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 distributions, 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 individual 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 dynamic treatment regimes 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

In the real world, making informed decisions is crucial. We constantly strive to take actions that lead to desirable outcomes, whether it's 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]. 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 the concept of *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 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. It allows 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 all the factors that influence a patient's response to treatment. 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. Imagine a doctor deciding on the best treatment for a patient. The optimal treatment might depend on various factors like the patient's 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.

Causal formulation To make the causal dependence of the action on the outcome explicit, we can reformulate the contextual bandits setting using the *potential outcomes framework* proposed by Robins [1986]. Here, for a given action $a \in \mathcal{A}$ we denote the random variable $Y(a)$ (also known as the potential outcome), as the outcome that *would* occur if the action a is chosen. As random variables, $Y(a)$ may also depend on the initial covariates X and additional randomness which is not explicitly modelled. Moreover, this notation can be reconciled with the conventional contextual bandits setting above by noting that, in the real world where an agent chooses action $A \in \mathcal{A}$ for context X, the outcome Y observed is explicitly expressed as $Y(A)$, i.e. $Y = Y(A)$ in the real world. This corresponds to the standard consistency assumption in causal inference [Hernán and Robins, 2020] and intuitively means that the potential outcome $Y(a)$ is observed in the data when the agent actually chose $A = a$.

Contextual bandits encapsulate the single-decision regimes where, for each observed context, we make 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 concept of dynamic treatment regimes (DTRs) [Tsiatis et al., 2019] in the following section. DTRs extend 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.

1.2 Dynamic treatment regimes

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 $\mathcal{X}_t := \mathbb{R}^{d_t}$. 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 setting describes dynamic treatment regimes, of which the contextual bandits are a special case when $T = 1$.

Causal formulation As before, we make the causal dependence of past actions on future observations explicit by modelling the dynamics of the real-world process via the longitudinal potential outcomes framework proposed [Rubin, 1974, 2005]. To streamline notation, we will index the 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 $a_1 \in \mathcal{A}_1, \ldots, a_t \in \mathcal{A}_t$.

In this formulation, for each action sequence $a_{1:T} \in \mathcal{A}_{1:T}$, 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 potential outcomes $X_{1:T}(a_{1:T})$ are also referred to as *interventional outcomes* under the intervention $a_{1:T}$.

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.

Just like in the contextual bandits setting, this causal notation can be reconciled with the conventional (non-causal) notation above by noting that, in the real world where an agent chooses action $A_{1:t} \in \mathcal{A}_{1:t}$, the observation at time t, X_t, is explicitly expressed as $X_t(A_{1:t})$, i.e. $X_t = X_t(A_{1:t})$ in the real world. Like before, this is 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 chose $A_{1:t} = a_{1:t}$.

Now that we have outlined the single and multiple decision frameworks, we are now equipped to formally define the off-policy evaluation setting which is fundamental to the problems considered in this thesis.

1.3 Off-policy evaluation

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., 2018, Su et al., 2019, Metelli et al., 2021, Liu et al., 2019, Sugiyama and Kawanabe, 2012, Swaminathan et al., 2017]. This is particularly valuable when conducting controlled experiments with the new policy is impractical or unethical. In what follows, we formally define the OPE problem in contextual bandits which will set up the challenges tackled in Chapters 2 and 3 of this thesis.

1.3.1 Off-policy evaluation in contextual bandits

Recall the standard contextual bandit setting, where $X \in \mathcal{X}$ is a context vector (e.g., user features), $A \in \mathcal{A}$ denotes an action (e.g., recommended website to the user), and $Y \in \mathcal{Y}$ denotes a scalar reward or outcome (e.g., whether the user clicks on the website).

Assumption 1.3.1 (No unmeasured confounding)**.** *In this setting, it is standard to assume that the agent's action A depends only on the context X and possibly additional randomness independent of everything else. This means that when choosing the action A, the agent does not rely on additional information relevant to the outcome Y which is not captured in the context X. To be concrete, 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] and can be technically outlined in the language of potential outcomes as follows:*

$$\{Y(a) \,|\, a \in \mathcal{A}\} \perp\!\!\!\perp A \mid X$$

Let $\mathcal{D} := \{(x_i, a_i, y_i)\}_{i=1}^n$ 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 $A \mid X = x \sim \pi^b(\cdot \mid x)$. Then, under Assumption 1.3.1, it is straightforward to show that the joint density of (X, A, Y) from which the logged data $\mathcal{D}$ is sampled, denoted by p_{π^b}, can be factorised as follows:

$$p_{\pi^b}(x, a, y) := p(y \mid x, a)\, \pi^b(a \mid x)\, p(x). \tag{1.1}$$

Likewise, the joint density of (X, A, Y) under the *target policy* π^* can be factorised as

$$p_{\pi^*}(x, a, y) := p(y \mid x, a)\, \pi^*(a \mid x)\, p(x). \tag{1.2}$$

Moreover, 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, i.e.,

$$\mathbb{E}_{\pi^b}[\,\cdot\,] := \mathbb{E}_{(X,A,Y)\sim p_{\pi^b}}[\,\cdot\,], \qquad \text{and} \qquad \mathbb{E}_{\pi^*}[\,\cdot\,] := \mathbb{E}_{(X,A,Y)\sim p_{\pi^*}}[\,\cdot\,].$$

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^*, i.e., $\mathbb{E}_{\pi^*}[Y]$, using only the logged data $\mathcal{D}$.

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 $\rho(a, x) := \frac{p_{\pi^*}(x,a,y)}{p_{\pi^b}(x,a,y)} = \frac{\pi^*(a|x)}{\pi^b(a|x)}$, given the factorizations in Eqns. (1.1) and (1.2). This leads to the commonly used *Inverse Probability Weighting (IPW)* [Horvitz and Thompson, 1952] estimator:

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

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:

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

where

$$\hat{\eta}(\pi^*) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a' \in \mathcal{A}} \hat{\mu}(a', x_i) \pi^*(a' \mid x_i) \approx \mathbb{E}_{\pi^*}[\hat{\mu}(A, X)]. \tag{1.3}$$

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.

1.4 Limitations of existing OPE methods

1.4.1 High variance

The conventional off-policy value estimators (including IPW and DR estimators) use policy ratios $\rho(a,x) := \pi^*(a \mid x)/\pi^b(a \mid x)$ 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. Even though the DR estimator uses control variates for variance reduction, it 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 size of the action and context spaces grows [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.

Weight clipping and normalization 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. However, these approaches introduce a bias-variance trade-off, as clipping the weights can introduce bias into the estimates. 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:

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

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 $\hat{\eta}$ is the Direct Method (DM) as defined in (1.3). Like weight clipping, this approach can also increase the bias, since the DM can have a high bias.

Marginalization-based techniques Several works explore marginalisation techniques for variance reductions. For example, Saito and Joachims [2022] propose 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 (as in IPW). While this approach reduces the variance associated with IPW, 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. [2019], 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.

1.4.2 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, given observational dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$, Chandak et al. [2021] proposed a non-parametric Weighted Importance Sampling (WIS) estimator for the empirical CDF of Y under π^*,

$$\hat{F}_{\text{WIS}}(t) := \frac{\sum_{i=1}^n \hat{\rho}(a_i, x_i)\mathbb{1}(y_i \leq t)}{\sum_{i=1}^n \hat{\rho}(a_i, x_i)}$$

where $\hat{\rho}(a, x) := \frac{\pi^*(a|x)}{\hat{\pi}^b(a|x)}$ are the importance weights. Chandak et al. [2021] show that $\hat{F}_{\text{WIS}}(t)$ is a uniformly consistent estimator of the off-policy CDF, $F_{\pi^*}(t) := \mathbb{P}_{\pi^*}(Y \leq t)$. Therefore, we can use $\hat{F}_{\text{WIS}}$ to construct predictive intervals on the outcome Y under target policy π^*. 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.

1.4.3 Assumption of no unmeasured confounding

The standard OPE methodologies assume no unmeasured confounding (formalised in Assumption 1.3.1) 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 A chosen, but also the outcome Y observed subsequently. This can happen when the real-world agent has access to more information than is available in the context X captured in the data. In such circumstances, the causal effect of a given action $a \in \mathcal{A}$ 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.

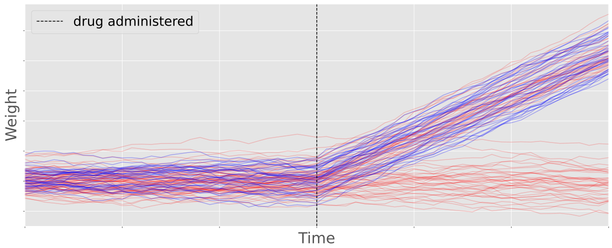

Figure 1.1: The discrepancy between observational data and interventional behaviour in the presence of unmeasured confounding: the range of outcomes observed in the data for patients who were administered the drug (blue) differs from what *would* be observed if the drug were administered to the general 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 dynamic treatment regimes (DTRs). This topic forms the central focus of Chapter 4 of this thesis.

Observational data Recall that in this setting, our observational data comprises 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:

$$X_0, A_1, X_1(A_1), \ldots, A_T, X_T(A_{1:T}). \tag{1.4}$$

As outlined in the previous section, the standard off-policy evaluation methods, both in contextual bandits and DTRs, assume no unmeasured confounding in the observational dataset. Informally, this assumption 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. 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]. The following foundational result from the causal inference literature (often referred to as the *fundamental problem of causal inference* [Holland, 1986]) formalises this.

Theorem 1.4.1

If $\mathbb{P}(A_{1:t} \neq a_{1:t}) > 0$, then the expectation $\mathbb{E}[X_t(a_{1:t})]$ is not uniquely identified by the distribution of the data in (1.4) without further assumptions.

The expectation $\mathbb{E}[X_t(a_{1:t})]$ in Theorem 1.4.1 denotes the value at time t of a deterministic target policy, which always chooses a given action sequence $a_{1:t}$ regardless of everything else. Therefore, this result shows that the value of a deterministic target policy is unidentifiable in general, even if we have access to an infinitely large observational dataset. This also implies that the distribution of outcomes under this deterministic target policy, $\text{Law}[X_{0:t}(a_{1:t})]$, is unidentifiable from the observational data distribution in general. We can also derive an analogous result for general target policies which choose actions at time $t' \in \{1, \dots, t\}$ dynamically depending on past observations [Tsiatis et al., 2019].

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 confounding models impose assumptions 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 dynamic treatment regimes. We propose a set of novel bounds on the causal effects in sequential decision-making settings, which remain valid in the presence of unmeasured confounding and rely on minimal assumptions making them highly applicable to a wide variety of real-world settings.

1.5 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 1.4.1, 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 focuse 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 1.4.2, 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 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 dynamic treatment regimes (DTRs), where available observational data may suffer from unmeasured confounding. As mentioned in Section 1.4.3, fundamental results from causal inference mean that in this setting the interventional behaviour of outcomes, $\text{Law}[X_t(a_{1:t})]$, 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, $X_t(a_{1:t})$. 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 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.

1.6 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.

1.6.1 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. Under review at *Biometrika*. [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]

1.6.2 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 these problems, 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 offmanifold 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

3
Conformal Off-Policy Prediction in Contextual Bandits

4

Causal Falsification of Digital Twins

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 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 presented 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. Howoever, 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.

Bibliography

Jason Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *Journal of Machine Learning Research*, 20(91):1–39, 2019.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68, 11 2019. doi: 10.1287/opre.2019.1902.

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.

Rob Cornish, Muhammad Faaiz Taufiq, Arnaud Doucet, and Chris Holmes. Causal falsification of digital twins, 2023. URL `https://arxiv.org/abs/2301.07210`.

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`.

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 2018. URL `https://proceedings.mlr.press/v80/farajtabar18a.html`.

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.

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`.

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

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`.

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.

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`.

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.

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.

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.

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.

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.

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`.

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`.

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`.

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`.

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. *Partial Identification of Probability Distributions*. Springer, 2003.

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`.

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`.

Paul R. Rosenbaum. *Observational Studies*. Springer, New York, NY, 2002.

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`.

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, 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.

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`.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. *International Conference on Representation Learning*, 2022.

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 2019. 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 *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3635–3645, Red Hook, NY, USA, 2017. 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, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *AAAI Conference on Artificial Intelligence*, 2015.

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

Vladimir Vovk, 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., 2019. 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.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *International Conference on Artificial Intelligence and Statistics*, pages 2751–2768. PMLR, 2023.