 1$ for $i \leq n$ and $w_{n+1}(x,y) = w(x,y) = \mathrm{d}P^{\pi^*}_{X,Y}/\mathrm{d}P^{\pi^b}_{X,Y}(x,y)$. Next, just as in Tibshirani et al. [2019] we can view:

$$V_{n+1} = v_i | E_z \sim \sum_{i=1}^{n+1} p_i^w(z_{n+1})\delta_{v_i} \tag{B.4}$$

which implies that:

$$\mathbb{P}\{V_{n+1} \leq \text{Quantile}_\beta(\sum_{i=1}^{n+1} p_i^w(z_{n+1})\delta_{v_i}) | E_z\} \geq \beta.$$

This is equivalent to

$$\mathbb{P}\{V_{n+1} \leq \text{Quantile}_\beta(\sum_{i=1}^{n+1} p_i^w(Z_{n+1})\delta_{v_i}) | E_z\} \geq \beta$$

and, after marginalizing, one has

$$\mathbb{P}\{V_{n+1} \leq \text{Quantile}_\beta(\sum_{i=1}^{n+1} p_i^w(Z_{n+1})\delta_{v_i})\} \geq \beta$$

This is equivalent to the claim in Proposition 3.4.1. □

B.1.2 Proof of Proposition 3.4.2

The following proof is an adaptation of [Lei and Candès, 2021, Proposition 1] to our setting.

Before detailing the main proof, we introduce a preliminary result which will be used in the proof of Proposition 3.4.2.

Lemma B.1.2

Let $\hat{w}(x,y)$ be an estimate of the weights $w(x,y) = \mathrm{d}P^{\pi^*}_{X,Y}/\mathrm{d}P^{\pi^b}_{X,Y}(x,y)$, and

$$(\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)^r])^{1/r} \leq M_r < \infty$$

for some $r \geq 2$. Let $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} P^{\pi^b}_{X,Y}$ and $\mathcal{A}$ denote the event that

$$\sum_{i=1}^{n} \hat{w}(X_i, Y_i) \leq n/2.$$

Then,

$$\mathbb{P}(\mathcal{A}) \leq \frac{c_1 M_r^2}{n}$$

where c_1 is an absolute constant, and the probability is taken over $\{X_i, Y_i\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P^{\pi^b}_{X,Y}$.

Proof of Lemma B.1.2

The condition $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)^r] < \infty \implies \mathbb{P}_{(X,Y)\sim P^{\pi^b}_{X,Y}}(\hat{w}(X,Y) < \infty) = 1$ and $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)] < \infty$. WLOG assume $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)] = 1$. Recall that $p_i^{\hat{w}}(x,y) \coloneqq \frac{\hat{w}(X_i,Y_i)}{\sum_{i=1}^n \hat{w}(X_i,Y_i)+\hat{w}(x,y)}$, and therefore, $p_i^{\hat{w}}(x,y)$ are invariant to weight scaling. Since $\mathbb{E}_{(X_i,Y_i)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X_i,Y_i)]^2 \leq M_r^2$ and $\mathbb{E}_{(X_i,Y_i)\sim P^{\pi^b}_{X,Y}}(\hat{w}(X_i,Y_i)) = 1$, using Chebyshev's inequality

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^n \hat{w}(X_i,Y_i) \leq n/2\right) =& \mathbb{P}\left(\sum_{i=1}^n (\hat{w}(X_i,Y_i)-1) \leq -n/2\right) \\
\leq& \mathbb{P}\left(|\sum_{i=1}^n (\hat{w}(X_i,Y_i)-1)| \geq n/2\right) \\
\leq& \frac{4}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n \hat{w}(X_i,Y_i) - \mathbb{E}[\hat{w}(X_i,Y_i)]\right)^2\right] \\
=& \frac{4}{n^2}\left\{n\mathbb{E}|\hat{w}(X_1,Y_1) - \mathbb{E}[\hat{w}(X_1,Y_1)]|^2\right\} \qquad \text{(B.5)} \\
\leq& \frac{16}{n^2} n\mathbb{E}|\hat{w}(X_1,Y_1)|^2 \qquad \text{(B.6)} \\
\leq& \frac{c_1 M_r^2}{n}
\end{aligned}$$

where to get from (B.5) to (B.6) we use:

$$\begin{aligned}
\mathbb{E}|\hat{w}(X_1,Y_1) - \mathbb{E}[\hat{w}(X_1,Y_1)]|^2 &\leq 2\mathbb{E}\left[\hat{w}(X_1,Y_1)^2 + \mathbb{E}[\hat{w}(X_1,Y_1)]^2\right] \\
&\leq 4\mathbb{E}[\hat{w}(X_1,Y_1)^2].
\end{aligned}$$

□

We can now prove Proposition 3.4.2.

Proof. The condition $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)^r] < \infty \implies \mathbb{P}_{(X,Y)\sim P^{\pi^b}_{X,Y}}(\hat{w}(X,Y) < \infty) = 1$ and $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)] < \infty$. WLOG assume $\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)] = 1$. Let $\tilde{P}^{\pi^*}_{X,Y}$ be a probability measure with

$$\mathrm{d}\tilde{P}^{\pi^*}_{X,Y}(x,y) \coloneqq \hat{w}(x,y)\mathrm{d}P^{\pi^b}_{X,Y}(x,y)$$

and $(\tilde{X},\tilde{Y}) \sim \tilde{P}^{\pi^*}_{X,Y}$ that is independent of the data. By Hölder's inequality,

$$\begin{aligned}
\mathbb{E}_{(\tilde{X},\tilde{Y})\sim\tilde{P}^{\pi^*}_{X,Y}}[\hat{w}(\tilde{X},\tilde{Y})] =& \int_{\tilde{x},\tilde{y}} \frac{\mathrm{d}\tilde{P}^{\pi^*}(\tilde{x},\tilde{y})}{\mathrm{d}P^{\pi^b}(\tilde{x},\tilde{y})}\mathrm{d}\tilde{P}^{\pi^*}(\tilde{x},\tilde{y}) \\
=& \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[\hat{w}(X,Y)^2] \leq M_r^2 < \infty
\end{aligned}$$

Note using Proposition 3.4.1 with $(\tilde{X}, \tilde{Y})$ denoting (X_{n+1}, Y_{n+1}) for simplicity

$$\begin{aligned}&\mathbb{P}(\tilde{Y} \in \hat{C}(\tilde{X}, \tilde{Y})) \\ &\quad = \mathbb{E}_{(\tilde{X},\tilde{Y})\sim P^{\pi^*}_{X,Y}}\left[\mathbb{P}\left(s(\tilde{X}, \tilde{Y}) \leq \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n} p_i^{\hat{w}}(\tilde{X}, \tilde{Y})\delta_{V_i} + p_{n+1}^{\hat{w}}(\tilde{X}, \tilde{Y})\delta_\infty\right) \mid \mathcal{E}(\tilde{V})\right)\right]\end{aligned} \tag{B.7}$$

where $\mathcal{E}(\tilde{V})$ denotes the unordered set of $V_1, \ldots, V_{n+1}$. Marginalising over $\{(X_i, Y_i)\}_{i=1}^n$, we obtain

$$(B.7) \leq \mathbb{E}\left(1 - \alpha + \max_{i\in[n+1]} p_i^{\hat{w}}(\tilde{X}, \tilde{Y})\right) \tag{B.8}$$

where the expectation is over $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P^{\pi^b}_{X,Y}$ and $(\tilde{X}, \tilde{Y}) \sim \tilde{P}^{\pi^*}_{X,Y}$. Let $\mathcal{A}$ denote the event that

$$\sum_{i=1}^{n} \hat{w}(X_i, Y_i) \leq n/2.$$

using Lemma B.1.2 and $\mathbb{E}[\hat{w}(\tilde{X}, \tilde{Y})] \leq M_r^2$, we get that

$$\begin{aligned}\mathbb{E}\left[\max_{i\in[n+1]} p_i^{\hat{w}}(\tilde{X}, \tilde{Y})\right] =& \mathbb{E}\left[\frac{\max\{\hat{w}(\tilde{X}, \tilde{Y}), \max_i \hat{w}(X_i, Y_i)\}}{\hat{w}(\tilde{X}, \tilde{Y}) + \sum_{i=1}^n \hat{w}(X_i, Y_i)}\right] \\ \leq& \mathbb{E}\left[\frac{\max\{\hat{w}(\tilde{X}, \tilde{Y}), \max_i \hat{w}(X_i, Y_i)\}}{\hat{w}(\tilde{X}, \tilde{Y}) + \sum_{i=1}^n \hat{w}(X_i, Y_i)} \mathbb{1}_{\mathcal{A}^C}\right] + \mathbb{P}(\mathcal{A}) \\ \leq& \mathbb{E}\left[\frac{2\max\{\hat{w}(\tilde{X}, \tilde{Y}), \max_i \hat{w}(X_i, Y_i)\}}{n} \mathbb{1}_{\mathcal{A}^C}\right] + \frac{c_1 M_r^2}{n} \\ \leq& \frac{2}{n}\left(\mathbb{E}[\hat{w}(\tilde{X}, \tilde{Y})] + \mathbb{E}\max_i \hat{w}(X_i, Y_i)\right) + \frac{c_1 M_r^2}{n} \\ \leq& \frac{2}{n}\left(\mathbb{E}[\hat{w}(\tilde{X}, \tilde{Y})] + \left(\sum_{i=1}^n \mathbb{E}[\hat{w}(X_i, Y_i)^r]\right)^{1/r}\right) + \frac{c_1 M_r^2}{n} \\ \leq& \frac{2}{n}\left(M_r^2 + n^{1/r} M_r\right) + \frac{c_1 M_r^2}{n}.\end{aligned}$$

This implies that

$$\mathbb{P}_{(X,Y)\sim\tilde{P}^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) \leq 1 - \alpha + cn^{1/r-1}$$

for some constant c that only depends on M_r and r. Note that

$$|\mathbb{P}_{(X,Y)\sim\tilde{P}^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) - \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}(X))| \leq d_{\text{TV}}(\tilde{P}^{\pi^*}, P^{\pi^*}) \tag{B.9}$$

where d_{TV} is the total variation norm which satisfies

$$\begin{aligned}d_{\text{TV}}(\tilde{P}^{\pi^*}, P^{\pi^*}) =& \frac{1}{2}\int |\hat{w}(x, y)\mathrm{d}P^{\pi^b}(x, y) - \mathrm{d}P^{\pi^*}(x, y)| \\ =& \frac{1}{2}\int |\hat{w}(x, y)\mathrm{d}P^{\pi^b}(x, y) - w(x, y)\mathrm{d}P^{\pi^b}(x, y)| \\ =& \frac{1}{2}\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[|\hat{w}(X, Y) - w(X, Y)|] = \Delta_w.\end{aligned} \tag{B.10}$$

Putting together (B.9) and (B.10), we get

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) \leq 1 - \alpha + \Delta_w + cn^{1/r-1}. \tag{B.11}$$

For the lower bound, using Proposition 3.4.1 we get that

$$\begin{aligned}\mathbb{P}_{(\tilde{X},\tilde{Y})\sim \tilde{P}^{\pi^*}_{X,Y}}(\tilde{Y} \in \hat{C}(\tilde{X},\tilde{Y})) =& \mathbb{P}\left(s(\tilde{X},\tilde{Y}) \leq \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{n} p_i^{\hat{w}}(\tilde{X},\tilde{Y})\delta_{V_i} + p_{n+1}^{\hat{w}}(\tilde{X},\tilde{Y})\delta_\infty\right)\right) \\ \geq& 1 - \alpha. \end{aligned} \tag{B.12}$$

Using (B.9) we thus obtain

$$\begin{aligned}\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) \geq& \mathbb{P}_{(X,Y)\sim \tilde{P}^{\pi^*}_{X,Y}}(Y \in \hat{C}(X)) - d_{TV}(\tilde{P}^{\pi^*}, P^{\pi^*}) \\ \geq& 1 - \alpha - \Delta_w. \end{aligned} \tag{B.13}$$

□

B.1.3 Proof of Proposition 3.4.3

For notational convenience, we suppress the subscripts m and n in $\hat{q}, \hat{w}, \hat{C}$. Moreover, we use $\hat{w}_i$ to denote $\hat{w}(X_i, Y_i)$ and $\eta(x, y)$ to denote $\text{Quantile}_{1-\alpha}(\sum_{i=1}^{n} \hat{p}_i(x, y)\delta_{V_i} + \hat{p}_{n+1}(x, y)\delta_\infty)$.

Proof. We use $(\tilde{X}, \tilde{Y}) \sim P^{\pi^*}_{X,Y}$ in place of (X_{n+1}, Y_{n+1}) and let $\epsilon < r/2$. By the definition of $\hat{C}(\tilde{X})$, we directly have

$$\begin{aligned}\mathbb{P}(\tilde{Y} \in \hat{C}(\tilde{X}) \mid \tilde{X}) =& \mathbb{P}(s(\tilde{X},\tilde{Y}) \leq \eta(\tilde{X},\tilde{Y}) \mid \tilde{X}) \\ \geq& \mathbb{P}(s^*(\tilde{X},\tilde{Y}) \leq \eta(\tilde{X},\tilde{Y}) - H(\tilde{X}) \mid \tilde{X}) \end{aligned} \tag{B.14}$$

where $s^*(\tilde{X}, \tilde{Y}) \coloneqq \max\{\tilde{Y} - q_{\alpha_{hi}}(\tilde{X}), q_{\alpha_{lo}}(\tilde{X}) - \tilde{Y}\}$ and the probability is taken over

$\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{X,Y}^{\pi^b}$ and $\tilde{Y} \sim P_{Y|X=\tilde{X}}^{\pi^*}$. We then get

$$\begin{aligned}
\text{(B.14)} \geq & \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq -\epsilon - H(\tilde{X}) \mid \tilde{X}) - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \\
\geq & \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq -\epsilon - H(\tilde{X}) \mid \tilde{X}) \Big(\mathbb{1}(H(\tilde{X}) \leq \epsilon) + \mathbb{1}(H(\tilde{X}) > \epsilon)\Big) - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \qquad \text{(B.15)} \\
\geq & \Big(\mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq 0 \mid \tilde{X}) - b_2\{\epsilon + H(\tilde{X})\}\Big) \mathbb{1}(H(\tilde{X}) \leq \epsilon) \\
& + \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq -\epsilon - H(\tilde{X}) \mid \tilde{X}) \mathbb{1}(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \qquad \text{(B.16)} \\
\geq & \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq 0 \mid \tilde{X}) \mathbb{1}(H(\tilde{X}) \leq \epsilon) - b_2\{\epsilon + H(\tilde{X}) \mathbb{1}(H(\tilde{X}) \leq \epsilon)\} \\
& + \Big(\mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq 0 \mid \tilde{X}) - \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \in (-\epsilon - H(\tilde{X}), 0))\Big) \mathbb{1}(H(\tilde{X}) > \epsilon) \\
& - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \\
\geq & \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq 0 \mid \tilde{X}) - b_2\{\epsilon + H(\tilde{X}) \mathbb{1}(H(\tilde{X}) \leq \epsilon)\} - \mathbb{1}(H(\tilde{X}) > \epsilon) \\
& - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \qquad \text{(B.17)}
\end{aligned}$$

where, to get from (B.15) to (B.16), we use the condition $2\epsilon < r$ and Assumption 2

$$\begin{aligned}
\text{(B.17)} \geq & \mathbb{P}(s^*(\tilde{X}, \tilde{Y}) \leq 0 \mid \tilde{X}) - b_2\{\epsilon + H(\tilde{X})\} - \mathbb{1}(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \qquad \text{(B.18)} \\
= & 1 - \alpha - b_2\{\epsilon + H(\tilde{X})\} - \mathbb{1}(H(\tilde{X}) > \epsilon) - \mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}). \qquad \text{(B.19)}
\end{aligned}$$

Next, we derive an upper bound on $\mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X})$. Let G denote the CDF of the random distribution $\sum_{i=1}^n \hat{p}_i(x, y)\delta_{V_i} + \hat{p}_{n+1}(x, y)\delta_\infty$. Then, $\eta(\tilde{X}, \tilde{Y}) < -\epsilon$ implies $G(-\epsilon) \geq 1 - \alpha$ and thus $\mathbb{P}(\eta(\tilde{X}, \tilde{Y}) < -\epsilon \mid \tilde{X}) \leq \mathbb{P}(G(-\epsilon) \geq 1 - \alpha \mid \tilde{X})$ a.s. Moreover, we have

$$\begin{aligned}
\mathbb{P}(G(-\epsilon) \geq 1 - \alpha \mid \tilde{X}) = & \mathbb{P}\left(\frac{\sum_{i=1}^n \hat{w}_i \mathbb{1}(V_i \leq -\epsilon)}{\sum_{i=1}^n \hat{w}_i + \hat{w}(\tilde{X}, \tilde{Y})} \geq 1 - \alpha \mid \tilde{X}\right) \\
\leq & \mathbb{P}\left(\frac{\sum_{i=1}^n \hat{w}_i \mathbb{1}(V_i \leq -\epsilon)}{\sum_{i=1}^n \hat{w}_i} \geq 1 - \alpha \mid \tilde{X}\right) \qquad \text{(B.20)} \\
= & \mathbb{P}\left(\frac{\sum_{i=1}^n \hat{w}_i \mathbb{1}(V_i \leq -\epsilon)}{\sum_{i=1}^n \hat{w}_i} \geq 1 - \alpha\right) \qquad \text{(B.21)}
\end{aligned}$$

where, to get from (B.20) to (B.21) we use the independence of $\{(X_i, Y_i)\}_{i=1}^n$ and $\tilde{X}$. Now we observe that

$$\frac{\sum_{i=1}^n \hat{w}_i \mathbb{1}(V_i \leq -\epsilon)}{n} = \frac{\sum_{i=1}^n (\hat{w}_i - w_i) \mathbb{1}(V_i \leq -\epsilon)}{n} + \frac{\sum_{i=1}^n w_i \mathbb{1}(V_i \leq -\epsilon)}{n}.$$

As $n \to \infty$, the strong law of large numbers yields

$$\begin{aligned}\left|\frac{\sum_{i=1}^n (\hat{w}_i - w_i)\mathbb{1}(V_i \leq -\epsilon)}{n}\right| &\xrightarrow{a.s.} \left|\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}\left[(\hat{w}(X,Y) - w(X,Y))\mathbb{1}(s(X,Y) \leq -\epsilon)\right]\right| \\ &\leq \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}\left[|\hat{w}(X,Y) - w(X,Y)|\mathbb{1}(s(X,Y) \leq -\epsilon)\right] \\ &\leq \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}\left[|\hat{w}(X,Y) - w(X,Y)|\right] \overset{m\to\infty}{\longrightarrow} 0 \end{aligned} \tag{B.22}$$

from Assumption 1 and

$$\frac{\sum_{i=1}^n w_i \mathbb{1}(V_i \leq -\epsilon)}{n} \xrightarrow{a.s.} \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[w(X,Y)\mathbb{1}(s(X,Y) \leq -\epsilon)] = \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(s(X,Y) \leq -\epsilon). \tag{B.23}$$

Using the triangle inequality,

$$\begin{aligned}\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(s(X,Y) \leq -\epsilon) &\leq \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(s^*(X,Y) \leq -\epsilon/2) + \mathbb{P}(H(X) \geq \epsilon/2) \quad \text{(B.24)} \\ &\leq \mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(s^*(X,Y) \leq 0) - \epsilon b_1/2 + 2^k\mathbb{E}[H^k(X)]/\epsilon^k \quad \text{(B.25)} \\ &= 1 - \alpha - \epsilon b_1/2 + 2^k\mathbb{E}[H^k(X)]/\epsilon^k \overset{m\to\infty}{\longrightarrow} 1 - \alpha - \epsilon b_1/2. \end{aligned}$$

To get from (B.24) to (B.25), we use Assumption 2 and Markov's inequality. Similarly, we have

$$\frac{\sum_{i=1}^n \hat{w}_i}{n} = \frac{\sum_{i=1}^n (\hat{w}_i - w_i)}{n} + \frac{\sum_{i=1}^n w_i}{n}$$

so, as $n \to \infty$,

$$\begin{aligned}\left|\frac{\sum_{i=1}^n (\hat{w}_i - w_i)}{n}\right| &\xrightarrow{a.s.} \left|\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}\left[(\hat{w}(X,Y) - w(X,Y))\right]\right| \\ &\leq \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}\left[|\hat{w}(X,Y) - w(X,Y)|\right] \overset{m\to\infty}{\longrightarrow} 0, \end{aligned} \tag{B.26}$$

and

$$\frac{\sum_{i=1}^n w_i}{n} \xrightarrow{a.s.} \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[w(X,Y)] = 1. \tag{B.27}$$

Putting this all together using the continuous mapping theorem, we get that, almost surely,

$$\lim_{m\to\infty}\lim_{n\to\infty} \frac{\sum_{i=1}^n \hat{w}_i \mathbb{1}(V_i \leq -\epsilon)}{\sum_{i=1}^n \hat{w}_i} = \lim_{m\to\infty}\lim_{n\to\infty} \frac{\sum_{i=1}^n \hat{w}_i \mathbb{1}(V_i \leq -\epsilon)/n}{\sum_{i=1}^n \hat{w}_i/n} = 1 - \alpha - \epsilon b_1/2. \tag{B.28}$$

Since convergence almost surely implies convergence in probability, we have

$$\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{P}\left(\frac{\sum_{i=1}^{n}\hat{w}_i\mathbb{1}(V_i\leq-\epsilon)}{\sum_{i=1}^{n}\hat{w}_i}\geq 1-\alpha\right)=0. \tag{B.29}$$

This implies that, for any $\epsilon>0$, $\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{P}(\eta(\tilde{X},\tilde{Y})<-\epsilon\mid\tilde{X})=0$ almost surely.

Using Markov's inequality and Assumption 3

$$\mathbb{P}(H(X)>\epsilon)\leq\mathbb{E}[H^k(X)]/\epsilon^k\overset{m\to\infty}{\longrightarrow}0. \tag{B.30}$$

So as $m\to\infty$, $H(X)\xrightarrow{\mathcal{P}}0$. Similarly, $\mathbb{1}(H(X)>\epsilon)\xrightarrow{\mathcal{P}}0$ as $m\to\infty$.

Recall (using B.19) that, for any $\epsilon\in(0,r/2)$, almost surely,

$$\mathbb{P}(\tilde{Y}\in\hat{C}(\tilde{X})\mid\tilde{X})-(1-\alpha-b_2\epsilon)\geq-b_2H(\tilde{X})-\mathbb{1}(H(\tilde{X})>\epsilon)-\mathbb{P}(\eta(\tilde{X},\tilde{Y})<-\epsilon\mid\tilde{X}). \tag{B.31}$$

For given $t>0$, pick $\epsilon<\min(r/2,t/2b_2)$. Then,

$$\mathbb{P}(\tilde{Y}\in\hat{C}(\tilde{X})\mid\tilde{X})-(1-\alpha-t/2)\geq-b_2H(\tilde{X})-\mathbb{1}(H(\tilde{X})>\epsilon)-\mathbb{P}(\eta(\tilde{X},\tilde{Y})<-\epsilon\mid\tilde{X}). \tag{B.32}$$

Each term on the right hand side of (B.32) converges in probability to 0 as $m,n\to\infty$, and therefore using continuous mapping theorem

$$b_2H(\tilde{X})+\mathbb{1}(H(\tilde{X})>\epsilon)+\mathbb{P}(\eta(\tilde{X},\tilde{Y})<-\epsilon\mid\tilde{X})\xrightarrow{\mathcal{P}}0.$$

This implies

$$\begin{aligned}&\mathbb{P}(\mathbb{P}(\tilde{Y}\in\hat{C}(\tilde{X})\mid\tilde{X})\leq 1-\alpha-t)\\&\quad\leq\mathbb{P}(b_2H(\tilde{X})+\mathbb{1}(H(\tilde{X})>\epsilon)+\mathbb{P}(\eta(\tilde{X},\tilde{Y})<-\epsilon\mid\tilde{X})\geq t/2)\to 0.\end{aligned}$$

Therefore,

$$\lim_{m\to\infty}\lim_{n\to\infty}\mathbb{P}(\mathbb{P}(\tilde{Y}\in\hat{C}(\tilde{X})\mid\tilde{X})\leq 1-\alpha-t)=0. \tag{B.33}$$

□

B.2 Conformal Off-Policy Prediction (COPP)

B.2.1 Further comments on the differences between Lei and Candès [2021] and COPP

In this subsection, we elaborate on the differences between our work and Lei and Candès [2021].

Firstly, Lei and Candès [2021] consider a setup in which the distribution of X is shifted, and construct intervals on the outcome under a specific (deterministic) action, i.e. $Y(a)$. In contrast, we consider a setup in which the distribution of $Y|X$ is shifted due to a change in the policy which is non trivial, and construct bounds on the outcome under this new policy (which could be stochastic). Additionally, since the theory in our methodology relies on the ratio of the joint distribution $P_{X,Y}$, our framework can be straightforwardly extended to the case where both, the conditional $P_{Y|X}$ and the covariate distribution P_X shift.

Secondly, as already mentioned in section 3.5, Lei and Candès [2021] can only be applied to the case where we have a deterministic target policy and a discrete action space, whereas COPP generalizes to the stochastic policy and continuous action space. This limitation of Lei and Candès [2021] can be partially addressed by employing the "*union method*" as described in the main text, which consists of constructing CP intervals for each action separately before taking the union of the intervals. However, we showed in our experiments that this leads to overly conservative intervals i.e. coverage above the required $1 - \alpha$ in Table 3.1a. This is because the predictive interval does not depend on the target policy, since every action is treated identically when taking the union. This approach is moreover unsuitable for continuous action spaces, whereas COPP applies without modification.

Thirdly, as stated in in section 3.5, even in the case when we only consider deterministic target policies, there is an important methodological difference between COPP and Lei and Candès [2021]. Lei and Candès [2021] construct the intervals on $Y(a)$ by only using calibration data with $A = a$ (see eq. 3.4 in Lei and Candès [2021]). In contrast, it can be shown that COPP uses the entire calibration data when constructing intervals on $Y(a)$. This is a consequence of integrating out the actions in the weights $w(x, y)$ (sec 3.3.1). This empirically leads to smaller variance in coverage compared to Lei and Candès [2021] as evidenced by the experimental results in B.2.2.

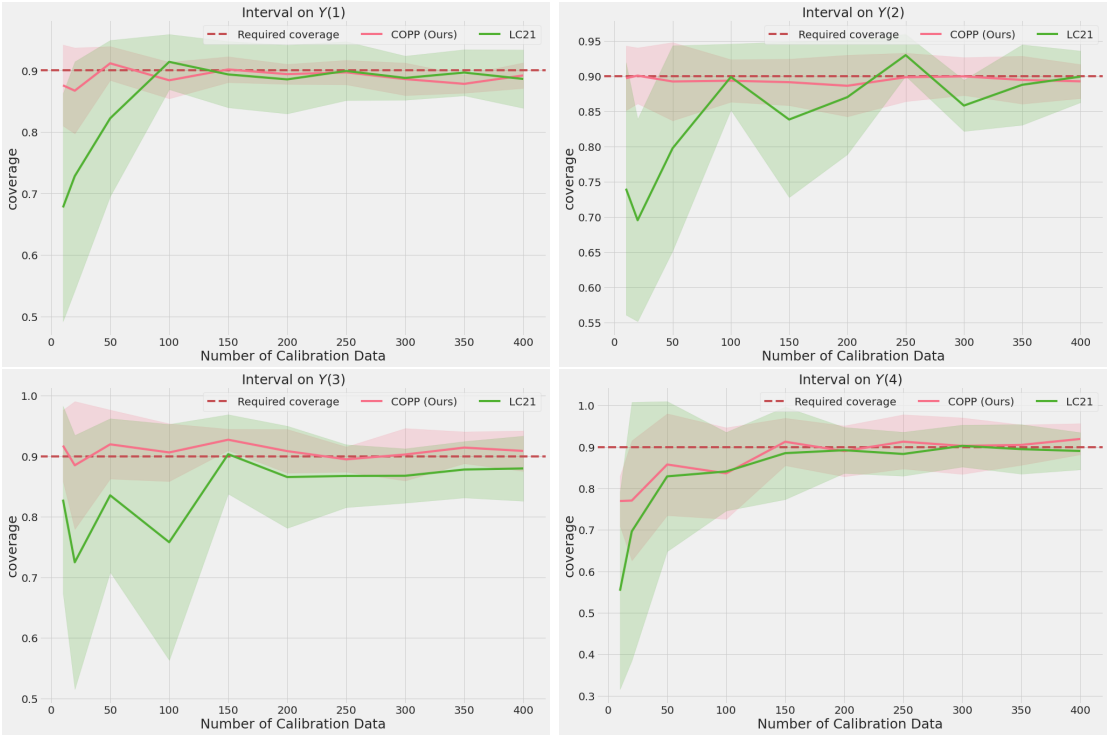


Figure B.1: Results for synthetic data experiment with $\pi^b = \pi_{0.3}$ and deterministic target policies.

Finally, in our paper we are *not* interested in a linear combination of the $Y(a)$ as in Lei and Candès [2021], who consider the linear combination of the form $Y(1) - Y(0)$. Instead, as described in section 3.1.1, we are interested in the outcome Y under the new target policy π^* (sometimes denoted as $Y(\pi^*)$ in the literature), which cannot be expressed as a linear combination of $Y(a)$. As a result, there does not appear to be a straightforward application of [Lei and Candès, 2021, Section 4.3] to our setup which relies on the linear combination assumption to be applicable.

B.2.2 Comparison with Lei and Candès [2021] on deterministic target policies.

In order to further clarify the distinction between COPP and Lei and Candès [2021], we conducted additional experiments when the target policy is deterministic i.e. $\pi^*(A|X) = \mathbb{1}\{A = a\}$. In the main text we modified Lei and Candès [2021] to our setting of stochastic policies by constructing the conformal intervals through the union of the CP sets across the actions. Here we aim to apply COPP to the setting of Lei and Candès [2021], i.e. deterministic target policy.

As mentioned in in the main text, given that we are integrating out the action in Eq. 3.7, we are essentially able to use the full dataset when constructing the CP intervals. To see this explicitly, consider the case where $Y \mid X, A$ is a normal random variable (as in our toy experiment). In this case, it can be straightforwardly shown that the weights $w(x_i, y_i)$ will be non-zero, and therefore, when constructing the COPP intervals using (3.5), we are able to use all the calibration datapoints.

This is contrary to Lei and Candès [2021], who only consider calibration data with $A = a$, when constructing the CP intervals for $Y(a)$. Below, we use the same experimental setup as our toy experiment in section 3.6.1 (see section B.4.1 for more details) with the difference here that we now consider deterministic target policies. In figure B.1 we plot the coverage for given deterministic target policies against the number of calibration datapoints. In this figure, we refer to the methodology of Lei and Candès [2021] as *LC21*. Here, we use the behavioural policy $\pi_{0.3}$ and a deterministic target policy which takes a single fixed action $a \in \{1, 2, 3, 4\}$ at test time. In the title of each subfigure, $Y(a)$ corresponds to the outcome for the target policy $\pi^*(A = a \mid X) = \mathbb{1}(A = a)$.

Results: We first note in Figure B.1 that the coverage of COPP intervals has a lower variance than Lei and Candès [2021]. Given that COPP is able to use all the data when constructing the CP intervals, as opposed to Lei and Candès [2021] which only uses a subset, our bounds have lower variance while also attaining the coverage guarantees. We observe this difference particularly in the case when we have little calibration data. Given that Lei and Candès [2021] have to split the data into 4 different splits (we have 4 different actions), the calibration data for each action is relatively small, whereas we are able to use the whole dataset to construct our CP intervals.

B.2.3 Motivation of using stochastic policies for bandits

One of the key difference between our method and that of Lei and Candès [2021] is that our method can be applied to the setting where the target policy is stochastic. In many settings, deterministic target policies might not be applicable such as in the settings of recommendation systems or RL where exploration is needed [Swaminathan et al., 2017a, Su et al., 2020]. For example, COPP can be used to compare different recommendation systems given some logged data. We explore this application in our MSR experiments

where the target policies correspond to different recommendation systems which are, by default, stochastic. Other applications which also make use of stochastic policies bandit problems can be found in Su et al. [2020], Farajtabar et al. [2018a].

B.2.4 COPP for Group-balanced coverage

As Angelopoulos and Bates [2021] point out, we may want predictive intervals that have same error rates across multiple different groups. Using our example of a recommendation system, we may want the predictive intervals to have same coverage across male and female users.

Formally, this problem can be expressed as follows. Let $\Omega = \{\Omega_1, \cdots, \Omega_k\}$ be subsets of $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}((X,Y) \in \Omega_j) > 0$ for $j \in \{1, \ldots, k\}$. We would like to construct predictive intervals $\hat{C}^{\Omega}_n$ which satisfy

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}^{\Omega}_n(X) \mid (X,Y) \in \Omega_j) \geq 1 - \alpha \text{ for all } j \in \{1, \ldots, k\}.$$

CP offers us the ability to construct such intervals $\hat{C}^{\Omega}_n$, by simply running algorithm 1 (main text) on each group separately. This has been visualized in figure B.2.

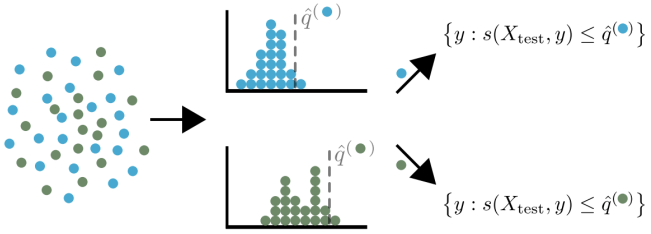


Figure B.2: Figure taken from Angelopoulos and Bates [2021]. To achieve group-balanced coverage, we simply run conformal prediction separately on each group.

Formally, this procedure can be described as follows. We group scores into different groups according to each subset.

$$\{(X_i^{\Omega_j}, Y_i^{\Omega_j})\}_{i=1}^{n_j} \coloneqq \{(X_i, Y_i) : (X_i, Y_i) \in \Omega_j\}_{i=1}^{n} \text{ and,}$$
$$V_i^{\Omega_j} \coloneqq (X_i^{\Omega_j}, Y_i^{\Omega_j})$$

Then, within each subset, we calculate the conformal quantile,

$$\eta^{\Omega_j}(x, y) \coloneqq \text{Quantile}_{1-\alpha}(\hat{F}_n^{\Omega_j}(x, y))$$

where,

$$\hat{F}_n^{\Omega_j}(x,y) \coloneqq \sum_{i=1}^{n_j} p_i^{\Omega_j}(x,y)\delta_{V_i^{\Omega_j}} + p_{n+1}^{\Omega_j}(x,y)\delta_\infty \text{ where,}$$
$$p_i^{\Omega_j}(x,y) \coloneqq \frac{w(X_i^{\Omega_j}, Y_i^{\Omega_j})}{\sum_{i=1}^{n_j} w(X_i^{\Omega_j}, Y_i^{\Omega_j}) + w(x,y)}$$
$$p_{n+1}^{\Omega_j}(x,y) \coloneqq \frac{w(x,y)}{\sum_{i=1}^{n_j} w(X_i^{\Omega_j}, Y_i^{\Omega_j}) + w(x,y)}$$

Next, we construct the set $\hat{C}_n^\Omega$ as follows:

$$\hat{C}_n^\Omega(x^{test}) \coloneqq \bigcup_{j=1}^{k} \hat{C}_n^{\Omega_j}(x^{test}) \text{ where,}$$
$$\hat{C}_n^{\Omega_j}(x^{test}) \coloneqq \{y : (x^{test}, y) \in \Omega_j \text{ and } s(x^{test}, y) \leq \eta^{\Omega_j}(x^{test}, y)\}. \tag{B.34}$$

Proposition B.2.1 (Coverage guarantee for class-balanced conformal prediction)

Let $\Omega = \{\Omega_1, \cdots, \Omega_k\}$ be subsets of $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{P}_{(X,Y)\sim P_{X,Y}^{\pi^*}}((X,Y) \in \Omega_j) > 0$ for $j \in \{1, \ldots, k\}$. Then, the set $\hat{C}_n^\Omega$ defined above satisfies the coverage guarantee

$$\mathbb{P}_{(X,Y)\sim P_{X,Y}^{\pi^*}}(Y \in \hat{C}_n^\Omega(X) \mid (X,Y) \in \Omega_j) \geq 1-\alpha \text{ for all } j \in \{1, \ldots, k\}.$$

Proof of Proposition B.2.1

$$\begin{aligned}
&\mathbb{P}_{(X,Y)\sim P_{X,Y}^{\pi^*}}(Y \in \hat{C}_n^\Omega(X) \mid (X,Y) \in \Omega_j) \\
&\geq \mathbb{P}_{(X,Y)\sim P_{X,Y}^{\pi^*}}(Y \in \hat{C}_n^{\Omega_j}(X) \mid (X,Y) \in \Omega_j) \\
&\geq \mathbb{P}_{(X,Y)\sim P_{X,Y}^{\pi^*}}((X,Y) \in \Omega_j : s(X,Y) \leq \eta^{\Omega_j}(X,Y) \mid (X,Y) \in \Omega_j)
\end{aligned} \tag{B.35}$$

Define the measure $P_{X,Y}^j$ by restricting $P_{X,Y}^{\pi^*}$ to Ω_j, i.e.

$$P_{X,Y}^j(x,y) \propto P_{X,Y}^{\pi^*}(x,y)\mathbb{1}((x,y) \in \Omega_j)$$

Then, (B.35) can be written as

$$\text{(B.35)} = \mathbb{P}_{(X,Y)\sim P_{X,Y}^j}(s(X,Y) \leq \eta^{\Omega_j}(X,Y)) \tag{B.36}$$

Moreover, for $(x, y) \in \Omega_j$ we have

$$w(x,y) = \frac{P^{\pi^*}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)} \propto \frac{P^{j}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)}$$

Since $p_i^{\Omega_j}(x,y)$ is invariant to scaling of weights $w(x,y)$, replacing the weights by $\tilde{w}(x,y) = \frac{P^{j}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)}$ keeps the conformal sets unchanged.

Therefore, using Proposition 3.4.1, the conformal sets constructed will provide coverage guarantees under the measure $P^j_{X,Y}$, i.e.

$$\mathbb{P}_{(X,Y)\sim P^j_{X,Y}}(s(X,Y) \leq \eta^{\Omega_j}(X,Y)) \geq 1-\alpha$$

Using (B.36), we get that

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}^{\Omega}_n(X) \mid (X,Y) \in \Omega_j) \geq \mathbb{P}_{(X,Y)\sim P^j_{X,Y}}(s(X,Y) \leq \eta^{\Omega_j}(X,Y)) \geq 1-\alpha$$

□

COPP for class-balanced coverage

Algorithm 2: COPP for class-balanced coverage

Inputs: Observational data $\mathcal{D}_{obs} = \{X_i, A_i, Y_i\}_{i=1}^{n_{obs}}$, conf. level α, a score function $s(x,y) \in \mathbb{R}$, new data point x^{test}, target policy π^* ;

Output: $\hat{C}^{\mathcal{Y}}_n(x^{test})$ with coverage guarantee (B.37);

Split $\mathcal{D}_{obs}$ into training data ($\mathcal{D}_{tr}$) and calibration data ($\mathcal{D}_{cal}$) of sizes m and n respectively;

Use $\mathcal{D}_{tr}$ to estimate weights $\hat{w}(\cdot,\cdot)$;

for $y \in \mathcal{Y}$ **do**

Let $\{X^y_j, Y^y_j\}_{j=1}^{n_y}$ be the following subset of calibration data: $\{(X_i, Y_i) : Y_i = y\}$;

Let $V^y_j := s(X^y_j, Y^y_j)$, for $j = 1, \ldots, n_y$;

Define $\hat{F}^{x,y}_n = \sum_{i=1}^{n_y} p^w_i(x,y)\delta_{V^y_i} + p^w_{n+1}(x,y)\delta_\infty$;

where, $p^w_i(x,y) := \frac{w(X^y_i, Y^y_i)}{\sum_{i=1}^{n_y} w(X^y_i, Y^y_i) + w(x,y)}$, $p^w_{n+1}(x,y) := \frac{w(x,y)}{\sum_{i=1}^{n_y} w(X^y_i, Y^y_i) + w(x,y)}$;

$\eta(x,y) := \text{Quantile}_{1-\alpha}(\hat{F}^{x,y}_n)$

end

Define $\hat{C}^{\mathcal{Y}}_n(x^{test}) := \{y : s(x^{test}, y) \leq \eta(x^{test}, y)\}$;

Return $\hat{C}^{\mathcal{Y}}_n(x^{test})$

In the case when Y is discrete, we construct predictive sets, $\hat{C}^{\mathcal{Y}}_n(x)$, which offer label conditioned coverage guarantees using the methodology described above,

$$\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}^{\mathcal{Y}}_n(X) \mid Y = y) \geq 1-\alpha, \text{ for all } y \in \mathcal{Y} \tag{B.37}$$

This is a strictly stronger guarantee than marginal coverage, i.e. $\mathbb{P}_{(X,Y)\sim P^{\pi^*}_{X,Y}}(Y \in \hat{C}_n(X)) \geq 1-\alpha$. To understand what (B.37) means, consider our running example of recommendation systems, where the outcome Y is whether the recommendation is relevant (0) or not (1) to the user. Then, Eq. (B.37) ensures that out of the users who received irrelevant recommendations, the predictive sets contain 'not relevant' (1) at least $100 \cdot (1-\alpha)\%$ of the times. This can be thought of as controlling the false negative rate of irrelevant recommendations at $100 \cdot \alpha\%$. The same is true for users who receive relevant recommendations. This is particularly useful when data is imbalanced, for example, when the majority of the users in observational receive relevant recommendations.

B.2.5 Weights estimation $\hat{w}(x, y)$

Consistent estimation of the weights does not imply consistent estimation of $\hat{P}(y|x, a)$

In Proposition 3.4.1, we assume to have consistent estimator of $w(x, y)$ which begs the following question: In general, does a consistent estimate of $w(x, y)$ imply that we also obtain a consistent estimate of $P(y|x, a)$? In particular, one could then just use the estimate of $\hat{P}(y|x, a)$ to construct the predictive interval. However, we answer the above question with the negative by supplying a counter-example.

Counter-example Let $X \in [1, +\infty), a \in \mathbb{R}$ s.t. $|a| < K$ for $K \in \mathbb{R}_{>0}$.

Let $Y|X, a \sim \mathcal{N}((KX^2 + a)^{0.5}, (KX^2 - a))$.

We have $\mathbb{E}[Y^2|X, a] = Var(Y|X, a) + \mathbb{E}[Y|X, a]^2 = KX^2 + a + KX^2 - a = 2KX^2$ (independent of a)

Next let

$$\hat{P}(y|x, a) \coloneqq \frac{y^2 P(y|x, a)}{2Kx^2}. \tag{B.38}$$

Recall that

$$w(x, y) = \frac{\int P(y|x, a)\pi^*(a|x)\mathrm{d}a}{\int P(y|x, a)\pi^b(a|x)\mathrm{d}a} \tag{B.39}$$

Using the above definition of $\hat{P}(y|x,a)$ we have:

$$\begin{aligned}\hat{w}(x,y) &= \frac{\int \hat{P}(y|x,a)\pi^*(a|x)\mathrm{d}a}{\int \hat{P}(y|x,a)\pi^b(a|x)\mathrm{d}a} \\ &= \frac{\int P(y|x,a)\frac{Y^2}{2KX^2}\pi^*(a|x)\mathrm{d}a}{\int P(y|x,a)\frac{Y^2}{2KX^2}\pi^b(a|x)\mathrm{d}a} \\ &= w(x,y).\end{aligned}$$

Hence, $w(x,y) \equiv \hat{w}(x,y) \implies \hat{P}(y|x,a) \equiv P(y|x,a)$. □

More generally, if there exists a function $\Phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ such that

1. $\Phi(x,y)$ is not constant in y
2. $0 < \mathbb{E}[\Phi(X,Y) \mid X, A] < \infty$, and does not depend on A

Then, we can define $\tilde{P}(y|x,a) \coloneqq P(y|x,a)\Phi(x,y)/\mathbb{E}[\Phi(X,Y) \mid X,A]$, and the weights computed using $\tilde{P}(y|x,a)$ will be the equal to $w(x,y)$ even though $\tilde{P}(y|x,a) \neq P(y|x,a)$.

Alternative ways to estimate $\hat{w}(x,y)$ without estimating $\hat{P}(y|x,a)$

In this section, we show how we could estimate $w(x,y)$ without having to estimate $\hat{P}(y|x,a)$. One way to obtain an estimate $\hat{w}(x,y)$ is by taking a closer look at the definition of $w(x,y)$ and rewriting the ratio.

$$\begin{aligned}w(x,y) &= \frac{P^{\pi^*}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)} \\ &= \int \frac{P^{\pi^*}_{X,A,Y}(x,a,y)}{P^{\pi^b}_{X,A,Y}(x,a,y)} P^{\pi^b}_{A|X,Y}(a|x,y)\mathrm{d}a \\ &= \int \frac{\pi^*(a|x)}{\pi^b(a|x)} P^{\pi^b}_{A|X,Y}(a|x,y)\mathrm{d}a \\ &= \mathbb{E}_{A\sim P^{\pi^b}_{A|X=x,Y=y}}\left[\frac{\pi^*(A|x)}{\pi^b(A|x)}\right]. \end{aligned} \tag{B.40}$$

Lemma B.2.1

Let $w(x,y) = \frac{P^{\pi^*}_{X,Y}(x,y)}{P^{\pi^b}_{X,Y}(x,y)}$, then

$$w(x,y) = \arg\min_f \mathbb{E}_{X,A,Y\sim P^{\pi^b}_{X,A,Y}}\left[\left|\left|\frac{\pi^*(A|X)}{\pi^b(A|X)} - f(X,Y)\right|\right|^2\right]. \tag{B.41}$$

Proof of Lemma B.2.1 This follows directly from the identity (B.40). We prove it here for sake of completeness.

$$\begin{aligned}
&\mathbb{E}_{X,A,Y\sim P^{\pi^b}_{X,A,Y}}\left[\left|\left|\frac{\pi^*(A|X)}{\pi^b(A|X)} - f(X,Y)\right|\right|^2\right] \\
&= \mathbb{E}_{X,Y\sim P^{\pi^b}_{X,Y}}\left[\mathbb{E}_{A\sim P^{\pi^b}_{A|X,Y}}\left|\left|\frac{\pi^*(A|X)}{\pi^b(A|X)} - f(X,Y)\right|\right|^2\right] \\
&= \mathbb{E}_{X,Y\sim P^{\pi^b}_{X,Y}}\left[\mathrm{Var}_{A\sim P^{\pi^b}_{A|X,Y}}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\right] + \left(\mathbb{E}_{A\sim P^{\pi^b}_{A|X,Y}}\left[\frac{\pi^*(A|X)}{\pi^b(A|X)}\right] - f(X,Y)\right)^2\right]. \quad \text{(B.42)}
\end{aligned}$$

Where, (B.42) is minimized if $f(x,y) = \mathbb{E}_{A\sim P^{\pi^b}_{A|X=x,Y=y}}\left[\frac{\pi^*(A|x)}{\pi^b(A|x)}\right] = w(x,y)$. □

Using Lemma B.2.1, we can thus approximate $w(x,y)$ by minimizing the loss

$$\hat{w}(x,y) = \arg\min_{f_\theta} \mathbb{E}_{X,A,Y\sim P^{\pi^b}_{X,A,Y}}\left[\left|\left|\frac{\pi^*(A|X)}{\pi^b(A|X)} - f_\theta(X,Y)\right|\right|^2\right] \quad \text{(B.43)}$$

Hence we see that the ratio estimation problem can be rewritten as a regression problem where $f_\theta(x,y)$ is for example a neural network. This allows one to estimate directly, without the need for estimating $\hat{P}(y \mid x,a)$ first.

B.3 Estimation of the quantiles of the target distribution

As mentioned in Section 3.4.2, we present here a way to estimate the quantiles of the target distribution $P^{\pi^*}_{X,Y}$ consistently when the ground truth weight function $w(x, y)$ is known. As we are interested in the quantiles, we will be using the pinball loss to train our model $\hat{f}_\theta$ defined by

$$L_\alpha(\theta, x, y) = \begin{cases} \alpha(\hat{f}_\theta(x) - y) & \text{if } (\hat{f}_\theta(x) - y) > 0, \\ (1-\alpha)(y - \hat{f}_\theta(x)) & \text{if } (\hat{f}_\theta(x) - y) < 0. \end{cases}$$

Then we have the following objective to optimize:

$$\begin{aligned} \mathbb{E}_{(X,Y)\sim P^{\pi^*}_{X,Y}}[L_\alpha(\theta, X, Y)] &= \int_{X,Y} L_\alpha(\theta, x, y) P^{\pi^*}_{X,Y}(\mathrm{d}x, \mathrm{d}y) \\ &= \int_{X,Y} L_\alpha(\theta, x, y) \frac{\mathrm{d}P^{\pi^*}_{X,Y}(x, y)}{\mathrm{d}P^{\pi^b}_{X,Y}(x, y)} P^{\pi^b}_{X,Y}(\mathrm{d}x, \mathrm{d}y) \\ &= \int_{X,Y} L_\alpha(\theta, x, y) w(x, y) P^{\pi^b}_{X,Y}(\mathrm{d}x, \mathrm{d}y) \\ &= \mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[L_\alpha(\theta, X, Y) w(X, Y)]. \end{aligned}$$

The above holds true if the true weight function is known. However in the case where we only have a consistent estimator of $w(x, y)$, it remains to be proven that the above objective will also yield a consistent estimator of the quantiles under π^*. We leave this for future work to prove as we are simply providing a possible avenue to relax the assumptions in Proposition 3.4.2.

B.4 Experiments

The code for our experiments is available at `https://anonymous.4open.science/r/COPP-75F5` and we ran all our experiments on Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz with 8GB RAM per core. We were able to use 100 CPUs in parallel to iterate over different configurations and seeds. However, we would like to note that our algorithms only requires 1 CPU and at most 10 mins to run, as our networks are relatively small.

B.4.1 Toy Experiment

Synthetic data experiments setup

Model. The observational data distribution is defined as follows:

$$
\begin{aligned}
&X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 9) \\
&A_i \mid x_i \sim \pi^b(\cdot \mid x_i) \text{ where } \pi^b \text{ has been defined below} \\
&Y_i \mid x_i, a_i \sim \mathcal{N}(a_i * x_i, 1)
\end{aligned}
$$

Behaviour and Target Policies. We define a family of policies $\pi_\epsilon(a \mid x)$ as follows:

$$
\pi_\epsilon(a|x) := \begin{cases} \epsilon\mathbb{1}(a \in \{1,2,3\}) + (1-3\epsilon)\mathbb{1}(a=4) & \text{if } |x| \in (3,\infty) \\ \epsilon\mathbb{1}(a \in \{1,2,4\}) + (1-3\epsilon)\mathbb{1}(a=3) & \text{if } |x| \in (2,3] \\ \epsilon\mathbb{1}(a \in \{1,3,4\}) + (1-3\epsilon)\mathbb{1}(a=2) & \text{if } |x| \in (1,2] \\ \epsilon\mathbb{1}(a \in \{2,3,4\}) + (1-3\epsilon)\mathbb{1}(a=1) & \text{if } |x| \in [0,1] \end{cases}
$$

We use the parameter $\epsilon \in (0, 1/3)$ to control the policy shift between target and behaviour policies. For the behaviour policy π^b, we use $\epsilon^b = 0.3$, and for target policies π^*, we use $\epsilon^* \in \{0.1, 0.2, 0.3\}$. Here we use $m = 1000$ training datapoints.

Neural Network Architectures

- To approximate the behaviour policy π^b, we use a neural network with 2 hidden layers and 16 nodes in each hidden layer, and ReLU activation function.
- To approximate $P(y|x, a)$, we use $\mathcal{N}(\mu(x, a), \sigma(x, a))$, where μ and σ are neural networks with one-hidden layer, 32 nodes in the hidden layer, and ReLU activation function.

- For the score function, we train the quantiles $\hat{q}_{\alpha/2}$ and $\hat{q}_{1-\alpha/2}$ using quantile regression, each of which are modelled using neural networks with one-hidden layer, 32 nodes in the hidden layer, and ReLU activation functions.

Results: Coverage as a function of increase calibration data As mentioned in the main text, we have also performed experiments to investigate how much calibration data is needed for COPP as well as other methods to converge to the required 90% coverage. In the below figure B.3 we have plotted the coverage as a function of n calibration data points. Our proposed method is converging much faster to the required coverage compared to the competing methods.

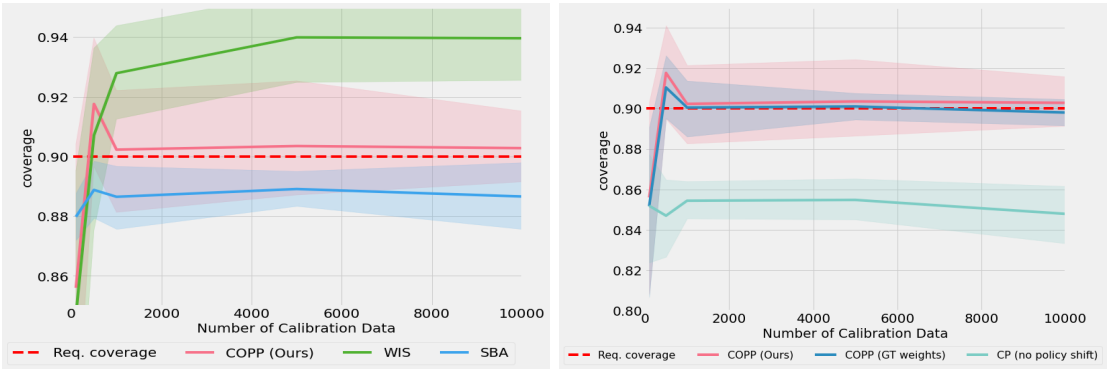


Figure B.3: Results for synthetic data experiment with $\pi^b = \pi_{0.3}$ and the target policy is $\pi^* = \pi_{0.1}$. **Left:** our proposed method is able to converge to the required coverage rather quickly compared to the competing methods. **Right:** here we see that our method is on par with using the GT weights. Due to estimation error, COPP with estimated weights has slightly higher variance in terms of coverage

Additional experimental baseline using weighted quantile regression. In order to add an additional baseline that is also covariate dependent, we have added some experiments using the weighted quantile regression (WQR) as described in Sec. B.3 on our toy experiments from Sec. 3.6 in the main text. Below in Table B.1 and Table B.2 we see the complete coverage table with the respective interval lengths. Note also that WQR does not seem to perform well as it does not have any statistical guarantees and heavily relies on good estimation of the ratio. We have added these experiments here in the appendix for completeness and did not add it in the main text as the results were not comparable to other baselines.

Table B.1: Mean Coverage as a function of policy shift with 2 standard errors over 10 runs. We have added weighted quantile regression (WQR) for completeness and note that it does not seem to perform well.

Coverage	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$
COPP (Ours)	**0.90 ± 0.01**	**0.90 ± 0.01**	**0.91 ± 0.01**
WIS	**0.89 ± 0.01**	**0.91 ± 0.02**	0.94 ± 0.02
SBA	**0.90 ± 0.01**	0.88 ± 0.01	0.87 ± 0.01
COPP (GT weights Ours)	**0.90 ± 0.01**	**0.90 ± 0.01**	**0.90 ± 0.01**
CP (no policy shift)	**0.90 ± 0.01**	0.87 ± 0.01	0.85 ± 0.01
CP (union)	0.96 ± 0.01	0.96 ± 0.01	0.96 ± 0.01
WQR	0.82 ± 0.04	0.76 ± 0.03	0.70 ± 0.03

Table B.2: Mean Interval Length as a function of policy shift with 2 standard errors over 10 runs. We have added weighted quantile regression (WQR) for completeness and note that it does not seem to perform well.

Interval Lengths	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$
COPP (Ours)	9.08 ± 0.10	9.48 ± 0.22	9.97 ± 0.38
WIS	24.14 ± 0.30	32.96 ± 1.80	43.12 ± 3.49
SBA	8.78 ± 0.12	8.94 ± 0.10	8.33 ± 0.09
COPP (GT weights Ours)	8.91 ± 0.09	9.25 ± 0.12	9.59 ± 0.20
CP (no policy shift)	9.00 ± 0.10	9.00 ± 0.10	9.00 ± 0.10
CP (union)	10.66 ± 0.18	11.04 ± 0.2	11.4 ± 0.26
WQR	8.55 ± 0.50	8.61 ± 0.52	8.70 ± 0.55

Experiments with continuous action space

As mentioned in the main text and also in Sec. B.2.1, our proposed method, contrary to the work of Lei and Candès [2021] is able to also handle continuous action space. Given that we are integrating out the actions when computing the weights in Eq. 3.7 our method trivially extends to the continuous action space, whereas Lei and Candès [2021] is only applicable for discrete action spaces, as they compute conformal intervals conditioned on a given action.

Model. The observational data distribution is defined as follows:

$$X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4)$$
$$A_i \mid x_i \sim \mathcal{N}(x_i/4, 1)$$
$$Y_i \mid x_i, a_i \sim \mathcal{N}(a_i + x_i, 1)$$

Target Policies. We define a family of policies $\pi_\epsilon(a \mid x)$ as follows:

$$\pi_\epsilon(a \mid x) = \mathcal{N}(x/4 + \epsilon, 1). \tag{B.44}$$

In our experiments, for the target policy π^*, we use $\pi^* = \pi_{\epsilon^*}$ for $\epsilon^* \in \{0, 0.5, 1, 1.5, 2, 2.5\}$.

Results. Table B.3 shows the coverages of different methods as the policy shift ϵ^* increases. The behaviour policy $\pi^b = \pi_0$ is fixed and we use $n = 5000$ calibration datapoints and $m = 1000$ training points, across 10 runs. Table B.3 shows, how COPP stays very close to the required coverage of 90% across all target policies with $\epsilon^* \leq 2.0$, compared to WIS and SBA. Both, WIS intervals and SBA intervals suffer from under-coverage i.e. below the required coverage. These results again support our hypothesis from Sec. 3.3.1, which stated that COPP is less sensitive to estimation errors of $\hat{P}(y|x, a)$ compared to directly using $\hat{P}(y|x, a)$ for the intervals i.e. SBA.

Next, Table B.4 shows the mean interval lengths and even though WIS intervals are under-covered, the average interval length is huge compared to COPP. Additionally, for $\epsilon^* \in \{0, 0.5, 1, 1.5\}$, COPP with estimated weights produces results which are close to COPP intervals with ground truth weights. This shows that when the behaviour and target policies have reasonable overlap, the effect of weights estimation error on COPP results is limited. However, as ϵ^* increases to 2.0 and 2.5, the overlap between behaviour and target policies becomes low. We empirically note that this leads to high weights estimation error and consequently under-coverage in COPP with estimated weights. In contrast, COPP with ground truth weights still achieves required coverage, even though it becomes conservative when the overlap is low. Figure B.4 visualises how the overlap between target and behaviour policies decreases with increasing ϵ^*. It can be seen that $\epsilon^* \in \{2, 2.5\}$ leads to very low overlap between the behaviour and target data.

Table B.3: Mean Coverage as a function of policy shift with 2 standard errors over 10 runs.

Coverage	$\epsilon^* = 0.0$	$\epsilon^* = 0.5$	$\epsilon^* = 1.0$	$\epsilon^* = 1.5$	$\epsilon^* = 2.0$	$\epsilon^* = 2.5$
COPP (Ours)	**0.90 ± 0.01**	**0.91 ± 0.01**	0.92 ± 0.01	**0.91 ± 0.01**	**0.89 ± 0.02**	0.85 ± 0.02
WIS	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.01	0.87 ± 0.02	**0.89 ± 0.02**	0.83 ± 0.02
SBA	0.86 ± 0.01	0.86 ± 0.01	0.86 ± 0.01	0.86 ± 0.01	**0.89 ± 0.02**	0.83 ± 0.02
COPP (GT Weights Ours)	**0.90 ± 0.01**	**0.91 ± 0.01**	**0.91 ± 0.01**	**0.90 ± 0.01**	0.96 ± 0.02	0.93 ± 0.02
CP (no policy shift)	**0.90 ± 0.01**	0.88 ± 0.01	0.82 ± 0.01	0.73 ± 0.01	0.60 ± 0.01	0.46 ± 0.01

Table B.4: Mean Interval Length as a function of policy shift with 2 standard errors over 10 runs.

Interval Lengths	$\epsilon^* = 0.0$	$\epsilon^* = 0.5$	$\epsilon^* = 1.0$	$\epsilon^* = 1.5$	$\epsilon^* = 2.0$	$\epsilon^* = 2.5$
COPP (Ours)	4.75 ± 0.04	5.08 ± 0.09	5.89 ± 0.14	6.92 ± 0.18	7.82 ± 0.41	8.45 ± 0.44
WIS	9.55 ± 0.1	9.56 ± 0.12	9.56 ± 0.27	9.44 ± 0.38	9.40 ± 0.59	9.08 ± 0.64
SBA	4.38 ± 0.03	4.37 ± 0.03	4.36 ± 0.04	4.34 ± 0.07	4.31 ± 0.1	4.28 ± 0.14
COPP (GT Weights Ours)	4.73 ± 0.05	5.07 ± 0.09	5.87 ± 0.14	6.82 ± 0.13	7.57 ± 0.19	8.07 ± 0.22
CP (no policy shift)	4.70 ± 0.05	4.70 ± 0.05	4.70 ± 0.05	4.70 ± 0.05	4.70 ± 0.05	4.70 ± 0.05

B.4.2 Experiments on Microsoft Ranking Dataset

Dataset details. The dataset contains relevance scores for websites recommended to different users, and comprises of $30,000$ user-website pairs. For a user i and website j, the data contains a 136-dimensional feature vector u_i^j, which consists of user i's attributes corresponding to website j, such as length of stay or number of clicks on the website. Furthermore, for each user-website pair, the dataset also contains a relevance score, i.e. how relevant the website was to the user.

First, given a user i we sample (with replacement) 5 websites, $\{u_i^j\}_{j=1}^5$, from the data. Next, we reformulate this into a contextual bandit where $A \in \{1, 2, 3, 4, 5\}$ corresponds to the website we recommend to a user. For a user i, we define X by combining the 5 feature vectors corresponding to the user, i.e. $X \in \mathbb{R}^{5\times 136}$, where $x_i = (u_i^1, u_i^2, u_i^3, u_i^4, u_i^5)$. In addition, $Y \in \{0, 1, 2, 3, 4\}$ corresponds to the relevance score for the A'th website, i.e. the recommended website. The goal is to construct prediction sets that are guaranteed to contain the true relevance score with a probability of 90%. Here we use $m = 5000$ training data points.

Behaviour and Target Policies. We first train a Neural Network (NN) classifier model mapping each 136-dimensional feature vector to the softmax scores for each relevance score class, $\hat{f}_\theta : \mathcal{U} \to [0, 1]^5$. We use this trained model $\hat{f}_\theta$ to define a family of policies such that we pick the most relevant website as predicted by $\hat{f}_\theta$ with probability ϵ and the rest uniformly with probability $(1 - \epsilon)/4$. Formally, this has been expressed as follows. We use $\hat{f}_\theta^{\text{label}}$ to denote the relevance class predicted by $\hat{f}_\theta$, i.e. $\hat{f}_\theta^{\text{label}}(u) \coloneqq \arg\max_i \{\hat{f}_\theta(u)_i\}$.

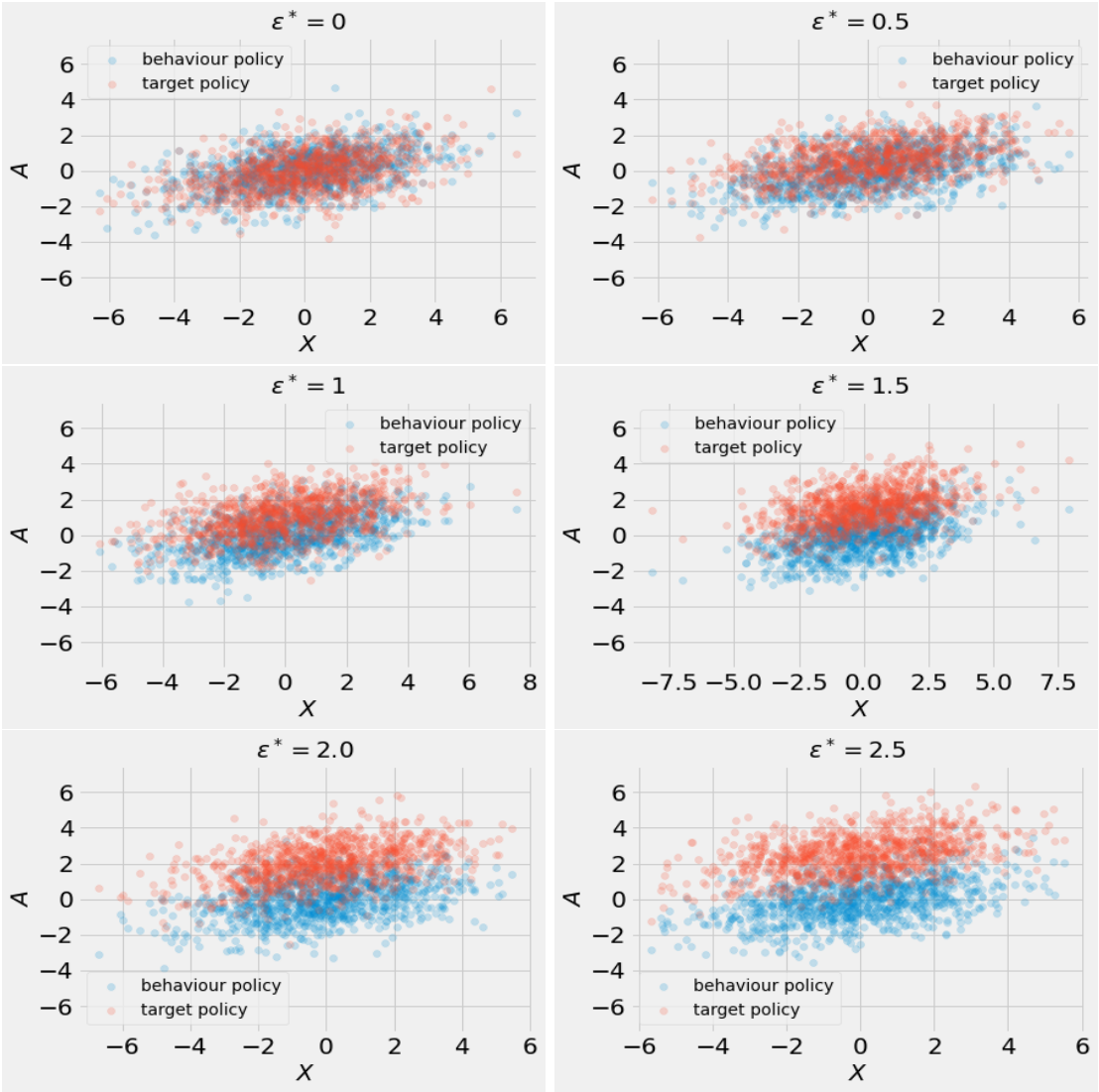


Figure B.4: Plots of A against X, where $X \sim \mathcal{N}(0, 4)$ and $A \mid X$ is sampled from behaviour and target policies. Here, target policies are defined in (B.44) for $\epsilon^* \in \{0, 0.5, 1, 1.5, 2, 2.5\}$.

Then,

$$\begin{aligned}\pi_\epsilon(a \mid X = (u^1, u^2, u^3, u^4, u^5)) :=& \epsilon \mathbb{1}(a = \arg\max_j \{\hat{f}_\theta^{\text{label}}(u^j)\}) \\ &+ (1-\epsilon)/4 \mathbb{1}(a \neq \arg\max_j \{\hat{f}_\theta^{\text{label}}(u^j)\})\end{aligned}$$

Estimation of ratios, $\hat{w}(X, Y)$. To estimate the $\hat{P}(y \mid x, a)$ we use the trained model $\hat{f}_\theta$ as follows:

$$\hat{P}(y \mid x = (u^1, u^2, u^3, u^4, u^5), a) = \hat{f}_\theta(u^a)_y$$

where $\hat{f}_\theta(u^a)_y$ corresponds to the softmax prediction of u^a for label y under the model $\hat{f}_\theta$. To estimate the behaviour policy $\hat{\pi}^b$, we train a classifier model $\mathcal{X} \to \mathcal{A}$ using a neural

network. We use (3.7) to estimate the weights $\hat{w}(x, y)$.

Neural Network Architectures

- To approximate the behaviour policy, we use a neural network with 2 hidden layers and 25 nodes in each hidden layer, ReLU activations and softmax output.
- To approximate $\hat{f}_\theta$, we use a neural network with 2 hidden layers with 64 nodes each and ReLU activations.

Results: Coverage as a function of increase calibration data. As mentioned in the main text, we have also performed experiments to investigate how much calibration data is needed for COPP as well as other methods to converge to the required 90% coverage. In the below plot we have plotted the coverage as a function of n calibration data points. We observe that our proposed method is converging much faster to the required coverage compared to the competing methods.

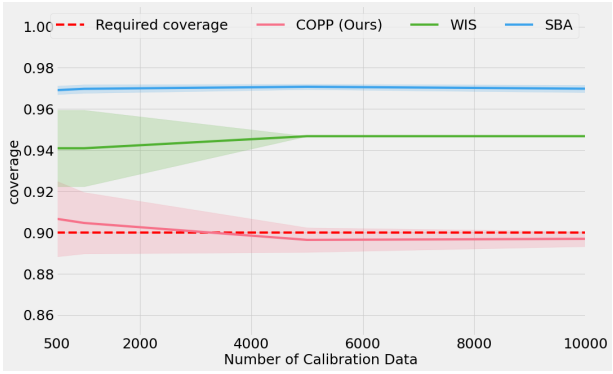


Figure B.5: Results of Microsoft Ranking Dataset experiment with behaviour policy $\pi^b = \pi_{0.5}$ and the target policy is $\pi^* = \pi_{0.2}$. Our proposed method is able to converge to the required coverage rather quickly compared to the competing methods

Table B.5: Coverages for COPP with and without label conditioned coverage, $\hat{C}_n^{\mathcal{Y}}$ and $\hat{C}_n$ respectively. Overall coverage refers to marginal coverage while $Y = y$ refers to coverage conditioned on $Y = y$. Here n_{test} corresponds to the number of test data points ($\sim P^{\pi^*}$).

	n_{test}	$\hat{C}_n$ Cov	$\hat{C}_n^{\mathcal{Y}}$ Cov
Overall	5000	0.896 ± 0.005	0.941 ± 0.003
$Y = 0$	266	0.700 ± 0.020	1.000 ± 0.000
$Y = 1$	293	0.526 ± 0.019	1.000 ± 0.000
$Y = 2$	228	0.772 ± 0.018	0.990 ± 0.029
$Y = 3$	320	0.852 ± 0.015	0.964 ± 0.035
$Y = 4$	3893	0.950 ± 0.006	0.928 ± 0.003

Results: COPP for Class-balanced coverage

Table B.5 shows the coverages of COPP predictive sets ($\hat{C}_n$ with marginal coverage guarantee constructed using algorithm 1) and COPP intervals with label conditioned coverage ($\hat{C}_n^{\mathcal{Y}}$ satisfying (B.37) constructed using algorithm 2). Extensions of WIS and SBA to the conditional case are not straightforward and hence have not been included. For $\hat{C}_n$, while the overall coverage is very close to the required coverage of 90%, we see that there is under-coverage for $Y = 0, 1, 2, 3$. This can be explained by the data imbalance – the number of test data points with $Y = 0, 1, 2, 3$ is significantly lower than $Y = 4$.

This under-coverage problem disappears in $\hat{C}_n^{\mathcal{Y}}$. Instead, in cases where number of data points is small, ($Y = 0, 1, 2, 3$), the predictive sets $\hat{C}_n^{\mathcal{Y}}$ are conservative (i.e. have coverage $> 90\%$). As a result, the overall coverage increases to 0.941. This is a price to be paid for label conditioned coverage – the overall coverage may increase, however, being conservative in safety-critical settings is better than being overly optimistic.

B.4.3 UCI Dataset experiments

Following Huang et al. [2021], Dudík et al. [2014a], Wang et al. [2017a] we apply COPP on UCI classification datasets. We can pose classification as contextual bandits by defining the covariates $\mathcal{X}$ as the features, the action space $\mathcal{A} = \mathcal{K}$, where $\mathcal{K}$ is the set of labels, and the outcomes are binary, i.e. $\mathcal{Y} = \{0, 1\}$, defined by $Y \mid X, A = \mathbb{1}(X \text{ belongs to class } A)$. Here we use $m = 1000$ training data points.

Behaviour and Target Policies. First we train a neural network classifier mapping each covariate to the softmax scores for each class, $\hat{f}_\theta : \mathcal{X} \to [0,1]^{|\mathcal{K}|}$. We use this trained model $\hat{f}_\theta$ to define a family of policies such that we pick the most likely label as predicted by $\hat{f}_\theta$ with probability ϵ and the rest uniformly with probability. Formally, this can be expressed as follows:

$$\pi_\epsilon(a \mid x) \coloneqq \epsilon \mathbb{1}(a = \arg\max_{k \in \mathcal{K}} \{\hat{f}_\theta(x)_k\}) + (1-\epsilon)/(|\mathcal{K}|-1)\mathbb{1}(a \neq \arg\max_{k \in \mathcal{K}} \{\hat{f}_\theta(x)_k\})$$

Like other experiments, we use ϵ to control the shift between behaviour and target policies. For π^b, we use $\epsilon^b = 0.5$ and for $\epsilon^* \in \{0.05, 0.3, 0.4, 0.5, 0.6, 0.7, 0.95\}$. Using this behaviour policy π^b, we generate an observational dataset $\mathcal{D}_{obs} = \{x_i, a_i, y_i\}_{i=1}^{n_{obs}}$ which is then split into training $\mathcal{D}_{tr}$ and calibration datasets $\mathcal{D}_{cal}$, of sizes m and n respectively.

Estimation of ratios, $\hat{w}(X,Y)$. To estimate the $\hat{P}(y \mid x, a)$ we use the trained model $\hat{f}_\theta$ as follows:

$$\hat{P}(Y = 1 \mid x, a) = \hat{f}_\theta(x)_a$$

where $\hat{f}_\theta(x)_a$ corresponds the softmax prediction of x for label a under the model $\hat{f}_\theta$. To estimate the behaviour policy $\hat{\pi}^b$, we train a classifier model $\mathcal{X} \to \mathcal{A}$ using a neural network. We use (3.7) in main text to estimate weights $\hat{w}(x,y)$.

Score. We define $\hat{P}^{\pi^b}(y \mid x) = \sum_{i \in \mathcal{K}} \hat{\pi}^b(A = i|x)\hat{P}(y|x, A = i)$. Using similar formulation as in Angelopoulos and Bates [2021], we define the score as

$$s(x,y) = \sum_{y'=0,1} \hat{P}^{\pi^b}(y' \mid x)\mathbb{1}(\hat{P}^{\pi^b}(y' \mid x) \geq \hat{P}^{\pi^b}(y \mid x))$$

Neural Network Architectures

- To approximate the behaviour policy, we use a neural network with 2 hidden layers and 64 nodes in each hidden layer, ReLU activations and softmax output.
- To approximate $\hat{f}_\theta$, we use a neural network with 2 hidden layers with 64 nodes each and ReLU activations.

Results. Tables B.6-B.11 show the coverages across varying target policies for different classification datasets. The behaviour policy $\pi^b = \pi_{0.5}$ is fixed and we use $n = 5000$ calibration datapoints, across 10 runs with $m = 5000$ training data. The tables show that COPP is able to provide the required coverage of 90% across all target policies. Moreover, compared to COPP, SBA and WIS are overly conservative. WIS estimates are not adaptive w.r.t. X, and as a result, the predictive sets produced are uninformative (i.e. contain all outcomes) in these experiments where the outcome is binary.

We have also included a comparison of COPP using estimated behaviour policy with COPP using GT behaviour policy. The latter provides more accurate coverage, and using estimated behaviour policy provides slightly over-covered predictive sets comparatively in most cases. This can be explained by policy estimation error. Additionally, we observe that using standard CP leads to predictive sets which are not adaptive to policy shift. As a result, the standard CP predictive sets get overly conservative (optimistic) as Δ_ϵ becomes more negative (positive).

Table B.6: Yeast dataset results

	$\Delta_\epsilon = -0.45$	$\Delta_\epsilon = -0.2$	$\Delta_\epsilon = -0.1$	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.45$
COPP (Ours)	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.91±0.00
WIS	0.99±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
SBA	0.98±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
COPP (GT behav policy)	0.91±0.00	0.91±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00
CP (no policy shift)	0.97±0.00	0.93±0.00	0.92±0.00	0.90±0.00	0.89±0.00	0.87±0.00	0.83±0.00

Table B.7: Ecoli dataset results

	$\Delta_\epsilon = -0.45$	$\Delta_\epsilon = -0.2$	$\Delta_\epsilon = -0.1$	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.45$
COPP (Ours)	0.92±0.00	0.91±0.00	0.91±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00
WIS	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
SBA	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
COPP (GT behav policy)	0.91±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.01
CP (no policy shift)	0.92±0.00	0.91±0.00	0.91±0.00	0.90±0.00	0.90±0.00	0.89±0.00	0.88±0.00

Table B.8: Letter dataset results

	$\Delta_\epsilon = -0.45$	$\Delta_\epsilon = -0.2$	$\Delta_\epsilon = -0.1$	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.45$
COPP (Ours)	0.95±0.00	0.93±0.00	0.93±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.91±0.00
WIS	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
SBA	0.97±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
COPP (GT behav policy)	0.92±0.00	0.91±0.00	0.91±0.00	0.90±0.00	0.89±0.00	0.89±0.00	0.88±0.00
CP (no policy shift)	0.99±0.00	0.94±0.00	0.92±0.00	0.90±0.00	0.88±0.00	0.86±0.00	0.81±0.00

Table B.9: Optdigits dataset results

	$\Delta_\epsilon = -0.45$	$\Delta_\epsilon = -0.2$	$\Delta_\epsilon = -0.1$	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.45$
COPP (Ours)	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00	0.93±0.00
WIS	0.99±0.01	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
SBA	0.97±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00
COPP (GT behav policy)	0.91±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.89±0.00	0.89±0.00
CP (no policy shift)	0.97±0.00	0.93±0.00	0.91±0.00	0.90±0.00	0.88±0.00	0.87±0.00	0.83±0.00

Table B.10: Pendigits dataset results

	$\Delta_\epsilon = -0.45$	$\Delta_\epsilon = -0.2$	$\Delta_\epsilon = -0.1$	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.45$
COPP (Ours)	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.92±0.00	0.91±0.00
WIS	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
SBA	0.97±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00
COPP (GT behav policy)	0.91±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.89±0.00	0.89±0.00
CP (no policy shift)	0.99±0.00	0.94±0.00	0.92±0.00	0.90±0.00	0.88±0.00	0.86±0.00	0.81±0.00

Table B.11: Satimage dataset results

	$\Delta_\epsilon = -0.45$	$\Delta_\epsilon = -0.2$	$\Delta_\epsilon = -0.1$	$\Delta_\epsilon = 0.0$	$\Delta_\epsilon = 0.1$	$\Delta_\epsilon = 0.2$	$\Delta_\epsilon = 0.45$
COPP (Ours)	0.92±0.00	0.91±0.00	0.91±0.00	0.91±0.00	0.91±0.00	0.91±0.00	0.91±0.00
WIS	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
SBA	0.98±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	0.99±0.00
COPP (GT behav policy)	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.90±0.00	0.89±0.00
CP (no policy shift)	0.97±0.00	0.93±0.00	0.92±0.00	0.90±0.00	0.88±0.00	0.87±0.00	0.83±0.00

B.5 How the miscoverage depends on $\hat{P}(y \mid x, a)$

Proposition B.5.1

Let

$$\tilde{w}(x, y) \coloneqq \frac{\int \hat{P}(y \mid x, a)\pi^*(a \mid x)\mathrm{d}a}{\int \hat{P}(y \mid x, a)\pi^b(a \mid x)\mathrm{d}a}.$$

Assume that $\hat{P}(y \mid x, a)/P(y \mid x, a) \in [1/\Gamma, \Gamma]$ for some $\Gamma \geq 1$. Then,

$$\Delta_w \coloneqq \tfrac{1}{2}\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}} \mid \tilde{w}(X, Y) - w(X, Y) \mid \leq \Gamma^2 - 1.$$

Proof. In this proof, we investigate the error of the weights as a function of the error in $\hat{P}(y \mid x, a)$. Therefore, to isolate this effect we ignore the Monte Carlo error, and assume known behavioural policy π^b.

Under the assumption above, we have that

$$\begin{aligned}\frac{1/\Gamma \int P(y \mid x, a)\pi^*(a \mid x)\mathrm{d}a}{\Gamma \int P(y \mid x, a)\pi^b(a \mid x)\mathrm{d}a} \leq &\tilde{w}(x, y) \leq \frac{\Gamma \int P(y \mid x, a)\pi^*(a \mid x)\mathrm{d}a}{1/\Gamma \int P(y \mid x, a)\pi^b(a \mid x)\mathrm{d}a}. \\ \implies \frac{1}{\Gamma^2} w(x, y) \leq &\tilde{w}(x, y) \leq \Gamma^2 w(x, y)\end{aligned}$$

This means that,

$$\left(\frac{1}{\Gamma^2} - 1\right) w(x, y) \leq \tilde{w}(x, y) - w(x, y) \leq (\Gamma^2 - 1)w(x, y)$$

So,

$$\mid \tilde{w}(x, y) - w(x, y) \mid \leq (\Gamma^2 - 1)w(x, y)$$

And therefore,

$$\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}} \mid \tilde{w}(X, Y) - w(X, Y) \mid \leq (\Gamma^2 - 1)\mathbb{E}_{(X,Y)\sim P^{\pi^b}_{X,Y}}[w(X, Y)] = \Gamma^2 - 1$$

□

C

Causal Falsification of Digital Twins

Contents

C.1 Notation

$z_{t:t'}$	The sequence of elements $(z_t, \dots, z_{t'})$ (or the empty sequence when $t > t'$)
$\mathcal{Z}_{t:t'}$ (where each $\mathcal{Z}_i$ is a set)	The cartesian product $\mathcal{Z}_t \times \cdots \times \mathcal{Z}_{t'}$ (or the empty set when $t > t'$)
$Z_{t:t'}(a_{1:t'})$	The sequence of potential outcomes $Z_t(a_{1:t}), \dots, Z_{t'}(a_{1:t'})$ (or the empty sequence when $t > t'$)
$\mathrm{Law}[Z]$	The distribution of the random variable Z
$\mathrm{Law}[Z \mid M]$	The conditional distribution of Z given M, where M is either an event or a random variable
$Z \overset{\text{a.s.}}{=} Z'$	The random variables Z and Z' are almost surely equal, i.e. $\mathbb{P}(Z = Z') = 1$
$Z \perp\!\!\!\perp Z'$	The random variables Z and Z' are independent
$Z \perp\!\!\!\perp Z' \mid Z''$	The random variables Z and Z' are conditionally independent given the random variable Z''
$\mathbb{1}(E)$	Indicator function of some event E

C.2 Proof of Proposition 4.2.1 (unconditional form of interventional correctness)

Proof. Fix any choice of $a_{1:T} \in \mathcal{A}_{1:T}$. By disintegrating $\mathrm{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})]$ and $\mathrm{Law}[X_{0:T}(a_{1:T})]$ along their common $\mathcal{X}_0$-marginal (which is namely $\mathrm{Law}[X_0]$), it holds that

$$\mathrm{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})] = \mathrm{Law}[X_{0:T}(a_{1:T})] \tag{C.1}$$

if and only if

$$\mathrm{Law}[\widehat{X}_{1:T}(X_0, a_{1:T}) \mid X_0 = x_0] = \mathrm{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0] \tag{C.2}$$

for $\mathrm{Law}[X_0]$-almost all $x_0 \in \mathcal{X}_0$. But now, our definition of $\widehat{X}_{1:T}(x_0, a_{1:T})$ in terms of h_t and $U_{1:t}$ means we can write $\widehat{X}_{1:T}(X_0, a_{1:T}) = \boldsymbol{h}(X_0, a_{1:T}, U_{1:T})$, where

$$\boldsymbol{h}(x_0, a_{1:T}, u_{1:T}) \coloneqq (h_1(x_0, a_1, u_1), \dots, h_T(x_0, a_{1:T}, u_{1:T})).$$

For all $x_0 \in \mathcal{X}_0$ and measurable $B_{1:T} \subseteq \mathcal{X}_{1:T}$, we then have

$$\begin{aligned}\mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T}) &= \mathbb{E}[\mathbb{1}(\boldsymbol{h}(x_0, a_{1:T}, U_{1:T}) \in B_{1:T})] \\ &= \int \mathbb{1}(\boldsymbol{h}(x_0, a_{1:T}, u_{1:T}) \in B_{1:T}) \,\mathrm{Law}[U_{1:T}](\mathrm{d}u_{1:T}).\end{aligned}$$

It is standard to show that the right-hand side is a Markov kernel in x_0 and $B_{1:T}$. Moreover, for any measurable $B_0 \subseteq \mathcal{X}_0$, we have

$$\begin{aligned}
&\int_{B_0} \mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T})\,\mathrm{Law}[X_0](\mathrm{d}x_0) \\
&\quad= \int_{B_0} \left[\int \mathbb{1}(\boldsymbol{h}(x_0, a_{1:T}, u_{1:T}) \in B_{1:T})\,\mathrm{Law}[U_{1:T}](\mathrm{d}u_{1:T})\right] \mathrm{Law}[X_0](\mathrm{d}x_0) \\
&\quad= \int \mathbb{1}(x_0 \in B_0, \boldsymbol{h}(x_0, a_{1:T}, u_{1:T}) \in B_{1:T})\,\mathrm{Law}[X_0, U_{1:T}](\mathrm{d}x_0, \mathrm{d}u_{1:T}) \\
&\quad= \mathrm{Law}[X_0, \widehat{X}_{1:T}(X_0, a_{1:T})](B_{0:T}),
\end{aligned}$$

where the second step follows because $X_0 \perp\!\!\!\perp U_{1:T}$. It therefore follows that $(x_0, B_{1:T}) \mapsto \mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})](B_{1:T})$ is a regular conditional distribution of $\widehat{X}_{1:T}(X_0, a_{1:T})$ given X_0, i.e.

$$\mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})] = \mathrm{Law}[\widehat{X}_{1:T}(X_0, a_{1:T}) \mid X_0 = x_0] \qquad \text{for } \mathrm{Law}[X_0]\text{-almost all } x_0 \in \mathcal{X}_0.$$

Substituting this into (C.2), we see that (C.1) holds if and only if

$$\mathrm{Law}[\widehat{X}_{1:T}(x_0, a_{1:T})] = \mathrm{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0]$$

for $\mathrm{Law}[X_0]$-almost all $x_0 \in \mathcal{X}_0$. The result now follows since $a_{1:T}$ was arbitrary. □

C.3 Online prediction

C.3.1 Correctness in the online setting

A distinguishing feature of many digital twins is their ability to integrate real-time information obtained from sensors in their environment [Barricelli et al., 2019]. It is therefore relevant to consider a setting in which a twin is used repeatedly to make a sequence of predictions over time, each time taking all previous information into account. One way to formalize this is to instantiate our model for the twin at each timestep. For example, we could represent the predictions made by the twin at $t = 0$ after observing initial covariates x_0 as potential outcomes $(\widehat{X}^1_{1:T}(x_0, a_{1:T}) : a_{1:T} \in \mathcal{A}_{1:T})$, similar to what we did in the main text. We could then represent the predictions made by the twin after some action a_1 is taken and an additional observation x_1 is made via potential outcomes $(\widehat{X}^2_{2:T}(x_{0:1}, a_{1:T}) : a_{2:T} \in \mathcal{A}_{2:T})$. More generally, for $t \in \{1, \ldots, T\}$, we could introduce potential outcomes $(\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T}) : a_{t:T} \in \mathcal{A}_{t:T})$ to represent the predictions

that the twin would make at time t after the observations $x_{0:t-1}$ are made and the actions $a_{1:t-1}$ are taken.

This extended model requires a new definition of correctness than our Definition 4.2.1 from the main text. A natural approach is to say that the twin is correct in this new setting if

$$\text{Law}[\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})] = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}] \tag{C.3}$$

for all $t \in \{1, \ldots, T\}$, $a_{1:T} \in \mathcal{A}_{1:T}$, and $\text{Law}[X_{0:t-1}(a_{1:t-1})]$-almost all $x_{0:t-1} \in \mathcal{X}_{0:t-1}$. A twin with this property would at each step be able to accurately simulate the future in light of previous information, use this to choose a next action to take, observe the result of doing so, and then repeat. It is possible to show that (C.3) holds if and only if we have

$$\begin{aligned}
\text{Law}[\widehat{X}^1_{1:T}(x_0, a_{1:T})] &= \text{Law}[X_{1:T}(a_{1:T}) \mid X_0 = x_0] \\
\text{Law}[\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})] &= \text{Law}[\widehat{X}^1_{t:T}(x_0, a_{1:T}) \mid \widehat{X}^1_{1:t-1}(x_0, a_{1:t-1}) = x_{1:t-1}]
\end{aligned}$$

for all $t \in \{1, \ldots, T\}$, $a_{1:T} \in \mathcal{A}_{1:T}$, $\text{Law}[X_0]$-almost all $x_0 \in \mathcal{X}_0$, and $\text{Law}[\widehat{X}^1_{1:t-1}(x_0, a_{1:t-1})]$-almost all $x_{1:t-1} \in \mathcal{X}_{1:t-1}$. The first condition here says that $\widehat{X}^1_{1:T}(x_0, a_{1:T})$ must be interventionally correct in the sense of Definition 4.2.1 from the main text. The second condition says that the predictions made by the twin across different timesteps must be internally consistent with each other insofar as their conditional distributions must align. This holds automatically in many circumstances, such as if the predictions of the twin are obtained from a Bayesian model (for example), and otherwise could be checked numerically given the ability to run simulations from the twin, without the need to obtain data or refer to the real-world process in any way. As such, the problem of assessing the correctness of the twin in this new sense primarily reduces to the problem of assessing the correctness of $\widehat{X}^1_{1:T}(x_0, a_{1:T})$ in the sense of Definition 4.2.1 in the main text, which motivates our focus on that condition.

C.3.2 Alternative notions of online correctness

An important and interesting subtlety arises in this context that is worth noting. In general it does not follow that a twin correct in the sense of (C.3) satisfies

$$\text{Law}[\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})] = \text{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}, A_{1:t-1} = a_{1:t-1}] \tag{C.4}$$

for all $a_{1:T} \in \mathcal{A}_{1:T}$, and $\mathrm{Law}[X_{0:t-1}(a_{1:t-1}) \mid A_{1:t-1} = a_{1:t-1}]$-almost all $x_{0:t-1} \in \mathcal{X}_{0:t-1}$, since in general it does not hold that

$$\mathrm{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}] = \mathrm{Law}[X_{t:T}(a_{1:T}) \mid X_{0:t-1}(a_{1:t-1}) = x_{0:t-1}, A_{1:t-1} = a_{1:t-1}].$$

for all $a_{1:T} \in \mathcal{A}_{1:T}$ and $\mathrm{Law}[X_{0:t-1}(a_{1:t-1}) \mid A_{1:t-1} = a_{1:t-1}]$-almost all $x_{0:t-1} \in \mathcal{X}_{0:t-1}$ unless the actions $A_{1:t-1}$ are unconfounded. (Here as usual $A_{1:T}$ denotes the actions of a behavioural agent; see Section 4.3 of the main text.) In other words, a twin that is correct in the sense of (C.3) will make accurate predictions at time t when every action taken before time t was unconfounded (as occurs for example when the twin is directly in control of the decision-making process), but in general not when certain taken actions before time t were chosen by a behavioural agent with access to more context than is available to the twin (as may occur for example when the twin is used as a decision-support tool). However, should it be desirable, our framework could be extended to encompass the alternative condition in (C.4) by relabelling the observed history $(X_{0:t-1}(A_{1:t-1}), A_{1:t-1})$ as X_0, and then assessing the correctness of the potential outcomes $\widehat{X}^t_{t:T}(x_{0:t-1}, a_{1:T})$ in the sense of Definition 4.2.1 from the main text.

Overall, the "right" notion of correctness in this online setting is to some extent a design choice. We believe our causal approach to twin assessment provides a useful framework for formulating and reasoning about these possibilities, and consider the investigation of assessment strategies for additional usage regimes to be an interesting direction for future work.

C.4 Proof of Theorem 4.3.1 (interventional distributions are not identifiable)

It is well-known in the causal inference literature that the interventional behaviour of the real-world process cannot be uniquely identified from observational data. For completeness, we now provide a self-contained proof of this result in our notation. Our statement here is lengthier than Theorem 4.3.1 in the main text in order to clarify what is meant by "uniquely identified": intuitively, the idea is that there always exist distinct families of potential outcomes whose interventional behaviours differ and yet give rise to the same observational data.

Theorem C.4.1

Suppose we have $a_{1:T} \in \mathcal{A}_{1:T}$ such that $\mathbb{P}(A_{1:T} \neq a_{1:T}) > 0$. Then there exist potential outcomes $(\tilde{X}_{0:T}(a'_{1:T}) : a'_{1:T} \in \mathcal{A}_{1:T})$ such that

$$(\tilde{X}_{0:T}(A_{1:T}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), A_{1:T}). \tag{C.5}$$

but for which $\text{Law}[\tilde{X}_{0:T}(a_{1:t})] \neq \text{Law}[X_{0:T}(a_{1:t})]$.

Proof. Our assumption that $\mathbb{P}(A_{1:T} \neq a_{1:T}) > 0$ means there must exist some $t \in \{1, \dots, T\}$ such that $\mathbb{P}(A_{1:t} \neq a_{1:t}) > 0$. Since $\mathcal{X}_t = \mathbb{R}^{d_t}$, we may also choose some $x_t \in \mathcal{X}_t$ with $\mathbb{P}(X_t(a_{1:t}) = x_t \mid A_{1:t} \neq a_{1:t}) \neq 1$. Then, for each $s \in \{0, \dots, T\}$ and $a'_{1:s} \in \mathcal{A}_{1:s}$, define

$$\tilde{X}_s(a'_{1:s}) \coloneqq \begin{cases} \mathbb{1}(A_{1:t} = a_{1:t})\, X_t(a_{1:t}) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, x_t & \text{if } s = t \text{ and } a'_{1:s} = a_{1:t} \\ X_s(a'_{1:s}) & \text{otherwise,} \end{cases}$$

It is then easily checked that (C.5) holds, but

$$\begin{aligned} \text{Law}[\tilde{X}_t(a_{1:t})] &= \text{Law}[\tilde{X}_t(a_{1:t}) \mid A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t}) + \text{Law}[\tilde{X}_t(a_{1:t}) \mid A_{1:t} \neq a_{1:t}]\, \mathbb{P}(A_{1:t} \neq a_{1:t}) \\ &= \text{Law}[X_t(a_{1:t}) \mid A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t}) + \text{Dirac}(x_t)\, \mathbb{P}(A_{1:t} \neq a_{1:t}) \\ &\neq \text{Law}[X_t(a_{1:t}) \mid A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t}) + \text{Law}[X_t(a_{1:t}) \mid A_{1:t} \neq a_{1:t}]\, \mathbb{P}(A_{1:t} \neq a_{1:t}) \\ &= \text{Law}[X_t(a_{1:t})], \end{aligned}$$

from which the result follows. □

C.5 Deterministic potential outcomes are unconfounded

In this section we expand on our earlier claim that, if the real-world process is deterministic, then the observational data is unconfounded. We first make this claim precise. By "deterministic", we mean that there exist measurable functions g_t for $t \in \{1, \dots, T\}$ such that

$$X_t(a_{1:t}) \overset{\text{a.s.}}{=} g_t(X_{0:t-1}(a_{1:t-1}), a_{1:t}) \qquad \text{for all } t \in \{1, \dots, T\} \text{ and } a_{1:t} \in \mathcal{A}_{1:t}. \tag{C.6}$$

By "unconfounded", we mean that the *sequential randomisation assumption (SRA)* introduced by Robins [Robins, 1986] holds, i.e.

$$(X_s(a_{1:s}) : s \in \{1, \dots, T\}, a_{1:s} \in \mathcal{A}_{1:s}) \perp\!\!\!\perp A_t \mid X_{0:t-1}(A_{1:t-1}), A_{1:t-1} \qquad \text{for all } t \in \{1, \dots, T\}, \tag{C.7}$$

where $\perp\!\!\!\perp$ denotes conditional independence. Intuitively, this says that, apart from the historical observations $(X_{0:t-1}(A_{1:t-1}), A_{1:t-1})$, any additional factors that influence the agent's choice of action A_t are independent of the behaviour of the real-world process. The SRA provides a standard formulation of the notion of unconfoundedness in longitudinal settings such as ours (see [Tsiatis et al., 2019, Chapter 5] for a review).

It is now a standard exercise to show that (C.6) implies (C.7). We include a proof below for completeness. Key to this is the following straightforward Lemma.

Lemma C.5.1

Suppose U and V are random variables such that, for some measurable function g, it holds that $U \overset{\text{a.s.}}{=} g(V)$. Then, for any other random variable W, we have

$$U \perp\!\!\!\perp W \mid V.$$

Proof. By standard properties of conditional expectations, for any measurable sets S_1 and S_2, we have almost surely

$$\begin{aligned}
\mathbb{P}(U \in S_1, W \in S_2 \mid V) &= \mathbb{E}[\mathbb{1}(g(V) \in S_1)\, \mathbb{1}(W \in S_2) \mid V] \\
&= \mathbb{1}(g(V) \in S_1)\, \mathbb{E}[\mathbb{1}(W \in S_2) \mid V] \\
&= \mathbb{E}[\mathbb{1}(U \in S_1) \mid V]\, \mathbb{P}(W \in S_2 \mid V) \\
&= \mathbb{P}(U \in S_1 \mid V)\, \mathbb{P}(W \in S_2 \mid V),
\end{aligned}$$

which gives the result. □

It is now easy to see that (C.6) implies (C.7). Indeed, by recursive substitution, it is straightforward to show that there exist measurable functions $\tilde{g}_t$ for $t \in \{1, \ldots, T\}$ such that

$$X_t(a_{1:t}) \overset{\text{a.s.}}{=} \tilde{g}_t(X_0, a_{1:t}) \qquad \text{for all } t \in \{1, \ldots, T\} \text{ and } a_{1:t} \in \mathcal{A}_{1:t},$$

and so

$$(X_s(a_{1:s}) : s \in \{1, \ldots, T\}, a_{1:s} \in \mathcal{A}_{1:s}) = (\tilde{g}_t(X_0, a_{1:s}) : s \in \{1, \ldots, T\}, a_{1:s} \in \mathcal{A}_{1:s}).$$

The right-hand side is now seen to be a measurable function of X_0 and hence certainly of $(X_{0:t-1}(A_{1:t-1}), A_{1:t-1})$, so that the result follows by Lemma C.5.1.

C.6 Motivating toy example

We provide here a toy scenario that illustrates intuitively the pitfalls that may arise when assessing twins using observational data without properly accounting for causal considerations (including unmeasured confounding in particular). Suppose a digital twin has been designed for a particular make of car, e.g. to facilitate autonomous driving [Allamaa et al., 2022]. The twin simulates how quantities such as the velocity and fuel consumption of the car respond as certain inputs are applied to it, such as braking, acceleration, steering, etc. We wish to assess the accuracy of this twin using a dataset obtained from a fleet of the same make. The braking performance of these vehicles is significantly affected by the age of their brake pads: if these are fairly new, then an aggressive braking strategy will stop the car, while if these are old, then the same aggressive strategy will send the car into a skid that will reduce braking efficacy. Brake pad age is not recorded in the data we have obtained, but *was* known to the drivers who operated these vehicles (e.g. perhaps they were aware of how recently their car was serviced), and so the drivers of cars with old brake pads tended to avoid braking aggressively out of safety concerns.

A naive approach to twin assessment in this situation would directly compare the outputs of the twin with the data and conclude the twin is accurate if these match closely. However, in this scenario, the data contains a spurious relationship between braking strategy and the performance of the car: since aggressive braking is only observed for cars with new brake pads, the data appears to show that aggressive braking is effective at stopping the car, while in fact this is not the case for cars with older brake pads. As such, the naive assessment approach would yield misleading information about the twin: a twin that captures only the behaviour of cars with newer brake pads would appear to be correct, while a twin that captures the full range of possibilities (i.e. regardless of brake pad age) would deviate from the observational data and appear therefore less accurate. Figure C.1 illustrates this pictorially under a toy model for this scenario.

In the causal inference literature, any unmeasured quantity (e.g. brake pad age) that affects both some choice of action taken in the data (e.g. aggressive braking) and the resulting observation (e.g. speed) is referred to as an *unmeasured confounder*. In general, whenever an unmeasured confounder is present, a potential discrepancy arises between how the real-world process was observed to behave in the dataset and how it *would* behave

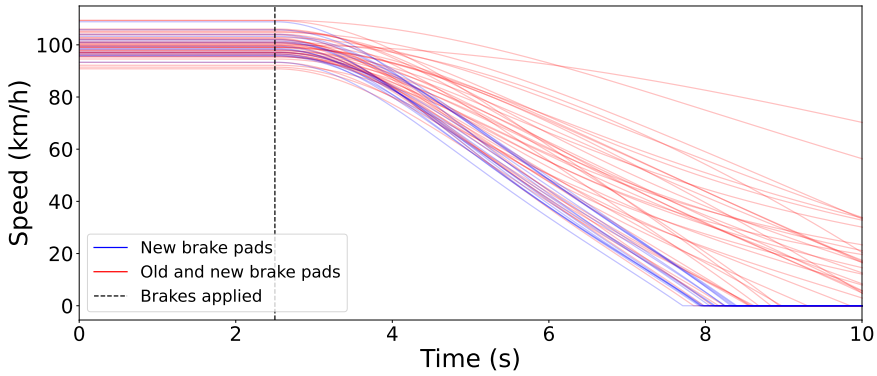


Figure C.1: The discrepancy between observational data and interventional behavior. The data only show the effect of aggressive braking on cars with new brake pads (blue). This differs from what *would* be observed if aggressive braking were applied to the entire fleet of cars, encompassing both those with old and new brake pads (red).

under certain interventions. An obvious approach towards mitigating this possibility is to measure additional quantities that may affect the outcome of interest. For example, if brake pad age were included in the data in the scenario above, then it would be possible to adjust for its effect on braking performance. However, in many cases, gathering additional data may be costly or impractical. Moreover, even if this strategy is pursued, it is rarely possible to rule out the possibility of unmeasured confounding altogether, especially for complicated real-world problems [Tsiatis et al., 2019]. For example, in the scenario above, it is very conceivable that some other factor such as weather conditions could play a similar confounding role as brake pad age, and so would need also to be included in the data, and so on. Analogous scenarios are also easily forthcoming for other application domains such as medicine and economics [Manski, 1995, Tsiatis et al., 2019, Hernán and Robins, 2020]. As such, rather than attempting to sidestep the issue of unmeasured confounding, we instead propose a methodology for assessing twins using data that is robust to its presence.

C.7 Causal bounds

C.7.1 Proof of Theorem 4.4.1

Proof. We prove the lower bound; the upper bound is analogous. It is easily checked that

$$\begin{aligned}
&\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \\
&\quad = \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) \\
&\qquad + \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} \neq a_{1:t}]\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}). \quad \text{(C.8)}
\end{aligned}$$

If $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) > 0$, then

$$\begin{aligned}\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}] &= \mathbb{E}[Y(A_{1:t}) \mid X_{0:t}(A_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t}] \\ &= \mathbb{E}[Y(A_{1:N}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}],\end{aligned}$$

where the second step follows because $\mathbb{P}(N = t \mid A_{1:t} = a_{1:t}) = 1$. Similarly, if $\mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) > 0$, then (4.5) implies

$$\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} \neq a_{1:t}] \geq y_{\text{lo}}.$$

Substituting these results into (C.8), we obtain

$$\begin{aligned}&\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \\ &\quad \geq \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) \\ &\qquad + y_{\text{lo}}\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}). \quad \text{(C.9)}\end{aligned}$$

Now observe that the right-hand side of (C.9) is a convex combination with mixture weights $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t})$ and $\mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t})$. We can bound

$$\begin{aligned}\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:t}(a_{1:t}) \in B_{0:t}) &= \frac{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t})} \\ &\geq \frac{\mathbb{P}(X_{0:t}(a_{1:t}) \in B_{0:t}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(a_{1:N}) \in B_{0:N})} \\ &= \frac{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t})}{\mathbb{P}(X_{0:N}(A_{1:N}) \in B_{0:N})} \\ &= \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}), \quad \text{(C.10)}\end{aligned}$$

where the inequality holds because $t \geq N$ almost surely, and the second equality holds because the definition of N means

$$X_{0:N}(a_{1:N}) \overset{\text{a.s.}}{=} X_{0:N}(A_{1:N}).$$

As such, we can bound the convex combination in (C.9) from below by replacing its mixture weights with $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ and $\mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:N}(a_{1:N}) \in B_{0:N})$, which shifts weight from the $\mathbb{E}[Y(A_{1:t}) \mid A_{1:t} = a_{1:t}, X_{0:N}(A_{1:N}) \in B_{0:N}]$ term onto the y_{lo}

term. This yields

$$\begin{aligned}
&\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}] \\
&\qquad \geq \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(a_{1:N}) \in B_{0:N}) \\
&\qquad\qquad + y_{\mathrm{lo}}\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:N}(a_{1:N}) \in B_{0:N}) \\
&\qquad = \mathbb{E}[Y(A_{1:t})\, \mathbb{1}(A_{1:t} = a_{1:t}) + y_{\mathrm{lo}}\, \mathbb{1}(A_{1:t} \neq a_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}] \\
&\qquad = \mathbb{E}[Y_{\mathrm{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}].
\end{aligned}$$

□

C.7.2 Proof of Proposition 4.4.1

Proof. From the definition of Y_{up}, we have straightforwardly

$$\begin{aligned}
Q_{\mathrm{up}} = \mathbb{E}[Y(A_{1:t}) \mid X_{0:N}(A_{1:N}) \in B_{0:N}, A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}) \\
+ y_{\mathrm{up}}\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}).
\end{aligned}$$

A similar expression holds for Q_{lo}. Subtracting these two expressions yields

$$Q_{\mathrm{up}} - Q_{\mathrm{lo}} = (y_{\mathrm{up}} - y_{\mathrm{lo}})\, (1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})).$$

Similar manipulations show that

$$\mathbb{E}[Y_{\mathrm{up}}] - \mathbb{E}[Y_{\mathrm{lo}}] = (y_{\mathrm{up}} - y_{\mathrm{lo}})\, (1 - \mathbb{P}(A_{1:t} = a_{1:t})),$$

and the result now follows.

□

C.7.3 Proof of Proposition 4.4.2 and discussion

Proof. We consider the case of the lower bound; the case of the upper bound is analogous. Choose $x_{1:T} \in B_{1:T}$ arbitrarily. (Certainly some choice is always possible, since each B_s has positive measure and is therefore nonempty.) Define

$$\begin{aligned}
\tilde{X}_0 &\coloneqq X_0 \\
\tilde{X}_s(a'_{1:s}) &\coloneqq \mathbb{1}(A_{1:s} = a'_{1:s})\, X_s(a'_{1:s}) + \mathbb{1}(A_{1:s} \neq a'_{1:s})\, x_s \qquad \text{for each } s \in \{0, \ldots, T\} \text{ and } a'_{1:s} \in \mathcal{A}_{1:s},
\end{aligned}$$

and similarly let

$$\tilde{Y}(a'_{1:t}) = \mathbb{1}(A_{1:t} = a'_{1:t})\, Y(a'_{1:t}) + \mathbb{1}(A_{1:t} \neq a'_{1:t})\, y_{\text{lo}} \qquad \text{for all } a'_{1:t} \in \mathcal{A}_{1:t}.$$

It is easy to check that $(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T})$. But now we have directly $\tilde{Y}(a_{1:t}) = Y_{\text{lo}}$. Moreover, it is easily checked from the definition of N and $\tilde{X}_{0:t}(a_{1:t})$ that

$$\tilde{X}_{0:t}(a_{1:t}) \overset{\text{a.s.}}{=} (X_{0:N}(A_{1:N}), x_{N+1:t}),$$

so that

$$\begin{aligned}\mathbb{1}(\tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}) &\overset{\text{a.s.}}{=} \mathbb{1}(\tilde{X}_{0:N}(a_{1:N}) \in B_{0:N}, x_{N+1:t} \in B_{N+1:t}) \\ &\overset{\text{a.s.}}{=} \mathbb{1}(X_{0:N}(A_{1:N}) \in B_{0:N})\end{aligned}$$

since each $x_s \in B_s$. Consequently,

$$\begin{aligned}\mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}] &= \mathbb{E}[Y_{\text{lo}} \mid \tilde{X}_{0:t}(a_{1:t}) \in B_{0:t}] \\ &= \mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}],\end{aligned}$$

which gives the result. □

C.7.4 Bounds on the conditional expectation given specific covariate values

Theorem 4.4.1 provides a bound on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t}) \in B_{0:t}]$, i.e. the conditional expectation given the *event* $\{X_{0:t}(a_{1:t}) \in B_{0:t}\}$, which is assumed to have positive probability. We consider here the prospect of obtaining bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$, i.e. the conditional expectation given the *value* of $X_{0:t}(a_{1:t})$. For falsification purposes, this would provide a means for determining that twin is incorrect when it outputs specific values of $\widehat{X}_{0:t}(a_{1:t})$, rather than just that it is incorrect on average across all runs that output values $\widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}$.

When $X_{0:t}(a_{1:t})$ is discrete, Theorem 4.4.1 yields measurable functions $g_{\text{lo}}, g_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ such that

$$g_{\text{lo}}(X_{0:t}(a_{1:t})) \leq \mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})] \leq g_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely.} \tag{C.11}$$

In particular, $g_{\text{lo}}(x_{0:t})$ is obtained as the value of $\mathbb{E}[Y_{\text{lo}} \mid X_{0:N}(A_{1:N}) \in B_{0:N}]$ for $B_{0:t} \coloneqq \{x_{0:t}\}$, and similarly for $g_{\text{up}}(x_{0:t})$. Moreover, since the constants $y_{\text{lo}}, y_{\text{up}} \in \mathbb{R}$ in Theorem

4.4.1 were allowed to depend on $B_{0:t}$, and hence here on each choice of $x_{0:t} \in \mathcal{X}_{0:t}$, we may think of these now as measurable functions $y_{\text{lo}}, y_{\text{up}} : \mathcal{X}_{0:t} \to \mathbb{R}$ satisfying

$$y_{\text{lo}}(X_{0:t}(a_{1:t})) \leq Y(a_{1:t}) \leq y_{\text{up}}(X_{0:t}(a_{1:t})) \qquad \text{almost surely.} \tag{C.12}$$

In other words, when $X_{0:t}(a_{1:t})$ is discrete, Theorem 4.4.1 provides bounds on the conditional expectation of $Y(a_{1:t})$ given the value of $X_{0:t}(a_{1:t})$ whenever we have y_{lo} and y_{up} such that (C.12) holds.

When $\mathbb{P}(X_{1:t}(a_{1:t}) \in B_{1:t}) > 0$, a fairly straightforward modification of the proof of Theorem 4.4.1 yields bounds of the following form:

$$\begin{aligned}\mathbb{E}[Y_{\text{lo}} \mid X_0, X_{1:N}(A_{1:N}) \in B_{1:N}] &\leq \mathbb{E}[Y(a_{1:t}) \mid X_0, X_{1:t}(a_{1:t}) \in B_{1:t}] \\ &\leq \mathbb{E}[Y_{\text{up}} \mid X_0, X_{1:N}(A_{1:N}) \in B_{1:N}] \qquad \text{almost surely.}\end{aligned} \tag{C.13}$$

In particular, this holds regardless of whether or not X_0 is discrete. In turn, if $X_{1:t}(a_{1:t})$ is discrete, then by a similar argument as was given in the previous subsection, this yields almost sure bounds on $\mathbb{E}[Y(a_{1:t}) \mid X_{0:t}(a_{1:t})]$ of the form in (C.11), provided (C.12) holds. Alternatively, by taking $B_{1:t} \coloneqq \mathcal{X}_{1:t}$, (C.13) yields bounds of the form

$$\mathbb{E}[Y_{\text{lo}} \mid X_0] \leq \mathbb{E}[Y(a_{1:t}) \mid X_0] \leq \mathbb{E}[Y_{\text{up}} \mid X_0].$$

If the action sequence $a_{1:t}$ is thought of as a single choice of an action from the extended action space $\mathcal{A}_{1:t}$, then this recovers the bounds originally proposed by Manski [1990], which allowed conditioning on potentially continuous pre-treatment covariates corresponding to our X_0.

C.7.5 Proof of Theorem 4.4.2 and discussion

Proof. Suppose we have a permissible g_{lo}. (The case of g_{up} is analogous). Choose $x_{1:T} \in \mathcal{X}_{1:T}$ arbitrarily, and define new potential outcomes

$$\begin{aligned}\tilde{X}_0 &\coloneqq X_0 \\ \tilde{X}_r(a'_{1:r}) &\coloneqq \mathbb{1}(A_{1:r} = a'_{1:r})\, X_r(a'_{1:r}) + \mathbb{1}(A_{1:r} \neq a'_{1:r})\, x_r \qquad \text{for } r \in \{1, \dots, T\} \text{ and } a'_{1:r} \in \mathcal{A}_{1:r}.\end{aligned}$$

Similarly, define

$$\tilde{Y}(a'_{1:t}) \coloneqq \mathbb{1}(A_{1:t} = a'_{1:t})\, Y(a'_{1:t}) + \mathbb{1}(A_{1:t} \neq a'_{1:t})\, y_{\text{lo}}(\tilde{X}_{0:t}(a'_{1:t})) \qquad \text{for all } a'_{1:t} \in \mathcal{A}_{1:t}.$$

It immediately follows that

$$(\tilde{X}_{0:T}(A_{1:T}), \tilde{Y}(A_{1:t}), A_{1:T}) \overset{\text{a.s.}}{=} (X_{0:T}(A_{1:T}), Y(A_{1:t}), A_{1:T}).$$

Moreover, it is easily checked that

$$y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \leq \tilde{Y}(a_{1:t}) \leq y_{\text{up}}(\tilde{X}_{0:t}(a_{1:t})) \qquad \text{almost surely.}$$

As such, since g_{lo} is permissible, we must have, almost surely,

$$\begin{aligned} g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) &\leq \mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t})] \\ &= \mathbb{E}[\tilde{Y}(A_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}), A_{1:t} = a_{1:t}]\, \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) \\ &\quad + \underbrace{\mathbb{E}[\tilde{Y}(a_{1:t}) \mid \tilde{X}_{0:t}(a_{1:t}), A_{1:t} \neq a_{1:t}]}_{= y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t}))}\, \mathbb{P}(A_{1:t} \neq a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})). \end{aligned} \tag{C.14}$$

Now, by our definition of $\tilde{X}_{0:t}(a_{1:t})$, we have almost surely

$$\begin{aligned} \mathbb{1}(A_1 \neq a_1)\, \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) &= \mathbb{1}(A_1 \neq a_1, \tilde{X}_s(a_{1:s}) = x_s)\, \mathbb{P}(A_{1:t} = a_{1:t} \mid \tilde{X}_{0:t}(a_{1:t})) \\ &= \mathbb{1}(A_1 \neq a_1)\, \mathbb{E}[\mathbb{1}(A_{1:t} = a_{1:t}, \tilde{X}_s(a_{1:s}) = x_s) \mid \tilde{X}_{0:t}(a_{1:t})] \\ &= \mathbb{1}(A_1 \neq a_1)\, \mathbb{E}[\mathbb{1}(A_{1:t} = a_{1:t}, X_s(A_{1:s}) = x_s) \mid \tilde{X}_{0:t}(a_{1:t})] \\ &= 0, \end{aligned}$$

where the last step follows by our assumption that $\mathbb{P}(X_s(A_{1:s}) = x_s) = 0$. Combining this with (C.14), we get, almost surely,

$$\begin{aligned} \mathbb{1}(A_1 \neq a_1)\, g_{\text{lo}}(X_0, x_{1:t}) &= \mathbb{1}(A_1 \neq a_1)\, g_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \\ &\leq \mathbb{1}(A_1 \neq a_1)\, y_{\text{lo}}(\tilde{X}_{0:t}(a_{1:t})) \\ &= \mathbb{1}(A_1 \neq a_1)\, y_{\text{lo}}(X_0, x_{1:t}). \end{aligned} \tag{C.15}$$

Now let $x_0 \in \mathcal{X}_0$ be the value such that $\mathbb{P}(X_0 = x_0) = 1$. Using our assumption that $\mathbb{P}(A_1 \neq a_1) > 0$ and the fact that $x_{1:t}$ was arbitrary, we obtain

$$g_{\text{lo}}(x_{0:t}) \leq y_{\text{lo}}(x_{0:t}) \qquad \text{for all } x_{1:t} \in \mathcal{X}_{1:t}.$$

The result now follows. □

To gain intuition for the phenomenon underlying Theorem 4.4.2, consider a simplified model consisting of $\mathcal{X}$-valued potential outcomes $(X(a') : a \in \mathcal{A})$, $\mathbb{R}$-valued potential outcomes $(Y(a') : a \in \mathcal{A})$, and an $\mathcal{A}$-valued random variable A representing the choice of action. (This constitutes a special case of our setup with $T = 1$ and $\mathcal{X}_0$ taken to be a singleton set.) Suppose moreover that the following conditions hold:

$$\mathbb{P}(X(A) = x) = 0 \qquad \text{for all } x \in \mathcal{X}$$
$$\mathbb{P}(A = a) < 1.$$

We then have

$$\mathbb{E}[Y(a) \mid X(a)] \overset{\text{a.s.}}{=} \mathbb{E}[Y(A) \mid X(A), A = a]\,\mathbb{P}(A = a \mid X(a)) + \mathbb{E}[Y(a) \mid X(a), A \neq a]\,\mathbb{P}(A \neq a \mid X(a)). \tag{C.16}$$

But now, since the behaviour of $X(a)$ is only observed on $\{A = a\}$, for any given value of $x \in \mathcal{X}$, we cannot rule out the possibility that

$$X(a) = \mathbb{1}(A = a)\,X(A) + \mathbb{1}(A \neq a)\,x \qquad \text{almost surely.}$$

In turn, since $\mathbb{P}(A = a) > 0$, this would imply $\mathbb{P}(X(a) = x) > 0$, and, since $\mathbb{P}(X(A) = x) = 0$, that $\mathbb{P}(A = a \mid X(a) = x) = 0$. From (C.16), this would yield

$$\mathbb{E}[Y(a) \mid X(a) = x] = \mathbb{E}[Y(a) \mid X(a) = x, A \neq a].$$

But now, since the behaviour of $Y(a)$ is unobserved on $\{A \neq a\}$, intuitively speaking, the observational distribution does not provide any information about the value of the right-hand side, and therefore about the behaviour of $\mathbb{E}[Y(a) \mid X(a)]$ more generally since $x \in \mathcal{X}$ was arbitrary.

C.8 Hypothesis testing methodology

C.8.1 Validity of testing procedure

We show here that our procedure for testing $\widehat{Q} \geq Q_{\text{lo}}$ based on the one-sided confidence intervals R^{α}_{lo} and $\widehat{R}^{\alpha}$ has the correct probability of type I error, provided R^{α}_{lo} and $\widehat{R}^{\alpha}$ have the correct coverage probabilities. In particular, the result below (which applies a standard union bound argument) shows that if $\widehat{Q} \geq Q_{\text{lo}}$, then our test rejects (i.e. $\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}$) with

probability at most α. An analogous result is easily proven for testing $\widehat{Q} \leq Q_{\text{up}}$ also, with R^{α}_{lo} replaced by a one-sided upper $(1-\alpha/2)$-confidence interval for Q_{up}, and $\widehat{R}^{\alpha}$ replaced by a one-sided lower $(1-\alpha/2)$-confidence interval for $\widehat{Q}$.

Proposition C.8.1

Suppose that for some $\alpha \in (0,1)$ we have random variables $\widehat{R}^{\alpha}$ and R^{α}_{lo} satisfying

$$\mathbb{P}(Q_{\text{lo}} \geq R^{\alpha}_{\text{lo}}) \geq 1 - \frac{\alpha}{2} \tag{C.17}$$

$$\mathbb{P}(\widehat{Q} \leq \widehat{R}^{\alpha}) \geq 1 - \frac{\alpha}{2}. \tag{C.18}$$

If $\widehat{Q} \geq Q_{\text{lo}}$, then $\mathbb{P}(\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}) \leq \alpha$.

Proof. If $\widehat{Q} \geq Q_{\text{lo}}$, then we have

$$\{\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}\} \subseteq \{\widehat{Q} > \widehat{R}^{\alpha}\} \cup \{Q_{\text{lo}} < R^{\alpha}_{\text{lo}}\}.$$

To see this, note that

$$(\{\widehat{Q} > \widehat{R}^{\alpha}\} \cup \{Q_{\text{lo}} < R^{\alpha}_{\text{lo}}\})^{c} = \{\widehat{Q} > \widehat{R}^{\alpha}\}^{c} \cap \{Q_{\text{lo}} < R^{\alpha}_{\text{lo}}\}^{c} = \{\widehat{Q} \leq \widehat{R}^{\alpha}\} \cap \{Q_{\text{lo}} \geq R^{\alpha}_{\text{lo}}\} \subseteq \{R^{\alpha}_{\text{lo}} \leq \widehat{R}^{\alpha}\}.$$

As such,

$$\mathbb{P}(\widehat{R}^{\alpha} < R^{\alpha}_{\text{lo}}) \leq \mathbb{P}(\{\widehat{Q} > \widehat{R}^{\alpha}\} \cup \{Q_{\text{lo}} < R^{\alpha}_{\text{lo}}\}) \leq \mathbb{P}(\widehat{Q} > \widehat{R}^{\alpha}) + \mathbb{P}(Q_{\text{lo}} < R^{\alpha}_{\text{lo}}) \leq \alpha/2 + \alpha/2 = \alpha.$$

□

C.8.2 Unbiased sample mean estimates of Q_{lo}, $\widehat{Q}$, and Q_{up}

We use our data to obtain one-sided confidence intervals R^{α}_{lo} and $\widehat{R}^{\alpha}$ satisfying (C.17) and (C.18) as required by our procedure for testing $\widehat{Q} \geq Q_{\text{lo}}$. We use an analogous procedure to obtain confidence intervals for testing $\widehat{Q} \leq Q_{\text{up}}$. We tried two techniques for this: an exact method based on Hoeffding's inequality, and an approximate method based on bootstrapping. Conceptually, both are based on obtaining unbiased sample mean estimates of Q_{lo} and $\widehat{Q}$, which we describe now, before giving the particulars of each method in the next two subsections.

We begin with our sample mean estimator of Q_{lo}. Recall that we assume access to a dataset $\mathcal{D}$ consisting of i.i.d. copies of observational trajectories of the form

$$X_0, A_1, X_1(A_1), \ldots, A_T, X_T(A_{1:T}).$$

Let $\mathcal{D}(a_{1:t}, B_{0:t})$ be the subset of trajectories in $\mathcal{D}$ for which $X_{0:N}(A_{1:N}) \in B_{0:N}$. Obtaining $\mathcal{D}(a_{1:t}, B_{0:t})$ is possible since the only random quantity that $N = \max\{0 \leq s \leq t \mid A_{1:s} = a_{1:s}\}$ depends on is $A_{1:t}$, which is included in the data. We denote the cardinality of $\mathcal{D}(a_{1:t}, B_{0:t})$ by $n \coloneqq |\mathcal{D}(a_{1:t}, B_{0:t})|$. We then denote by $Y_{\text{lo}}^{(i)}$ for $i \in \{1, \ldots, n\}$ the corresponding values of

$$Y_{\text{lo}} = \mathbb{1}(A_{1:t} = a_{1:t})\, f(X_{0:t}(A_{1:t})) + \mathbb{1}(A_{1:t} \neq a_{1:t})\, y_{\text{lo}}$$

obtained from each trajectory in $\mathcal{D}(a_{1:t}, B_{0:t})$. This is again possible since both terms only depends on the observational quantities $(X_{0:t}(A_{1:t}), A_{1:t})$. It is easily seen that the values of $Y_{\text{lo}}^{(i)}$ are i.i.d. and satisfy $\mathbb{E}[Y_{\text{lo}}^{(i)}] = Q_{\text{lo}}$. As a result, the sample mean

$$\mu_{\text{lo}} \coloneqq \frac{1}{n} \sum_{i=1}^{n} Y_{\text{lo}}^{(i)} \tag{C.19}$$

is an unbiased estimator of Q_{lo}.

We obtain an unbiased sample mean estimate of $\widehat{Q}$ in a similar fashion as for Q_{lo}. Recall that we assume access to a dataset $\widehat{\mathcal{D}}(a_{1:t})$ consisting of i.i.d. copies of

$$X_0, \widehat{X}_1(X_0, a_1), \ldots, \widehat{X}_t(X_0, a_{1:t}).$$

Let $\widehat{\mathcal{D}}(a_{1:t}, B_{0:t})$ denote the subset of twin trajectories in $\widehat{\mathcal{D}}(a_{1:t})$ for which $(X_0, \widehat{X}_t(X_0, a_{1:t})) \in B_{0:t}$, and denote its cardinality by $\widehat{n} \coloneqq |\widehat{\mathcal{D}}(a_{1:t}, B_{0:t})|$. Then denote by $\widehat{Y}^{(i)}$ for $i \in \{1 \ldots, \widehat{n}\}$ the corresponding values of

$$\widehat{Y} = f(X_0, \widehat{X}_{1:t}(X_0, a_{1:t}))$$

obtained from each trajectory in $\widehat{\mathcal{D}}(a_{1:t}, B_{0:t})$. It is easily seen that the values $\widehat{Y}^{(i)}$ are i.i.d. (since the entries of $\widehat{\mathcal{D}}(a_{1:t})$ are) and satisfy $\mathbb{E}[\widehat{Y}^{(i)}] = \widehat{Q}$. As a result, the sample mean

$$\widehat{\mu} \coloneqq \frac{1}{\widehat{n}} \sum_{i=1}^{\widehat{n}} \widehat{Y}^{(i)}$$

is an unbiased estimator of $\widehat{Q}$.

C.8.3 Exact confidence intervals via Hoeffding's inequality

Recall that we assume in Section 4.5.1 that $Y(a_{1:t})$ has the form $Y(a_{1:t}) = f(X_{0:t}(a_{1:t}))$, and that moreover

$$y_{\text{lo}} \leq f(x_{0:t}) \leq y_{\text{up}} \qquad \text{for all } x_{0:t} \in B_{0:t}. \tag{C.20}$$

This means $\widehat{Y}^{(i)}$ is almost surely bounded in $[y_{\text{lo}}, y_{\text{up}}]$, and so $\widehat{\mu}$ gives rise to one-sided confidence intervals via an application of Hoeffding's inequality. The exact form of these confidence intervals is as follows:

Proposition C.8.2

If (C.20) holds, then for each $\alpha \in (0,1)$, letting

$$\Delta := (y_{\text{up}} - y_{\text{lo}}) \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \qquad \text{and} \qquad \widehat{\Delta} := (y_{\text{up}} - y_{\text{lo}}) \sqrt{\frac{1}{2\widehat{n}} \log \frac{2}{\alpha}},$$

and similarly

$$R^{\alpha}_{\text{lo}} := \mu_{\text{lo}} - \Delta \qquad \text{and} \qquad \widehat{R}^{\alpha} := \widehat{\mu} + \widehat{\Delta},$$

it follows that

$$\mathbb{P}(Q_{\text{lo}} \geq R^{\alpha}_{\text{lo}}) \geq 1 - \frac{\alpha}{2} \qquad \text{and} \qquad \mathbb{P}(\widehat{Q} \leq \widehat{R}^{\alpha}) \geq 1 - \frac{\alpha}{2}.$$

Proof. We only prove the result for R^{α}_{lo}; the other statement can be proved analogously. Recall that μ_{lo} is the empirical mean of i.i.d. samples $Y^{(i)}_{\text{lo}}$ for $i \in \{1, \ldots, n\}$ with $\mathbb{E}[Y^{(i)}_{\text{lo}}] = Q_{\text{lo}}$ (see (C.19)). Moreover, by (C.20), $Y^{(i)}_{\text{lo}}$ is almost surely bounded in $[y_{\text{lo}}, y_{\text{up}}]$. Hoeffding's inequality then implies that

$$\mathbb{P}(\mu_{\text{lo}} - Q_{\text{lo}} > \Delta) \leq \exp\left(-\frac{2n\Delta^2}{(y_{\text{up}} - y_{\text{lo}})^2}\right).$$

In turn, some basic manipulations yield

$$\begin{aligned}
\mathbb{P}(Q_{\text{lo}} \geq R^{\alpha}_{\text{lo}}) &= 1 - \mathbb{P}(Q_{\text{lo}} < \mu_{\text{lo}} - \Delta) \\
&\geq 1 - \exp\left(-\frac{2n\Delta^2}{(y_{\text{up}} - y_{\text{lo}})^2}\right) \\
&= 1 - \frac{\alpha}{2}.
\end{aligned}$$

□

C.8.4 Approximate confidence intervals via bootstrapping

While Hoeffding's inequality yields the probability guarantees in (C.17) and (C.18) exactly, the confidence intervals obtained can be conservative. Consequently, our testing procedure may have lower probability of falsifying certain twins that in fact do not satisfy the causal bounds. To address this, we also consider an approximate approach based on bootstrapping that can produce tighter confidence intervals. While other schemes are possible, bootstrapping provides a general-purpose approach that is straightforward to implement and works well in practice.

At a high level, our approach here is again to construct one-sided level $1-\alpha/2$ confidence intervals via bootstrapping [Efron, 1979] on Q_{lo} and $\widehat{Q}$. Many bootstrapping procedures for obtaining confidence intervals have been proposed in the literature [Tibshirani and Efron, 1993, Davison and Hinkley, 1997, Hesterberg, 2015]. Our results reported below were obtained via the *reverse percentile* bootstrap (see Hesterberg [2015] for an overview). (We also tried the *percentile* bootstrap method, which obtained nearly indistinguishable results.) In particular, this method takes

$$R^{\alpha}_{\text{lo}} \coloneqq 2\mu_{\text{lo}} - \Delta \qquad \widehat{R}^{\alpha} \coloneqq 2\widehat{\mu} - \widehat{\Delta},$$

where Δ and $\widehat{\Delta}$ correspond to the approximate $1-\alpha/2$ and $\alpha/2$ quantiles of the distributions of

$$\frac{1}{n}\sum_{i=1}^{n} Y_{\text{lo}}^{(i^*)} \qquad \text{and} \qquad \frac{1}{\widehat{n}}\sum_{i=1}^{\widehat{n}} \widehat{Y}^{(i^*)},$$

where each $Y_{\text{lo}}^{(i^*)}$ and $Y^{(i^*)}$ is obtained by sampling uniformly with replacement from among the values of $Y_{\text{lo}}^{(i)}$ and $Y^{(i)}$. In our case study, as is typically done in practice, we approximated Δ and $\widehat{\Delta}$ via Monte Carlo sampling. It can be shown that the confidence intervals produced in this way obtain a coverage level that approaches the desired level of $1-\alpha/2$ as n and $\widehat{n}$ grow to infinity under mild assumptions [Hall, 1988].

C.9 Experimental Details

C.9.1 MIMIC preprocessing

For data extraction and preprocessing, we re-used the same procedure as Komorowski et al. [2018] with minor modifications. For completeness, we describe the pre-processing steps applied in Komorowski et al. [2018] and subsequently outline our modifications to these.

Following Komorowski et al. [2018], we extracted adult patients fulfilling the sepsis-3 criteria [Singer et al., 2016]. Sepsis was defined as a suspected infection (as indicated by prescription of antibiotics and sampling of bodily fluids for microbiological culture) combined with evidence of organ dysfunction, defined by a SOFA score ≥ 2 [Singer et al., 2016, Seymour et al., 2016].

Following Komorowski et al. [2018], we excluded patients for whom any of the following was true: their age was less than 18 years old at the time of ICU admission; their mortality not documented; their IV fluid/vasopressors intake was not documented; their treatment was withdrawn.

We made the following modifications to the preprocessing code of Komorowski et al. [2018] for our experiment. First, instead of extracting physiological quantities (e.g. heart rate) every 4 hours, we extracted these every hour. Additionally, we excluded patients with any missing hourly vitals during the first 4 hours of their ICU stay.

We then extracted a total of 19 quantities of interest listed in Table C.1. Of these, 17 were physiological quantities associated with the patient, including static demographic quantities (e.g. age), patient vital signs (e.g. heart rate), and patient lab values (e.g. potassium blood concentration). All of these were continuous values, apart from sex. These were chosen as the subset of physiological quantities extracted from MIMIC by Komorowski et al. [2018] that are also modelled by Pulse, and were used to define our observation spaces $\mathcal{X}_t$ as described next. The remaining 2 quantities (intravenous fluids and vasopressor doses) were chosen since they correspond to treatments that the patient received, and were used to define our action spaces $\mathcal{A}_t$ as described below.

C.9.2 Sample splitting

Before proceeding further, we randomly selected 5% of the extracted our trajectories (583 trajectories, denoted as $\mathcal{D}_0$) to use for preliminary tasks such as choosing the parameters of our hypotheses. We reserved the remaining 95% (11,094 trajectories, denoted as $\mathcal{D}$) for the actual testing. By a standard sample splitting argument [Cox, 1975], the statistical guarantees of our testing procedure established above continue to apply even when our hypotheses are defined in this data-dependent way.

Category	Physiological quantity
Demographic	Age
	Sex
	Weight
Vital Signs	Heart rate (HR)
	Systolic blood pressure (SysBP)
	Diastolic blood pressure (DiaBP)
	Mean blood pressure (MeanBP)
	Respiratory Rate (RR)
	Skin Temperature (Temp)
Lab Values	Potassium Blood Concentration (Potassium)
	Sodium Blood Concentration (Sodium)
	Chloride Blood Concentration (Chloride)
	Glucose Blood Concentration (Glucose)
	Calcium Blood Concentration (Calcium)
	Bicarbonate Blood Concentration (HCO_3)
	Arterial O_2 Pressure (PaO_2)
	Arterial CO_2 Pressure ($PaCO_2$)
Treatments	Intravenous fluid (IV) dose
	Vasopressor dose

Table C.1: Physiological quantities and treatments extracted from MIMIC

C.9.3 Observation spaces

Our $\mathcal{X}_0$ consisted of the following features: age, sex, weight, heart rate, systolic blood pressure, diastolic blood pressure and respiration rate. We chose $\mathcal{X}_0$ in this way because, out of the 17 physiological quantities we extracted from MIMIC, these were the quantities that can be initialised to user-provided values before starting a simulation in the version of Pulse we considered (4.x). (In contrast, Pulse initialises the other 10 features to default values.) For the remaining observation spaces, we used the full collection of the 17 physiological quantities we extracted to define $\mathcal{X}_1 = \cdots = \mathcal{X}_4$. We encoded all features in $\mathcal{X}_t$ numerically, i.e. $\mathcal{X}_0 = \mathbb{R}^7$, and $\mathcal{X}_t = \mathbb{R}^{17}$ for $t \in \{1, 2, 3, 4\}$.

C.9.4 Action spaces

Following Komorowski et al. [2018], we constructed our action space using 2 features obtained from MIMIC, namely intravenous fluid (IV) and vasopressor doses. To obtain discrete action spaces suitable for our framework, we used the same discretization procedure for these quantities as was used by Komorowski et al. [2018]. Specifically, we divided the hourly doses of intravenous fluids and vasopressors into 5 bins each, with the first bin corresponding to zero drug dosage, and the remaining 4 bins based on the quartiles

		Vasopressor dose (μg/kg/min)				
		0	0.0 - 0.061	0.061 - 0.15	0.15 - 0.313	> 0.313
IV dose (mL/h)	0	16659	329	256	152	145
	0 - 20	5840	428	351	244	145
	20 - 75	6330	297	378	383	309
	75 - 214	6232	176	175	197	273
	> 214	5283	347	488	544	747

Table C.2: Action space with frequency of occurrence in observational data

of the non-zero drug dosages in our held-out observational dataset $\mathcal{D}_0$. From this we obtained action spaces $\mathcal{A}_1 = \cdots = \mathcal{A}_4$ with $5 \times 5 = 25$ elements. Table C.2 shows the dosage bins constructed in this way, as well as the frequency of each bin's occurrence in the observational data.

C.9.5 Hypothesis parameters

We used our held-out observational dataset $\mathcal{D}_0$ to obtain a collection of hypothesis parameters $(t, f, a_{1:t}, B_{0:t})$. Specifically, for each physiological quantity of interest (e.g. heart rate) in the list of 'Vital Signs' and 'Lab Values' given in Table C.1, we did the following. First, for each $t \in \{0, \ldots, 4\}$, we obtained 16 choices of B_t by discretizing the patient space $\mathcal{X}_t$ into 16 subsets based on the values of certain features as follows: 2 bins corresponding to sex; 4 bins corresponding to the quartiles of the ages of patients in $\mathcal{D}_0$; 2 bins corresponding to whether or not the value of the chosen physiological quantity of interest at time t was above or below its median value in $\mathcal{D}_0$.

Next, for each $t \in \{1, \ldots, 4\}$, $a_{1:t} \in \mathcal{A}_{1:t}$, and sequence $B_{0:t}$ with each $B_{t'}$ as defined in the previous step, let $\mathcal{D}_0(t, a_{1:t}, B_{0:t})$ denote the subset of $\mathcal{D}_0$ corresponding to $(t, a_{1:t}, B_{0:t})$, i.e.

$$\mathcal{D}_0(t, a_{1:t}, B_{0:t}) \coloneqq \{X_{0:t}(A_{1:t}) \mid (X_{0:T}(A_{1:T}), A_{1:T}) \in \mathcal{D}_0 \text{ with } A_{1:t} = a_{1:t} \text{ and } X_{0:t}(A_{1:T}) \in B_{0:t}\}.$$

We then selected the set of all triples $(t, a_{1:t}, B_{0:t})$ such that $\mathcal{D}_0(t, a_{1:t}, B_{0:t})$ contained at least one trajectory. This meant the number of combinations of hypotheses parameters that we considered was limited to a tractable quantity, which had benefits both computationally, and also by ensuring that we did not sacrifice too much power when adjusting for multiple testing.

Finally, for each selected triple $(t, a_{1:t}, B_{0:t})$, we chose a corresponding f as follows. First, we let $i \in \{1, \ldots, d_t\}$ denote the index of the physiological quantity of interest in

$\mathcal{X}_t = \mathbb{R}^{d_t}$. We then set $y_{\text{lo}}, y_{\text{up}}$ to be the .2 and the .8 quantiles of the values in

$$\{(X_t(A_{1:t}))_i \mid X_{0:t}(A_{1:t}) \in \mathcal{D}_0(t, a_{1:t}, B_{0:t})\}$$

We then obtained $f : \mathcal{X}_{0:t} \to \mathbb{R}$ as the function that extracts the physiological quantity of interest from $\mathcal{X}_t$ and clips its value to between y_{lo} and y_{up}, i.e. $f(x_{0:t}) \coloneqq \min(\max(x_t)_i, y_{\text{lo}}), y_{\text{up}})$.

Overall, accounting for all physiological quantities of interest, we obtained 721 distinct choices of $(t, f, a_{1:t}, B_{0:t})$ in this way. Figure C.2 shows the amount of non-held out observational and twin data that we subsequently used for testing each hypothesis, i.e. the values of n and $\widehat{n}$ as defined in Section C.8.2 above. (We describe how we generated our dataset of twin trajectories in Section C.9.6.)

C.9.6 Generating twin trajectories using the Pulse Physiology Engine

The Pulse Physiology Engine is an open source comprehensive human physiology simulator that has been used in medical education, research, and training. The core engine of Pulse is C++ based with APIs available in different languages, including python. Detailed documentation is available at: pulse.kitware.com. Pulse allows users to initialize patient trajectories with given age, sex, weight, heart rate, systolic blood pressure, diastolic blood pressure and respiration rate and medical conditions such as sepsis, COPD, ARDS, etc. Once initialised, users have the ability to advance patient trajectories by a given time step (one hour in our case), and administer actions (e.g. administer a given dose of IV fluids or vasopressors).

In Algorithm 3 we describe how we generated the twin data to test the chosen hypotheses. Note that we sampled X_0 without replacement as it ensures that each X_0 is chosen at most once and consequently twin trajectories in $\widehat{\mathcal{D}}(a_{1:t})$ are i.i.d. Additionally, Algorithm 3 can be easily parallelised to improve efficiency. Figure C.2 shows histograms of the number of twin trajectories $\widehat{n}$ (as defined in Section C.8.2 above) obtained in this way across all hypotheses.

C.9.7 Bootstrapping details

In addition to Hoeffding's inequality, we also used reverse percentile bootstrap method (see e.g. Hesterberg [2015]) to obtain our confidence intervals on Q_{lo} and Q_{up} as described in Section C.8.2. We used 100 bootstrap samples for each confidence interval. To avoid

Algorithm 3: Generating Twin data $\widehat{\mathcal{D}}(a_{1:t})$.

Inputs: Action sequence $a_{1:t}$; Observational dataset $\mathcal{D}$.
Output: Twin data $\widehat{\mathcal{D}}(a_{1:t})$ of size m.
for $i = 1, \ldots, m$ **do**
  Sample X_0 without replacement from $\mathcal{D}$;
  $\widehat{X_0} \leftarrow X_0$ i.e., initialize the Pulse trajectory with the information of X_0;
  for $t' = 1, \ldots, t$ **do**
    Administer the median doses of IV fluids and vasopressors in action bin $a_{t'}$;
    if $t' \equiv 0 \pmod 3$ **then**
      Virtual patient in Pulse consumes nutrients and water, and urinates;
    end
    Advance the twin trajectory by one hour;
  end
  Add the trajectory $\widehat{X}_{0:t}(a_{1:t})$ to $\widehat{\mathcal{D}}(a_{1:t})$;
end
Return $\widehat{\mathcal{D}}(a_{1:t})$

	Ours		Manski	
Physiological quantity	Rejs.	Hyps.	Rejs.	Hyps.
Chloride Blood Concentration (Chloride)	24	94	1	46
Sodium Blood Concentration (Sodium)	21	94	9	46
Potassium Blood Concentration (Potassium)	13	94	0	46
Skin Temperature (Temp)	10	86	9	46
Calcium Blood Concentration (Calcium)	5	88	0	46
Glucose Blood Concentration (Glucose)	5	96	1	46
Arterial CO_2 Pressure ($paCO_2$)	3	70	0	46
Bicarbonate Blood Concentration (HCO_3)	2	90	1	46
Systolic Arterial Pressure (SysBP)	2	154	0	46
Arterial O_2 Pressure (paO_2)	0	78	1	46
Arterial pH (Arterial_pH)	0	80	0	46
Diastolic Arterial Pressure (DiaBP)	0	72	0	46
Mean Arterial Pressure (MeanBP)	0	92	0	46
Respiration Rate (RR)	0	172	0	46
Heart Rate (HR)	0	162	0	46

Table C.3: Total hypotheses (Hyps.) and rejections (Rejs.) per physiological quantity obtained using Hoeffding's inequality

bootstrapping on small numbers of data points, we did not reject any hypothesis where either the number of observational trajectories n or twin trajectories $\widehat{n}$ was less than 100, and returned a p-value of 1 in each such case.

Table C.4 shows the number of rejected hypotheses for each physiological quantity using this approach. We observed a similar trend as in our results obtained using Hoeffding's inequality (Table C.3). For example, we obtained high number of rejections for Sodium, Chloride and Potassium blood concentrations but few rejections for Arterial Pressure and

Physiological quantity	Ours		Manski	
	Rejs.	Hyps.	Rejs.	Hyps.
Chloride Blood Concentration (Chloride)	47	94	1	46
Sodium Blood Concentration (Sodium)	46	94	12	46
Potassium Blood Concentration (Potassium)	33	94	0	46
Skin Temperature (Temp)	43	86	13	46
Calcium Blood Concentration (Calcium)	44	88	0	46
Glucose Blood Concentration (Glucose)	19	96	0	46
Arterial CO_2 Pressure ($paCO_2$)	13	70	0	46
Bicarbonate Blood Concentration (HCO_3)	8	90	0	46
Systolic Arterial Pressure (SysBP)	8	154	0	46
Arterial O_2 Pressure (paO_2)	4	78	1	46
Arterial pH (Arterial_pH)	0	80	0	46
Diastolic Arterial Pressure (DiaBP)	0	72	0	46
Mean Arterial Pressure (MeanBP)	3	92	0	46
Respiration Rate (RR)	12	172	0	46
Heart Rate (HR)	1	162	0	46

Table C.4: Total hypotheses (Hyps.) and rejections (Rejs.) per physiological quantity obtained using the reverse percentile bootstrap

Heart Rate. Overall, bootstrapping increased the number of rejected hypotheses by a factor of roughly 3.3 compared with Hoeffding's inequality (281 vs. 85 rejections in total). Like we described for Hoeffding's inequality in the main text, we also ran this analysis with each hypothesis obtained using the unconditional bounds of Manski [1990], and again obtained substantially fewer rejections compared with our approach based on Theorem 4.4.1.

C.9.8 Tightness of bounds and number of data points per hypothesis

In this section, we show empirically how both the tightness of the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ and the number of data points per hypothesis relate to the number of falsifications obtained in our case study. Recall that the tightness of $[Q_{\text{lo}}, Q_{\text{up}}]$ is determined by the value of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, since we have

$$\frac{Q_{\text{up}} - Q_{\text{lo}}}{y_{\text{up}} - y_{\text{lo}}} = 1 - \mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}). \tag{C.21}$$

Here the left-hand side is a number in $[0, 1]$ that quantifies the tightness of the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ relative to the trivial worst-case bounds $[y_{\text{lo}}, y_{\text{up}}]$, with smaller values meaning tighter bounds. The equation above shows that the higher the value of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$, the tighter the bounds are.

Figure C.5 shows the bounds are often informative in practice, with $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ being reasonably large (and hence the bounds tight, by (C.21) above)

for a significant number of hypotheses we consider. However, rejections still occur even when the bounds are reasonably loose (e.g. $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N}) \approx 0.3$), which shows our method can still yield useful information even in this case. We moreover observe rejections across a range of different numbers of observational data points used to test each hypothesis, which shows that our method is not strongly dependent on the size of the dataset obtained.

C.9.9 Sensitivity to y_{lo} and y_{up}

We investigated the sensitivity of our methodology with respect to our choices of the values y_{lo} and y_{up}. Specifically, we repeated our procedure with the intervals $[y_{\text{lo}}, y_{\text{up}}]$ replaced with $[y_{\text{lo}}\,(1 - \Delta/2), y_{\text{up}}\,(1 + \Delta/2)]$ for a range of different values of $\Delta \in \mathbb{R}$. Figure C.6 plots the number of rejections for different values of Δ. We observe that for significantly larger $[y_{\text{lo}}, y_{\text{up}}]$ intervals, we do obtain fewer rejections, although this is to be expected since the widths of our both the bounds $[Q_{\text{lo}}, Q_{\text{up}}]$ and our confidence intervals R^{α}_{lo} and R^{α}_{up} obtained using Hoeffding's inequality (see Proposition C.8.2) grow increasingly large as the width of $[y_{\text{lo}}, y_{\text{up}}]$ grows. However, we observe that the number of rejections per outcome is stable for a moderate range of widths of $[y_{\text{lo}}, y_{\text{up}}]$, which indicates that our method is reasonably robust to the choice of $y_{\text{lo}}, y_{\text{up}}$ parameters.

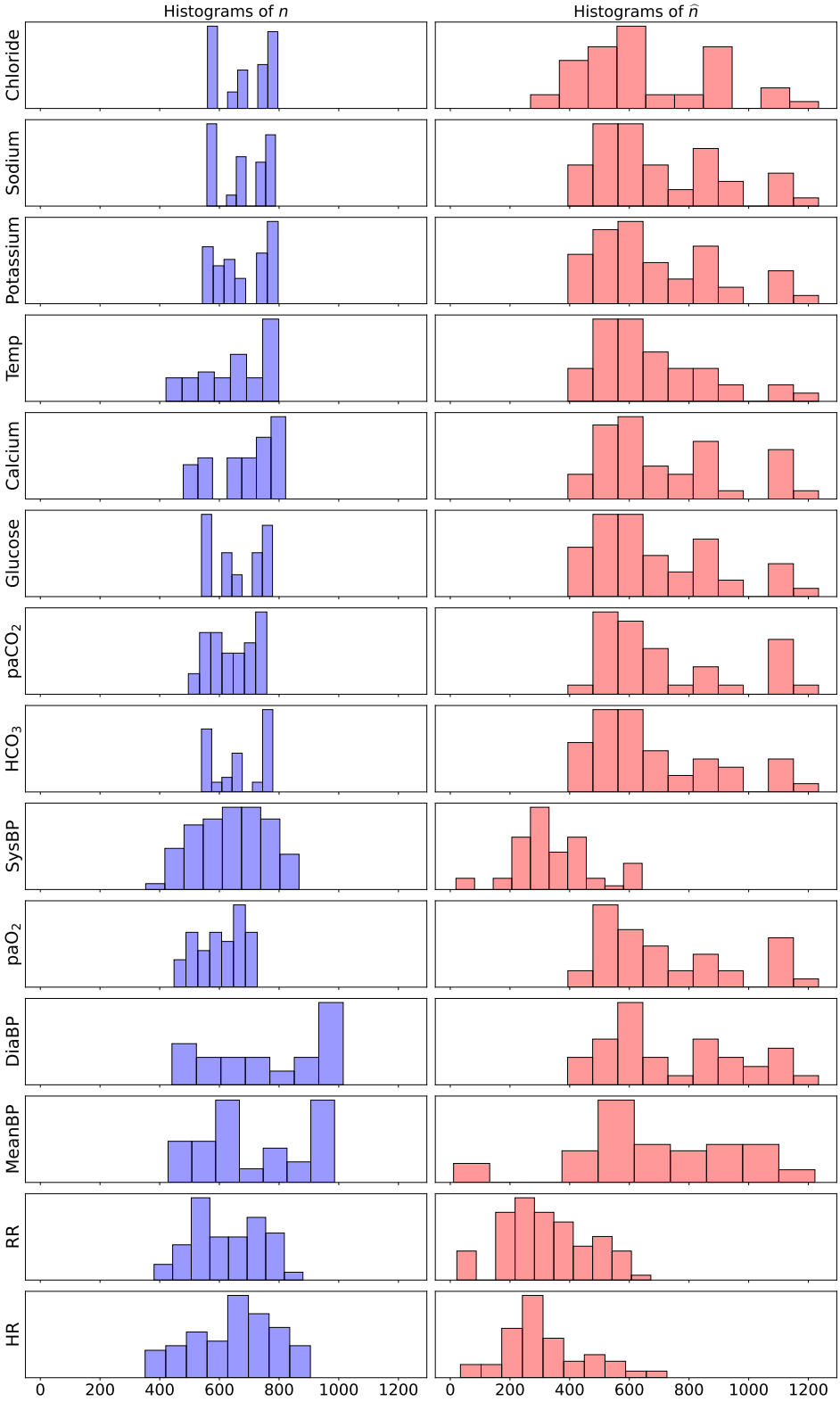


Figure C.2: Histograms of n and $\widehat{n}$ (as defined in Section C.8.2) across all hypothesis parameters corresponding to each physiological quantity of interest.

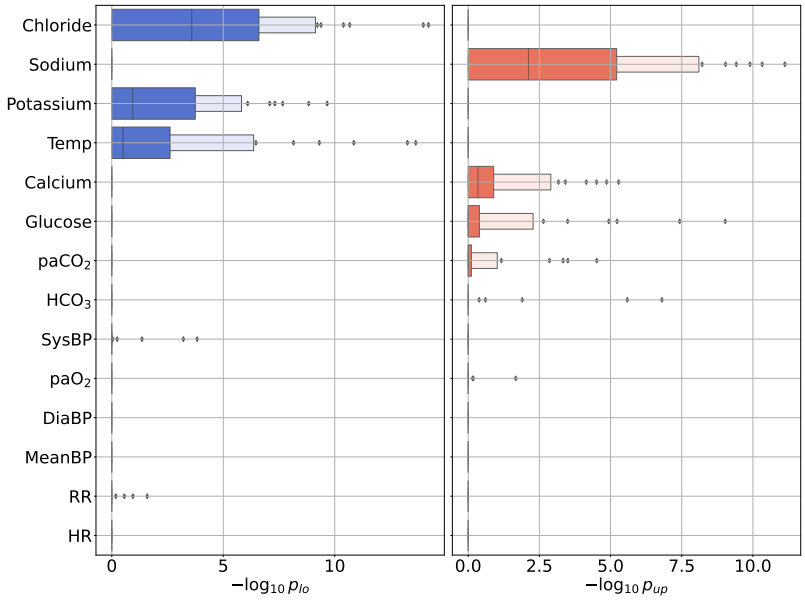
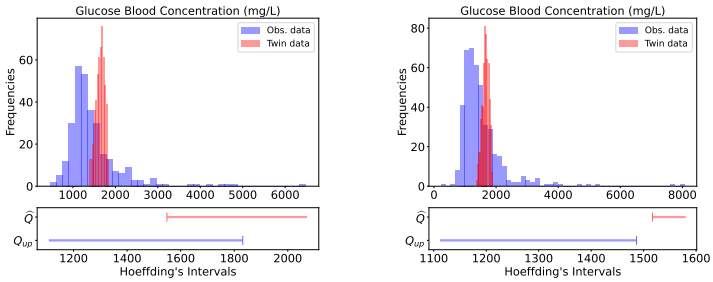
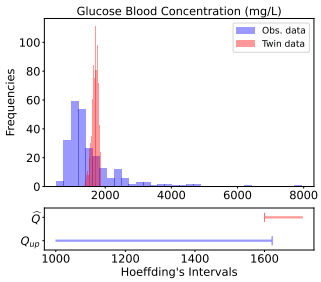


Figure C.3: Boxenplots showing distributions of $-\log_{10} p_{\text{lo}}$ and $-\log_{10} p_{\text{up}}$ for different physiological quantities obtained via Hoeffding's inequality. Higher values indicate greater evidence in favour of rejection.

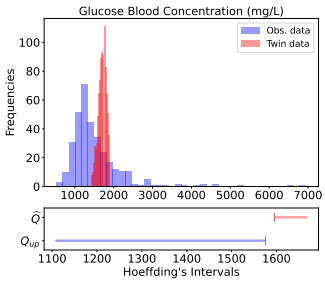


(a) Not rejected

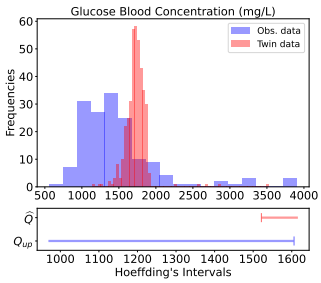
(b) Rejected



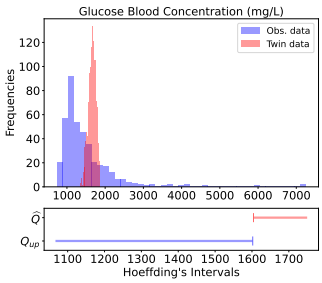
(c) Not rejected



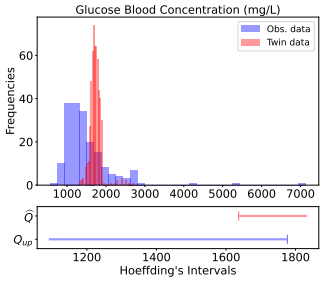
(d) Rejected



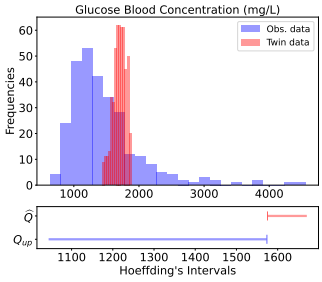
(e) Not rejected



(f) Rejected



(g) Not rejected



(h) Rejected

Figure C.4: Raw observational data values conditional on $A_{1:t} = a_{1:t}$ and $X_{0:t}(A_{1:t}) \in B_{0:t}$, and from the output of the twin conditional on $\widehat{X}_{0:t}(a_{1:t}) \in B_{0:t}$. Each row shows two distinct choices of $(B_{0:t}, a_{1:t})$. Below each figure are shown 95% Hoeffding confidence intervals for $\widehat{Q}$ and Q_{up}. Unlike Figure 4.3 from the main text, the horizontal axes of the histograms are not truncated, and the first row is in particular an untruncated version of Figure 4.3 from the main text. Note however that the scales of the horizontal axes of the confidence intervals differ from those of the histograms, since it is visually more difficult to determine whether or not the confidence intervals overlap when fully zoomed out.

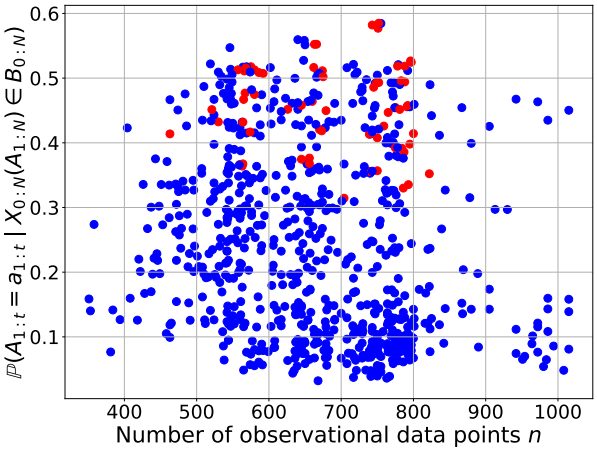


Figure C.5: Sample mean estimate of $\mathbb{P}(A_{1:t} = a_{1:t} \mid X_{0:N}(A_{1:N}) \in B_{0:N})$ for each pair of hypotheses $(\mathcal{H}_{\text{lo}}, \mathcal{H}_{\text{up}})$ corresponding to the same set of parameters $(t, f, a_{1:t}, B_{0:t})$ that we tested, along with the corresponding number of observational data points used to test each hypothesis. Red points indicate that either $\mathcal{H}_{\text{lo}}$ or $\mathcal{H}_{\text{up}}$ were rejected, while blue points indicate that both $\mathcal{H}_{\text{lo}}$ and $\mathcal{H}_{\text{up}}$ were not rejected.

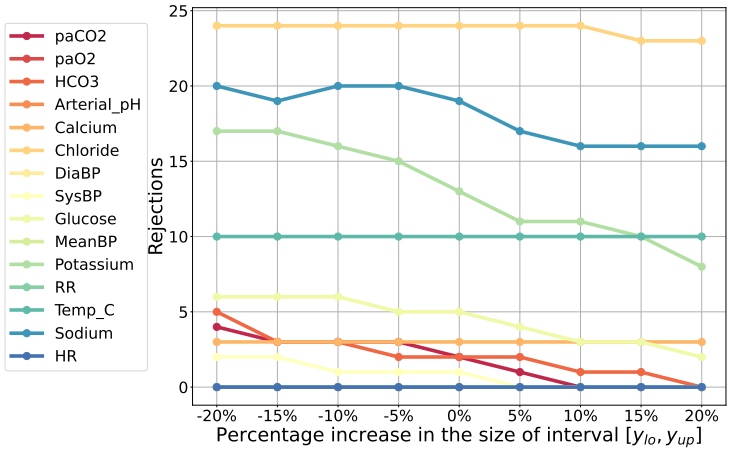


Figure C.6: Rejections obtained as the width of the $[y_{\text{lo}}, y_{\text{up}}]$ interval changes. Here, the interval is increased (or decreased) symmetrically on each side.

Bibliography

Jean Pierre Allamaa, Panagiotis Patrinos, Herman Van der Auweraer, and Tong Duy Son. Sim2real for autonomous vehicle control using executable digital twin. *IFAC-PapersOnLine*, 55(24):385–391, 2022. ISSN 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2022.10.314. URL `https://www.sciencedirect.com/science/article/pii/S2405896322023461`. 10th IFAC Symposium on Advances in Automotive Control AAC 2022.

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

AMSE. *Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices*. AMSE, 2018.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Barbara Rita Barricelli, Elena Casiraghi, and Daniela Fogli. A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE access*, 7:167653–167671, 2019.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68, 11 2019. doi: 10.1287/opre.2019.1902.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458, 2017.

Nicholas Bellinger, Eric J. Tuegel, Anthony R. Ingraffea, Thomas G. Eason, and S. Michael Spottswood. Reengineering aircraft structural life prediction using a digital twin. *International Journal of Aerospace Engineering*, 2011:154798, 2011. doi: 10.1155/2011/154798. URL `https://doi.org/10.1155/2011/154798`.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188, 2001. doi: 10.1214/aos/1013699998. URL `https://doi.org/10.1214/aos/1013699998`.

Alina Beygelzimer and John Langford. The offset tree for learning with partial labels. *CoRR*, abs/0812.4044, 2008. URL `http://arxiv.org/abs/0812.4044`.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Aaron Bray, Jeffrey B. Webb, Andinet Enquobahrie, Jared Vicory, Jerry Heneghan, Robert Hubal, Stephanie TerMaath, Philip Asare, and Rachel B. Clipp. Pulse Physiology Engine: an Open-Source Software Platform for Computational Modeling of Human Medical Simulation. *SN Comprehensive Clinical Medicine*, 1(5):362–377, 2019. doi: 10.1007/s42399-019-00053-w. URL `https://doi.org/10.1007/s42399-019-00053-w`.

Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324. URL `https://doi.org/10.1023/A:1010933404324`.

Yash Chandak, Scott Niekum, Bruno Castro da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *arXiv preprint arXiv:2104.12820*, 2021.

Genevieve Coorey, Gemma A Figtree, David F Fletcher, Victoria J Snelson, Stephen Thomas Vernon, David Winlaw, Stuart M Grieve, Alistair McEwan, Jean Yee Hwa Yang, Pierre Qian, et al. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *NPJ Digital Medicine*, 2022.

Rob Cornish, Muhammad Faaiz Taufiq, Arnaud Doucet, and Chris Holmes. Causal falsification of digital twins, 2023. URL `https://arxiv.org/abs/2301.07210`.

Jorge Corral-Acero, Francesca Margara, Maciej Marciniak, Cristobal Rodero, Filip Loncaric, Yingjing Feng, Andrew Gilbert, Joao F Fernandes, Hassaan A Bukhari, Ali Wajdan, et al. The 'digital twin' to enable the vision of precision cardiology. *European Heart Journal*, 41(48): 4556–4564, 2020.

David R Cox. A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2): 441–444, 1975.

Ulrich Richard Dahmen, Tobias Osterloh, and Heinz-Jürgen Roßmann. Verification and validation of digital twins and virtual testbeds. *International journal of advances in engineering sciences and applied mathematics*, 11(1):47–64, 2022. ISSN 0975-5616. doi: 10.11591/ijaas.v11.i1.pp47-64. URL `https://publications.rwth-aachen.de/record/843535`.

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. doi: 10.1017/CBO9780511802843.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4), 2014a.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014b. ISSN 08834237, 21688745. URL `http://www.jstor.org/stable/43288496`.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552. URL `https://doi.org/10.1214/aos/1176344552`.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456. PMLR, 10–15 Jul 2018a. URL `https://proceedings.mlr.press/v80/farajtabar18a.html`.

Mehrdad Farajtabar, Mohammad Ghavamzadeh, and Yinlam Chow. More robust doubly robust off-policy evaluation. 2018b.

Daniele Foffano, Alessio Russo, and Alexandre Proutiere. Conformal off-policy evaluation in markov decision processes. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 3087–3094. IEEE, 2023.

Rina Foygel Barber, Emmanuel J Candès, Aaditya Ramdas, and Ryan J Tibshirani. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Scott Fujimoto, David Meger, and Doina Precup. A deep reinforcement learning approach to marginalized importance sampling with the successor representation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3518–3529. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fujimoto21a.html.

Suran Galappaththige, Richard A Gray, Caroline Mendonca Costa, Steven Niederer, and Pras Pathmanathan. Credibility assessment of patient-specific computational modeling using patient-specific cardiac modeling as an exemplar. *PLoS computational biology*, 18(10): e1010541, 2022.

Michael Grieves and John Vickers. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In *Transdisciplinary Perspectives on Complex Systems*, pages 85–113. Springer, 2017.

Peter Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 16(3):927 – 953, 1988. doi: 10.1214/aos/1176350933. URL https://doi.org/10.1214/aos/1176350933.

André Hemmler, Brigitta Lutz, Günay Kalender, Christian Reeps, and Michael W Gee. Patient-specific in silico endovascular repair of abdominal aortic aneurysms: application and validation. *Biomechanics and Modeling in Mechanobiology*, 18(4):983–1004, 2019.

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

Tim C. Hesterberg. What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician*, 69(4):371–386, 2015. doi: 10.1080/00031305.2015.1089789. URL https://doi.org/10.1080/00031305.2015.1089789. PMID: 27019512.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL http://www.jstor.org/stable/2289064.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

Audrey Huang, Liu Leqi, Zachary C. Lipton, and Kamyar Azizzadenesheli. Off-policy risk assessment in contextual bandits. *arXiv preprint arXiv:2104.08977*, 2021.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, pages 601–608, 2007.

Guido W Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.

Guido W Imbens and Charles F Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Melanie Jans-Singh, Kathryn Leeming, Ruchi Choudhary, and Mark Girolami. Digital twin of an urban-integrated hydroponic farm. *Data-Centric Engineering*, 1:e20, 2020. doi: 10.1017/dce.2020.21.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/jiang16.html`.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *arXiv preprint arXiv:2111.12161*, 2021.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35. URL `https://doi.org/10.1038/sdata.2016.35`.

David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, 2020.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res.*, 21(1), jun 2022. ISSN 1532-4435.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/3a09a524440d44d7f19870070a5ad42f-Paper.pdf`.

Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67, 10 2020. doi: 10.1287/mnsc.2020.3699.

Nathan Kallus, Yuta Saito, and Masatoshi Uehara. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*, pages 5247–5256. PMLR, 2021.

Michael G Kapteyn, Jacob VR Pretorius, and Karen E Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1(5): 337–347, 2021.

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

Adnan Khan, Martin Dahl, Petter Falkman, and Martin Fabian. Digital twin for legacy systems: Simulation model testing and validation. In *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2018.

Brendan Kochunas and Xun Huan. Digital twin concepts with uncertainty for nuclear power applications. *Energies*, 14(14):4235, 2021.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5. URL `https://doi.org/10.1038/s41591-018-0213-5`.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Tom Kuipers, Renukanandan Tumu, Shuo Yang, Milad Kazemi, Rahul Mangharam, and Nicola Paoletti. Conformal off-policy prediction for multi-agent systems. *arXiv preprint arXiv:2403.16871*, 2024.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 03 2014. ISSN 0006-3444. doi: 10.1093/biomet/ast066. URL `https://doi.org/10.1093/biomet/ast066`.

Ilja Kuzborskij, Claire Vernade, András György, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648, 2021.

Jianfa Lai, Manyun Xu, Rui Chen, and Qian Lin. Generalization ability of wide neural networks on $\mathbb{R}$, 2023. URL `https://arxiv.org/abs/2302.05933`.

Amos Lal, Guangxi Li, Edin Cubro, Sarah Chalmers, Heyi Li, Vitaly Herasevich, Yue Dong, Brian Pickering, Kilickaya Oguz, and Ognjen Gajic. Development and verification of a digital twin patient model to predict treatment response in sepsis. *Critical Care Medicine*, 49: 611–611, 01 2021. doi: 10.1097/01.ccm.0000730744.82258.38.

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*, 2022. https://huggingface.co/blog/rlhf.

Ignacio Larrabide, Minsuok Kim, Luca Augsburger, Maria Cruz Villa-Uriol, Daniel Rüfenacht, and Alejandro F Frangi. Fast virtual deployment of self-expandable stents: method and in vitro evaluation for intracranial aneurysmal stenting. *Medical Image Analysis*, 16(3):721—730, April 2012. ISSN 1361-8415. doi: 10.1016/j.media.2010.04.009. URL `https://doi.org/10.1016/j.media.2010.04.009`.

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

Philip W Lavori and Ree Dawson. Dynamic treatment regimes: practical design considerations. *Clinical trials*, 1(1):9–20, 2004.

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

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series* B, pages 911–938, 2021.

Fan Li, Laine E Thomas, and Fan Li. Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*, 188(1):250–257, 09 2018. ISSN 0002-9262. doi: 10.1093/aje/kwy201. URL https://doi.org/10.1093/aje/kwy201.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 661–670, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605587998. doi: 10.1145/1772690.1772758. URL https://doi.org/10.1145/1772690.1772758.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over hilbert spaces. *Applied and Computational Harmonic Analysis*, 48(3):868–890, 2020. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2018.09.009. URL https://www.sciencedirect.com/science/article/pii/S1063520318300174.

Anqi Liu, Hao Liu, Anima Anandkumar, and Yisong Yue. Triply robust off-policy evaluation, 2019. URL https://arxiv.org/abs/1911.05811.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/dda04f9d634145a9c68d5dfe53b21272-Paper.pdf.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024. URL https://arxiv.org/abs/2308.05374.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Ben London and Ted Sandler. Bayesian counterfactual risk minimization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4125–4133. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/london19a.html.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Nan Lu, Tianyi Zhang, Tongtong Fang, Takeshi Teshima, and Masashi Sugiyama. Rethinking importance weighting for transfer learning. *CoRR*, abs/2112.10157, 2021. URL https://arxiv.org/abs/2112.10157.

Yuqian Lu, Chao Liu, I Kevin, Kai Wang, Huiyue Huang, and Xun Xu. Digital twin-driven smart manufacturing: Connotation, reference model, applications and research issues. *Robotics and Computer-Integrated Manufacturing*, 61:101837, 2020.

Charles F. Manski. Anatomy of the selection problem. *The Journal of Human Resources*, 24(3): 343–360, 1989. ISSN 0022166X. URL http://www.jstor.org/stable/145818.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990. ISSN 00028282. URL http://www.jstor.org/stable/2006592.

Charles F Manski. *Identification Problems in the Social Sciences*. Harvard University Press, 1995.

Charles F Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Joseph Masison, Jonathan Beezley, Yu Mei, Henrique Assis Lopes Ribeiro, Adam C Knapp, L Sordo Vieira, Bandita Adhikari, Yogesh Scindia, Michael Grauer, Brian Helba, et al. A modular computational framework for medical digital twins. *Proceedings of the National Academy of Sciences*, 118(20):e2024287118, 2021.

Matthew McDaniel and Austin Baird. A Full-Body Model of Burn Pathophysiology and Treatment Using the BioGears Engine. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 261–264, 2019. doi: 10.1109/EMBC.2019.8857686.

Matthew McDaniel, Jonathan M. Keller, Steven White, and Austin Baird. A Whole-Body Mathematical Model of Sepsis Progression and Treatment Designed in the BioGears Physiology Engine. *Frontiers in Physiology*, 10:1321, 2019. ISSN 1664-042X. doi: 10.3389/fphys.2019.01321. URL https://www.frontiersin.org/article/10.3389/fphys.2019.01321.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8119–8132. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4476b929e30dd0c4e8bdbcc82c6ba23a-Paper.pdf.

S. A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481, 2005. doi: https://doi.org/10.1002/sim.2022. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2022.

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

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18819–18831. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/da21bae82c02d1e2b8168d57cd3fbab7-Paper.pdf.

Whitney K. Newey and James R. Robins. Cross-fitting and fast remainder rates for semiparametric estimation, 2018. URL `https://arxiv.org/abs/1801.09138`.

Steven A Niederer, Michael S Sacks, Mark Girolami, and Karen Willcox. Scaling digital twins from the artisanal to the industrial. *Nature Computational Science*, 1(5):313–320, 2021.

Muhammad Osama, Dave Zachariah, and Peter Stoica. Learning robust decision policies from observational data. *arXiv preprint arXiv:2006.02355*, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Karl Popper. *The Logic of Scientific Discovery*. Routledge, 2005.

Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *arXiv preprint arXiv:1306.2597*, 2013.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9):1393–1512, 1986. ISSN 0270-0255. doi: https://doi.org/10.1016/0270-0255(86)90088-6. URL `https://www.sciencedirect.com/science/article/pii/0270025586900886`.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3543–3553, 2019.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J. Candès. With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Science Review*, 2 (2), 4 2020.

Paul R. Rosenbaum. *Observational Studies*. Springer, New York, NY, 2002.

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. ISSN 00063444. URL `http://www.jstor.org/stable/2335942`.

Mark Rowland, Anna Harutyunyan, Hado Hasselt, Diana Borsa, Tom Schaul, Rémi Munos, and Will Dabney. Conditional importance sampling for off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 45–55. PMLR, 2020.

Christopher J. Roy and William L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2011.03.016. URL `https://www.sciencedirect.com/science/article/pii/S0045782511001290`.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974. doi: https://doi.org/10.1037/h0037350.

Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331, 2005. doi: 10.1198/016214504000001880. URL `https://doi.org/10.1198/016214504000001880`.

Noveen Sachdeva, Yi Su, and Thorsten Joachims. Off-policy bandits with deficient support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 965–975, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403139. URL https://doi.org/10.1145/3394486.3403139.

Rafael Sacks, Ioannis Brilakis, Ergo Pikas, Haiyan Xie, and Mark Girolami. Construction with digital twin information systems. *Data-Centric Engineering*, 1, 2020.

Yuta Saito and Thorsten Joachims. Off-policy evaluation for large action spaces via embeddings. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19089–19122. PMLR, 2022.

Yuta Saito, Aihara Shunsuke, Matsutani Megumi, and Narita Yusuke. Open bandit dataset and pipeline: Towards realistic and reproducible off-policy evaluation. *arXiv preprint arXiv:2008.07146*, 2020.

Yuta Saito, Takuma Udagawa, Haruka Kiyohara, Kazuki Mogi, Yusuke Narita, and Kei Tateno. Evaluating the robustness of off-policy evaluation. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 114–123, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384582. doi: 10.1145/3460231.3474245. URL https://doi.org/10.1145/3460231.3474245.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, Clifford S. Deutschman, Gabriel J. Escobar, and Derek C. Angus. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):762–774, 02 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0288. URL https://doi.org/10.1001/jama.2016.0288.

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

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4. URL https://www.sciencedirect.com/science/article/pii/S0378375800001154.

Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, 315(8):801–810, 02 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.0287. URL https://doi.org/10.1001/jama.2016.0287.

Arjun Sondhi, David Arbour, and Drew Dimmery. Balanced off-policy evaluation in general action spaces. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2413–2423. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/sondhi20a.html.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. *International Conference on Representation Learning*, 2022.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. *CoRR*, abs/1907.09623, 2019a. URL `http://arxiv.org/abs/1907.09623`.

Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. CAB: Continuous adaptive blending for policy evaluation and learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6005–6014. PMLR, 09–15 Jun 2019b. URL `https://proceedings.mlr.press/v97/su19a.html`.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 2012. ISBN 9780262017091. URL `http://www.jstor.org/stable/j.ctt5hhbtm`.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5):985–1005, 2007.

Masashi Sugiyama, Shinichi Nakajima, Hisashi Kashima, Paul von Buenau, and Motoaki Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems 20*, pages 1433–1440, 2008.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 814–823. JMLR.org, 2015a.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b. URL `https://proceedings.neurips.cc/paper_files/paper/2015/file/39027dfad5138c9ca0c474d71db915c3-Paper.pdf`.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52): 1731–1755, 2015c.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015d.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Advances in Neural Information Processing Systems*, 2017a.

Adith Swaminathan, Akshay Krishnamurthy, Alekh Agarwal, Miroslav Dudík, John Langford, Damien Jose, and Imed Zitouni. Off-policy evaluation for slate recommendation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*,

NIPS'17, page 3635–3645, Red Hook, NY, USA, 2017b. Curran Associates Inc. ISBN 9781510860964.

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, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=IfgOWI5v2f.

Muhammad Faaiz Taufiq, Patrick Blöbaum, and Lenon Minorics. Manifold restricted interventional shapley values. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5079–5106. PMLR, 25–27 Apr 2023a. URL https://proceedings.mlr.press/v206/taufiq23a.html.

Muhammad Faaiz Taufiq, Arnaud Doucet, Rob Cornish, and Jean-Francois Ton. Marginal density ratio for off-policy evaluation in contextual bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=noyleECBam.

Muhammad Faaiz Taufiq, Jean-Francois Ton, and Yang Liu. Achievable fairness on your data with utility guarantees, 2024. URL https://arxiv.org/abs/2402.17106.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning, 2020.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2139–2148. JMLR.org, 2016.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, 2015.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, 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.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning, 2022. URL https://arxiv.org/abs/2212.06355.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=IsK8iKbL-I.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine, editors, *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, 2005.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, page 3589–3597, 2017a.

Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3589–3597. JMLR.org, 2017b.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/4ffb0d2ba92f664c2281970110a2e071-Paper.pdf.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper/2019/file/4ffb0d2ba92f664c2281970110a2e071-Paper.pdf.

Liyuan Xu and Arthur Gretton. Kernel single proxy control for deterministic confounding, 2024.

Liyuan Xu, Heishiro Kanagawa, and Arthur Gretton. Deep proxy causal learning and its application to confounded bandit policy evaluation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=0FDxsIEv9G.

Xiao Xu, Fang Dong, Yanghua Li, Shaojian He, and Xin Li. Contextual-bandit based personalized recommendation with time-varying user interests. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:6518–6525, 04 2020. doi: 10.1609/aaai.v34i04.6125.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *The Annals of Statistics*, 50(5):2587 – 2615, 2022. doi: 10.1214/22-AOS2195. URL https://doi.org/10.1214/22-AOS2195.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *arXiv preprint arXiv:2112.03493*, 2021.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran

Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8252831b9fce7a49421e622c14ce0f65-Paper.pdf.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2751–2768. PMLR, 2023.