State-Action Balancing in Multi-Stage Causal Inference

Q. Du[⋆], G. Biau[⋆], F. Petit[†], and R. Porcher[†]

[⋆] *Sorbonne Université, CNRS, LPSM*
[†] *Université de Paris, CRESS, INSERM, INRA*

Abstract

We propose in this article new strategies for off-policy policy evaluation problem in multi-stage causal inference under finite horizon setting. Our methods are mainly based on a recursive implementation of the recently developed balancing techniques, widely used in static causal inference and other disciplines. The balancing aims at correcting the difference between distributions through re-weighting. In order to adapt to the dynamical setting, we introduce a new framework of multi-stage causal inference that emphasizes the measure flows and semigroup structures in the associated causal dynamics, especially in the change-of-policy procedure. This allows us to study different applications such as Reinforcement Learning and Dynamical Treatment Regimes in full generality, without worrying about the problems such as covariate shifts and continuous-valued actions spaces. We also provide a new theoretical foundation of general balancing methods through a Riesz representable measurable space argument. It reveals that several completely different balancing methods can be unified in a consistent manner, which also give birth to a new balancing implementation given by adversarial optimization. For the actual policy evaluation estimators, we study both the classical direct estimator and the doubly robust one, with separate assumptions. To illustrate the performance of our methods, several numerical tests are provided in the end.

Key words: Causal inference, Off-policy policy evaluation, Reinforcement learning, Dynamical treatment regimes, density ratio estimation, doubly robust estimation.

Contents

1 Introduction

1.1 Multi-stage causal inference

Causal inference and policy design is an important and active area in modern statistics. The most classical formulation is the potential outcomes framework formalized by Rubin (1974) (the book by Imbens and Rubin, 2015, provides a good summary to the domain). In a high level, the ultimate goal is to learn, in a one-stage trial, the optimal policy based on collected triples of covariates, outcomes, and treatment assignment indicators. In order to achieve this, the essential is to be able to provide high-quality policy evaluation methods. In the potential outcomes framework, this is often done through estimation of the Average Treatment Effects (ATE) or Conditional Average Treatment Effects (CATE) functions, depending on whether we consider personalized policy assignment. The idea is to estimate directly the potential outcomes or the associated advantages for given policies in order to support the decision making procedure.

One of the many difficulties for policy evaluation is that the data collecting procedure is often observational, namely, the treatment assignments may depend on the covariates. In this case, assuming *unconfoundedness*, i.e., the outcomes and treatment assignments are conditionally independent given the covariates, an important technique is to balance the sub-populations that receive different treatments. In this way, when conducting policy evaluation, one can efficiently exchange the information between different treatment groups (Manski, 2004; Kallus and Zhou, 2018; Chernozhukov et al., 2018; Kallus, 2020; Kallus et al., 2019; Kallus and Santacatterina, 2019b,a; Chernozhukov et al., 2020). For other topics on the recent developments of causal inference, the reader is referred to, e.g., Zhang et al. (2012); Dudík et al. (2014); Athey et al. (2017); Nie and Wager (2017); Kitagawa and Tetenov (2018); Künzel et al. (2019); Bertsimas and Kallus (2020).

On the other hand, many real-world problems require more complicated inference that involves multi-stage decision making. This is for example the typical situation in the clinical trials that deal with sequential treatments. In the sequel, such setting is referred to as dynamical or multi-stage causal inference. Similar to the static setting, the typical goal of dynamical causal inference is to estimate the optimal policy at each stage of the considered system. In contrast to the classical control theory, the decision making process encountered in dynamical causal inference often involves complex interactions with the environment and unknown stochastic mechanisms. Only observations under unknown sampling policy can be collected. In the dynamical setting, the most popular frameworks are offline Reinforcement Learning (RL, see, e.g., Sutton and Barto, 2018) and Dynamical Treatment Regimes (DTR, see, e.g., Chakraborty, 2013). From a mathematical point of view, off-policy policy evaluations (i.e., policy evaluation of a target policy under a different sampling policy) in RL and DTR share the same probability language. However, there are many differences that should be taken into consideration when it comes to the real-world applications. For example, offline RL often emphasizes long-term (or even infinite horizon, e.g., Nachum et al., 2019; Bennett et al., 2021) behaviors of the policy while the time horizon studied in DTR is usually much shorter. At the same time, each interaction with the environment is often assumed to be time-homogeneous and Markovian, in the usual sense, in the RL setting, while in the DTR context, the evolution is in general non-homogeneous and no Markovian properties are guaranteed. Such considerations give rise to different yet connected methods that deal with the policy evaluation problems in these two areas (e.g., Zhang et al., 2013; Zhao et al., 2015; Jiang and Li, 2016; Thomas and Brunskill, 2016; Nie et al., 2020; Liu et al., 2020).

One of the objectives of the present paper is to introduce a unified language that is suitable to conduct analysis for the policy evaluation problem from these two different angles. Unlike the traditional formalisms in the context of RL and DTR, we emphasize the measure flows and the associated partial semigroup structures encountered in the change of policy. This abstraction allows us to understand and study the multi-stage causal inference in full generality. For example,

our framework is insensitive to covariate shifts (Sugiyama et al., 2007; Shimodaira, 2000) of the target population and dynamics that involve continuous-value action spaces. Using these formalisms, we are able to transfer the balancing techniques (see, e.g., Kallus, 2020; Chernozhukov et al., 2020) from static problems to the dynamical setting.

The balancing techniques that we develop are in line with some recently introduced weight function estimation techniques in causal inference and other disciplines (Kallus, 2018; Kallus et al., 2019; Nachum et al., 2019; Kallus and Santacatterina, 2019b,a; Kallus, 2020; Reygner and Touboul, 2020). Roughly speaking, the goal of balancing is to correct the difference between two measures through re-weighting. To have an idea, in the context of potential outcomes framework, the well-known Inverse Probability Weighting (IPW) estimator can indeed be interpreted as a special form of such re-weighting. In this case, the difference between the sub-populations is corrected by the "inverse" of the nuisance propensity score estimators. It is known that due to this specific "inverse" construction, such weight estimation methods, despite their asymptotic behaviors, suffer from numerical instability when the overlap between different sub-populations is not significant in the data set. On the contrary, the general strategy of balancing methods minimize directly the difference between empirical measures, and no "inverse probability" manipulation is involved. This "end-to-end" nature of balancing also improves the robustness of policy evaluation.

Typically, the analysis of re-weighting strategies fall into the context of density ratio estimation (see, e.g., Sugiyama et al., 2012). We first show that in the static setting, one is able to use the idea of the density ratio estimation to construct generalized inverse probability weighting-type estimators—one estimates the re-weighting function by constructing an artificial supervised learning (0-1 classificaiton) problem. This makes the policy evaluation possible when the action is space is potentially of continuous values. Then, thanks to our new formulation that emphasized the measure flows of the causal dynamics, the generalized inverse probability weighting can also be applied in the dynamical setting. Such constructions serve as the baseline of our numerical tests.

Our formulation of balancing is mainly inspired by the language of generalized optimal matching studied by Kallus (2020) and a Riesz representer argument introduced in Chernozhukov et al. (2020). Interestingly, we show that these two static balancing methods, no matter how dissimilar they seem to be, share the same basic motivation through a Riesz representable measure space argument, which serves as the theoretical foundation of the balancing techniques. The contribution of the present paper can be summarized as follows:

- We propose a new language in dynamical causal inference that focus on the measure flows in the change-of-policy procedure, which enables us to study different applications in full generality without worrying about continuous action space, covariate shifts, etc. This is the topic of the first part of Section 2.
- We provide in Section 2 and Section 3 new insights into balancing methods and propose a recursive balancing strategy that can be applied in the dynamical setting, which, in turn, gives robust and efficient estimators compared to the classical inverse probability weighting-type construction.

Before starting, we provide in the next subsection the notation and conventions that will be useful throughout the article.

1.2 Notation and conventions

- If not mentioned otherwise, all random variables considered in the paper are defined on the same canonical probability space $(\Omega, \mathscr{F}, \mathbb{P})$. They take their values in Polish spaces—namely, separable completely metrizable topological spaces—and are assumed to be at least Borel measurable. Whenever a new space is defined as the product of two Polish spaces, say $\mathcal{E} \times \mathcal{E}'$, we suppose that the associated σ-algebra contains at least $\mathcal{B}(\mathcal{E}) \otimes \mathcal{B}(\mathcal{E}')$, where

$\mathcal{B}(\mathcal{E})$ and $\mathcal{B}(\mathcal{E}')$ denote respectively the Borel σ-algebra of $\mathcal{E}$ and $\mathcal{E}'$. This ensures that every projection in this article is measurable. All Polish spaces are denoted by calligraphy letters, e.g., $\mathcal{X}$. If not mentioned otherwise, the associated normal uppercase letter, i.e., X, denotes a random variables on $\mathcal{X}$. The corresponding lowercase letter, i.e., x, denotes an arbitrary point on $\mathcal{X}$.

- We let $\mathscr{M}(\mathcal{E})$, $\mathscr{M}_+(\mathcal{E})$, and $\mathscr{P}(\mathcal{E})$ be the set of all signed finite measures, the subset of all nonnegative finite measures, and the subset of all probability measures on $(\mathcal{E}, \mathscr{E})$ for some σ-algebra $\mathscr{E}$ that contains the Borel σ-algebra $\mathcal{B}(\mathcal{E})$. We denote by $\mathscr{B}_b(\mathcal{E})$ the collection of all bounded measurable functions from $(\mathcal{E}, \mathscr{E})$ to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ equipped with uniform norm $\|\cdot\|_\infty$, among which the constant function will be denoted by $\mathbf{1}_\mathcal{E}$ or simply 1 with a slight abuse of notation.
- For all $\mu \in \mathscr{M}(\mathcal{E})$ and for all test functions $f \in \mathscr{B}_b(\mathcal{E})$, $\mu(f)$ is the integral

$$\int_\mathcal{E} f(x)\mu(\mathrm{d}x).$$

A finite transition kernel Q from $(\mathcal{E}, \mathscr{E})$ to $(\mathcal{E}', \mathscr{E}')$ is a function

$$Q : \mathcal{E} \times \mathscr{E}' \mapsto \mathbb{R}_+.$$

More precisely, for all $x \in \mathcal{E}$, $Q(x, \cdot)$ is a finite nonnegative measure in $\mathscr{M}_+(\mathcal{E}')$ and for all $B \in \mathscr{E}'$, $x \mapsto Q(x, B)$ is a $\mathscr{E}$-measurable function. We say that Q is a Markov transition kernel if Q is a finite transition kernel and for all $x \in \mathcal{E}$, $Q(x, \cdot)$ is a probability measure in $\mathscr{P}(\mathcal{E}')$. For a signed measure $\mu \in \mathscr{M}(\mathcal{E})$ and a test function $f \in \mathscr{B}_b(\mathcal{E}')$, we let respectively $\mu Q \in \mathscr{M}(\mathcal{E})$ and $Q(f) \in \mathscr{B}_b(\mathcal{E})$ respectively defined as follows:

$$\forall B \in \mathscr{E}', \quad \mu Q(B) = \int_\mathcal{E} Q(x, B)\mu(\mathrm{d}x),$$

and

$$\forall x \in \mathcal{E}, \quad Q(f)(x) = \int_{\mathcal{E}'} Q(x, \mathrm{d}y)f(y).$$

Let Q_1 and Q_2 be two finite transition kernels respectively from $\mathcal{E}_0$ to $\mathcal{E}_1$ and from $\mathcal{E}_1$ to $\mathcal{E}_2$. When well-defined, we denote by $Q_1 \circ Q_2$, or simply Q_1Q_2, the transition kernel from $\mathcal{E}_0$ to $\mathcal{E}_2$, defined by

$$\forall (x, B) \in \mathcal{E}_0 \times \mathscr{B}(\mathcal{E}_2), \quad Q_1Q_2(x, B) = \int_{\mathcal{E}_1} Q_2(y, B)Q_1(x, \mathrm{d}y).$$

Note that there is no reason that Q_1Q_2 is still a finite transition kernel in general.

- For two test functions $f, g \in \mathscr{B}_b(\mathcal{E})$, we let

$$f \otimes g : \mathcal{E}^2 \ni (x, y) \mapsto f(x)g(y) \in \mathbb{R},$$

as well as

$$f \times g : \mathcal{E} \ni x \mapsto f(x)g(x) \in \mathbb{R}.$$

For two finite transition kernels Q and H from $(\mathcal{E}, \mathscr{E})$ to $(\mathcal{E}', \mathscr{E}')$, we denote by $Q \otimes H$ the finite transition kernel defined by, for all $(x, y) \in \mathcal{E} \times \mathcal{E}'$ and for all $(B, B') \in \mathscr{E} \times \mathscr{E}'$,

$$Q \otimes H\left((x, y), (B, B')\right) := Q(x, B) \times H(y, B').$$

Similarly, we let

$$Q^{\otimes 2} := Q \otimes Q.$$

For any $\mu \in \mathscr{M}(\mathcal{E})$ and for any finite transition kernel Q from $(\mathcal{E}, \mathscr{E})$ to $(\mathcal{E}', \mathscr{E}')$, we denote

$$\mu \otimes Q = \mu(\mathrm{d}x) \times Q(x, \mathrm{d}y).$$

- For some scalar function $G : \mathcal{E} \mapsto \mathbb{R}$, we consider the re-weighting transformation Ψ_G, when well-defined, given by
$$\Psi_G : \mathscr{M}(\mathcal{E}) \ni \xi \mapsto \xi(G \times \cdot) \in \mathscr{M}(\mathcal{E}).$$
The image $\Psi_G(\xi)$ is a weighted measure of ξ according to the weighting function G.

2 Dynamical causal inference

We introduce in this section the framework of policy evaluation in dynamical causal inference. Unlike the traditional formulation in the RL and DTR, we emphasize the transitions and measure flows in the dynamics, especially in the change-of-policy procedure. We start the section by providing the formal definition of discrete-time Markov Decision Process (MDP), emphasizing the partial semigroups therein. To study the structural properties in full generality (such as continuous-value action dynamics), the associated state and action spaces are assumed to be general Polish spaces. This in particular allows us to treat non-Markovian models as a special Markov case (see, e.g., Kallus and Uehara, 2020, Remark 8), where the states are simply replaced by the associated historical trajectories. For concrete applications in RL and DTR, we refer to, e.g., Murphy (2005) and Nie et al. (2020). Next, we study the measure flows in the change-of-policy procedure. After figuring out the essential measure flows that characterize the causal dynamics, we briefly introduce through an example the core idea of the recursive balancing strategy, postponing the theoretical formalism to Section 3. Finally, we study the actual estimators for policy evaluation. We show that the role of balancing is to provide the nuisance weight function estimation, and thoroughly discuss how the quality of balancing may affect the quality of the actual policy evaluation estimators.

2.1 Setting

We first reformulate the basic concept of MDP, the probability language used in dynamical causal inference. Denote by $T \in \mathbb{N}$ a fixed time horizon. Let $(\mathcal{X}_t; 0 \leq t \leq T+1)$ and $(\mathcal{A}_t; 0 \leq t \leq T)$ be respectively the sequence of Polish state spaces and actions spaces. We call $\pi = (\pi_t; 0 \leq t \leq T)$ the *sampling policy*, which is composed by a sequence of Markov kernels. More precisely, π_t is a Markov kernel from the state space $\mathcal{X}_t$ to the action space $\mathcal{A}_t$. The state-action spaces $(\mathcal{X}_t^\natural; 0 \leq t \leq T)$ are defined by $\mathcal{X}_t^\natural = \mathcal{X}_t \times \mathcal{A}_t$ at each time step $0 \leq t \leq T$. To find out the partial semigroup structure along with time t, we consider a natural state-action Markov transition kernel from $\mathcal{X}_t$ to the state-action space $\mathcal{X}_t^\natural$, defined by

$$\pi_t^\natural = \mathrm{id}_{\mathcal{X}_t} \times \pi_t,$$

where $\mathrm{id}_{\mathcal{X}_t}$ is the Dirac mass on $\mathcal{X}_t$. In order to highlight the role of state-action dynamics, any notation that involves state-action concept are equipped with a symbol "$\natural$". The interactions with environment are modeled by a sequence of transition kernels $(M_t; 1 \leq t \leq T+1)$. More precisely, the Markov kernel M_t evolves the system from the state-action space $\mathcal{X}_{t-1}^\natural$ to the next state space $\mathcal{X}_t$. The dynamics start from an initial distribution ξ_0 on the state space $\mathcal{X}_0$. This causal dynamics is illustrated in Figure 1, where $\mathrm{Proj}(\cdot|\mathcal{A}_t)$ denotes the projection on the space $\mathcal{A}_t$. Remark that this figure is essential to understand the mechanism of the strategy presented in the present paper.

The quality of a policy is assessed by a sequence of observation/reward functions $(r_t; 0 \leq t \leq T)$, where r_t is a measurable function on $\mathcal{X}_{t+1}$. Intuitively, starting from the state X_t, after

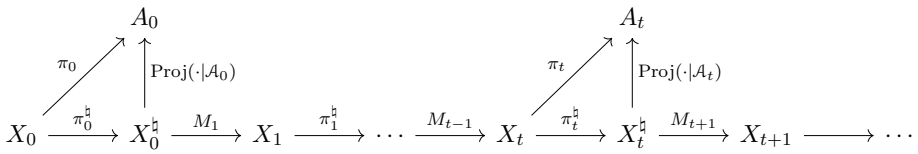


Figure 1: Markov structure of causal dynamics

making decision w.r.t. π_t, an interaction with environment happens, which is modeled by M_{t+1}. The function r_t then observes the outcomes from the state X_{t+1}. Roughly speaking, the quality of a policy is therefore represented by the sum of all such observations at each time step. The challenge is that the evolution mechanisms presented above are unknown. In practice, we only collect i.i.d. trajectories $\mathcal{D}_n = \{(X_t^{(i)}, A_t^{(i)}; 0 \leq t \leq T) : 1 \leq i \leq n\}$ under an unknown sampling policy π, together with the observations $\mathcal{R}_n = (r_t(X_{t+1}^{(i)}); 0 \leq t \leq T, 1 \leq i \leq n)$. The goal of policy evaluation is then to estimate the potential outcomes/rewards for a new user-prefixed policy $\mathring{\pi}$, using the collected data set $\mathcal{D}_n$ and the observation values $\mathcal{R}_n$ under sampling policy π. More precisely, we estimate the value $\mathcal{V}^{\mathring{\pi}}$ defined by

$$\mathcal{V}^{\mathring{\pi}} = \mathbb{E}\left[\sum_{t=0}^{T} r_t(\mathring{X}_{t+1})\right],$$

where $(\mathring{X}_t; 0 \leq t \leq T+1)$ is generated by the mechanism described in Figure 1, where the policy π is replaced by the target policy $\mathring{\pi}$.

The *covariate shifts* in the dynamical causal inference mean that the initial distribution ξ_0 also changes with the new target policy $\mathring{\pi}$. This is usually an important factor to be considered in the context of DTR. For example, it is in general difficult to ensure that the testing group of a clinical trial is identical, in terms of distribution, to the target population of the potential receivers of the considered treatment. This phenomenon is known as a lack of *external validity* (Shimodaira, 2000; Sugiyama et al., 2007; Pearl et al., 2014). Returning to the mathematical argument, the covariate shifts indicate that the initial distribution $\mathring{\xi}_0$ of $\mathring{X}_0$ is different from ξ_0, the initial distribution that generates the data set $\mathcal{D}_n$. In this case, we assume that we have access to the i.i.d. samples of the target population $\mathcal{I}_m = \{\mathring{X}_0^{(j)} : 1 \leq i \leq m\}$. However, we do not assume that $\mathcal{I}_m \perp\!\!\!\perp \mathcal{D}_n$. This in particular allows us to treat the case without covariate shifts as a special sub-case by taking $m = n$ and replacing $\mathring{X}_0^{(i)}$ with $X_0^{(i)}$. Now, it is time to provide the formal definition of the policy evaluation problem in dynamical causal inference.

Definition 1 (Policy evaluation)**.** Given a known target evaluation policy $\mathring{\pi}$, the i.i.d. trajectories of the causal dynamics $\mathcal{D}_n$, the observation values $\mathcal{R}_n$, and the i.i.d. target population samples $\mathcal{I}_m$ (without assuming $\mathcal{I}_m \perp\!\!\!\perp \mathcal{D}_n$), the goal of policy evaluation is to estimate the value $\mathcal{V}^{\mathring{\pi}}$. □

The challenge of policy evaluation problem lies in the fact that the data set $\mathcal{D}_n$ is collected under sampling policy π, and no observations under the target evaluation policy $\mathring{\pi}$ are available. Moreover, the objective value $\mathcal{V}^{\mathring{\pi}}$ can be regarded as the average reward observations under target policy $\mathring{\pi}$, while the collectable reward observations $\mathcal{R}_n$ is only available on the support of $X_t^{(i)}$, which is generated under sampling policy π. Therefore, the policy evaluation problem consists in how we can exploit $\mathcal{D}_n$ and $\mathcal{R}_n$ (as well as $\mathcal{I}_m$ when covariate shifts are involved) to approximate the average accumulated rewards $\mathcal{V}^{\mathring{\pi}}$ of a given target policy $\mathring{\pi}$.

Remark 2.1 (Observations values $\mathcal{R}_n$). In the RL context, the reward functions r_t are known and we usually also have access to the samples of the terminal state $(X_{T+1}^{(i)}; 1 \leq i \leq n)$. It is then obvious that $\mathcal{R}_n$ is available. In the DTR context, we usually have a single potential outcomes, say $Y^{(i)}$, for i-th trajectory in the data set $\mathcal{D}_n$. In this case, the random variable

$Y^{(i)}$ can indeed be characterized by r_T and $X_{T+1}^{(i)}$ under some regular assumptions such as *consistency* and *sequential ignorability* (Nie et al., 2020). For example, we consider $X_{T+1}^{(i)} = Y^{(i)}$ and $r_T = \mathrm{id}_{\mathcal{X}_{T+1}}$. Accordingly, we let $r_t = 0$ for all $0 \leq t \leq T-1$. This is also compatible with our current formulation. □

Partial semigroups in MDP To have a deeper understanding of the problem proposed in Definition 1, let us go one step further from the dynamics illustrated in Figure 1. If we let respectively

$$M_t^{\pi} = \pi_{t-1}^{\natural} \circ M_t \quad \text{and} \quad M_t^{\pi^{\natural}} = M_t \circ \pi_t^{\natural},$$

one gets a double partial semigroup structure illustrated in Figure 2. More precisely, we consider

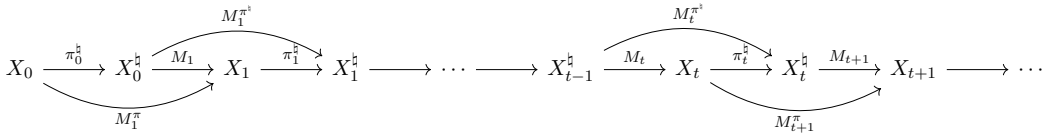


Figure 2: Double semigroups in causal dynamics

two partial semigroups defined respectively by

$$\forall s > t, \quad M_{t,s}^{\pi} = M_{t+1}^{\pi} \circ \cdots \circ M_s^{\pi}, \quad \text{with} \quad M_{t,t}^{\pi} = \mathrm{id}_{\mathcal{X}_t},$$

and

$$\forall s > t, \quad M_{t,s}^{\pi^{\natural}} = M_{t+1}^{\pi^{\natural}} \circ \cdots \circ M_s^{\pi^{\natural}}, \quad \text{with} \quad M_{t,t}^{\pi^{\natural}} = \mathrm{id}_{\mathcal{X}_t^{\natural}}.$$

They connect respectively the dynamics of state chain $(X_t; 0 \leq t \leq T+1)$ and state-action chain $(X_t^{\natural}; 0 \leq t \leq T)$. Considering the initial distributions $\xi_0^{\pi} = \xi_0$ and $\xi_0^{\pi^{\natural}} = \xi_0 \circ \pi_0^{\natural}$, we define the terminal measures ξ_t^{π} and $\xi_t^{\pi^{\natural}}$ respectively by

$$\xi_t^{\pi} = \xi_0^{\pi} M_{0,t}^{\pi} \quad \text{and} \quad \xi_t^{\pi^{\natural}} = \xi_0^{\pi^{\natural}} M_{0,t}^{\pi^{\natural}}.$$

When the policy is replaced by $\mathring{\pi}$, we also change the initial distributions respectively by $\mathring{\xi}_0^{\mathring{\pi}} = \mathring{\xi}_0$ and $\mathring{\xi}_0^{\mathring{\pi}^{\natural}} = \mathring{\xi}_0 \circ \mathring{\pi}_0^{\natural}$. Mutatis mutandis, we also define the terminal measures $\mathring{\xi}_t^{\mathring{\pi}}$ as well as $\mathring{\xi}_t^{\mathring{\pi}^{\natural}}$. The introduction of these terminal measures provides an alternative formulation of the objective, namely, the average accumulated rewards, in policy evaluation problem given in Definition 1, that is,

$$\mathcal{V}^{\mathring{\pi}} = \sum_{t=0}^{T} \mathring{\xi}_{t+1}^{\mathring{\pi}}(r_t) = \sum_{t=0}^{T} \mathring{\xi}_t^{\mathring{\pi}^{\natural}}(r_t^{\natural}), \tag{1}$$

where $r_t^{\natural} = M_{t+1}(r_t)$. The alternative representation (1) gives two interpretations of the policy evaluation problem, respectively in terms of terminal measures $\mathring{\xi}_{t+1}^{\mathring{\pi}}$ and $\mathring{\xi}_t^{\mathring{\pi}^{\natural}}$. Followed from the double partial semigroup structure revealed in Figure 2, these two formulations are closely related to the two types of estimators that will be discussed later in Section 2.3. From this point of view, the core problem of the policy evaluation is to approximate a sequence of terminal measures with the observable empirical measures available in $\mathcal{D}_n$ and the observation values $\mathcal{R}_n$. It is then crucial to understand the measure flows in the change-of-policy procedure, as the objective is given by the terminal measures ($\mathring{\xi}_{t+1}^{\mathring{\pi}}$ or $\mathring{\xi}_t^{\mathring{\pi}^{\natural}}$) under target evaluation policy $\mathring{\pi}$. This is the motivation of the following section.

2.2 Change of policy

Before diving into the mechanism of change of policy in causal dynamics, we recall some basic concepts of the change of measure, which is the cornerstone of the static balancing methods in general. Let $\mathcal{X}$ be a Polish space and let ξ and $\mathring{\xi}$ be two arbitrary finite measures in $\mathscr{M}(\mathcal{X})$. The fundamental idea of balancing is to find a weight function $\mathring{\eta}$ such that the re-weighted source measure $\Psi_{\mathring{\eta}}(\xi)$ is equal, or close, in some sense, to the target measure $\mathring{\xi}$. First, we recall that the re-weighting transformation is defined by

$$\Psi_{\cdot}(\xi) : L^1(\xi) \ni \eta \mapsto \Psi_\eta(\xi) = \xi(\eta \times \cdot) \in \mathscr{M}(\mathcal{X}).$$

To give more intuition, let us see a concrete example. Consider a standard Gaussian probability measure $\xi(\mathrm{d}x) = 1/\sqrt{2\pi}\exp(-x^2/2)\mathrm{d}x$. If the target measure $\mathring{\xi}$ is a centered Gaussian with variance $1/2$, it is sufficient to consider the weight function $\mathring{\eta}(x) = \sqrt{2}\exp(-x^2/2)$. Then, the weighted measure

$$\Psi_{\mathring{\eta}}(\xi)(\mathrm{d}x) = \sqrt{2}\exp(-x^2/2) \times 1/\sqrt{2\pi}\exp(-x^2/2)\mathrm{d}x = 1/\sqrt{\pi}\exp(-x^2)\mathrm{d}x$$

is exactly the target measure $\mathring{\xi}$. Theoretically speaking, assuming that $\mathring{\xi}$ is absolutely continuous w.r.t. ξ, it is legal to define the Radon-Nikodym derivative $\mathring{\eta}$ by

$$\mathring{\eta} = \frac{\mathrm{d}\mathring{\xi}}{\mathrm{d}\xi}.$$

Thanks to the Radon-Nikodym theorem, the function $\mathring{\eta}$ is uniquely determined in terms of equivalence class in $L^1(\xi)$. More practically, the actual source measure and the target measure considered in the present paper are some (potentially weighted) empirical measures associated to the observable data set, say ξ_n and $\mathring{\xi}_m$. In this case, there is in general no overlap between the source and target measures in terms of support. Hence, we cannot expect to perfectly transform the source measure to the target measure through re-weighting as illustrated in the ideal Gaussian case above. Instead, the balancing procedure is executed in order to minimize some difference between the re-weighted measure $\Psi_\eta(\xi_n)$ and the target measure $\mathring{\xi}_m$. For example, Reygner and Touboul (2020) studied the optimal transport costs, which recovers the well-known optimal matching method in the context of causal inference. In addition, since the support of ξ_n is discrete, we only need n values of the weight function η in order to transform the source measure ξ_n. Therefore, the functional optimization w.r.t. the weight function η can indeed be regarded as a finite-value optimization problem, which is usually tractable on the algorithmic level. With a slight abuse of language, we say that we compare the measures ξ_n and $\mathring{\xi}_m$ in order to get an estimation of the weight function $\mathring{\eta}$ on the support of ξ_n.

Returning to the causal dynamics, the core idea remains the same, except that we have now a sequence of measures, i.e., $\mathring{\xi}_t^{\mathring{\pi}}$ and/or $\mathring{\xi}_t^{\mathring{\pi}^\natural}$, to approximate. In fact, the data set $\mathcal{D}_n$ contains enough information to construct the empirical measures $\xi_{t,n}^\pi$ and $\xi_{t,n}^{\pi^\natural}$ w.r.t. ξ_t^π and $\xi_t^{\pi^\natural}$, respectively defined by

$$\xi_{t,n}^\pi = \frac{1}{n}\sum_{i=1}^n \delta_{X_t^{(i)}} \quad \text{and} \quad \xi_{t,n}^{\pi^\natural} = \frac{1}{n}\sum_{i=1}^n \delta_{X_t^{\natural(i)}} = \frac{1}{n}\sum_{i=1}^n \delta_{(X_t^{(i)}, A_t^{(i)})}. \tag{2}$$

Since the reward observations $\mathcal{R}_n$ are available on the support of each resource empirical terminal measure $\xi_{t,n}^\pi$, it is natural to consider estimating recursively the weight functions defined by

$$\mathring{\eta}_t = \frac{\mathrm{d}\mathring{\xi}_t^{\mathring{\pi}}}{\mathrm{d}\xi_t^\pi} \quad \text{and} \quad \mathring{\eta}_t^\natural = \frac{\mathrm{d}\mathring{\xi}_t^{\mathring{\pi}^\natural}}{\mathrm{d}\xi_t^{\pi^\natural}}.$$

In this way, the observations collected in $\mathcal{R}_n$ can be used to approximate $\mathcal{V}^{\mathring{\pi}}$ through re-weighting according to (1). This is exactly how the direct estimator is constructed. Therefore, following this idea, we now concentrate on the estimation of the weight functions $\mathring{\eta}_t$ and/or $\mathring{\eta}_t^\natural$ on the support of $\xi_{t,n}^{\pi}$ and/or $\xi_{t,n}^{\pi^\natural}$. Obviously, it is impossible to provide policy evaluation for an arbitrary new policy without limitation. For example, $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$ are not necessarily always well-defined. Hence, at least, we need the following assumption in order to guarantee the existence of the weight functions $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$.

(H1) *The target distribution $\mathring{\xi}_0$ is absolutely continuous w.r.t. the sampling distribution ξ_0, and the target evaluation policy $\mathring{\pi}$ is also absolutely continuous w.r.t. the sampling policy π, that is,*

$$\forall 0 \leq t \leq T,\ \forall x_t \in \mathcal{X}_t, \quad \pi_t(x_t, \cdot) \ll \mathring{\pi}_t(x_t, \cdot).$$

Next, to better understand the weight functions $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$, we define, thanks to **(H1)**,

$$\forall 1 \leq t \leq T,\ \forall x_t^\natural = (x_t, a_t) \in \mathcal{X}_t^\natural, \quad \mathring{e}_t^\natural(x_t^\natural) = \frac{\mathrm{d}\mathring{\pi}_t(x_t, \cdot)}{\mathrm{d}\pi_t(x_t, \cdot)}(a_t).$$

For $t = 0$, we let, taking into account the covariate shifts,

$$\forall x_0^\natural = (x_0, a_0) \in \mathcal{X}_0^\natural, \quad \mathring{e}_0^\natural(x_0^\natural) = \frac{\mathrm{d}\mathring{\xi}_0}{\mathrm{d}\xi_0}(x_0)\frac{\mathrm{d}\mathring{\pi}_0(x_0, \cdot)}{\mathrm{d}\pi_0(x_0, \cdot)}(a_0).$$

The following Proposition 1 gives the characterization of the weight functions $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$.

Proposition 1. *Assuming **(H1)**, the weight functions $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$ are well-defined respectively in $L^1(\mathcal{X}_t)$ and $L^1(\mathcal{X}_t^\natural)$. In addition, for any $0 \leq t \leq T$, we have*

$$\mathring{\eta}_t^\natural(\cdot) = \mathbb{E}\left[\prod_{s=0}^{t} \mathring{e}_s^\natural(X_s^\natural) \,\middle|\, X_t^\natural = \cdot\right] \quad \textit{and} \quad \mathring{\eta}_{t+1}(\cdot) = \mathbb{E}\left[\mathring{\eta}_t^\natural(X_t^\natural) \,\middle|\, X_{t+1} = \cdot\right].$$

An important observation from Proposition 1 is that $\mathring{\eta}_{t+1}$ is somehow a "smoothed version" of $\mathring{\eta}_t^\natural$ given an observable (through the data set $\mathcal{D}_n$) random variable X_{t+1}. This is in fact the very property that makes the recursive balancing strategy tractable. To present this idea in a more detailed manner, let us continue to see the measure flows in the change of policy.

Measure flows in change of policy By definition of the weight functions $\mathring{\eta}_t^\natural$ and $\mathring{\eta}_{t+1}$, the causal dynamics with the change of policy can be illustrated in the diagram given in Figure 3. Let us give a brief description on the tractability of the recursive balancing strategy. For

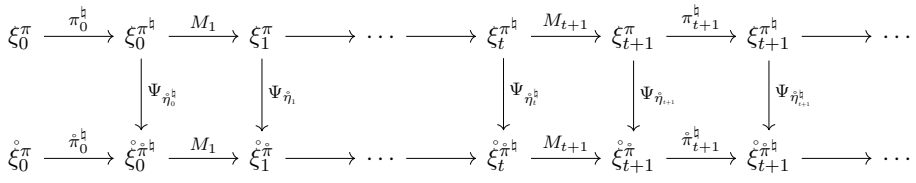


Figure 3: Measure flows in the change of policy

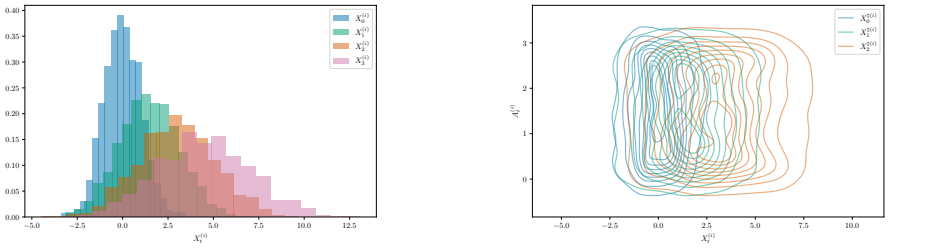
the initial part, we first notice that both $\xi_0^{\pi^\natural}$ and $\mathring{\xi}_0^{\mathring{\pi}}$ are sampled respectively in $\mathcal{D}_n$ and in $\mathcal{I}_m$. Moreover, as the target evaluation policy $\mathring{\pi}_0$ is known, it is then sufficient to compare the empirical versions of $\xi_0^{\pi^\natural}$ and $\xi_0^{\mathring{\pi}^\natural}$ in order to get access to the weight function $\mathring{\eta}_0^\natural$. Assume now

that we have access to the weight function $\mathring{\eta}_t^\natural$ at $t \geq 0$. Since X_{t+1} is sampled in the data set $\mathcal{D}_n$, then, according to Proposition 1, the weight function $\mathring{\eta}_{t+1}$ can be computed using some smoothing mechanism. Finally, according to the measure flows given in Figure 3, one can thus get access to the terminal measure $\xi_{t+1}^{\mathring{\pi}\natural}$ by the given policy $\mathring{\pi}_{t+1}^\natural$. In this way, one compares the observable measure $\xi_{t+1}^{\pi^\natural}$ (see (2)) with the empirical counterpart of $\mathring{\xi}_{t+1}^{\mathring{\pi}\natural}$ to get the weight function $\mathring{\eta}_{t+1}^\natural$. In the following paragraph, this conceptual idea is developed with more details in terms of implementation.

Recursive balancing strategy through a concrete example The motivation of this paragraph is to show how to approximate the weight functions $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$ at each data point $X_t^{(i)}$. To better illustrate the core mechanism, we provide a simplified conceptual example. Assuming the lack of covariate shifts, we let the initial distribution ξ_0^π to be a standard Gaussian on $\mathbb{R}$. At each time step, we consider an action that takes values in $[0, 3]$. The state spaces $\mathcal{X}_t$ are assumed to be homogeneous, i.e., $\mathcal{X}_t = \mathbb{R}$. The state-actions spaces at each time step t is characterized by $\mathbb{R} \times [0, 3]$. For the transition kernels M_t, we consider a Gaussian kernel $M_t(x_t^\natural, \mathrm{d}x_{t+1}) = 1/\sqrt{2\pi}\exp(-(x_{t+1} - x_t - a_t)^2/2)\mathrm{d}x_{t+1}$. To give more intuition, the action can be regarded as a force that increases the values of X_t, and the environment gives a standard Gaussian perturbation through Markov kernels M_t each time the dynamics interact with the environment, that is,

$$X_{t+1} = X_t + A_t + G_t, \quad \text{with } G_t \sim \mathcal{N}(0, 1) \text{ and } G_t \perp\!\!\!\perp (X_t, A_t).$$

For simplicity, we rule out the dependence between the policy and the state-action variable $X_t^\natural$. More precisely, the sampling policy π is modeled by a uniform distribution on $[0, 3]$ for each individual and for each iteration t, and the target evaluation policy $\mathring{\pi}$ is a uniform distribution on $[0, 1]$. We consider a data set of size $n = 1000$ with time horizon $T = 2$. For the data set, the state-action trajectories $(X_0^{(i)}, A_0^{(i)}, X_1^{(i)}, A_1^{(i)}, X_2^{(i)}, A_2^{(i)}, X_3^{(i)}; 1 \leq i \leq 1000)$ under sampling policy π are collected. Figure 4 illustrates the empirical terminal measures $\xi_{t,n}^\pi$ and $\xi_{t,n}^{\pi^\natural}$.



(a) Histogram of $\xi_{t,n}^\pi$

(b) Standard kernel density estimation illustration of $\xi_{t,n}^{\pi^\natural}$

Figure 4: Illustration of empirical terminal measures under sampling policy π

In the dynamical setting, the recursive strategy can be informally summarized by the following steps:

(i) **Initial balancing:** Consider the empirical measure $\mathring{\xi}_0^{\mathring{\pi}}$ of the target population given by

$$\mathring{\xi}_{0,m}^{\mathring{\pi}} = \frac{1}{m}\sum_{i=1}^{m} \delta_{\mathring{X}_0^{(i)}},$$

where $\mathring{X}_0^{(i)}$ is sampled in the data set $\mathcal{I}_m$. Fixing an arbitrary large number $N > 0$, and

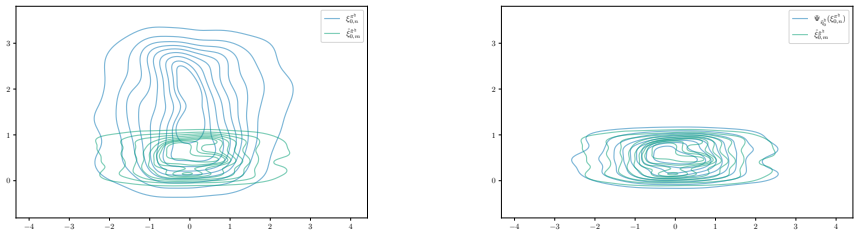
for each (j,ℓ) in $\{1,\dots,m\}\times\{1,\dots,N\}$, we let $\mathring{X}_0^{\natural(j,\ell)} = (\mathring{X}_0^{(j)}, \mathring{A}_0^{((j-1)N+\ell)})$ with

$$\forall 1 \le k \le mN, \quad \mathring{A}_0^{(k)} \sim \mathring{\pi}_0\left(\mathring{X}_0^{(\lfloor k/N \rfloor+1)}, \cdot\right).$$

Basically, for each data point $\mathring{X}_0^{(j)}$, we independently sample N copies of action variable $\mathring{A}_0$ given $\mathring{X}_0^{(j)}$. Then, we compare

$$\mathring{\xi}^{\mathring{\pi}^\natural}_{0,mN} = \frac{1}{mN}\sum_{j=1}^{m}\sum_{\ell=1}^{N}\delta_{\mathring{X}_0^{\natural(j,\ell)}}$$

to $\xi_{0,n}^{\pi^\natural}$ (cf. equation (2)) by balancing method, which outputs the values $\{\hat{\eta}_0^\natural(X_0^{\natural(i)}) : 1 \le i \le n\}$. After balancing, the measure $\Psi_{\hat{\eta}_0^\natural}(\xi_{0,n}^{\pi^\natural})$ becomes an approximation of the terminal measure $\mathring{\xi}_0^{\mathring{\pi}^\natural}$. Figure 5 with $N = 1500$ provides an illustration.



(a) Before initial balancing

(b) After initial balancing

Figure 5: Illustration of initial balancing

(ii) **Weight smoothing:** After getting $\{\hat{\eta}_t^\natural(X_t^{\natural(i)}) : 1 \le i \le n\}$ at each time step $t \ge 0$, we perform a transductive learning to get the estimation $\{\hat{\eta}_{t+1}(X_{t+1}^{(i)}) : 1 \le i \le n\}$. For example, this can be done through an ordinary regression of $\hat{\eta}_t^\natural(X_t^{\natural(i)})$ against $X_{t+1}^{(i)}$ according to Proposition 1. In practice, one considers a K-fold splitting of the data set in order to remove the dependence of the estimated weight function $\hat{\eta}_{t+1}^{(i)}$ and the evaluation data point $X_{t+1}^{(i)}$. More precisely, for each i, the weight function $\hat{\eta}_{t+1}^{(i)}$ is trained on the $K-1$ folds that excludes the data point $X_{t+1}^{(i)}$. In fact, it is also possible to omit this step, namely, let $\hat{\eta}_{t+1}^{(i)}(\cdot) = \mathrm{Id}_{\mathcal{X}_{t+1}}$ for each i, without influencing the convergence rate of the actual policy evaluation estimator. However, this leads to an additional variance term to the weight function estimation, which makes it impossible to obtain the semiparametric efficiency, as we will see in Section 2.3.

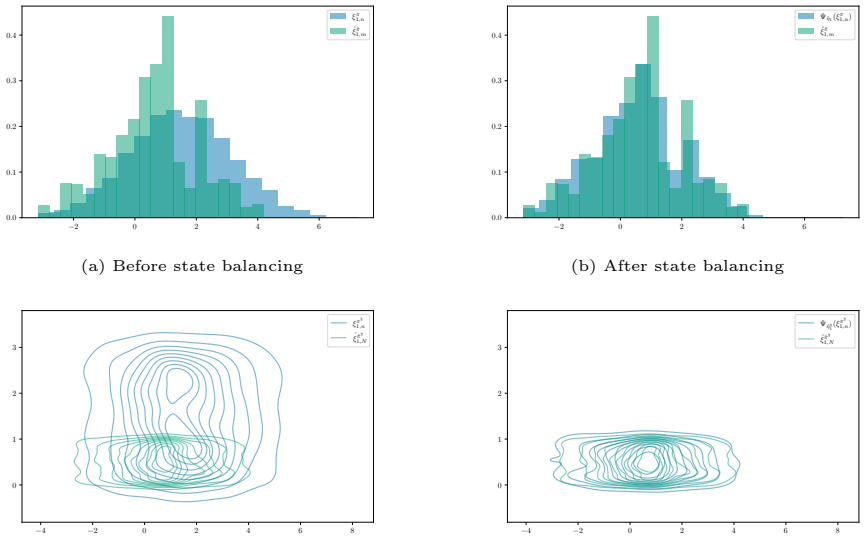
(iii) **Update balancing:** When $\{\hat{\eta}_{t+1}(X_{t+1}^{(i)}) : 1 \le i \le n\}$ is available at time $t \ge 0$, one considers the approximation $\Psi_{\hat{\eta}_{t+1}}(\xi_{t+1,n}^{\pi})$ of $\mathring{\xi}_{t+1}^{\mathring{\pi}}$. Then, similar to the initial step, we consider a bootstrap sampler of size N. For each (j,ℓ) in $\{1,\dots,n\}\times\{1,\dots,N\}$, we let $\mathring{X}_{t+1}^{\natural(j,\ell)} = (X_{t+1}^{(j)}, \mathring{A}_{t+1}^{((j-1)N+\ell)})$ with

$$\forall 1 \le k \le nN, \quad \mathring{A}_{t+1}^{(k)} \sim \mathring{\pi}_{t+1}\left(X_{t+1}^{(\lfloor k/N \rfloor+1)}, \cdot\right).$$

Then, as is conducted in the initial balancing step, we compare

$$\mathring{\xi}^{\mathring{\pi}\natural}_{t+1,nN} = \frac{1}{nN}\sum_{j=1}^{n}\sum_{\ell=1}^{N}\hat{\eta}^{(i)}_{t+1}(X^{(i)}_{t+1})\delta_{\mathring{X}^{\natural(j,\ell)}_{t+1}}$$

and $\xi^{\pi\natural}_{t+1,n}$ which updates the weight function estimation $\{\hat{\eta}^{\natural}_{t+1}(X^{\natural(i)}_{t+1}) \mid 1 \le i \le n\}$ at time $t+1$. Figure 6 provides an illustration of the update balancing at step $t = 1$.



(a) Before state balancing

(b) After state balancing

(c) Before state-action balancing

(d) After state-action balancing

Figure 6: Illustration of update balancing

Remark 2.2 (Alternative sampling of $\mathring{\xi}^{\mathring{\pi}\natural}_{t,N}$). The sampling methods described in step (i) and step (iii) may be used to any target policy $\mathring{\pi}$. On the other hand, when some additional assumption on the target policy $\mathring{\pi}$ is available, it is possible to avoid this bootstrap mechanism so that the terminal measure $\mathring{\xi}^{\mathring{\pi}\natural}_{t}$ can be approximated directly by $\mathring{\xi}^{\mathring{\pi}}_{t,n} \circ \mathring{\pi}^{\natural}_{t}$ without dealing with the choice of N and the associated additional randomness. There are two typical situations:

(i) When the policy $\mathring{\pi}_t$ is deterministic, namely, $\mathring{\pi}_t = \delta_{f_t(x_t)}$ with f_t a measurable function from $\mathcal{X}_t$ to $\mathcal{A}_t$. Here, we consider

$$\forall t \ge 1, \quad \mathring{\xi}^{\mathring{\pi}\natural}_{t,n} = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_t(X^{(i)}_t)\delta_{(X^{(i)}_t, f_t(X^{(i)}_t))},$$

as well as

$$\mathring{\xi}^{\mathring{\pi}\natural}_{0,m} = \frac{1}{m}\sum_{i=1}^{m}\delta_{(\mathring{X}^{(i)}_0, f_0(X^{(i)}_0))}.$$

(ii) When the action space $\mathcal{A}_t$ is of finite values, say, $\mathcal{A}_t = \{1, 2, \ldots, L\}$. In this case, we consider

$$\forall t \ge 1, \quad \mathring{\xi}^{\mathring{\pi}\natural}_{t,n} = \frac{1}{n}\sum_{i=1}^{n}\sum_{\ell=1}^{L}\hat{\eta}_t(X^{(i)}_t)\mathring{\pi}_t(X^{(i)}_t, \ell)\delta_{(X^{(i)}_t, \ell)},$$

as well as

$$\mathring{\xi}^{\mathring{\pi}^\natural}_{0,m} = \frac{1}{m}\sum_{i=1}^{m}\sum_{\ell=1}^{L}\mathring{\pi}_0(\mathring{X}_0^{(i)},\ell)\delta_{(\mathring{X}_t^{(i)},\ell)}.$$

□

2.3 Policy evaluation estimators

Following the discussion on the implementation of recursive balancing strategy, we focus now on the actual policy evaluation estimators. After we have obtained the estimation of weight function $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$, there are two typical ways to construct the policy evaluation estimators, namely, the direct estimator (DE) and the doubly robust estimator (DRE). The DE is constructed by the weighted average of collected reward observations $\mathcal{R}_n$, without involving extra regression step. Such an estimator, in particular their oracle counterparts (i.e., with true weight functions instead of their approximations), have already been intensely investigated in the literature (Thomas and Brunskill, 2016; Liu et al., 2018; Xie et al., 2019; Liu et al., 2020, etc). Obviously, the performance of DE is totally determined by the quality of weight function estimation. Unfortunately, the DE cannot achieve semiparametric efficiency in general according to Kallus and Uehara (2020) due to its relatively straightforward construction. On the other hand, the DRE requires external regressors (for estimating $r_t^\natural$ given in (1)). Therefore, the performance of DRE depends on both the quality of balancing *and* the reliability of the implemented regressors. In reward, the DRE is possible to achieve the semiparametric efficiency according to Kallus and Uehara (2020) under some convergence rate assumptions. Essentially, this is because the form of DRE is in fact deduced from the effective influence function according to semiparametric theory (see Jiang and Li, 2016; Kallus and Uehara, 2020 for details).

Direct estimator Following identity (1) and Proposition 1, we define the DE $\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}}$ by

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}} = \sum_{t=0}^{T}\Psi_{\hat{\eta}_{t+1}}(\xi^{\pi}_{t+1,n})(r_t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\hat{\eta}^{(i)}_{t+1}(X^{(i)}_{t+1})r_t(X^{(i)}_{t+1}).$$

According to the recursive balancing strategy, the measure $\Psi_{\hat{\eta}_{t+1}}(\xi^{\pi}_{t+1,n})$ is an approximation of the terminal measure $\mathring{\xi}^{\mathring{\pi}}_{t+1}$. Therefore, the construction of $\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}}$ is immediate by taking advantage of the alternative representation (1). Since the weight smoothing step requires additional regressors, we only consider $\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}}$ with the simplified mechanism where $\hat{\eta}^{(i)}_{t+1}(X^{(i)}_{t+1}) = \hat{\eta}^\natural_t(X^{\natural(i)}_t)$ if not mentioned otherwise. Thus, the performance of $\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}}$ is totally characterized by the quality of the balancing implemented at each time step. To study the behavior of $\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}}$, it is therefore inevitable to provide more details of the balancing mechanism instead of still treating it as a black box. The formal theoretical description will be provided later in Section 3. Under some regularity assumptions, the weight function estimation $\hat{\eta}^\natural_t$ is almost surely given by a minimax-type optimization of the form

$$(3)\quad \left(\hat{\eta}^\natural_t(X^{\natural(1)}_t),\ldots,\hat{\eta}^\natural_t(X^{\natural(n)}_t)\right) \in \underset{\left(\eta^\natural_t(X^{\natural(1)}_t),\ldots,\eta^\natural_t(X^{\natural(n)}_t)\right)\in\mathbb{S}_n\subset\mathbb{R}^n}{\arg\min}\ \sup_{\gamma_t\in\Gamma_t}\left|\Psi_{\eta^\natural_t}(\xi^{\pi^\natural}_{t,n})(\gamma_t) - \mathring{\xi}^{\mathring{\pi}^\natural}_{t,nN}(\gamma_t)\right|,$$

where $\mathbb{S}_n$ is some subset of $\mathbb{R}^n$ and Γ_t is a user-prefixed collection of test functions on $\mathcal{X}^\natural_t$. For example, when Γ_t is assumed to be the collection of L-Lipshitz functions (with $L < +\infty$), the objective of the minimization turns out to be the well-known 1-Wasserstein distance between probability measures thanks to Kantorovich-Rubinstein duality. In this case, the optimization (3) is explicitly solvable with an reverse nearest neighbor argument (see, e.g., Reygner and Touboul, 2020, Proposition 2.2). To make it clearer, the core idea of balancing in the DE context is the

worst-case-error minimization, i.e., a minimax game between $\eta_t^\natural$ and γ_t, where γ_t represents all the possible observations of the causal dynamics after time t. As a consequence, the choice of the test function collection Γ_t is an essential ingredient required to implement balancing. In this regard, we introduce the *well-specifiedness* of the causal dynamics, which is important to deduce the associated convergence results.

Definition 2 (Well-specifiedness of causal dynamics)**.** We say that the causal dynamics are well-specified by a sequence of collections of test functions $\Gamma = (\Gamma_t; 0 \leq t \leq T)$ if

$$\forall 0 \leq t \leq T, \quad r_t^\natural = M_{t+1}(r_t) \in \Gamma_t,$$

and

$$\forall 1 \leq t \leq T, \ \forall \gamma_t \in \Gamma_t, \quad M_t^{\mathring{\pi}^\natural}(\gamma_t) \in \Gamma_{t-1}.$$

□

Doubly robust estimator To simplify the presentation, we first assume that no covariate shifts are involved in the causal dynamics. In this case, according to the recent work of Kallus and Uehara (2020), the construction of the DRE can be understood as the natural estimator deduced by the efficient influence function according to semiparametric theory. Unlike the DE, we need to provide a nuisance estimator $\hat{r}_t^\natural$ of $r_t^\natural$ from the data set $\mathcal{D}_n$. In addition, data splitting is required in order to remove the dependence between the nuisance estimator $\hat{r}_t^\natural$ and the data points to be evaluated. Thus, we consider a K-fold splitting of the data set $\mathcal{D}_n$ with $K \geq 2$. The notation $\hat{r}_t^{\natural(i)}$ denotes the regressor of $r_t(X_{t+1}^{(i)})$ against $X_t^{\natural(i)}$, trained on the $K-1$ folds that excludes the one containing $X_t^{\natural(i)}$. With a slight abuse of notation, we may omit (i) and use directly the notation $\hat{r}_t^\natural$ in the following. Following Kallus and Uehara (2020), the DRE $\mathcal{V}_{\text{DRE}}^{\mathring{\pi}}$ is defined by

$$\hat{\mathcal{V}}_{\text{DRE}}^{\mathring{\pi}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{T}\left(\hat{\eta}_t^\natural(X_t^{\natural(i)})\left(r_t(X_{t+1}^{(i)}) - \hat{r}_t^{\natural(i)}(X_t^{\natural(i)})\right) + \hat{\eta}_{t-1}^\natural(X_{t-1}^{\natural(i)})\mathring{\pi}_t^\natural(\hat{r}_t^{\natural(i)})(X_t^{(i)})\right),$$

with the convention $\hat{\eta}_{-1}^\natural(X_{-1}^{\natural(i)})\mathring{\pi}_0^\natural(\hat{r}_0^{\natural(i)})(X_0^{(i)}) = \mathring{\pi}_0^\natural(\hat{r}_0^{\natural(i)})(\mathring{X}_0^{(i)})$, and we omit the difference between the sum over n and over m at the initial step for simplicity. To give more intuition, we observe that identity (1) can be rewritten as

$$\mathcal{V}^{\mathring{\pi}} = \sum_{t=0}^{T}\left(\mathring{\xi}_{t+1}^{\mathring{\pi}}(r_t) - \mathring{\xi}_t^{\mathring{\pi}^\natural}(r_t^\natural) + \mathring{\xi}_t^{\mathring{\pi}}(\mathring{\pi}^\natural(r_t^\natural))\right).$$

Similar to the construction of DE, we replace the unknown measures above by their approximations to get the form of DRE. This estimator is slightly more complicated than the DE. However, according to Kallus and Uehara (2020, Section 2), it benefits from the optimality in terms of asymptotic variance w.r.t. all the sub-parametric models when the optimal $\sqrt{n}$-rate is achieved. In order to get this favorable property, one needs to understand the balancing method from a different angle. In the construction of DE, the balancing can be regarded as a worst-case-error optimization, as shown in (3). However, for ensuring the convergence of DRE, we additionally need to provide the convergence of weight function estimation in an L^2 sense. Meanwhile, the balancing mechanism is almost unchanged and remains the same as in (3).

Unlike the worst-case-error interpretation discussed in the DE context, the L^2 perspective of balancing is not as straightforward. For example, it is not clear how the choice of Γ_t can influence the convergence of weight function estimation in a L^2 sense. To have a full understanding of the balancing optimization (3) from these two different angles, one needs to set up a solid theoretical

foundation for balancing method. This is the question we attack in the following section.

3 Balancing

Roughly speaking, balancing is a re-weighting technique that corrects the difference between the source measure and the target measure. Traditionally speaking, such a setting is usually referred to as density ratio estimation problem (see, e.g., Sugiyama et al., 2012), as the optimal weights are given by the ratio of the two density functions w.r.t. another reference measure (e.g., the Lebesgue measure). In this article, we approach this problem from two different perspectives. On the one hand, it is easy to see that the associated density ratio is but the Radon-Nikodym derivative of the target measure w.r.t. the source measure. In our context, we slightly limit our discussion to the case where the considered Radon-Nikodym derivative is square integrable w.r.t. the source measure. Thus, the identification of the optimal re-weighting function falls into the context of Riesz representers (see, e.g., Chernozhukov et al., 2020, 2018) based on the Hilbert structure of the associated L^2 spaces. Hence, as shown by Chernozhukov et al. (2020), a natural way of estimating the Riesz representers is to minimize the associated L^2-error. A universal balancing strategy based on an adversarial optimization strategy can thus be implemented. On the other hand, balancing can naturally be interpreted as a worst-case error minimization, as discussed in Section 2.3 in the context of the DE. This is also the motivation of general optimal matching method introduced by Kallus (2020) in the static setting. The aim of this section is to show that these two different interpretations can indeed be analyzed in a unified manner, through a *Riesz representable measure space* argument. This serves as the theoretical foundation of the balancing method presented in Section 3.2 and 3.3.

3.1 Riesz representable measure space

Let $\mathcal{X}$ be a Polish space. Denote by $\xi \in \mathscr{M}_+(\mathcal{X})$ and $\mathring{\xi} \in \mathscr{M}_+(\mathcal{X})$ respectively the source and the target measure, and keep in mind that the goal of balancing is to estimate the value $\mathring{\xi}(r^\natural)$ for some observation function $r^\natural$. In a nutshell, the Riesz representable measure space can be regarded as a weighted measure space w.r.t. the source measure, serving as the collection of the candidates of re-weighted measures, which is important when considering balancing optimization in an oracle sense. Apart from the absolute continuity similar to **(H1)**, we need an additional L^2-property.

Definition 3 (Riesz representable)**.** For two positive finite measures ξ and $\mathring{\xi}$ on a Polish space $\mathcal{X}$, we say that $\mathring{\xi}$ is Riesz representable by ξ if $\mathring{\xi}$ is absolutely continuous w.r.t. ξ and the associated Radon-Nikodym derivative $\mathring{\eta} = \mathrm{d}\mathring{\xi}/\mathrm{d}\xi$ is in $L^2(\xi)$.

Assuming that $\mathring{\xi}$ is Riesz representable by ξ, it is easy to verify that the linear functional $\mathring{\Gamma}$ defined by

$$\mathring{\Gamma} : L^2(\xi) \ni f \mapsto \mathring{\xi}(f) \in \mathbb{R}$$

is bounded. Therefore, thanks to Riesz representation theorem on the Hilbert space $L^2(\xi)$, one has

$$\forall f \in L^2(\xi), \quad \mathring{\Gamma}(f) = \Psi_{\mathring{\eta}}(\xi)(f).$$

Recall that $\Psi_{\mathring{\eta}}(\xi)(f)$ is given by $\xi(\mathring{\eta} \times f)$. With this in mind, let us define the *Riesz representable measure space* w.r.t. the source measure ξ by

$$\Xi(\xi) := \left\{ \Psi_\eta(\xi) : \eta \in L^2(\xi) \right\}.$$

It turns out that $\Xi(\xi)$ is endowed with a natural pseudo-metric structure. The Integral Probability Metric (IPM, see, e.g., Rachev, 1991; Müller, 1997) is a classical way to assess the difference

between two (probability) measures. The IPM is constructed by measuring the worst-case-error between two measures tested on a given family of test functions. Given a collection of test function $\Gamma \subset L^1(\xi)$, the IPM between two finite positive measures ζ and ζ' is defined by

$$\mathrm{IPM}_\Gamma(\zeta, \zeta') := \sup_{f \in \Gamma} |\zeta(f) - \zeta'(f)| \,. \tag{4}$$

It is easily checked that IPM_Γ is a pseudometric on $\mathscr{M}_+(\mathcal{X})$ and it becomes a metric when $\mathscr{F}$ is "rich" enough so that $\mathrm{IPM}_\Gamma(\xi, \xi') = 0$ is capable to ensure that $\xi = \xi'$. In general, it is not easy to deal with the IPM defined between finite positive measures, since it tends to be infinite as soon as the measures do not have the same mass. However, due to the Hilbert structure encoded in $\Xi(\xi)$, the topology induced by some specific IPM can be much simpler than one may think. Denote by $\mathscr{U}(L^2(\xi))$ the unit ball in $L^2(\xi)$. We have that $(\Xi(\xi), \mathrm{IPM}_{\mathscr{U}(L^2(\xi))}(\cdot,\cdot))$ is a metric space since all the indicator functions are contained in $\mathscr{U}(L^2(\xi))$. Meanwhile, it is easily verified that

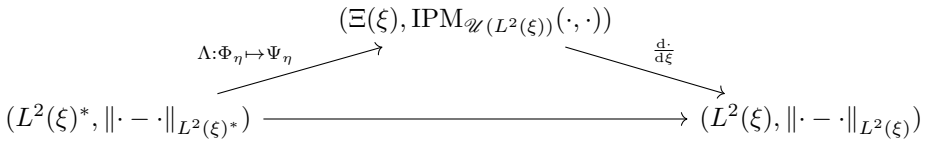
$$\forall \zeta, \zeta' \in \Xi(\xi), \quad \mathrm{IPM}_{\mathscr{U}(L^2(\xi))}(\zeta, \zeta') \leq \left\| \mathfrak{R}_{\Xi(\xi)}(\zeta) \right\|_{L^2(\xi)} + \left\| \mathfrak{R}_{\Xi(\xi)}(\zeta') \right\|_{L^2(\xi)} ,$$

where $\mathfrak{R}_{\Xi(\xi)}(\zeta)$ denotes the Riesz representer of ζ in $\Xi(\xi)$. Moreover, the following Proposition 2 gives more intuition of the structure the Riesz representable measure space equipped with unit ball IPM.

Proposition 2. *Let ξ be a positive finite measure on $\mathcal{X}$. We have the following isometric isomorphism between the metric spaces $L^2(\xi)$ and $\Xi(\xi)$:*

$$(L^2(\xi), \|\cdot - \cdot\|_{L^2(\xi)}) \underset{\frac{\mathrm{d}\cdot}{\mathrm{d}\xi}}{\overset{\Psi_\cdot(\xi)}{\rightleftarrows}} (\Xi(\xi), \mathrm{IPM}_{\mathscr{U}(L^2(\xi))}(\cdot,\cdot))$$

Remark 3.1. In fact, a more complete commutative diagram can be proposed as:



In the diagram above, $(L^2(\xi)^*, \|\cdot - \cdot\|_{L^2(\xi)^*})$ denotes the dual Hilbert space of $(L^2(\xi), \|\cdot - \cdot\|_{L^2(\xi)})$ equipped with operator norm. The functional Φ_η is given by $\Phi_\eta : L^2(\xi) \ni f \mapsto \xi(\eta \times f)$. The Riesz representable measure space equipped with unit ball IPM is but an explicit measure representation of the associated dual Hilbert space. This illustrates the idea that the motivation behind the L^2-based Riesz representation learning (Chernozhukov et al., 2020) and the IPM-based balancing that involves worst-case-error argument (Kallus, 2020) are, interestingly, equivalent in this ideal case. □

Proposition 2 seems to provide an equivalence between the worst-case-error and L^2 interpretations of balancing. However, the unit ball of an L^2 space is far too rich in a general setting. In practice, such IPM can only be computed in a tabular/parametric case, where all the state spaces and transitions considered in the causal dynamics are of finite values. Before answering the question on how the choice of IPM interact with L^2 aspect of balancing in a more practical situation, we continue to investigate the static balancing problem in the empirical setting. To move forward, the next subsection introduces the static balancing problem through a Riesz representable measure space point of view. The goal is to obtain a full understanding on the worst-case-error perspectives, as well as the L^2 aspects of the balancing method.

3.2 Static balancing

In order to better understand the role of balancing in our framework, we focus in this subsection in the static setting. We note that this can be regarded as a reformulation of the generalized (with general state and/or action spaces, as well as covariate shifts) Potential Outcomes framework under unconfoundedness and consistency assumptions (Rubin, 1974; Imbens and Rubin, 2015). Before starting, we remark that our reformulation can also be treated as a dynamical model with $T = 1$. Therefore, the extension to the dynamical setting, presented in Section 3.3, will be very natural.

Since there is not dynamics w.r.t. time in the static setting, the notation is slightly simplified compared to Section 2. Let X and $\mathring{X}$ be respectively the source and target state random variable on the state space $\mathcal{X}$, with associated probability measures denoted by ξ and $\mathring{\xi}$. Let π be the sampling policy and let $\mathring{\pi}$ be the target policy. They are both Markov kernels from the state space $\mathcal{X}$ to the action space $\mathcal{A}$. Similar to the dynamical setting, we denote respectively

$$\pi^\natural = \mathrm{id}_\mathcal{X} \times \pi \quad \text{and} \quad \mathring{\pi}^\natural = \mathrm{id}_\mathcal{X} \times \mathring{\pi}$$

the policy transition from state space $\mathcal{X}$ to the state-action space $\mathcal{X}^\natural$. The state-action measures under π and $\mathring{\pi}$ are respectively defined by

$$\xi^{\pi^\natural} = \xi \circ \pi^\natural \quad \text{and} \quad \mathring{\xi}^{\mathring{\pi}^\natural} = \mathring{\xi} \circ \mathring{\pi}^\natural.$$

We assume that $\mathring{\xi}^{\mathring{\pi}^\natural}$ is Riesz representable by $\xi^{\pi^\natural}$ (see Defintion 3), and we denote by $\mathring{\eta}^\natural = \mathrm{d}\mathring{\xi}^{\mathring{\pi}^\natural}/\mathrm{d}\xi^{\pi^\natural}$ the associated weight function. The interaction with the environment is modeled by another Markov kernel M from the state-action space $\mathcal{X}^\natural = \mathcal{X} \times \mathcal{A}$ to an underlying state space $\mathcal{Z}$, as presented in Section 2. Remark that M remains the same when the policy is changed. Finally, the system is observed by a measurable function $r : \mathcal{Z} \mapsto \mathbb{R}$. We let $r^\natural$ be the conditional expectation observation function w.r.t. M, i.e., $r^\natural = M(r)$. The causal Markov structure is illustrated in Figure 7, which is but a static version (with $T = 1$) of Figure 1.

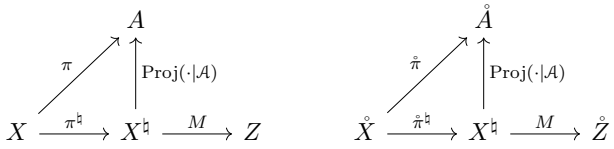


Figure 7: Markov structure of static causal model

Assume that the i.i.d. samples $\{X^{(i)} : 1 \le i \le n\}$ and $\{\mathring{X}^{(j)} : 1 \le j \le m\}$ of the state variables are collected, as well as the associated actions under the sampling policy $\{A^{(i)} : 1 \le i \le n\}$. Remark that $X^{(i)}$ and $\mathring{X}^{(j)}$ are not necessarily independent. For example, in the situation where covariate shifts are ruled out, we may consider $m = n$ and $\mathring{X}^{(i)} = X^{(i)}$. Denote by $X^{\natural(i)} = (X^{(i)}, A^{(i)})$ for $1 \le i \le n$ the state-action samples. It is supposed that we have also access to the observation values $\{r(Z^{(i)}) : 1 \le i \le n\}$. Our goal is to estimate the average effect defined by $\mathcal{V}^{\mathring{\pi}} = \mathring{\xi}^{\mathring{\pi}^\natural} M(r) = \mathring{\xi}^{\mathring{\pi}^\natural}(r^\natural)$.

Let us now consider the measures to be compared through static balancing. The state-action measure $\xi^{\pi^\natural}$ can indeed be sampled in the data set by

$$\xi_n^{\pi^\natural} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^{\natural(i)}}.$$

Since the target policy $\mathring{\pi}$ is known, one is able to sample the empirical version of the state-action

measure under target policy $\mathring{\xi}^{\mathring{\pi}^\natural}$ by

$$\mathring{\xi}^{\mathring{\pi}^\natural}_{mN} = \frac{1}{mN}\sum_{j=1}^{m}\sum_{\ell=1}^{N}\delta_{\mathring{X}^{\natural(j,\ell)}},$$

where for each (j,ℓ) in $\{1,\dots,m\}\times\{1,\dots,N\}$, $\mathring{X}^{\natural(j,\ell)} = (\mathring{X}^{(j)}, \mathring{A}^{((j-1)N+\ell)})$ with

$$\forall 1 \le k \le mN, \quad \mathring{A}^{(k)} \sim \mathring{\pi}\left(\mathring{X}^{(\lfloor k/N \rfloor+1)}, \cdot\right).$$

Basically, for each data point $\mathring{X}^{(j)}$, we independently sample N copies of action variable $\mathring{A}$ given $\mathring{X}^{(j)}$. As discussed in Remark 2.2, when addition information of the target policy is available, the construction of $\mathring{\xi}^{\mathring{\pi}^\natural}_{mN}$ (i.e., the choice of N) can be simplified.

The balancing method outputs the values of the weight function $\{\hat{\eta}^\natural(X^{\natural(i)}) : 1 \le i \le n\}$ on the support of $\xi_n^{\pi^\natural}$. In a formal way, the static balancing method can be summarized as the optimization

$$\hat{H} = \underset{\eta^\natural \in H}{\arg\min}\ \mathrm{IPM}_\Gamma\left(\Psi_{\eta^\natural}(\xi_n^{\pi^\natural}), \mathring{\xi}^{\mathring{\pi}^\natural}_{mN}\right), \tag{5}$$

where H is some bounded subset (in terms of L^2 norm) of $L^2(\xi^{\pi^\natural})$ that contains the candidates of weight functions and $\hat{H} \subset L^2(\xi)$ denotes the solution(s) of the optimization problem given above. Since the balancing (5) is implemented between two empirical measures, one notices that for each $\hat{\eta}^\natural \in \hat{H}$, any $\hat{\eta}^{\natural'} \in H$ that coincides with $\hat{\eta}^\natural$ on the support of $\xi_n^{\pi^\natural}$ can also be identified in $\hat{H}$. Said differently, the optimization problem in (5) is equivalent to solving a finite dimensional optimization

$$\underset{\left(\eta^\natural(X^{\natural(1)}),\dots,\eta^\natural(X^{\natural(n)})\right)\in\mathbb{S}_n\subset\mathbb{R}^n}{\arg\min}\ \mathrm{IPM}_\Gamma\left(\Psi_{\eta^\natural}(\xi_n^{\pi^\natural}), \mathring{\xi}^{\mathring{\pi}^\natural}_{mN}\right), \tag{6}$$

where $\mathbb{S}_n$ is some subset of $\mathbb{R}^n$ that is characterized by the choice of H. Once we have obtained the estimated values $\{\hat{\eta}^\natural(X^{\natural(i)}) : 1 \le i \le n\}$, there are two typical ways to construct the actual policy evaluation estimators. As discussed in Section 2.3, we may first consider the DE $\hat{\mathcal{V}}^{\mathring{\pi}}_{\mathrm{DE}}$, defined by

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\mathrm{DE}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}^\natural(X^{\natural(i)})r(Z^{(i)}).$$

We note that the DE given above is identical to $\hat{\mathcal{V}}^{\mathring{\pi}}_{\mathrm{DE}}$ introduced in Section 2.3, with $T = 1$ and some minor differences in terms of notation. To better understand the worst-case-error estimation essential of balancing method, we need to introduce an instrumental estimator given by

$$\Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})(r^\natural) = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}^\natural(X^{\natural(i)})r^\natural(X^{\natural(i)}).$$

Note that such an estimator is not available from the data set since no observation of the conditional expectation function $r^\natural$ is collected. On the one hand, the error of the DE contains a sampling error w.r.t. the observation of r, which is given by the following Proposition 3.

Proposition 3. *Assume that* $\xi^{\pi^\natural}M(r^2) < +\infty$. *We have*

$$\mathbb{E}\left[\left|\hat{\mathcal{V}}^{\mathring{\pi}}_{\mathrm{DE}} - \Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})(r^\natural)\right|^2\right] \le \frac{1}{n}\sup_{\eta^\natural\in H}\left\|\eta^\natural\right\|^2_{L^2(\xi^{\pi^\natural})}\cdot\left(\xi^{\pi^\natural}M(r^2) - (\xi^{\pi^\natural}M(r))^2\right).$$

Proposition 3 shows that $\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}}$ and $\Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})(r^\natural)$ is very close (up to a $\mathcal{O}_{\mathbb{P}}(1/\sqrt{n})$ stochastic error) under mild assumptions. On the other hand, the following Proposition 4 shows that the quality of the instrumental estimator $\Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})(r^\natural)$ is totally determined by some properties of the chosen IPM.

Proposition 4. *Assume that $\mathring{\xi}^{\mathring{\pi}^\natural}$ is Riesz representable by $\xi^{\pi^\natural}$ and that $r^\natural \in \Gamma$. Then we have, almost surely,*

$$\left|\Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})(r^\natural) - \mathring{\xi}^{\mathring{\pi}^\natural}(r^\natural)\right| \leq \text{IPM}_\Gamma\left(\Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural}), \Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})\right) + 2\,\text{IPM}_\Gamma\left(\mathring{\xi}^{\mathring{\pi}^\natural}, \mathring{\xi}^{\mathring{\pi}^\natural}_{mN}\right)$$

The right-hand-side of the inequality of Proposition 4 can be regarded as the *sample complexity* of the chosen IPM, which is determined by the richness of the family of test functions Γ. With standard probability calcualtions, we can show that this rate cannot be better than $\mathcal{O}_{\mathbb{P}}(\sqrt{\frac{n+mN}{nmN}})$ due to the worst-case-error construction of IPM. It turns that for some typical IPMs, which are well-studied in the literature, the optimization given in (5) is tractable. For example, when the IPM is assumed to be the 1-Wasserstein distance (Reygner and Touboul, 2020) or the Mean Maximum Dispersion (MMD, Gretton et al., 2007), the associated test function collections Γ are respectively the 1-Lipschitz functions and the unit ball of some RKHS. In particular, the optimal $\mathcal{O}_{\mathbb{P}}(\sqrt{\frac{n+mN}{nmN}})$-rate can be achieved when Γ is, e.g., a unit ball of some RKHS (Gretton et al., 2007) in the MMD context.

Remark 3.2. It is very important to remark that the worst-case-error interpretation of balancing method does not require specific regularity assumptions on the weight function $\eta^\natural$, except for the L^2 boundedness as is illustrated in Proposition 3. Meanwhile, we *only* need to ensure that the target evaluation function $r^\natural$ is included in the test function family Γ, so that the associated IPM is exactly an upper bound of the associated error. Thanks to the Lagrange multiplier method, for the optimization given in (6), one may consider the unconstrained optimization in $\mathbb{R}_n$ with a L^2-type penalty, i.e.,

$$\underset{\left(\eta^\natural(X^{\natural(1)}),\ldots,\eta^\natural(X^{\natural(n)})\right)\in\mathbb{R}^n}{\arg\min} \left(\text{IPM}_\Gamma\left(\Psi_{\eta^\natural}(\xi_n^{\pi^\natural}), \mathring{\xi}^{\mathring{\pi}^\natural}_{mN}\right)^2 + \frac{\lambda_w}{n}\sum_{i=1}^n \eta^\natural(X^{\natural(i)})^2\right). \tag{7}$$

Interestingly, the L^2 penalty term can be interpreted in two different ways. On the one hand, it is needed to ensure that the weight function is bounded in an L^2 sense as the upper bound given in Proposition 3 involves the L^2-norm of the candidates of the weight functions. On the other hand, λ_w can be understood as a uniform upper bound of the "variance" term, whose oracle version is also given in Proposition 3. In fact, one can show that the regularized objective function above can indeed be regarded as an upper bound of the conditional quadratic error—with bias and variance decomposition—in the same spirit of Kallus (2020).

□

Following the logic of Section 2.3, we may also investigate the DRE in the static setting, defined by

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}} = \frac{1}{n}\sum_{i=1}^n \hat{\eta}^\natural(X^{\natural(i)})(r(Z^{(i)}) - \hat{r}^{\natural(i)}(X^{\natural(i)})) + \frac{1}{m}\sum_{j=1}^m \mathring{\pi}^\natural(\hat{r}^{\natural(j)})(\mathring{X}^{(j)}),$$

where we consider a K-folds splitting such that $\hat{r}^{\natural(i)}$ is trained with the $K-1$ folds of data that excludes $X^{(i)}$. Remark that the construction of the DRE requires that $\mathring{\pi}^\natural(\hat{r}^{\natural(j)})$ can be evaluated at each data point $\mathring{X}^{(j)}$, which corresponds to the ideal case where the number of the sampled actions N is infinity. In practice, such a property is available with finite numbers of sampled actions when some additional information of $\mathring{\pi}$ is provided, as discussed in Remark 2.2. Our

objective is to have a similar results as shown in Proposition 3 and Proposition 4. First, we define the oracle version of the DRE by

$$\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{ODRE}} = \frac{1}{n}\sum_{i=1}^{n} \mathring{\eta}^{\natural}(X^{\natural(i)})(r(Z^{(i)}) - r^{\natural}(X^{\natural(i)})) + \frac{1}{m}\sum_{j=1}^{m} \mathring{\pi}^{\natural}(r^{\natural})(\mathring{X}^{(j)}),$$

where the nuisance estimators $(\hat{\eta}^{\natural}, \hat{r}^{\natural})$ are replaced by the corresponding $(\mathring{\eta}^{\natural}, r^{\natural})$. The Proposition 5 below shows that the oracle DRE is close to the DRE in an L^2 sense.

Proposition 5. *When $(\hat{\eta}^{\natural}, \hat{r}^{\natural})$ are both consistent estimators of $(\mathring{\eta}^{\natural}, r^{\natural})$, we have*

$$\left|\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}} - \hat{\mathcal{V}}^{\mathring{\pi}}_{\text{ODRE}}\right| \leq \left\|\hat{\eta}^{\natural} - \mathring{\eta}^{\natural}\right\|_{L^2(\xi_n^{\pi^{\natural}})} \left\|\hat{r}^{\natural} - r^{\natural}\right\|_{L^2(\xi_n^{\pi^{\natural}})} + o_{\mathbb{P}}\left(\sqrt{\frac{n+m}{nm}}\right).$$

Next, the following Proposition 6 shows that the oracle DRE always converges to $\mathring{\xi}^{\mathring{\pi}^{\natural}}(r^{\natural})$ at the optimal rate $\mathcal{O}_{\mathbb{P}}\left(\sqrt{\frac{n+m}{nm}}\right)$.

Proposition 6. *We have*

$$\mathbb{E}\left[\left|\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{ODRE}} - \mathring{\xi}^{\mathring{\pi}^{\natural}}(r^{\natural})\right|^2\right] = \sigma^2_{\text{ODRE}} \leq \frac{n+m}{nm} C,$$

where $C > 0$ is a constant that does not involve n and m.

According to Kallus and Uehara (2020), the error term σ^2_{ODRE} is optimal in terms of semi-parametric when no covariate shifts are involved. Obviously, σ^2_{ODRE} only depends on the double roubust structure of the DRE and has nothing to do with the implemented balancing methods. At the same time, Proposition 5 shows that the role of balancing is to provide a L^2-type convergence of the weight function estimation in the DRE context, so that the asymptotic behaviors of the DRE is similar to the oracle counterpart. To achieve this convergence, let us provide more details on how the L^2 distance between Riesz representers can be recovered when Γ is not the unit ball of L^2 space as required in Proposition 2.

A sub-optimal solution is to consider a "dual" version (in a Fenchel sense) of the adversarial balancing method introduced in Chernozhukov et al. (2020), which connects naturally to our IPM/Riesz representable measure space argument. We denote by ∂H the difference collection defined by $\partial H = \{\eta - \eta' : \eta, \eta' \in H\}$, as well as αH the star hull of H given by $\alpha H = \{\rho\eta : \rho \in [0, \alpha],\ \eta \in H\}$. Returning to the balancing optimization (5), we have the following inequality.

Lemma 1. *If $\mathring{\eta}^{\natural} \in H$ and there exists $\alpha > 0$ such that $\alpha(\partial H) \subset \Gamma$, then we have almost surely*

$$\forall \hat{\eta}^{\natural} \in \hat{H}, \quad \left\|\hat{\eta}^{\natural} - \mathring{\eta}^{\natural}\right\|^2_{L^2(\xi_n^{\pi^{\natural}})} \leq \frac{2}{\alpha}\left(\text{IPM}_{\Gamma}\left(\Psi_{\mathring{\eta}^{\natural}}(\xi^{\pi^{\natural}}), \Psi_{\mathring{\eta}^{\natural}}(\xi_n^{\pi^{\natural}})\right) + \text{IPM}_{\Gamma}\left(\mathring{\xi}^{\mathring{\pi}^{\natural}}, \mathring{\xi}^{\mathring{\pi}^{\natural}}_{mN}\right)\right).$$

Lemma 1 recovers L^2 control by the associated sampling complexity of IPM. However, unlike the case for the IPM induced by the L^2 unit ball, such construction only recovers "half" of the convergence rate of the sampling complexity. For example, assume that $n = m$ for simplicity. If the sampling complexity of the associated IPM admits a $\sqrt{n}$-rate, the associated L^2 error of Riesz representer estimation thus converges with a $n^{1/4}$-rate. The isometric isomorphism given in Proposition 2 is therefore not available anymore.

A natural question is therefore to ask if it is possible to remove the square at the left hand side of Lemma 1 and fully recovers the L^2 convergence rate in terms of sampling complexity of IPM. The answer is yes, with a simple modification of the original idea of Chernozhukov et al. (2020). Denote by U the unit ball $\mathscr{U}(L^2(\xi^{\pi^{\natural}}))$ of $L^2(\xi^{\pi^{\natural}})$.

Lemma 2. *If $\mathring{\eta}^\natural \in H$ and there exists $\alpha > 0$ and $\beta > 0$ such that $\alpha(\partial H) \cap \beta U \subset \Gamma$, then we have almost surely*

$$\forall \hat{\eta}^\natural \in \hat{H}, \quad \left\|\hat{\eta}^\natural - \mathring{\eta}^\natural\right\|_{L^2(\xi_n^{\pi^\natural})} \leq \frac{2}{\min(\alpha,\beta)} \left(\mathrm{IPM}_\Gamma\left(\Psi_{\hat{\eta}^\natural}(\xi^{\pi^\natural}), \Psi_{\hat{\eta}^\natural}(\xi_n^{\pi^\natural})\right) + \mathrm{IPM}_\Gamma\left(\mathring{\xi}^{\mathring{\pi}^\natural}, \mathring{\xi}^{\mathring{\pi}^\natural}_{mN}\right)\right).$$

Intuitively speaking, the large values of α and/or β require richer structure of the test function family Γ, which naturally gives smaller constant w.r.t. IPM upper bound. The crucial idea of the improvement of Lemma 2 is an additional L^2 regularization w.r.t. βU. In practice, one simply considers an additional regularization term similar to (7) in the DE context.

Remark 3.3. To achieve the L^2 control as is given in Lemma 2, the actual optimization to be conducted is indeed different to the DE case. As stated in Lemma 1 and Lemma 2, one needs to optimize in a prefixed functional space H, rather than simply in a L^2 ball as is in the DE setting. Therefore, it is not obvious to directly optimize the values of weight function on the support of the source measure. Moreover, another difference of the worst-case-error interpretation is that one does not need any assumptions on the evaluation function $r^\natural$. On the contrary, one is required to ensure that the real weight function $\mathring{\eta}^\natural$ is included in the prefixed candidate functional space H. □

After we have seen the core of the balancing mechanism in the static setting, it is time to return to the dynamical case.

3.3 Dynamical balancing

In the dynamical setting, the key is to focus on how the errors of the actual estimators accumulate w.r.t. time t. First, we need the following assumption to ensure that the target terminal measures $\mathring{\xi}_t^{\mathring{\pi}^\natural}$ is Riesz representable by the observable measures $\xi_t^{\pi^\natural}$ under sampling policy.

(H2) *For each time horizon t, we have respectively $\mathring{\eta}_t \in L^2(\xi_t^\pi)$ and $\mathring{\eta}_t^\natural \in L^2(\xi_t^{\pi^\natural})$.*

In fact, **(H2)** is essentially an assumption on the weight function $\mathring{e}_t^\natural$ (as well as $\mathring{\xi}_0$ when covariate shifts are involved) by considering the characterization of $\mathring{\eta}_t$ and $\mathring{\eta}_t^\natural$ given in Proposition 1. A standard sufficient condition is to assume that $\mathring{e}_t^\natural$ is uniformly bounded, as is required in many applications in the dynamical causal inference such as Kallus and Uehara (2020) and Nie et al. (2020). Nevertheless, **(H2)** requires that the target evaluation policy is "close", in some L^2 sense, to the sampling policy.

Following the brief introduction of recursive balancing strategy in Section 2.2, we now give the formal description of the balancing optimization conducted at each iteration. Denote by H_t the underlying candidates collection of weight functions at time t, and consider

$$\hat{H}_t = \operatorname*{arg\,min}_{\eta_t^\natural \in H_t} \mathrm{IPM}_{\Gamma_t}\left(\Psi_{\eta_t^\natural}(\xi_{t,n}^{\pi^\natural}), \mathring{\xi}_{t,nN}^{\mathring{\pi}^\natural}\right),$$

where $\hat{H}_t$ is the solution(s) of the optimization and we have omitted the difference of n and m at $t = 0$. To better present the error decompositions, let us define the error terms w.r.t. IPM sampling complexity as well as weight smoothing mechanism.

$$\forall t \geq 1, \quad \sigma_t^{\mathrm{IPM}}(n) = \mathrm{IPM}_{\Gamma_t}\left(\Psi_{\mathring{\eta}_t^\natural}(\xi_{t,n}^{\pi^\natural}), \Psi_{\mathring{\eta}_t^\natural}(\xi_t^{\pi^\natural})\right).$$

With a slight abuse of notation, we omit m and N at time 0, i.e.,

$$\sigma_0^{\mathrm{IPM}}(n) = \sigma_0^{\mathrm{IPM}}(n,m,N) = \mathrm{IPM}_{\Gamma_0}\left(\Psi_{\mathring{\eta}_0^\natural}(\xi_{0,n}^{\pi^\natural}), \Psi_{\mathring{\eta}_0^\natural}(\xi_0^{\pi^\natural})\right) + \mathrm{IPM}_{\Gamma_0}\left(\mathring{\xi}_0^{\mathring{\pi}^\natural}, \mathring{\xi}_{0,mN}^{\mathring{\pi}^\natural}\right).$$

For the error term given by weight smoothing, we denote

$$\sigma_t^{\text{WS}}(n) = \|\hat{\eta}_{t+1} - \mathring{\eta}_{t+1}\|_{L^2(\xi^{\pi}_{t+1,n})}.$$

Note that when no weight smoothing is implemented, we have

$$\sigma_t^{\text{WS}}(n) \leq \frac{C}{\sqrt{n}}.$$

We first give the formal convergence result of the DE in the dynamical setting.

Theorem 1. *Assume **(H2)**. If the causal dynamics are well-specified by* $\Gamma = (\Gamma_t; 0 \leq t \leq T)$*, then we have*

$$\left|\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DE}} - \mathcal{V}^{\mathring{\pi}}\right| \leq C\left(\sum_{t=0}^{T}(T-t+1)\left(\sigma_t^{\text{IPM}}(n) + \frac{1}{\sqrt{N}}\right)\right),$$

where $C > 0$ *is a constant that is independent to* n, m, N *and* T*.*

For the convergence results of the DRE, we denote the error of the additional regression of the average reward function $r^{\natural}$ by

$$\forall 0 \leq t \leq T, \quad \sigma_t^{\text{REG}}(n) = \left\|\hat{r}_t^{\natural} - r_t^{\natural}\right\|_{L^2(\xi^{\pi^{\natural}}_{t,n})}.$$

Theorem 2. *Assume **(H2)**. If the causal dynamics satisfy*

$$\forall 0 \leq t \leq T, \qquad \exists \alpha_t, \beta_t > 0, \quad \alpha_t(\partial H_t) \cap \beta_t U_t \subset \Gamma_t,$$

then we have

$$\left|\hat{\mathcal{V}}^{\mathring{\pi}}_{\text{DRE}} - \mathcal{V}^{\mathring{\pi}}\right| \leq C\left(\sum_{t=0}^{T}(T-t+1)\left(\sigma_t^{\text{IPM}}(n) + \frac{1}{\sqrt{N}} + \sigma_t^{\text{WS}}(n)\right)\sigma_t^{\text{REG}}(n)\right),$$

where $C > 0$ *is a constant that is independent from* $n, m, N,$ *and* T*.*

References

Athey, S., Wager, S., et al. (2017). Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 78.

Bennett, A., Kallus, N., Li, L., and Mousavi, A. (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 1999–2007. PMLR.

Bertsimas, D. and Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044.

Chakraborty, B. (2013). *Statistical Methods for Dynamic Treatment Regimes Reinforcement Learning, Causal Inference, and Personalized Medicine*. Statistics for Biology and Health, 76. Springer New York, New York, NY, 1st ed. 2013. edition.

Chernozhukov, V., Newey, W., and Singh, R. (2018). De-biased machine learning of global and local parameters using regularized riesz representers. *arXiv e-prints*, pages arXiv–1802.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. (2020). Adversarial estimation of riesz representers. *arXiv e-prints*.

Dudík, M., Erhan, D., Langford, J., Li, L., et al. (2014). Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2007). A kernel method for the two-sample-problem. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., 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. PMLR.

Kallus, N. (2018). Balanced policy evaluation and learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8909–8920, Red Hook, NY, USA. Curran Associates Inc.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54.

Kallus, N., Pennicooke, B., and Santacatterina, M. (2019). More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions.

Kallus, N. and Santacatterina, M. (2019a). Kernel optimal orthogonality weighting: A balancing approach to estimating effects of continuous treatments.

Kallus, N. and Santacatterina, M. (2019b). Optimal estimation of generalized average treatment effects using kernel optimal matching.

Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63.

Kallus, N. and Zhou, A. (2018). Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*.

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

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Liu, Q., Li, L., Tang, Z., and Zhou, D. (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. *arXiv preprint arXiv:1810.12429*.

Liu, Y., Bacon, P.-L., and Brunskill, E. (2020). Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6184–6193. PMLR.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

Murphy, S. A. (2005). A generalization error for q-learning. *Journal of Machine Learning Research*, 6(37):1073–1097.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Nie, X., Brunskill, E., and Wager, S. (2020). Learning when-to-treat policies. *Journal of the American Statistical Association*, 0(0):1–18.

Nie, X. and Wager, S. (2017). Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.

Pearl, J., Bareinboim, E., et al. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595.

Rachev, S. (1991). *Probability Metrics and the Stability of Stochastic Models*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley.

Reygner, J. and Touboul, A. (2020). Reweighting samples under covariate shift using a wasserstein distance criterion.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Sugiyama, M., Nakajima, S., Kashima, H., Von Buenau, P., and Kawanabe, M. (2007). Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, volume 7, pages 1433–1440. Citeseer.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thomas, P. and Brunskill, E. (2016). Data-efficient off-policy policy evaluation for reinforcement learning. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, New York, New York, USA. PMLR.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694.

Zhao, Y.-Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598. PMID: 26236062.