

# Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach

Soroush Saghafian<sup>a</sup>

<sup>a</sup>Harvard Kennedy School, Harvard University, Cambridge, Massachusetts 02138

Contact: [soroush\\_saghafian@hks.harvard.edu](mailto:soroush_saghafian@hks.harvard.edu),  <https://orcid.org/0000-0002-9781-6561> (SS)

Received: March 21, 2022

Revised: December 9, 2022

Accepted: May 26, 2023

Published Online in Articles in Advance:  
October 4, 2023

<https://doi.org/10.1287/mnsc.2022.00883>

Copyright: © 2023 INFORMS

**Abstract.** A main research goal in various studies is to use an observational data set and provide a new set of counterfactual guidelines that can yield causal improvements. Dynamic Treatment Regimes (DTRs) are widely studied to formalize this process and enable researchers to find guidelines that are both personalized and dynamic. However, available methods in finding optimal DTRs often rely on assumptions that are violated in real-world applications (e.g., medical decision making or public policy), especially when (a) the existence of unobserved confounders cannot be ignored, and (b) the unobserved confounders are time varying (e.g., affected by previous actions). When such assumptions are violated, one often faces ambiguity regarding the underlying causal model that is needed to be assumed to obtain an optimal DTR. This ambiguity is inevitable because the dynamics of unobserved confounders and their causal impact on the observed part of the data cannot be understood from the observed data. Motivated by a case study of finding superior treatment regimes for patients who underwent transplantation in our partner hospital (Mayo Clinic) and faced a medical condition known as new-onset diabetes after transplantation, we extend DTRs to a new class termed Ambiguous Dynamic Treatment Regimes (ADTRs), in which the causal impact of treatment regimes is evaluated based on a “cloud” of potential causal models. We then connect ADTRs to Ambiguous Partially Observable Markov Decision Processes (APOMDPs) proposed by Saghafian (2018), and consider unobserved confounders as latent variables but with ambiguous dynamics and causal effects on observed variables. Using this connection, we develop two reinforcement learning methods termed Direct Augmented V-Learning (DAV-Learning) and Safe Augmented V-Learning (SAV-Learning), which enable using the observed data to effectively learn an optimal treatment regime. We establish theoretical results for these learning methods, including (weak) consistency and asymptotic normality. We further evaluate the performance of these learning methods both in our case study (using clinical data) and in simulation experiments (using synthetic data). We find promising results for our proposed approaches, showing that they perform well even compared with an imaginary oracle who knows both the true causal model (of the data-generating process) and the optimal regime under that model. Finally, we highlight that our approach enables a *two-way personalization*; obtained treatment regimes can be personalized based on both patients’ characteristics and physicians’ preferences.

**History:** Accepted by David Simchi-Levi, data science.

**Supplemental Material:** The data files and online appendix are available at <https://doi.org/10.1287/mnsc.2022.00883>.

**Keywords:** observational data • dynamic treatment regimes • unobserved confounders • APOMDPs • reinforcement learning • precision medicine

## 1. Introduction

In a variety of applications in public policy, governance, medicine, economics, education, energy, and e-commerce, a main goal is to make better decisions that are both *personalized* and *dynamic*. This requires learning from a data set which actions to choose and when to apply them given the dynamic conditions of each subject (e.g., an individual). One of the main factors that makes this learning process challenging is that one needs to estimate the impact of an alternative sequence of actions that *could have been* used in order to improve outcomes. This requires *causal reasoning*, as the estimand—the effect of an

alternative sequence of actions—is a *counterfactual* quantity (see, e.g., Murphy et al. 2001, Murphy 2003, Namkoong et al. 2020).

Dynamic Treatment Regimes (DTRs) have been widely studied for this goal, enabling the finding of effective alternative policies from observational data (Robins 1986, 1997, 2004; Murphy et al. 2001; Murphy 2003; Zhao et al. 2015; Wang et al. 2018; Zhang et al. 2018; Kosorok and Laber 2019; Tsiatis et al. 2019; Luckett et al. 2020; Leqi and Kennedy 2021; Nie et al. 2021). A DTR is, in essence, a set of rules that prescribe an individualized sequence of actions by mapping a subject’s history to a series of

recommended treatments (Murphy 2003, Chakraborty and Murphy 2014, Tsiatis et al. 2019, Luckett et al. 2020, Xu et al. 2020).

Using available results in finding effective DTRs, however, requires making strong assumptions that might not hold in real-world applications, especially when the data in hand are *observational*. Notably, one needs to assume *sequential ignorability*<sup>1</sup> (Robins 1986, 1997, 2004; Murphy et al. 2001; Murphy 2003), meaning that the data are rich enough, and hence, unobserved/latent/unmeasured confounding variables either do not exist or their effects can be ignored. When using observational data sets, this assumption is often violated in many real-world applications. Even in some secondary analyses of experimental data sets (e.g., those obtained under microrandomized trials in some mobile health studies where the goal is to study the effect of users following a treatment regime and not just being assigned to it), various practical challenges (e.g., user habituation, user engagement, and/or user compliance) may lead to unobserved confounding (see, e.g., Saghafian and Murphy 2021 for some discussions on scientific challenges in mobile health applications). Furthermore, unobserved confounders are time-varying in most applications: they are themselves affected by the previous actions taken. Adjusting for them, thus, is a perplexing task, making standard approaches for adjustment of confounding erroneous (see, e.g., Robins et al. 2000).

Correctly adjusting for unobserved time-varying confounding can be managed if one assumes a specific causal model for the data-generating process.<sup>2</sup> Assuming such a model can allow for estimating a distribution for potential trajectories under any alternative decision-making policy (i.e., treatment regime), which is central to estimating its effect. However, because time-varying confounders are often unobserved, estimating and assuming any such model is subject to significant misspecifications (also known as model ambiguity). We address this challenge by extending the analyses of DTRs to a new class termed *Ambiguous Dynamic Treatment Regimes* (ADTRs), in which the impact of any sequence of actions is evaluated based on a “cloud” of potential data-generating models as opposed to a single one. Specifically, we allow for nonprobabilistic ambiguity (also known as Knightian uncertainty) about the true data-generating model, while (similar to the literature on DTRs) we assume that under any given potential model, there is a certain probabilistic understanding of how data are generated (see, e.g., Chapter 11 of Manski 2007 and Saghafian 2018 for further discussions, Stoy 2011 for an axiomatic treatment of statistical decision-making under these conditions, and Saghafian and Tomlin 2016 for an information entropy view of data-driven decision-making under ambiguity).<sup>3</sup> This allows for (a) directly taking into account potential model misspecifications when estimating causal impacts, and (b)

distinguishing between *ambiguity* (lack of knowledge about the true model) and *risk* (probabilistic consequences of decisions under a known model).<sup>4</sup>

In extending DTRs to ADTRs, we are particularly motivated by our various collaborations with our partner hospital, the Mayo Clinic. In various studies (see, e.g., Bolori et al. 2015, 2020; Munshi et al. 2020a, b, 2021), we have collected data sets from our partner hospital and have examined clinical decisions for patients who undergo a solid organ transplantation and develop what is known as *New Onset Diabetes After Transplantation* (NODAT). In practice, physicians often use an intensive amount of an immunosuppressive drug (e.g., tacrolimus) to reduce the risk of organ rejection post-transplant (see, e.g., Bolori et al. 2015, 2020). Because of a well-established effect known as the diabetogenic effect, this can increase the risk of NODAT, which prompts physicians to use a glucose control drug (e.g., insulin). Learning better ways to prescribe these drugs (e.g., tacrolimus and insulin) in both a personalized and dynamic way to jointly control risks of NODAT and organ rejection is not an easy endeavor; the available data sets are only observational, the main health states are hidden (see, e.g., Bolori et al. 2020), and the existence of unobserved confounders that are time varying disallows using existing methods.

Our approach in extending DTRs to ADTRs and analyzing them involves the following three main steps. (1) We make use of a utility function that is appropriate under model ambiguity (instead of the expected value of outcomes widely used in the literature). (2) We generalize traditional importance sampling methods to accommodate model ambiguity. (3) We connect ADTRs to *Ambiguous Partially Observable Markov Decision Processes* (APOMDPs) proposed by Saghafian (2018) by showing that ADTRs can be studied via APOMDPs, which in turn, enables us to develop reinforcement learning (RL) algorithms capable of learning optimal treatment regimes from the observed data in effective ways.

The utility function we use is based on a generalization of the traditional *Maximin Expected Utility* (MEU) theory (also known as Wald’s or robust optimization criterion). The MEU theory assumes that outcomes should be obtained by maximizing utility with respect to the worst possible member of the ambiguity set (cloud of potential causal models in our setting). In most applications, using the MEU approach yields overly conservative decisions (for related discussions, see, e.g., Saghafian 2018 and the references therein), and furthermore, it does not allow for representing meaningful human choices, such as those of ambiguity-seeking individuals established in some behavioral studies (see, e.g., Heath and Tversky 1991, Bhidé 2000, Ahn et al. 2014). This was also recognized in the seminal work of Savage (1951), who wrote that this criterion is “ultrapessimistic” and “can lead to absurd conclusion[s].” The generalization we use

is known as  $\alpha$ -Maximin Expected Utility ( $\alpha$ -MEU), which allows for both optimistic and pessimistic views of the world (Hurwicz 1951a, b; Arrow and Hurwicz 1977; Ghiradato et al. 2004; Saghafian 2018). Unlike studies that use the MEU criterion, using the  $\alpha$ -MEU criterion avoids overly conservative decisions by allowing for a controllable *pessimism level* (denoted by the parameter  $\alpha$ ) that can take values in  $[0, 1]$ .

Within the utility theory literature, early studies (see, e.g., Hurwicz 1951a, b; Arrow and Hurwicz 1977) provided four axioms that a choice operator must satisfy. These axioms allowed such studies to show that, under complete ignorance, one can focus merely on two extreme cases: the best-case and the worst-case. Later studies (see, e.g., Marinacci 2002, Ghiradato et al. 2004) further axiomatized preferences under the  $\alpha$ -MEU criterion and also highlighted another importance of using the  $\alpha$ -MEU criterion in decision making: it allows for differentiating between the *inherent ambiguity* (a property related to the true causal model) and *ambiguity attitude* (a property related to the decision maker). In our study, using the  $\alpha$ -MEU criterion not only allows us to provide an alternative for the expectation operator—the conventional measure of performance used in the literature surrounding DTRs<sup>5</sup>—but also allows for finding treatment regimes that are tailored to the preferences and attitudes of the decision-maker.

Importantly, this means that our work enables a *two-way personalization*; treatment regimes can be personalized based on both the subject's and the decision-maker's characteristics. This is important in various domains, such as medicine, where not only the treatment plan needs to be customized for each patient but also, the physician in charge should be given the ability to include his or her preferences in providing the best course of treatment. Incorporating a physician's preferences is important for many reasons, including the fact that several behavioral challenges often make it difficult for the physician to follow treatment decisions that are personalized to the patient but not him or her (see, e.g., Frank and Zeckhauser 2007). When using our framework, the physician's preferences in dealing with ambiguous outcomes can be incorporated in various ways. For example, questionnaires similar to those used in preference elicitation methods<sup>6</sup> can be designed to first understand the preferences and attitudes of the physician toward ambiguous outcomes, thereby obtaining a small interval (if not a specific value) for  $\alpha$ . Using these values of  $\alpha$ , a small set of corresponding optimal treatment regimes can be presented to the physician for further consideration. Alternatively, when it is crucial for the physician to follow the treatment regime that provides the maximum robustness and/or efficacy, one can use our framework to find the best treatment regime across all values of  $\alpha$  in  $[0, 1]$ . Finally, when other factors besides efficacy or robustness (e.g., cost,

availability, or patient consent) need to be considered, one can simply present the set of all treatment regimes that are optimal as  $\alpha$  ranges in  $[0, 1]$ , allowing the provider to inspect a broader set of treatment regimes.

We start our analyses by showing how a generalization of importance sampling methods (also known as inverse probability weighting) widely used in the literature (see, e.g., Precup et al. 2000, Robins et al. 2000, Murphy 2005, Tsiatis et al. 2019) can be utilized to find optimal regimes for ADTRs without requiring the dynamics of observed or unobserved variables to be memoryless (i.e., satisfy the Markov property).<sup>7</sup> Specifically, we start by generalizing importance sampling methods by allowing sampling across a *cloud* of potential data-generating models (also known as ambiguity set). We show that under some conditions, the resulted method, which we term *Generalized Sequential Importance Sampling* (GSIS), provides a baseline for estimating the causal impact of any dynamic treatment regime, and hence, finding the optimal one.

When the dynamics of variables satisfy the Markov property, we connect ADTRs to APOMDPs recently introduced by Saghafian (2018). APOMDPs generalize traditional Partially Observable Mark Decision Processes (POMDPs) by allowing model ambiguity. APOMDPs, however, were proposed without any causal inference application in mind. In this paper, for the first time, we make use of them through a causal inference lens. Notably, by connecting ADTRs to APOMDPs, we consider time-varying unobserved confounders as dynamic latent states, and form dynamic belief distributions over them, while allowing ambiguity regarding the true (data-generating) causal model.<sup>8</sup> We then make use of known structural results for APOMDPs (e.g., piecewise linearity and continuity of the value function) established in the literature (Saghafian 2018) and develop two RL approaches that can provide effective treatment regimes. In developing these RL approaches, as is common, we view the problem of finding an effective treatment regime as an off-policy RL problem. However, in contrast to main RL methods, such as Q-Learning (an approximate dynamic programming approach that uses regression to learn the “quality” function) and A-Learning (which tries to learn the “advantage” function), our approaches try to learn the value function directly. Thus, roughly speaking they are within the V-Learning methods (see, e.g., Lockett et al. 2020, Xu et al. 2020). We term our proposed learning algorithms *Direct Augmented V-Learning* (DAV-Learning) and *Safe Augmented V-Learning* (SAV-Learning) as they augment the V-Learning methods by (a) making use of the structural properties of the value function, and (b) incorporating model ambiguity (in a direct and safe way, respectively).<sup>9</sup>

For our proposed learning approaches, we establish important theoretical results, including weak consistency and asymptotic normality of both the estimated optimal treatment regime and the associate overall gain.

To establish these results, we require specific but relatively common “regularity” conditions, including conditions on (a) basic “complexity” properties of the class of allowable policies (measured by entropy-based versions of the Donsker theorems with bracketing integrals) and (b) absolute regularity of the underlying empirical processes.

We also examine the performance of our proposed approaches by applying them to a clinical data set of over 63,000 observations made of patients who underwent kidney transplantation in our partner hospital and faced NODAT. We find promising results, indicating that using DAV-Learning and SAV-Learning yields notable improvements over the treatment regime used in practice; depending on the decision maker’s pessimism level, these improvements are in the ranges (10%, 42%) and (10%, 32%) for DAV-Learning and SAV-Learning, respectively. Furthermore, we observe that the performance of the SAV-Learning regime is much more robust to the value of the pessimism level (parameter  $\alpha$ ) than that of DAV-Learning, and hence, a decision-maker who uses SAV-Learning does not need to be worried about the value of  $\alpha$  he or she uses in obtaining an optimal treatment regime. We further investigate the performance of our proposed approaches using simulations experiments (synthetic data). Our results show that DAV-Learning and SAV-Learning can improve the observed regime by an amount that ranges in (1%, 37%) and (1%, 8%), respectively. Furthermore, we make use of our simulation experiments to quantify the robustness of our approaches to model ambiguity, and find that DAV-Learning and the SAV-Learning are able to strongly shield against model ambiguity; the gain loss under these approaches compared with an imaginary oracle who knows both the true data-generating model and the optimal treatment regime under that model is very low (below 0.6%), regardless of the value of  $\alpha$ . Thus, a decision-maker who is facing model ambiguity can make use of our proposed approaches and obtain a treatment policy that has a similar performance to that of an imaginary decision-maker who knows both the true data-generating model and the optimal policy under that model. Finally, our results show that the gain loss compared with such an imaginary decision-maker has a U-shaped curve in the pessimism level, and hence, the minimum losses for both DAV-Learning and SAV-Learning are obtained at a midvalue of  $\alpha$ . This implies that (a) using extreme cases of  $\alpha = 0$  (a maximax view) or  $\alpha = 1$  (a maximin view) is almost never *robustness-maximizing*, and (b) by viewing  $\alpha$  as a tuning parameter (when needed) in our proposed approaches, one can obtain a treatment regime that performs best across all possible pessimism levels.

In closing this section, we note that our work in incorporating model ambiguity a priori in the analyses not only provides robustness to potential misspecifications

but more broadly, can bridge the gap between two philosophical views of decision-making using causal inference: *model based* and *model free*. The former postulates that any sensible causal reasoning for decision-making needs to be based on a specific model and set of assumptions in addition to data, whereas the latter advocates that it needs to rely only on data. We hope that our work in taking a middle ground and considering a cloud of models can serve as a step for future research in trying to further bridge the gap between the two. The importance of doing so has its roots in seminal work in statistical decision theory (see, e.g., Wald 1939, 1945, 1950) but has also been highlighted in various more recent studies. For example, Manski (2021) emphasizes that “models can at most approximate actualities” and highlights that statistical inference for decision-making needs to be performed across all feasible models. Similarly, referring to the famous quote from Box (1979), Watson and Holmes (2016) state that “statisticians are taught from an early stage that essentially all models are wrong, but some are useful,” and stress that decision-making needs to rely on a set of models that are misspecified (hence, “wrong”) but useful in that they can be “helpful for aiding actions (taking decisions).”

## 2. The Framework

Throughout the paper, the notation “ $\triangleq$ ” is used to differentiate between definitions and equations. For a set  $\mathcal{T} \triangleq \{1, 2, 3, \dots, T\}$ , the notations  $(X_t)_{t \in \mathcal{T}}$  and  $\mathcal{T}_{\leq t}$  are used to represent the vector  $(X_1, X_2, \dots, X_T)$  and the set  $\mathcal{T} \setminus \{t+1, t+2, \dots, T\}$ , respectively. All vectors are considered to be in the column format (e.g.,  $(X_t)_{t \in \mathcal{T}}$  is  $|\mathcal{T}| \times 1$ ). For any finite set  $\Xi \subset \mathbb{R}$ , we let  $\Delta_{\Xi}$  denote the probability simplex induced by  $\Xi$ . The notations  $\xrightarrow{p}$  and  $\rightarrow$  denote convergence in probability and distribution, respectively. The set  $\mathcal{I}$  represents the interval  $[0, 1]$ .

We let the observed data be a collection of  $n \in \mathbb{N}$  i.i.d. realizations (called trajectories) of the vector of variables  $(O_t, A_t)_{t \in \mathcal{T}}$ . For a realized trajectory,  $(o_t, a_t)_{t \in \mathcal{T}}$ ,  $o_t \in \mathcal{O}$  is the observation made about a subject (e.g., a patient’s observed covariates or an observed health state serving as a summary of them) at time  $t \in \mathcal{T}$ , and  $A_t \in \mathcal{A}$  denotes the action/treatment assigned at time  $t \in \mathcal{T}$ , where  $\mathcal{T}$  is the set of time periods (e.g., patients’ visits/follow-ups).<sup>10</sup> For example, in our study of NODAT patients, observations made about each patient ( $O_t$ ) include various test results, demographic information, and other observed risk factors, such as diabetes history, body mass index, blood pressure, triglyceride, uric acid, and lipoprotein information (see Table 1). Actions taken ( $A_t$ ) include low-dose (nonaggressive) or high-dose (aggressive) tacrolimus prescriptions as well as information on whether insulin has been used (see Table 3). Finally,  $\mathcal{T} \triangleq \{1, 2, 3, \dots, 12\}$  because patient follow-ups are monthly for a year after transplantation.

Besides the observed data, there are often unobserved variables that might have affected what is observed in the data. Let  $S_t$  denote a summary of them at time  $t$ , and let  $\mathcal{S}$  be the support of  $S_t$ . For example, in mHealth applications,  $S_t$  might include information relating to the patient’s habituation level (see, e.g., Saghafian and Murphy 2021) and/or patient true health state, both of which are often unobserved. In our case study of NODAT patients,  $S_t$  is a nine-level variable that summarizes the unobserved health state of the patient in terms of both transplantation and diabetes conditions (see Table 2). We denote the observable history up to each time  $t \in \mathcal{T}$  by  $\mathbf{H}_t^o \triangleq (O_1, A_1, O_2, A_2, \dots, O_t)$  and let  $\mathcal{H}_t^o$  be the support of  $\mathbf{H}_t^o$ . Similarly, we denote the (partially) unobservable history up to each time  $t \in \mathcal{T}$  by  $\mathbf{H}_t^u \triangleq (S_1, O_1, A_1, S_2, O_2, A_2, \dots, S_t, O_t)$  and let  $\mathcal{H}_t^u$  be the support of  $\mathbf{H}_t^u$ . It is important to note that, in general, both variables  $S_t$  and  $O_t$  depend on the previous treatments. However, for notational simplicity, we suppress the dependency of  $S_t$  and  $O_t$  on the vector  $(a_t)_{t \in \mathcal{T}_{\leq t-1}} \triangleq (a_1, a_2, \dots, a_{t-1})$ .

We assume the latent state summaries  $(S_t)_{t \in \mathcal{T}}$  are such that the immediate gain in each decision epoch depends on the history only through them. This can always be achieved with appropriate definition of variables  $(S_t)_{t \in \mathcal{T}}$  (see, e.g., Xu et al. 2020). For example, in our case study, the immediate gains are based on predefined *quality of life* (QoL) scores that depend only on patient summaries defined by  $S_t$  (see Table 4). Thus, we denote the immediate gain at time  $t$  through  $G_t \triangleq g(S_t, A_t) \in \mathbb{R}$ , where  $g$  is a known function.<sup>11</sup> The set of all possible immediate gains can be denoted by  $\mathcal{G} \triangleq \{\mathbf{g}^a \in \mathbb{R}^{|\mathcal{S}|} : a \in \mathcal{A}\}$ , where  $\mathbf{g}^a \triangleq (g(s, a))_{s \in \mathcal{S}}$ .

A treatment regime (hereafter also “policy” for simplicity)  $\boldsymbol{\lambda} \triangleq (\lambda_t)_{t \in \mathcal{T}}$  in this setting is a vector of time-dependent mappings from the available history at each time  $t$  to the probability simplex induced by actions,  $\Delta_{\mathcal{A}}$ . It defines the probability of assigning each action/treatment at each decision epoch given the available history up to that point. Policies are compared using the overall gain they generate. The overall gain of a policy  $\boldsymbol{\lambda}$  is defined by the discounted sum of immediate gains it generates, which we denote by

$$\Gamma_T(\boldsymbol{\lambda}) \triangleq \sum_{t \in \mathcal{T}} \beta^{t-1} G_t^\lambda, \quad (1)$$

where  $\beta \in \mathcal{I} \setminus \{1\}$  is a discount factor. Similarly, the long-run impact of  $\boldsymbol{\lambda}$  can be analyzed using  $\Gamma_\infty(\boldsymbol{\lambda}) \triangleq \lim_{T \rightarrow \infty} \Gamma_T(\boldsymbol{\lambda})$ .<sup>12</sup> Here, we shall note that  $G_t^\lambda$ , and hence,  $\Gamma_T(\boldsymbol{\lambda})$  should be viewed with a potential outcomes lens (for more discussions, see, e.g., Robins 1986, 1997; Rubin 1986; Angrist et al. 1996; Murphy et al. 2001); equivalently, in the language of do calculus,  $G_t^\lambda$  and  $\Gamma_T(\boldsymbol{\lambda})$  should be viewed as  $G_t | do(\boldsymbol{\lambda})$  and  $\Gamma_T | do(\boldsymbol{\lambda})$ , respectively (see, e.g., Pearl 2009). In addition to this, which is implicit in our notation, our notation also implicitly implies

*consistency*,<sup>13</sup> which is a standard assumption in the causal inference literature with time-varying variables (see, e.g., Robins 1997, Murphy et al. 2001) and holds in our motivating study of NODAT patients. In settings we consider, however, the distribution of  $\Gamma_T(\boldsymbol{\lambda})$  cannot be solely identified from the observed data alone. In fact, there are often a variety of plausible data-generating models all agreeing with the observed part of the data but with different implications about the distribution of  $\Gamma_T(\boldsymbol{\lambda})$ . We let  $\mathcal{M}$  denote the set of all such models (also known as an ambiguity set).<sup>14</sup> In this view, each given model  $m \in \mathcal{M}$  can be viewed as a rule that, given  $\boldsymbol{\lambda}$ , imposes a specific probability distribution over the full history  $\mathbf{H}_T^u$ . Thus, each given model  $m \in \mathcal{M}$  implies a distribution for  $\Gamma_T(\boldsymbol{\lambda})$ , which we denote by  $f_m \in \mathcal{F}$ , where  $\mathcal{F} \triangleq \{f_m : m \in \mathcal{M}\}$ .

Finally, because the distribution of  $\Gamma_T(\boldsymbol{\lambda})$  varies across the models in  $\mathcal{M}$ , we define a utility function that allows us to compare the performance of different policies. To this end, we make use of  $\alpha$ -MEU, which is suitable for decision-making under ambiguity (see, e.g., Marinacci 2002, Ghiradato et al. 2004, Saghafian 2018). Specifically, by considering  $Y(\boldsymbol{\lambda}) \triangleq \Gamma_T(\boldsymbol{\lambda})$  or  $Y(\boldsymbol{\lambda}) \triangleq \Gamma_\infty(\boldsymbol{\lambda})$  as our main outcome variable of interest, we make use of

$$MEU_\alpha[Y(\boldsymbol{\lambda})] \triangleq \alpha \inf_{f^m \in \mathcal{F}} \mathbb{E}^{f^m}[Y(\boldsymbol{\lambda})] + (1 - \alpha) \sup_{f^m \in \mathcal{F}} \mathbb{E}^{f^m}[Y(\boldsymbol{\lambda})] \\ \alpha \in \mathcal{I} \quad (2)$$

as the utility of  $Y(\boldsymbol{\lambda})$ , where  $\alpha$  represents the pessimism level and  $\mathbb{E}^{f^m}$  denotes the expectation operator with respect to the distribution  $f^m$ . For example, at  $\alpha = 1$  (100% pessimism level), policies are compared with respect to their worst-case performance. At  $\alpha = 0$  (0% pessimism level), on the other hand, policies are compared with respect to their best-case performance. Of note, when  $|\mathcal{M}| = 1$ ,  $MEU_\alpha[Y(\boldsymbol{\lambda})]$  returns the expected value of  $Y(\boldsymbol{\lambda})$ , and hence, the utility function in (2) provides a generalization for the traditional expectation operator that is widely used in the causal inference literature.

We say that the effect of treatment policy  $\boldsymbol{\lambda}$  is “ $\alpha$ -MEU identifiable,” if  $|MEU_\alpha[Y(\boldsymbol{\lambda})]| < \infty$  and  $MEU_\alpha[Y(\boldsymbol{\lambda})]$  can be identified given  $\mathcal{M}$ . Because a main goal is to learn the optimal policy, we next define the following notion of optimality in ADTRs, which is a generalization of the traditional notion of optimality used in analyzing DTRs.

**Definition 1** (Optimality). Let  $\Lambda$  be the set of all  $\alpha$ -MEU identifiable policies. We say that a policy  $\boldsymbol{\lambda}^* \in \Lambda$  is optimal if with  $Y(\boldsymbol{\lambda}) \triangleq \Gamma_T(\boldsymbol{\lambda})$ , we have

$$MEU_\alpha[Y(\boldsymbol{\lambda}^*)] \geq MEU_\alpha[Y(\boldsymbol{\lambda})] \quad \forall \boldsymbol{\lambda} \in \Lambda. \quad (3)$$

**Remark 1** (Fairness). When performance is evaluated under the traditional expectation operator, it is known

that (under some assumptions) the optimal policy assigns treatment (when  $|\mathcal{A}|=2$ ) only to those subjects who benefit from it: optimizing mean treatment and conditional mean treatment are equivalent. However, when using some other measures such as median, this no longer holds; the treatment decision for a given group might depend on outcomes from a different group, creating a “across-group fairness” concern (see, e.g., Leqi and Kennedy 2021). This concern is relatively mitigated when using the notion of optimality defined above. This is clear when  $|\mathcal{M}|=1$ , because  $MEU_\alpha[Y(\boldsymbol{\lambda})]$  returns the expected value of  $Y(\boldsymbol{\lambda})$ . More broadly, it can be seen that, under some conditions, the optimal policy  $\boldsymbol{\lambda}^* \in \Lambda$  defined in Definition 1 also optimizes  $MEU_\alpha[Y(\boldsymbol{\lambda})]$  after conditioning on subject-specific observed variables. For example, if we let  $\underline{m}(\boldsymbol{\lambda}^*) \triangleq \inf_{f^m \in \mathcal{F}} \mathbb{E}^{f^m}[Y(\boldsymbol{\lambda}^*)]$  and  $\overline{m}(\boldsymbol{\lambda}^*) \triangleq \sup_{f^m \in \mathcal{F}} \mathbb{E}^{f^m}[Y(\boldsymbol{\lambda}^*)]$ , assume that both of these models are in  $\mathcal{M}$ , impose similar conditions to those needed when evaluating the mean treatment effect, and evaluate the conditional  $MEU_\alpha$  value after fixing these models (as the worse-case and best-case models, respectively), we can see that  $\boldsymbol{\lambda}^*$  satisfies a version of “across-group fairness.” Specifically,  $\boldsymbol{\lambda}^*$  treats only those subjects who benefit from it in terms of the conditional  $MEU_\alpha$  value, regardless of the outcomes of the other subjects. In addition, when needed, one can further restrict the set of allowable policies  $\Lambda$  to those that satisfy some desirable fairness attributes. However, studying fairness and what should be considered as “fair” is outside the scope of this work, especially because measuring fairness under *ambiguity* is more complex than that under *risk*; see the discussion in Section 1 that highlights the difference between the two as well as more in-depth discussions in prior work (see, e.g., Saghafian and Tomlin 2016, Saghafian 2018). In what follows, we simply focus on data-driven ways of finding a policy  $\boldsymbol{\lambda}^* \in \Lambda$  that is optimal based on Definition 1.

To perform our analyses, it is useful to differentiate between the policy under which the data have been generated (hereafter, the “behavior policy”) and the policy that we would like to evaluate and recommend (hereafter, the “evaluation policy”). The behavior policy denoted by  $\boldsymbol{\lambda}^b \triangleq (\lambda_t^b)_{t \in \mathcal{T}}$  is a vector of time-dependent mappings  $\lambda_t^b: \mathcal{H}_t^u \rightarrow \Delta_{\mathcal{A}}$ , whereas the evaluation policy denoted by  $\boldsymbol{\lambda}^e \triangleq (\lambda_t^e)_{t \in \mathcal{T}}$  is a vector of time-dependent mappings  $\lambda_t^e: \mathcal{H}_t^o \rightarrow \Delta_{\mathcal{A}}$ . An important difference between the evaluation and the behavior policies relates to a condition known as *sequential ignorability*<sup>15</sup> (see, e.g., Robins 1986, 1997, 2004; Murphy et al. 2001; Murphy 2003), which we define next.

**Definition 2** (Sequential Ignorability). For any policy  $\boldsymbol{\lambda} \triangleq (\lambda_t)_{t \in \mathcal{T}}$ , let  $\mathbf{H}_t^{o,m}(\boldsymbol{\lambda}) \triangleq (O_1^{m,\boldsymbol{\lambda}}, A_1^{m,\boldsymbol{\lambda}}, O_2^{m,\boldsymbol{\lambda}}, A_2^{m,\boldsymbol{\lambda}}, \dots, O_t^{m,\boldsymbol{\lambda}})$  denote the observable history up to time  $t \in \mathcal{T}$ ,

generated under  $\boldsymbol{\lambda}$  and model  $m \in \mathcal{M}$ . We say that  $\boldsymbol{\lambda}$  satisfies sequential ignorability under model  $m \in \mathcal{M}$ , if for all  $t \in \mathcal{T}$ , the action generated by  $\lambda_t$  is independent of the collection of potential outcomes  $(G_t^{m,\boldsymbol{\lambda}}, O_{t+1}^{m,\boldsymbol{\lambda}}, G_{t+1}^{m,\boldsymbol{\lambda}}, O_{t+2}^{m,\boldsymbol{\lambda}}, \dots, G_T^{m,\boldsymbol{\lambda}})_{\boldsymbol{\lambda}' \in \Lambda}$  conditional on  $\mathbf{H}_t^{o,m}(\boldsymbol{\lambda})$ .

In essence, this definition requires, for each time  $t \in \mathcal{T}$  and given  $\mathbf{H}_t^{o,m}(\boldsymbol{\lambda})$ , the action generated by  $\lambda_t$  to be independent of  $(G_t^m, O_{t+1}^m, G_{t+1}^m, O_{t+2}^m, \dots, G_T^m)$  values that can be obtained when following any feasible sequence of actions  $(a_t)_{t \in \mathcal{T}}$ . Both by this definition and naturally, any evaluation policy  $\boldsymbol{\lambda}^e \triangleq (\lambda_t^e)_{t \in \mathcal{T}}$  (where  $\lambda_t^e: \mathcal{H}_t^o \rightarrow \Delta_{\mathcal{A}}$ ) satisfies sequential ignorability under any model  $m \in \mathcal{M}$ , because it is only a function of the observed history and possible exogenous randomness (it maps  $\mathcal{H}_t^o$  to  $\Delta_{\mathcal{A}}$  by definition). In contrast, a behavior policy  $\boldsymbol{\lambda}^b \triangleq (\lambda_t^b)_{t \in \mathcal{T}}$  (where  $\lambda_t^b: \mathcal{H}_t^u \rightarrow \Delta_{\mathcal{A}}$ ) may or may not satisfy this condition, because it might depend on unobservable confounders (variables in  $(S_t)_{t \in \mathcal{T}}$  that affect both the gain and the actions selected by  $\boldsymbol{\lambda}^b$ ). In fully randomized experiments (e.g., microrandomized trials), the behavior policy may satisfy sequential ignorability. However, when the data are observational, it is often impossible to test whether the behavior policy satisfies this assumption, and in addition, it is highly likely that this assumption does not hold. Finally, we note that although the behavior policy is often known in randomized experiments, it might need to be estimated when using observational data (unless specific known treatment protocols or standards are fully followed).

## 2.1. Analyzing ADTRs via GSIS

We now show that, under some conditions, an optimal policy for an ADRT can be found using a generalized version of sequential importance sampling, which we term Generalized Sequential Importance Sampling (GSIS). Although allowing for *ambiguity* (not risk, which is the focus of the existing literature), GSIS assigns weights under each model and sequentially adjusts the trajectory probabilities that occur under a given evaluation policy compared with those observed in the data set. Of note, we use GSIS in this section to study ADTRs that do not satisfy any Markovian (also known as memoryless) property regarding the dynamics of the underlying variables. In the next section, we show how the analyses of ADTRs can be simplified when such dynamics satisfy a Markovian structure. Notably, the results presented in this section provide a building block for the algorithms and the theoretical findings established in the next sections.

To present GSIS, we first suppress the dependencies to the underlying model by assuming the model is fixed. Consider an evaluation policy  $\boldsymbol{\lambda}^e$ , and let  $\mathbf{H}_t^o(\boldsymbol{\lambda}^e)$  be the history that will be observed under  $\boldsymbol{\lambda}^e$  up to time  $t$ . Also, denote by  $\lambda_t^e(A_t | \mathbf{H}_t^o(\boldsymbol{\lambda}^e))$  the probability that action  $A_t$  is chosen under  $\boldsymbol{\lambda}^e$  when the observed history is  $\mathbf{H}_t^o(\boldsymbol{\lambda}^e)$ . Furthermore, although the behavior policy is not known

(e.g., because of its potential dependency on unobserved variables), we can observe the *marginalized* probabilities of action selection under the behavior policy, which we denote by  $\lambda_t^b(A_t | \mathbf{H}_t^o(\boldsymbol{\lambda}^b))$ . These allow us to define importance sampling weights

$$w_t(\boldsymbol{\lambda}^e) \triangleq \frac{\lambda_t^e(A_t | \mathbf{H}_t^o(\boldsymbol{\lambda}^e))}{\lambda_t^b(A_t | \mathbf{H}_t^o(\boldsymbol{\lambda}^b))} \quad \forall t \in \mathcal{T}. \quad (4)$$

Proposition 1 establishes that, under some conditions, the optimal policy for an ADTR governed by a set of models  $\mathcal{M}$  can be found via GSIS. Specifically, an  $MEU_\alpha$ -unbiased estimator of  $\Gamma_T(\boldsymbol{\lambda}^e)$  is obtained in Proposition 1, where the notion of  $MEU_\alpha$ -unbiased estimation is defined next.

**Definition 3** ( $MEU_\alpha$  Unbiasedness). An estimator  $\hat{Y}$  of an outcome variable of interest  $Y$  is said to be  $MEU_\alpha$  unbiased if, and only if,  $MEU_\alpha[\hat{Y}] = MEU_\alpha[Y]$  for any  $\alpha \in \mathcal{F}$ .

To establish an  $MEU_\alpha$ -unbiased estimator of  $\Gamma_T(\boldsymbol{\lambda}^e)$ , we also need to make sure that the evaluation and the behavior policies sufficiently *overlap*. Specifically, we need to ensure that these policies overlap almost surely (defined next).

**Definition 4** (Almost Sure Overlap). We say that the evaluation and the behavior policy almost surely overlap, if  $\lambda_t^b(a_t | \mathbf{H}_t^o(\boldsymbol{\lambda}^b)) > 0$  whenever  $\lambda_t^e(a_t | \mathbf{H}_t^o(\boldsymbol{\lambda}^e)) > 0$  a.s. over  $\mathbf{H}_t^o(\boldsymbol{\lambda}^b)$  and  $\mathbf{H}_t^o(\boldsymbol{\lambda}^e)$  for all  $t \in \mathcal{T}$  and  $a_t \in \mathcal{A}$ .

Intuitively, the evaluation and the behavior policy need to overlap to ensure that trajectories obtained under the behavior policy are to some extent informative about the trajectories under the evaluation policy. When the evaluation and the behavior policy almost surely overlap, the importance sampling weights defined in (4) are well defined for all  $t \in \mathcal{T}$  (except perhaps on histories that might happen with probability 0).

**Proposition 1** (GSIS). Suppose that the evaluation and behavior policies (a) satisfy sequential ignorability under all models  $m \in \mathcal{M}$ , and (b) almost surely overlap. Then, for any  $\alpha \in \mathcal{F}$ , we have

$$MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^e)] = MEU_\alpha \left[ \Gamma_T(\boldsymbol{\lambda}^b) \prod_{t \in \mathcal{T}} w_t(\boldsymbol{\lambda}^e) \right], \quad (5)$$

and hence,  $\hat{\Gamma}_T(\boldsymbol{\lambda}^e) \triangleq \Gamma_T(\boldsymbol{\lambda}^b) \prod_{t \in \mathcal{T}} w_t(\boldsymbol{\lambda}^e)$  is an  $MEU_\alpha$ -unbiased estimator of  $\Gamma_T(\boldsymbol{\lambda}^e)$ .

The proof of Proposition 1 is developed by understanding how the impact of an evaluation policy can be first analyzed for any given (a) sequence of actions, and (b) model  $m \in \mathcal{M}$  under which the data might be generated (see Lemma EC.1 in Online Appendix A).

Of note, Proposition 1 also provides a partial way of characterizing the optimal evaluation policy because it provides a way of estimating  $MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^e)]$  under any given evaluation policy. That is, using this proposition and optimizing  $MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^e)]$  over a given set of policies (which for use in practice, might be restricted to

those satisfying desirable attributes, such as fairness or interpretability) can shed light on the optimal evaluation policy. However, Proposition 1 provides only a partial way of characterizing the optimal evaluation policy because analyzing ADTRs often requires considering a behavior policy that might not satisfy sequential ignorability (at least under some models in  $\mathcal{M}$ ). Therefore, we next study scenarios in which the behavioral policy does not fully satisfy sequential ignorability but satisfies it *to some extent*. This entails limiting the impact of unobserved confounders (which make the probability of observing certain trajectories in the observed data biased compared with what would have happen if we could observe unobservables) on the behavior policy under each model. In limiting the impact of unobserved confounders on the behavior policy, we are mainly motivated by extending the analyses of confounding in causal inference (see, e.g., Rosenbaum 2002) from a traditional setting in which  $|\mathcal{T}| = 1$ , the treatment variable is binary  $|\mathcal{A}| = 2$ , and there is no model ambiguity, to ADTRs in which these restrictions are all relaxed. Two notable challenges in doing so are (1) because future actions depend on the history, a confounding decision/treatment in any period can make future decisions confounding as well, and (2) because the trajectory probabilities depend on the underlying model, the impact of unobserved confounding depends on the underlying model. We next introduce the notion of *bounded unobservable confoundedness*, which we define using the likelihood ratios of treatment propensities (functions  $\ell(\cdot)$  in the following definition). This, in turn, allows us to provide a version of GSIS under bounded unobservable confoundedness (Proposition 2).

**Definition 5** (Bounded Unobservable Confounding). We say that the behavioral policy satisfies bounded unobserved confounding (BUC) under a model  $m \in \mathcal{M}$  if there exist constants  $\eta_t^m \in [1, \infty)$  such that

$$(\eta_t^m)^{-1} \leq \frac{\ell(a_t, a'_t, \mathbf{H}_t^{o,m}, \mathbf{S}_t^m = \mathbf{s})}{\ell(a_t, a'_t, \mathbf{H}_t^{o,m}, \mathbf{S}_t^m = \mathbf{s}')} \leq \eta_t^m \quad (6)$$

a.s. over observable history  $\mathbf{H}_t^{o,m}$ , for all  $t \in \mathcal{T}$ ,  $a_t, a'_t \in \mathcal{A}$ ,  $\mathbf{s}, \mathbf{s}' \in \mathcal{S}^{(T)}$ , where  $\mathbf{S}_t^m \triangleq (\mathbf{S}_t^m)_{t \in \mathcal{T}_{\leq t}}$  and

$$\ell(a_t, a'_t, \mathbf{H}_t^{o,m}, \mathbf{S}_t^m = \mathbf{s}) \triangleq \frac{\lambda_t^b(a_t | \mathbf{H}_t^{o,m}, \mathbf{S}_t^m = \mathbf{s})}{\lambda_t^b(a'_t | \mathbf{H}_t^{o,m}, \mathbf{S}_t^m = \mathbf{s})}.$$

The definition bounds the impact of the vector of the unobservable confounder variables,  $\mathbf{S}_t^m$ , in each period. In Lemma EC.2 (Online Appendix B), we show that this definition results in

$$(\eta_t^m)^{-1} \leq \frac{\lambda_t^b(a_t | \mathbf{H}_t^{u,m})}{\lambda_t^b(a_t | \mathbf{H}_t^{o,m})} \leq \eta_t^m \quad \text{a.s.} \quad (7)$$

over  $\mathbf{H}_t^{o,m}$  and  $\mathbf{H}_t^{u,m} = (\mathbf{H}_t^{o,m}, \mathbf{S}_t^m)$  for all  $t \in \mathcal{T}$  and  $a \in \mathcal{A}$ . Thus, benefiting from the observed history (as opposed to the unobserved one) and making use of marginalized

propensities  $\lambda_t^b(a_t | \mathbf{H}_t^{o,m})$  as an estimate of the true treatment propensities  $\lambda_t^b(a_t | \mathbf{H}_t^{u,m})$  will not be unboundedly misleading. The results provided in the following proposition are analogous to *design sensitivity* analyses (see, e.g., Rosenbaum 2010) in static (i.e.,  $T=1$ ) settings, where the idea is to examine how much propensity odds need to vary so that the gained causal understanding becomes invalid (see also Kallus and Zhou 2020, 2021). This proposition can also be viewed as a generalization of some of the available bounds in the literature of DRTs (see, e.g., lemma 2 of Namkoong et al. 2020), mainly because such bounds can be obtained from our results under the special case of  $|\mathcal{M}| = 1$ . Furthermore, we note that although these bounds can be conservative (hence, not useful) for the purpose of estimating the mean performance, they are relatively suitable for estimating the  $MEU_\alpha$  value of it (see part (ii) of Proposition 2 as well as the learning approaches discussed in Section 6).

**Proposition 2** (GSIS Under Bounded Unobservable Confounding). *Suppose the behavior policy satisfies BUC under all models  $m \in \mathcal{M}$ . If the evaluation policy satisfies sequential ignorability under all models  $m \in \mathcal{M}$  and it overlaps with the behavior policy almost surely, then*

i. *Under each model  $m \in \mathcal{M}$ , we have*

$$\begin{aligned} \mathbb{E}^m \left[ \Gamma_T(\boldsymbol{\lambda}^b) \prod_{t \in \mathcal{T}} \underline{w}_t^m(\boldsymbol{\lambda}^e) \right] &\leq \mathbb{E}^m [\Gamma_T(\boldsymbol{\lambda}^e)] \\ &\leq \mathbb{E}^m \left[ \Gamma_T(\boldsymbol{\lambda}^b) \prod_{t \in \mathcal{T}} \overline{w}_t^m(\boldsymbol{\lambda}^e) \right], \end{aligned}$$

where

$$\underline{w}_t^m(\boldsymbol{\lambda}^e) \triangleq w_t(\boldsymbol{\lambda}^e) ((\eta_t^m)^{-1} \mathbb{1}_{\{\Gamma_T(\boldsymbol{\lambda}^b) > 0\}} + \eta_t^m \mathbb{1}_{\{\Gamma_T(\boldsymbol{\lambda}^b) < 0\}})$$

and

$$\overline{w}_t^m(\boldsymbol{\lambda}^e) \triangleq w_t(\boldsymbol{\lambda}^e) ((\eta_t^m)^{-1} \mathbb{1}_{\{\Gamma_T(\boldsymbol{\lambda}^b) < 0\}} + \eta_t^m \mathbb{1}_{\{\Gamma_T(\boldsymbol{\lambda}^b) > 0\}}).$$

ii. *For any  $\alpha \in \mathcal{F}$ , there exists  $\tilde{\alpha} \in \mathcal{F}$  such that  $MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^e)] = f(\tilde{\alpha})$ , where  $f(\tilde{\alpha}) \triangleq \tilde{\alpha} MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^b) \prod_{t \in \mathcal{T}} \underline{w}_t^m(\boldsymbol{\lambda}^e)] + (1 - \tilde{\alpha}) MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^b) \prod_{t \in \mathcal{T}} \overline{w}_t^m(\boldsymbol{\lambda}^e)]$ .*

Similar to Proposition 1, part (ii) of Proposition 2 provides a way of finding the optimal evaluation policy because it characterizes the causal impact of any such policy. Whereas Proposition 1 requires the behavior policy to satisfy sequential ignorability—an unrealistic assumption in most applications—Proposition 2 only requires the unobserved variables to have a bounded impact. Importantly, however, Proposition 1 directly provides an  $\alpha$ -MEU unbiased estimator, but Proposition 2 does so subject to a tuning parameter  $\tilde{\alpha}$ . Specifically, in part (ii) of Proposition 2, the function  $f$  can be computed using only observed data. This, in turn, resolves the issue that the outcome of interest under the evaluation policy as well as the time-varying confounders needed to estimate it are unobservable. However,

to use Proposition 2, part (ii), one needs to tune the parameter  $\tilde{\alpha}$ . Because  $f$  is a decreasing function and  $\tilde{\alpha} \in \mathcal{F}$ , tuning  $\tilde{\alpha}$  can be done in a structured way. For example, in practice, one is often interested in evaluating policies that are known to be better than the behavior policy. Thus, we have  $f(\tilde{\alpha}) \geq MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^b)]$ , implying that one can start tuning  $\tilde{\alpha}$  using the threshold value  $\tilde{\alpha}^* \triangleq \min\{f^{(-1)}(MEU_\alpha[\Gamma_T(\boldsymbol{\lambda}^b)]), 1\}$  and only consider values of  $\tilde{\alpha}$  that are in  $[0, \tilde{\alpha}^*]$ . More importantly, it should be noted that the parameters  $(\eta_t^m)_{t \in \mathcal{T}, m \in \mathcal{M}}$  are design sensitivity parameters. Specifically, for any  $\epsilon > 0$ , they can be chosen so that  $f(0) - f(\tilde{\alpha}^*) < \epsilon$ . Because  $f(0) - f(\tilde{\alpha}^*) \geq 0$ , this allows one to use any  $\tilde{\alpha}$  in  $[0, \tilde{\alpha}^*]$  and obtain an *approximate* unbiased  $MEU_\alpha$  estimator for  $\Gamma_T(\boldsymbol{\lambda}^e)$  with a guaranteed approximation error of  $\epsilon$ . Although this provides an approximation method, two limitations are noteworthy; (1) obtaining an exact value for  $\tilde{\alpha}$  (and hence, an exact unbiased  $MEU_\alpha$  estimator for  $\Gamma_T(\boldsymbol{\lambda}^e)$ ) can be challenging, and (2) the bounds in part (i) of Proposition 2 can be conservative, because in general they may diverge exponentially in  $T$ . Nonetheless, as described, one can obtain an  $\epsilon$ -approximate unbiased  $MEU_\alpha$  estimator for  $\Gamma_T(\boldsymbol{\lambda}^e)$ .

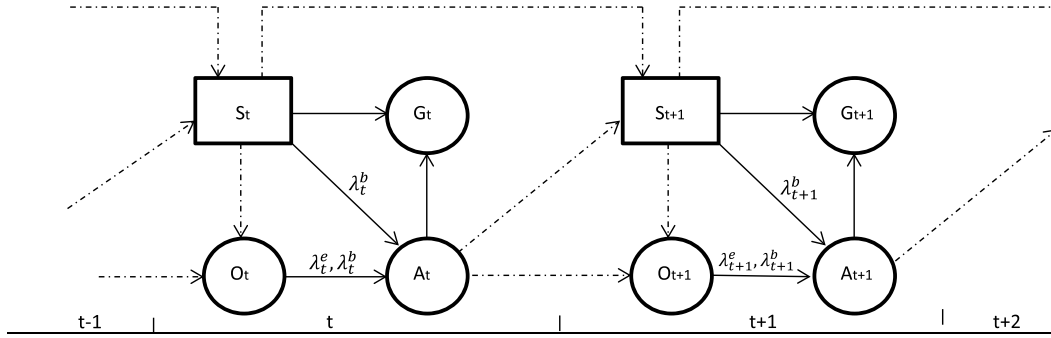
Finally, we note that one can extend Propositions 1 and 2 to provide doubly robust estimators<sup>16</sup> to account for the fact that, under each given model  $m \in \mathcal{M}$ , the variance of an importance sampling estimator can be high. Such an extension is, however, not that useful in our work because we are (a) directly allowing for a cloud of models, and (b) using  $MEU_\alpha$  of the outcome variable as opposed to its expected value (the criterion used in the studies related to doubly robust estimation). Instead, we next develop two RL methods based on our results and establish that they have suitable asymptotic behavior, including consistency and asymptotic normality. We also test their performance directly using both a clinical data set and simulation experiments, and we find that our proposed learning methods provide strong robustness to model ambiguity (see, e.g., Section 7.3).

### 3. Analyzing ADTRs via APOMDPs

In this section, we show that a tractable way of analyzing ADTRs is via APOMDPs. Specifically, analyzing ADTRs via APOMDPs enables (a) considering unobserved variables as latent time-varying states while allowing for model ambiguity, and (b) developing effective RL methods.<sup>17</sup>

An APOMDP (Saghafian 2018) can be represented via the directed acyclic graph (DAG) depicted in Figure 1. The ambiguous mechanisms in this figure represent causal relationships that cannot be quantified from the data alone. The main assumption needed to represent an ADTR via an APOMDP is that the dynamics of the variables is Markovian. In various applications, it is often possible to transform data so that this assumption holds

**Figure 1.** DAG Representation of APOMDPs



Notes. Circles indicate observable variables. Rectangles indicate unobservable variables. Solid arrows indicate unambiguous causal mechanisms. Dashed arrows indicate ambiguous causal mechanisms. Note that using  $\boldsymbol{\pi}_t \triangleq \nu_t(\mathbf{h}_t^o)$ —a belief distribution over the latent states—allows for working with an equivalent DAG representation where  $\boldsymbol{\pi}_t$  becomes the observed state at time  $t$ .

(see, e.g., Xu et al. 2020). Specifically, although the observed history  $\mathbf{H}_t^o$  grows over time, we can assume that there are summary functions  $\nu_t: \mathcal{H}_t^o \rightarrow \Delta_{\mathcal{S}}$  such that  $\boldsymbol{\pi}_t \triangleq \nu_t(\mathbf{h}_t^o)$  (a belief distribution over the latent states) is a sufficient statistic.<sup>18</sup>

Using the belief distribution  $\boldsymbol{\pi}_t$ , we can work with transformed policies; we can consider  $\boldsymbol{\mu}^e \triangleq (\mu_t^e(\boldsymbol{\pi}_t))_{t \in \mathcal{T}}$  and  $\boldsymbol{\mu}^b \triangleq (\mu_t^b(\boldsymbol{\pi}_t))_{t \in \mathcal{T}}$  as the evaluation and behavior policies, respectively, where  $\mu_t^e, \mu_t^b: \Delta_{\mathcal{S}} \rightarrow \Delta_{\mathcal{A}}$ . We denote the probability that an action  $a_t$  is applied at time  $t$  (when the belief distribution is  $\boldsymbol{\pi}_t$ ) under these transformed evaluation and behavior policies by  $\mu_t^e(a_t | \boldsymbol{\pi}_t)$  and  $\mu_t^b(a_t | \boldsymbol{\pi}_t)$ , respectively. In what follows, we first define the class of APOMDPs and then develop two RL algorithms that enable finding the optimal policy by effectively learning the causal impact of any given evaluation policy.

As defined in Saghafian (2018), a time-homogenous APOMDP is an extension of the classical POMDPs and can be defined by the tuple  $(\alpha, \beta, \mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{G}, \mathcal{P}, \mathcal{Q})$ . The notation used in the first part of this tuple is as introduced earlier. In addition,  $\mathcal{P}$  and  $\mathcal{Q}$  are the sets of possible transition probability matrices with respect to (latent) states and observations, respectively (Saghafian 2018). These sets define the ambiguous causal mechanisms depicted in Figure 1.

To simplify the analyses, we can index members of the set  $\mathcal{P} \times \mathcal{Q}$  using  $\mathcal{M}$  so that each  $m \in \mathcal{M}$  represents a specific (unambiguous) POMDP model. In particular, associated with each  $m \in \mathcal{M}$  is a set of the form  $P_m \times Q_m$ , with  $P_m \in \mathcal{P}$  and  $Q_m \in \mathcal{Q}$  denoting the set of state and observation transition probabilities under model  $m$ , respectively (Saghafian 2018). In this setting, (a)  $P_m \triangleq \{P_m^a: a \in \mathcal{A}\}$ , where for each  $a \in \mathcal{A}$ ,  $P_m^a \triangleq [p_{ij}^a(m)]_{i,j \in \mathcal{S}}$  is an  $|\mathcal{S}| \times |\mathcal{S}|$  matrix with  $p_{ij}^a(m) \triangleq \Pr\{j | i, a, m\}$  denoting the probability that the (latent) state moves to  $j$  from  $i$  under action  $a$  and model  $m$ , and (b)  $Q_m \triangleq \{Q_m^a: a \in \mathcal{A}\}$ , where for each  $a \in \mathcal{A}$ ,  $Q_m^a \triangleq [q_{jo}^a(m)]_{j \in \mathcal{S}, o \in \mathcal{O}}$  is an  $|\mathcal{S}| \times |\mathcal{O}|$  matrix with  $q_{jo}^a(m) \triangleq \Pr\{o | j, a, m\}$  denoting the probability of observing  $o$

under action  $a$  and model  $m$  when the (latent) state is  $j$  (Saghafian 2018).

If  $\mathcal{M}$  was a singleton with its only member being  $m$ , the optimal gain and policy for any  $t \in \mathcal{T}$  and  $\boldsymbol{\pi} \in \Delta_{\mathcal{S}}$  could be obtained by a traditional POMDP Bellman equation (along with the terminal condition  $V_0^m(\boldsymbol{\pi}) \triangleq 0$ ):

$$V_t^m(\boldsymbol{\pi}) = \max_{a \in \mathcal{A}} \left\{ \boldsymbol{\pi}' \mathbf{g}^a + \beta \sum_{o \in \mathcal{O}} \Pr\{o | \boldsymbol{\pi}, a, m\} V_{t-1}^m(T(\boldsymbol{\pi}, a, o, m)) \right\}, \quad (8)$$

where  $V_t^m(\boldsymbol{\pi})$  denotes the value function under model  $m$  when the belief distribution is  $\boldsymbol{\pi}$  and there are  $t$  periods to go,  $'$  represents the transpose operator,  $\Pr\{o | \boldsymbol{\pi}, a, m\} = \sum_i \sum_j \pi_i p_{ij}^a(m) q_{jo}^a(m)$ , and the belief updating operator  $T: \Delta_{\mathcal{S}} \times \mathcal{A} \times \mathcal{O} \times \mathcal{M} \rightarrow \Delta_{\mathcal{S}}$  is defined by the Bayes' rule (in the matrix form):

$$T(\boldsymbol{\pi}, a, o, m) = \frac{(\boldsymbol{\pi}' P_m^a Q_m^a(o))'}{\Pr\{o | \boldsymbol{\pi}, a, m\}}, \quad (9)$$

with  $Q_m^a(o) \triangleq \text{diag}(q_{1o}^a(m), q_{2o}^a(m), \dots, q_{no}^a(m))$  denoting the diagonal matrix made of the  $o$ th column of  $Q_m^a$  (Saghafian 2018).

Unlike POMDPs, in APOMDPs  $\mathcal{M}$  is not a singleton. However, it is shown in Saghafian (2018) that the APOMDP value function, a model independent function that we denote by  $V_t(\boldsymbol{\pi})$ , can still be obtained using dynamic programming. Furthermore, the underlying Bellman operator in the APOMDP is a contraction mapping with modulus  $\beta$  on a complete metric space (under some mild conditions), which in turn, allows for analyzing the APOMDP value function in infinite-horizon settings as the limit of its finite-horizon version. More importantly, Saghafian (2018) establishes some structural properties for the value function of the APOMDP (e.g, piecewise linearity and continuity in  $\boldsymbol{\pi}$ ). In the next section, we make use of these structural properties to develop effective RL approaches (termed augmented

V-Learning). We start our analyses by first developing suitable algorithms for learning the value function in POMDPs (i.e., when  $|\mathcal{M}| = 1$ ), and then, we show how they can be extended to learn the APOMDP value function.

## 4. Augmented V-Learning for POMDPs and APOMDPs

### 4.1. Augmented V-Learning for POMDPs

To develop our results, we require that the behavior policy,  $\mu^b$ , satisfies *positivity* defined here.

**Definition 6** (Positivity). We say that a policy  $\mu \triangleq (\mu_t)_{t \in \mathcal{T}}$  satisfies positivity if there exists a constant  $c_0 > 0$  such that  $\mu_t(a_t | \pi_t) \geq c_0$  for all  $t \in \mathcal{T}$ ,  $\pi_t \in \Delta_{\mathcal{G}}$ , and  $a_t \in \mathcal{A}$ .

Positivity implies that all actions have a positive chance of being selected (appear in the observed data) regardless of the belief. The behavior policy,  $\mu^b$ , automatically satisfies positivity when the data are collected based on a randomized trial. When using observational data, this assumption is sensible because inference involving treatment patterns (using action  $a_t$  when the belief is  $\pi_t$ ) that cannot occur in the observational study requires further knowledge and assumptions (Murphy et al. 2001). If the behavior satisfies positivity, we can establish the following result (see also Lemma 4.1 of Murphy et al. 2001 and Lemma 2.1 of Lockett et al. 2020 for related results in settings with fully observable states).

**Proposition 3** (Weight-Adjusted Bellman Equation). Suppose  $|\mathcal{M}| = 1$ , and denote the only member of  $\mathcal{M}$  by  $m$ . If  $\mu^b$  satisfies positivity and sequential ignorability, then for any policy  $\mu^e$ , the finite-horizon value function satisfies the weight-adjusted Bellman equation

$$V_{T-t+1}^{m, \mu^e}(\pi_t) = \mathbb{E}^m \left[ \frac{\mu_t^e(A_t | \Pi_t^m)}{\mu_t^b(A_t | \Pi_t^m)} \left[ G_t + \beta V_{T-t}^{m, \mu^e}(T(\Pi_t^m, A_t, O_t, m)) \right] \middle| \Pi_t^m = \pi_t \right] \quad (10)$$

for all  $t \in \mathcal{T}$  and  $\pi_t \in \Delta_{\mathcal{G}}$ , where  $\pi_t$  is considered as a realization (of a model dependent random variable denoted by  $\Pi_t^m$ ) and  $V_0^{m, \mu^e}(\pi) \triangleq 0$ . Therefore, for any function  $\phi$  defined on  $\Delta_{\mathcal{G}}$  and for all  $t \in \mathcal{T}$ , we have

$$\mathbb{E}^m \left[ \frac{\mu_t^e(A_t | \Pi_t^m)}{\mu_t^b(A_t | \Pi_t^m)} \left[ G_t + \beta V_{T-t}^{m, \mu^e}(T(\Pi_t^m, A_t, O_t, m)) - V_{T-t+1}^{m, \mu^e}(\Pi_t^m) \right] \phi(\Pi_t^m) \right] = 0. \quad (11)$$

The importance of Proposition 3 (which is built on the importance sampling results of the previous section) is that it allows us to empirically estimate the value function under any evaluation policy and hence,

learn the optimal policy. Specifically, using the data, we can make use of the sample-average version of (11):

$$\mathbb{E}^{\mathbb{P}} \left[ \sum_{t \in \mathcal{T}} \left[ \frac{\mu_t^e(A_t | \Pi_t^m)}{\mu_t^b(A_t | \Pi_t^m)} \left[ G_t + \beta V_{T-t}^{m, \mu^e}(T(\Pi_t^m, A_t, O_t, m)) - V_{T-t+1}^{m, \mu^e}(\Pi_t^m) \right] \phi(\Pi_t^m) \right] \right] = 0, \quad (12)$$

where  $\mathbb{E}^{\mathbb{P}}$  denotes average with respect to the empirical probability measure.<sup>19</sup> It is important to note that although we are using sample average in (12), the result still depends on the assumed  $m$  because although the sequence  $\{(A_t, O_t)\}_{t \in \mathcal{T}}$  is observable to us, to form the sequence  $\{\Pi_t^m\}_{t \in \mathcal{T}}$ , we need to have an assumed model. That is, because of the existence of unobserved variables, the empirical measure alone is *insufficient* for our goal.

**Remark 2** (Weight Adjustment Under BUC). The weight-adjusted results (10)–(12) are obtained assuming that  $\mu^b$  satisfies sequential ignorability. It should be noted that this assumption is milder than assuming that the original behavior policy satisfies sequential ignorability. This is especially the case because  $\mu^b \triangleq (\mu_t^b(\pi_t))_{t \in \mathcal{T}}$  is a *transformed* policy that takes advantage of the availability of information about the unobserved variables summarized in  $(\pi_t)_{t \in \mathcal{T}}$ . In particular, because in APOMDPs (Saghafian 2018), we have access to  $(\pi_t^m)_{t \in \mathcal{T}}$  under each model  $m \in \mathcal{M}$ , we might be able to make use of it to ensure that  $\mu^b$  satisfies sequential ignorability. For example, in many medical decision-making applications, including the case study with our partner hospital discussed in Section 7.1, the immediate gain depends on hidden states (confounders) only through  $(\pi_t^m)_{t \in \mathcal{T}}$ . This occurs, for example, when the gain in each period  $t$  is simply  $(\pi_t^m)' \mathbf{g}_t^a$ , where  $\mathbf{g}_t^a \triangleq (g(s_t, a_t))_{s_t \in \mathcal{G}}$ ; that is, when the gain in each period  $t$  is considered to be the average value of  $\mathbf{g}_t^a$  considering the belief distribution  $\pi_t^m$ .<sup>20</sup> (See Boloori et al. 2020 for more discussions about the immediate gain values in our medical decision-making application, and see Xu et al. 2020 for a similar assumption made in analyzing latent-state models.) However,  $\mu^b$  might not satisfy sequential ignorability in some other applications; see, for example, Saghafian and Murphy (2021) for related challenges in mHealth applications. Thus, in Section 6, we provide an extension of Proposition 3 by making use of the BUC results presented in Proposition 2. However, in what follows, we first focus on understanding and analyzing the cases where  $\mu^b$  satisfies sequential ignorability, as it provides a foundation for the extensions discussed in Section 6.

**Remark 3** (Effective Approximation). In using the results of Proposition 3, we only require an *approximate* solution to (12). Thus, how the solution to (12) is obtained is not

that restrictive. Indeed, there are many ways to obtain an approximate solution to (12). In what follows, however, we provide an effective way of estimating the optimal policy and optimal value function using (12). We do so by taking advantage of important structural properties of the optimal value function of POMDPs and APOMDPs. Specifically, the optimal value function of POMDPs is known to be piecewise linear and convex in  $\boldsymbol{\pi}$  under some mild conditions (Smallwood and Sondik 1973). Saghafian (2018) shows that, in general, the convexity does not hold in APOMDPs, and some additional conditions are needed (see Proposition 2 of Saghafian 2018). To be consistent, for both POMDP and APOMDP settings, we only assume *piecewise linearity* and *continuity*, but we do not impose any assumption on convexity. This, in turn, helps us in another way; although piecewise linear and continuous functions can be effectively learned from data, learning a function that is both piecewise linear and convex (i.e., is *point-wise maximum* of a set of linear functions) is much harder (see, e.g., Magnani and Boyd 2009 and the references therein).

Let  $\mathcal{V}$  denote the set of real-valued piecewise linear and continuous bounded functions defined on  $\Delta_{\mathcal{S}}$ , and assume  $V_t^{m, \mu^e} \in \mathcal{V}$ . To learn  $V_t^{m, \mu^e} \in \mathcal{V}$  using (12), we consider the parametric version of the value function:  $V_t^{m, \mu^e}(\boldsymbol{\pi}; \boldsymbol{\psi}_t) \triangleq (\mathbf{b}(\boldsymbol{\pi}))' \boldsymbol{\psi}_t$ , where  $\mathbf{b}(\boldsymbol{\pi}) \triangleq (\mathbf{b}_1(\boldsymbol{\pi}), \mathbf{b}_2(\boldsymbol{\pi}), \dots, \mathbf{b}_{d_t}(\boldsymbol{\pi}))'$  is a predefined *basis function* that allows us to ensure that the learned function is in  $\mathcal{V}$ , and  $\boldsymbol{\psi}_t \in \Psi_t \subseteq \mathbb{R}^{d_t}$  is the parameter.<sup>21</sup> This also enables us to set  $\phi(\boldsymbol{\pi}) \triangleq \mathbf{b}(\boldsymbol{\pi})$  in (12), because  $\mathbf{b}(\boldsymbol{\pi})$  can be thought of as the gradient of  $V_t^{m, \mu^e}(\boldsymbol{\pi}; \boldsymbol{\psi}_t)$  with respect to its parameter, which only depends on  $\boldsymbol{\pi}$  (and not the parameter) and is almost everywhere defined.

Furthermore, because  $\boldsymbol{\psi}_t$  can be high dimensional in some applications (especially when  $t$  is large), we estimate it using a regularized approach as follows (to avoid overfitting). Starting with  $V_0^m(\boldsymbol{\pi}) = 0$  and moving backward iteratively, having an estimation of  $T - t$  periods to go value function in hand ( $\hat{V}_{T-t}^{m, \mu^e}$ ), we define

$$\varphi^{m, \mu^e}(\boldsymbol{\psi}_t) \triangleq \mathbb{E}^{\mathbb{P}} \left[ \frac{\mu_t^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu_t^b(A_t | \boldsymbol{\Pi}_t^m)} \left[ G_t + \beta \hat{V}_{T-t}^{m, \mu^e} \right. \right. \\ \left. \left. (T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) - V_{T-t+1}^{m, \mu^e}(\boldsymbol{\Pi}_t^m; \boldsymbol{\psi}_t) \right] \mathbf{b}(\boldsymbol{\Pi}_t^m) \right]. \quad (13)$$

We then obtain the estimate

$$\hat{\boldsymbol{\psi}}_t^{\mu^e} = \arg \min_{\boldsymbol{\psi}_t \in \Psi_t} \{ (\varphi^{m, \mu^e}(\boldsymbol{\psi}_t))' \Omega \varphi^{m, \mu^e}(\boldsymbol{\psi}_t) + \theta_t \mathcal{P}(\boldsymbol{\psi}_t) \}, \quad (14)$$

where  $\Omega$  is an arbitrary positive definite matrix,  $\mathcal{P}(\cdot)$  is a penalty function, and  $\theta_t$  is a tuning parameter.<sup>22</sup> Consequently, we plug in  $\hat{\boldsymbol{\psi}}_t^{\mu^e}$  in  $V_{T-t+1}^{m, \mu^e}(\boldsymbol{\pi}; \boldsymbol{\psi}_t)$ , and thereby,

we obtain an estimate for the value function  $V_{T-t+1}^{m, \mu^e}(\boldsymbol{\pi}_t)$  and move to the next period (backward). This procedure, under a given model  $m \in \mathcal{M}$ , yields an estimator for the gain under  $\mu^e$ . That is,  $\hat{\Gamma}_T^m(\mu^e) \triangleq \int \hat{V}_T^{m, \mu^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$  can be used as an estimator for  $\Gamma_T^m(\mu^e) = \int V_T^{m, \mu^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ , where  $dF(\boldsymbol{\pi})$  is a given distribution on (starting) belief values. Because we have an estimator for the gain under any policy  $\mu^e$ , we can obtain  $\hat{\boldsymbol{\mu}}^{e*} \triangleq \arg \max_{\mu^e \in \Upsilon} \hat{\Gamma}_T^m(\mu^e)$  as an estimate for the optimal policy under model  $m$ , where  $\Upsilon$  is a given set of policies.<sup>23</sup> Finally, an estimate of the gain under the optimal policy is  $\hat{\Gamma}_T^m(\hat{\boldsymbol{\mu}}^{e*})$ .

In an infinite-horizon setting, the procedure described above simplifies. This is because in homogenous POMDPs (and APOMDPs), the value function with  $t$  periods to go converges to a stationary value function as  $t \rightarrow \infty$  (see, e.g., proposition 1 of Saghafian 2018). Therefore, in (12), we can replace both  $V_{T-t}^{m, \mu^e}(\cdot)$  and  $V_{T-t+1}^{m, \mu^e}(\cdot)$  with the same function. This removes the need for recursive calculations and allows us to follow a “one-shot” method. We discuss this further in the next sections and also study the asymptotic behavior of our proposed approach.

## 4.2. Augmented V-Learning for APOMDPs

Motivated by the results in the previous section, we now extend our approach to APOMDPs, where the condition  $|\mathcal{M}| = 1$  does not hold. We propose two approaches termed Direct Augmented V-Learning (DAV-Learning) and Safe Augmented V-Learning (SAV-Learning). As we will see, in DAV-Learning, we *directly* extend the approach presented in the previous section for POMDPs by first obtaining a value function separately for each POMDP model in  $\mathcal{M}$ . These values are then combined at the end of the horizon to provide an estimate of the value function for the APOMDP. In SAV-Learning, however, we make use of a *safe* estimation approach up front that takes into account ambiguity and removes the need to obtain a value function separately for each POMDP model in  $\mathcal{M}$ .

**4.2.1. DAV-Learning.** Recall that for each evaluation policy  $\mu^e$  and each given  $m \in \mathcal{M}$ , we can use the approach proposed for POMDPs in Section 4.1 to obtain an estimate for the value function  $V_T^{m, \mu^e}(\cdot)$ , which we denote by  $\hat{V}_T^{m, \mu^e}(\cdot)$ . Thus, we can first obtain an estimate for the APOMDP value function:

$$\hat{V}_T^{\mu^e}(\boldsymbol{\pi}) = MEU_{\alpha}[\hat{V}_T^{m, \mu^e}(\boldsymbol{\pi})] \triangleq \alpha \inf_{m \in \mathcal{M}} \hat{V}_T^{m, \mu^e}(\boldsymbol{\pi}) \\ + (1 - \alpha) \sup_{m \in \mathcal{M}} \hat{V}_T^{m, \mu^e}(\boldsymbol{\pi}). \quad (15)$$

Next, to estimate the optimal policy, we note that for any policy  $\mu^e$ , the estimator of the gain is  $\hat{\Gamma}_T(\mu^e) = \int \hat{V}_T^{\mu^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ , where  $dF(\boldsymbol{\pi})$  is a given distribution on

(starting) belief values. This means that we can obtain an estimate of the optimal policy as  $\hat{\boldsymbol{\mu}}^{e^*} \triangleq \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_T(\boldsymbol{\mu}^e)$ . Finally, the estimated optimal gain is  $\hat{\Gamma}_T(\hat{\boldsymbol{\mu}}^{e^*})$ .

This DAV-Learning approach for APOMDPs in the infinite-horizon case is presented in Algorithm 1. In presenting this algorithm, as is often the case, we assume that the data only include a finite number of periods for each subject, but the goal is to estimate the long-run performance of policies (see, e.g., Luckett et al. 2020, Xu et al. 2020). We also use subscript  $n$  to highlight the dependency of our estimators to the number of trajectories in the data set, which in turn allows us to investigate the behavior of our proposed learning algorithm as  $n \rightarrow \infty$  (see Section 5). Our estimation equations for the infinite-horizon gain are

$$\varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) \triangleq \mathbb{E}^{\mathbb{P}} \left[ \sum_{t \in \mathcal{T}} \left[ \frac{\mu^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu^b(A_t | \boldsymbol{\Pi}_t^m)} \left[ G_t + \beta V_{\infty}^{m, \boldsymbol{\mu}^e} \right. \right. \right. \\ \left. \left. \left. (T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) - V_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\Pi}_t^m)) \mathbf{b}(\boldsymbol{\Pi}_t^m) \right] \right] \right] \quad (16)$$

and

$$\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e} = \arg \min_{\boldsymbol{\psi} \in \Psi} \{ (\varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}))' \Omega \varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) + \theta_n \mathcal{P}(\boldsymbol{\psi}) \}, \quad (17)$$

where  $\Psi \subseteq \mathbb{R}^d$ . Similar to before, we make use of the piecewise linearity and continuity of the value function (i.e., the fact that  $V_{\infty}^{m, \boldsymbol{\mu}^e} \in \nu$ ) for all  $m \in \mathcal{M}$ . This allows us to use a predefined basis function to ensure that the learned function remains in  $\nu$  when we use the parametric form  $V_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}, \boldsymbol{\psi}) \triangleq (\mathbf{b}(\boldsymbol{\pi}))' \boldsymbol{\psi}$ .

Using (17), we then set  $\hat{V}_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) \triangleq V_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}; \hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e})$ . In addition, denoting the infinite-horizon gain under any policy  $\boldsymbol{\mu}^e$  and  $m \in \mathcal{M}$  by  $\Gamma_{\infty}^m(\boldsymbol{\mu}^e) \triangleq \int V_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ , we consider  $\hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e) \triangleq \int \hat{V}_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$  as an estimator for  $\Gamma_{\infty}^m(\boldsymbol{\mu}^e)$ . With estimated values under each model  $m$  in hand, we next define the estimated overall gain (a model independent value) as  $\hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e) \triangleq \alpha \inf_{m \in \mathcal{M}} \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e) + (1 - \alpha) \sup_{m \in \mathcal{M}} \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e)$ , which provides an estimation for the overall gain  $\Gamma_{\infty}(\boldsymbol{\mu}^e) \triangleq \alpha \inf_{m \in \mathcal{M}} \Gamma_{\infty}^m(\boldsymbol{\mu}^e) + (1 - \alpha) \sup_{m \in \mathcal{M}} \Gamma_{\infty}^m(\boldsymbol{\mu}^e)$ .

Finally, the estimated optimal policy and its infinite-horizon value for the APOMDP are obtained as  $\hat{\boldsymbol{\mu}}^{e^*} \triangleq \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e)$  and  $\hat{\Gamma}_{\infty}(\hat{\boldsymbol{\mu}}^{e^*}) = \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e)$ , respectively, where the latter provides an estimate for  $\Gamma_{\infty}(\boldsymbol{\mu}^{e^*}) \triangleq \max_{\boldsymbol{\mu}^e \in \Upsilon} \Gamma_{\infty}(\boldsymbol{\mu}^{e^*})$ . Similarly, under each model  $m$ , we denote the estimated optimal policy and its infinite-horizon value as  $\hat{\boldsymbol{\mu}}^{e^*, m} \triangleq \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e)$  and  $\hat{\Gamma}_{\infty}^m(\hat{\boldsymbol{\mu}}^{e^*, m}) = \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e)$ , respectively, where the latter provides an estimate for  $\Gamma_{\infty}^m(\boldsymbol{\mu}^{e^*, m}) \triangleq \max_{\boldsymbol{\mu}^e \in \Upsilon} \Gamma_{\infty}^m(\boldsymbol{\mu}^{e^*, m})$ .

#### Algorithm 1 (DAV-Learning)

- 1 **for** each observed trajectory and model  $m \in \mathcal{M}$  **do**
- 2   Initialize  $\boldsymbol{\pi}_0^m$  using a random draw from  $F(\boldsymbol{\pi})$ ;
- 3   set  $t = 1$ ;
- 4   **while**  $t + 1 \in \mathcal{T}$  **do**
- 5      $\boldsymbol{\pi}_{t+1}^m \leftarrow T(\boldsymbol{\pi}_t^m, a_t, o_t, m)$ ;
- 6   **for** any given  $\boldsymbol{\mu}^e \in \Upsilon$  and  $m \in \mathcal{M}$  **do**
- 7      $\varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) \leftarrow \mathbb{E}^{\mathbb{P}} \left[ \sum_{t \in \mathcal{T}} \left[ \frac{\mu^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu^b(A_t | \boldsymbol{\Pi}_t^m)} \left[ G_t + \beta V_{\infty}^{m, \boldsymbol{\mu}^e} \right. \right. \right. \right. \\ \left. \left. \left. (T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) - V_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\Pi}_t^m)) \mathbf{b}(\boldsymbol{\Pi}_t^m) \right] \right] \right]$ ;
- 8      $\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e} \leftarrow \arg \min_{\boldsymbol{\psi} \in \Psi} \{ (\varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}))' \Omega \varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) + \theta_n \mathcal{P}(\boldsymbol{\psi}) \}$ ;
- 9      $\hat{V}_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) \leftarrow (\mathbf{b}(\boldsymbol{\pi}))' \hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e}$ ;
- 10     $\hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e) \leftarrow \int \hat{V}_{\infty}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ ;
- 11   **for** any given  $\boldsymbol{\mu}^e \in \Upsilon$  **do**
- 12      $\hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e) \leftarrow \alpha \inf_{m \in \mathcal{M}} \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e) + (1 - \alpha) \sup_{m \in \mathcal{M}} \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e)$ ;
- 13     $\hat{\boldsymbol{\mu}}^{e^*} \leftarrow \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e)$ ;
- 14     $\hat{\Gamma}_{\infty}(\hat{\boldsymbol{\mu}}^{e^*}) \leftarrow \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e)$ .

**4.2.2. SAV-Learning.** The DAV-Learning algorithm presented in the previous section is a direct extension of the approach proposed for POMDPs (Section 4.1), in which “the curse of ambiguity” (Saghafian 2018) is overcome at the end. In contrast, in SAV-Learning, this curse is overcome up front via a “safe method” for estimating the underlying parameter  $\boldsymbol{\psi}_t$  and hence, the value function. To develop the SAV-Learning algorithm, similar to before, we first denote the APOMDP value function with  $t$  periods to go under policy  $\boldsymbol{\mu}^e$  (a model independent function) with  $V_t^{\boldsymbol{\mu}^e}$ , assume that  $V_t^{\boldsymbol{\mu}^e} \in \nu$ , and parameterize it via  $V_t^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}; \boldsymbol{\psi}_t) \triangleq (\mathbf{b}(\boldsymbol{\pi}))' \boldsymbol{\psi}_t$ . We then estimate its parameter as

$$\hat{\boldsymbol{\psi}}_t^{\boldsymbol{\mu}^e} \triangleq MEU_{\alpha}[\hat{\boldsymbol{\psi}}_t^{m, \boldsymbol{\mu}^e}] \triangleq \alpha \hat{\boldsymbol{\psi}}_t^{m, \boldsymbol{\mu}^e} + (1 - \alpha) \hat{\boldsymbol{\psi}}_t^{\bar{m}, \boldsymbol{\mu}^e}, \quad (18)$$

where  $\alpha \in \mathcal{F}$  can be viewed as a tuning parameter,  $\underline{m} \triangleq \arg \inf_{m \in \mathcal{M}} \|\hat{\boldsymbol{\psi}}_t^{m, \boldsymbol{\mu}^e}\|$ ,  $\bar{m} \triangleq \arg \sup_{m \in \mathcal{M}} \|\hat{\boldsymbol{\psi}}_t^{m, \boldsymbol{\mu}^e}\|$ ,<sup>24</sup> and

$$\hat{\boldsymbol{\psi}}_t^{\boldsymbol{\mu}^e} = \arg \min_{\boldsymbol{\psi}_t \in \Psi_t} \{ (\varphi_t^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}_t))' \Omega \varphi_t^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}_t) + \theta_t \mathcal{P}(\boldsymbol{\psi}_t) \}, \quad (19)$$

where  $\varphi_t^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}_t)$  is defined in (13). Consequently, we plug  $\hat{\boldsymbol{\psi}}_t^{\boldsymbol{\mu}^e}$  obtained in (18) in  $V_{T-t+1}^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}; \boldsymbol{\psi}_t)$ , which yields an estimate for the APOMDP value function  $V_{T-t+1}^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}_t)$ , and we move to the next period (backward) as before. This yields an estimated value function  $\hat{V}_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi})$ . Denoting the gain under any policy  $\boldsymbol{\mu}^e$  by  $\Gamma_T(\boldsymbol{\mu}^e) \triangleq \int V_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ , we use  $\hat{\Gamma}_T(\boldsymbol{\mu}^e) \triangleq \int \hat{V}_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$  as an estimator for  $\Gamma_T(\boldsymbol{\mu}^e)$ .

Finally, optimization over  $\boldsymbol{\mu}^e \in \Upsilon$  will provide the estimated optimal policy of the APOMDP under the SAV-Learning approach:  $\hat{\boldsymbol{\mu}}^{e^*} \triangleq \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_T(\boldsymbol{\mu}^e) = \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \int \hat{V}_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ . The estimated optimal gain under this approach is  $\hat{\Gamma}_T(\hat{\boldsymbol{\mu}}^{e^*}) = \max_{\boldsymbol{\mu}^e \in \Upsilon} \int \hat{V}_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ ,

which provides an estimate for  $\Gamma_T(\boldsymbol{\mu}^{e*}) \triangleq \max_{\boldsymbol{\mu}^e \in \Upsilon} \int V_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ . Similar to before, this procedure can also be used for the infinite-horizon case by noting that because both  $V_{T-t}(\cdot)$  and  $V_{T-t+1}(\cdot)$  become  $V_\infty(\cdot)$ , the calculations simplifies. The SAV-Learning approach for infinite-horizon case is presented in Algorithm 2. Besides their benefit in analyzing the long-run impact of different treatment regimes, both Algorithms 1 and 2 can also be used as *approximations* for learning policies that work well over a finite but long horizon.

#### Algorithm 2 (SAV-Learning)

- 1 for each observed trajectory and model  $m \in \mathcal{M}$  do
- 2     Initialize  $\boldsymbol{\pi}_0^m$  using a random draw from  $F(\boldsymbol{\pi})$ ;
- 3     set  $t = 1$ ;
- 4     while  $t + 1 \in \mathcal{T}$  do
- 5          $\boldsymbol{\pi}_{t+1}^m \leftarrow T(\boldsymbol{\pi}_t^m, a_t, o_t, m)$ ;
- 6     for any given  $\boldsymbol{\mu}^e \in \Upsilon$  and  $m \in \mathcal{M}$  do
- 7          $\varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) \leftarrow \mathbb{E}^{\mathbb{P}} \left[ \sum_{t \in \mathcal{T}} \left[ \frac{\mu^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu^b(A_t | \boldsymbol{\Pi}_t^m)} [G_t + \beta V_\infty^{m, \boldsymbol{\mu}^e}(T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) - V_\infty^{m, \boldsymbol{\mu}^e}(\boldsymbol{\Pi}_t^m)] \mathbf{b}(\boldsymbol{\Pi}_t^m)] \right]$ ;
- 8          $\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e} \leftarrow \arg \min_{\boldsymbol{\psi} \in \Psi} \{ (\varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}))' \Omega \varphi_n^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) + \theta_n \mathcal{P}(\boldsymbol{\psi}) \}$ ;
- 9     for any given  $\boldsymbol{\mu}^e \in \Upsilon$  do
- 10          $\underline{m} \leftarrow \arg \inf_{m \in \mathcal{M}} \|\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e}\|$ ;
- 11          $\overline{m} \leftarrow \arg \sup_{m \in \mathcal{M}} \|\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e}\|$ ;
- 12          $\hat{\boldsymbol{\psi}}_n^{\boldsymbol{\mu}^e} \leftarrow \alpha \hat{\boldsymbol{\psi}}_n^{\underline{m}, \boldsymbol{\mu}^e} + (1 - \alpha) \hat{\boldsymbol{\psi}}_n^{\overline{m}, \boldsymbol{\mu}^e}$ ;
- 13          $\hat{V}_\infty^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) \leftarrow (\mathbf{b}(\boldsymbol{\pi}))' \hat{\boldsymbol{\psi}}_n^{\boldsymbol{\mu}^e}$ ;
- 14          $\hat{\Gamma}_\infty(\boldsymbol{\mu}^e) \leftarrow \int \hat{V}_\infty^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ ;
- 15          $\hat{\boldsymbol{\mu}}^{e*} \leftarrow \arg \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_\infty(\boldsymbol{\mu}^e)$ ;
- 16          $\hat{\Gamma}_\infty(\hat{\boldsymbol{\mu}}^{e*}) \leftarrow \max_{\boldsymbol{\mu}^e \in \Upsilon} \hat{\Gamma}_\infty(\boldsymbol{\mu}^e)$ .

## 5. Performance Analyses: Theoretical Results

We now establish some theoretical results for the performance of our proposed approaches. Specifically, we demonstrate the asymptotic properties of the estimators under our main proposed algorithm, DAV-Learning (Algorithm 1). With some minor modifications, one can then also establish similar results for the estimators under the second proposed approach, SAV-Learning (Algorithm 2).<sup>25</sup>

The main results of this section are as follows. Under some conditions discussed, we first establish weak consistency and asymptotic normality of the estimators under any policy  $\boldsymbol{\mu}^e \in \Upsilon$  (Theorem 1). We then move to the estimators related to the optimal policy and establish weak consistency and asymptotic normality of both the estimated optimal policy and its estimated value (Theorem 2). To establish our results, we make use of arguments in *empirical processes* (specifically for the stationary process as opposed to i.i.d. ones) (see, e.g., Dedecker and

Louhichi 2002, Kosorok 2008) and think of each realization of the underlying stochastic process as a function in  $\ell^\infty(\Upsilon)$  (i.e., the set of real-valued bounded functions indexed by  $\boldsymbol{\mu}^e \in \Upsilon$ ).

We assume that  $\Omega$  in (17) is an arbitrary positive-definite matrix,  $\mathcal{P}(\cdot)$  is the squared norm penalty function, and  $\theta_n$  is a tuning parameter satisfying  $\theta_n = o_p(n^{-1/2})$ . We also assume that  $\mathbb{E}^m[\|\mathbf{b}(\boldsymbol{\Pi}_t)\|^2]$  and  $\mathbb{E}^m[G_t^2]$  are both finite for all  $m \in \mathcal{M}$  and  $t \in \mathcal{T}$ . Some other technical conditions are needed, mainly because of two broad sets of challenges in our setting that make establishing asymptotic results more involved; (1) the underlying process is not i.i.d over time, and (2) there is model ambiguity ( $|\mathcal{M}| \neq 1$ ). Specifically, we need the following “regularity” conditions on the parameter space, trajectories space, policy space, and model space.

**Condition 1.** For every  $\boldsymbol{\mu}^e \in \Upsilon$  and  $m \in \mathcal{M}$ , there exists a unique solution to  $\varphi^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) = 0$  denoted by  $\boldsymbol{\psi}_\diamond^{m, \boldsymbol{\mu}^e} \in \Psi \subseteq \mathbb{R}^d$ , where  $\sup_{\boldsymbol{\mu}^e \in \Upsilon} \|\boldsymbol{\psi}_\diamond^{m, \boldsymbol{\mu}^e}\| < \infty$ ,  $\boldsymbol{\psi}_\diamond^{m, \boldsymbol{\mu}^e}$  is an interior point of  $\Psi$  and  $\Psi$  is compact subset of  $\mathbb{R}^d$ .

**Condition 2.** There exists a  $2 < \rho < \infty$  such that for all  $m \in \mathcal{M}$ ,

- a. the class of policies ( $\Upsilon$ ) is either finite or its bracketing integral satisfies  $J_{[]}(\infty, \Upsilon, L_\rho(P^m)) < \infty$ , where  $P^m$  is the marginal stationary distribution of the sequence  $\{(\boldsymbol{\Pi}_t^m, A_t)\}_{t \geq 1}$ ,<sup>26</sup>
- b. the sequence  $\{(\boldsymbol{\Pi}_t^m, A_t)\}_{t \geq 1}$  is an absolutely regular stationary process with its  $\beta$ -mixing coefficients  $\zeta^m(t)$  satisfying  $\sum_{t=1}^\infty k^{2/(\rho-2)} \zeta^m(t) < \infty$ .<sup>27</sup>

**Condition 3.** There exists a constant  $c_1 > 0$  such that for all  $m \in \mathcal{M}$ ,  $t \in \mathcal{T}$ ,  $\boldsymbol{\mu}^e \in \Upsilon$ , and  $\mathbf{c} \in \mathbb{R}^d$ ,

$$\mathbf{c}' \mathbb{E}^m \left[ \begin{array}{c} \frac{\mu^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu^b(A_t | \boldsymbol{\Pi}_t^m)} \mathbf{b}(\boldsymbol{\Pi}_t^m) (\mathbf{b}(\boldsymbol{\Pi}_t^m) \\ - \beta \mathbf{b}(T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)))' \end{array} \right] \mathbf{c} \geq c_1 \|\mathbf{c}\|^2. \quad (20)$$

**Condition 4.**  $\boldsymbol{\mu}^{e*}$  is a unique and well-septated maximizer of  $\Gamma_\infty(\boldsymbol{\mu}^e)$ , and  $\boldsymbol{\mu}^{e*}$  is in the interior  $\Upsilon$ .

**Condition 5.** For every  $\boldsymbol{\mu}^e \in \Upsilon$ ,  $|\inf_{m \in \mathcal{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)| < \infty$ ,  $|\sup_{m \in \mathcal{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)| < \infty$ , and  $\mathcal{M}$  contains both  $\arg \inf_{m \in \mathcal{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)$  and  $\arg \sup_{m \in \mathcal{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)$ .

Assumptions related to these conditions are relatively common in the Z-estimation and M-estimation theories (see, e.g., Kosorok 2008). Some of these conditions are also assumed to hold in the *Generalized Method of Moments* (GMM) (for asymptotic properties of GMM, see, e.g., Hansen 1982). These conditions hold both in our case study of NODAT patients (Section 7.1) and in our simulation experiments (Section 7.2). Condition 1 is a regularity condition on the parameter space and ensures that the solutions obtained by solving  $\varphi^{m, \boldsymbol{\mu}^e}(\boldsymbol{\psi}) = 0$  are “well-behaved.” Condition 2(a) is a regularity condition on the policy space and requires that the set of policies under

consideration satisfies a minimum level of “complexity” (measured by an appropriate *entropy-based* metric). This condition clearly allows working with any finite set of policies but also holds for many infinite sets of policies (see, e.g., the parametric class of policies in Luckett et al. 2020). Condition 2(b) is a regulatory condition on the space of trajectories and allows for viewing their formation as a suitable stationary process. The  $\beta$ -mixing coefficients  $\zeta^m(t)$  quantify dependency of the observed values in the process  $t$  steps removed and are zero when there is no such dependency. Condition 3 ensures that the matrix  $\mathbf{C}^m(\boldsymbol{\mu}^e)$  defined in Theorem 1 is positive definite, and hence, invertible. One can empirically check whether Condition 3 holds by creating certain matrixes using data and testing whether they are positive definite. Condition 4 is needed to establish that the sequence of estimated optimal policies converges to the true optimal policy, which is a stronger result than just the gain of these policies converging to each other. Condition 5 is a regularity condition on the space of models,  $\mathcal{M}$ , which holds in most real-world applications because any set of models can be represented/ approximated with a finite set (with any required level of accuracy).

We first establish the asymptotic behavior of our estimators under any given policy  $\boldsymbol{\mu}^e \in \Upsilon$  by only requiring Conditions 1–3. The proof is based on some additional results provided in Online Appendix B (see Lemmas EC.3 and EC.4), which establish Donsker properties and asymptotic normality in  $\ell^\infty(\Upsilon)$  for the underlying absolutely regular stationary process in our setting.

**Theorem 1** (Asymptotic Behavior: Fixed Policy and Its Value). *Suppose Conditions 1–3 hold and the behavior policy satisfies positivity. Then, under DAV-Learning (Algorithm 1), for any  $\boldsymbol{\mu}^e \in \Upsilon$  and  $m \in \mathcal{M}$ , we have*

- i.  $\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e} \xrightarrow{P} \boldsymbol{\psi}_{\diamond}^{m, \boldsymbol{\mu}^e}$ .
- ii.  $\sqrt{n} [\hat{\boldsymbol{\psi}}_n^{m, \boldsymbol{\mu}^e} - \boldsymbol{\psi}_{\diamond}^{m, \boldsymbol{\mu}^e}] \xrightarrow{d} \mathbb{G}(\boldsymbol{\mu})$  in  $\ell^\infty(\Upsilon)$ , where  $\mathbb{G}(\boldsymbol{\mu})$  is a zero-mean and tight Gaussian process indexed by  $\boldsymbol{\mu} \in \Upsilon$  with the covariance function given by

$$\mathbb{E}[\mathbb{G}(\boldsymbol{\mu})\mathbb{G}(\tilde{\boldsymbol{\mu}})] = (\mathbf{C}^m(\boldsymbol{\mu}^e))^{-1} \tilde{\mathbf{C}}^m(\boldsymbol{\mu}^e, \tilde{\boldsymbol{\mu}}^e) ((\mathbf{C}^m(\boldsymbol{\mu}^e))^{-1})' \quad \forall \boldsymbol{\mu}, \tilde{\boldsymbol{\mu}} \in \Upsilon, \quad (21)$$

where

$$\mathbf{C}^m(\boldsymbol{\mu}^e) \triangleq \mathbb{E}^m \left[ \begin{array}{c} \frac{\mu^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu^b(A_t | \boldsymbol{\Pi}_t^m)} \mathbf{b}(\boldsymbol{\Pi}_t^m) (\mathbf{b}(\boldsymbol{\Pi}_t^m))' \\ - \beta \mathbf{b}(T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) \end{array} \right], \quad (22)$$

$$\tilde{\mathbf{C}}^m(\boldsymbol{\mu}^e, \tilde{\boldsymbol{\mu}}^e) \triangleq \mathbb{E}^m \left[ \begin{array}{c} \frac{\mu^e(A_t | \boldsymbol{\Pi}_t^m) \tilde{\mu}^e(A_t | \boldsymbol{\Pi}_t^m)}{\mu^b(A_t | \boldsymbol{\Pi}_t^m) \mu^b(A_t | \boldsymbol{\Pi}_t^m)} \boldsymbol{\theta}(\boldsymbol{\Pi}_t^m, \boldsymbol{\psi}_{\diamond}^{m, \boldsymbol{\mu}^e}) \\ \boldsymbol{\theta}(\boldsymbol{\Pi}_t^m, \boldsymbol{\psi}_{\diamond}^{\tilde{\boldsymbol{\mu}}^e}) \mathbf{b}(\boldsymbol{\Pi}_t^m) (\mathbf{b}(\boldsymbol{\Pi}_t^m))' \end{array} \right], \quad (23)$$

and

$$\boldsymbol{\theta}(\boldsymbol{\Pi}_t^m, \boldsymbol{\psi}_{\diamond}^{m, \boldsymbol{\mu}^e}) \triangleq G_t + [\beta \mathbf{b}(T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) - \mathbf{b}(\boldsymbol{\Pi}_t^m)] \boldsymbol{\psi}_{\diamond}^{m, \boldsymbol{\mu}^e}. \quad (24)$$

$$\text{iii. } \hat{\Gamma}_{\infty}^m(\boldsymbol{\mu}^e) \xrightarrow{P} \Gamma_{\infty}^m(\boldsymbol{\mu}^e).$$

$$\text{iv. } \hat{\Gamma}_{\infty}(\boldsymbol{\mu}^e) \xrightarrow{P} \Gamma_{\infty}(\boldsymbol{\mu}^e) \text{ assuming Condition 5 holds.}$$

We next establish the asymptotic properties of the optimal policy and the gain under it. The proof of the following theorem is based on an additional result provided in Online Appendix B (see Lemma EC.5), which in turn, relies on results from the  $M$ -estimation theory.

**Theorem 2** (Asymptotic Behavior: Optimal Policy and Its Value). *Suppose Conditions 1–5 hold and the behavior policy satisfies positivity. Then, considering a metric space  $(\Upsilon, d_{\Upsilon})$ , under DAV-Learning (Algorithm 1) we have*

$$\text{i. } d_{\Upsilon}(\hat{\boldsymbol{\mu}}^{e*, m}, \boldsymbol{\mu}^{e*, m}) \xrightarrow{P} 0 \text{ for all } m \in \mathcal{M}.$$

$$\text{ii. } d_{\Upsilon}(\hat{\boldsymbol{\mu}}^{e*}, \boldsymbol{\mu}^{e*}) \xrightarrow{P} 0.$$

$$\text{iii. } \hat{\Gamma}_{\infty}^m(\hat{\boldsymbol{\mu}}^{e*, m}) \xrightarrow{P} \Gamma_{\infty}^m(\boldsymbol{\mu}^{e*, m}).$$

$$\text{iv. } \hat{\Gamma}_{\infty}(\hat{\boldsymbol{\mu}}^{e*}) \xrightarrow{P} \Gamma_{\infty}(\boldsymbol{\mu}^{e*}).$$

## 6. Extension: Learning Under BUC

As discussed in Remark 2, in various applications, one might be able to ensure that  $\boldsymbol{\mu}^b$  satisfies sequential ignorability because it takes advantage of the availability of information about the unobserved variables (confounders) summarized in  $(\boldsymbol{\pi}_t)_{t \in \mathcal{T}}$ . The DAV-Learning and SAV-Learning approaches introduced earlier are based on the results of Proposition 3, which assumes that  $\boldsymbol{\mu}^b$  satisfies sequential ignorability. In this section, we show how such results can be extended to cases where  $\boldsymbol{\mu}^b$  does not satisfy sequential ignorability, but satisfies the BUC conditions introduced earlier (see, e.g., Definition 5). This, in turn, allows us to extend DAV-Learning and SAV-Learning and introduce their BUC counterparts, which we term DAV-Learning-BUC and SAV-Learning-BUC, respectively.

Similar to (7), which is based on Lemma EC.2 (Online Appendix B), assume for each model  $m \in \mathcal{M}$  that there exist constants  $\eta_t^m \in [1, \infty)$  such that

$$(\eta_t^m)^{-1} \leq \frac{\mu_t^b(a_t | \boldsymbol{\Pi}_t^m, \mathbf{S}_t^m)}{\mu_t^b(a_t | \boldsymbol{\Pi}_t^m)} \leq \eta_t^m \text{ a.s.} \quad (25)$$

over  $\boldsymbol{\Pi}_t^m$  and  $\mathbf{S}_t^m$  for all  $t \in \mathcal{T}$  and  $a \in \mathcal{A}$ . If  $\eta_t^m = 1$ ,  $\mu_t^b$  satisfies sequential ignorability. Furthermore, with  $\eta_t^m = 1$ , (25) implies that benefiting from  $\boldsymbol{\Pi}_t^m$  under each model  $m$  and making use of marginalized treatment propensities  $\mu_t^b(a_t | \boldsymbol{\Pi}_t^m)$  are enough for the goal of estimating the true treatment propensities  $\mu_t^b(a_t | \boldsymbol{\Pi}_t^m, \mathbf{S}_t^m)$ . More broadly, however, (25) ensures that this estimation exercise is not unboundedly misleading. Of note, because  $\mu_t^b(a_t | \boldsymbol{\Pi}_t^m) = \mathbb{E}_{\mathbf{S}_t^m}^m[\mu_t^b(a_t | \boldsymbol{\Pi}_t^m, \mathbf{S}_t^m)]$ , one can also view (25) as bounded variations from the average (under each model).

Assuming that (25) holds, we next extend Proposition 3 by relaxing the assumption that  $\boldsymbol{\mu}^b$  satisfies sequential ignorability. To this end, we introduce the following *weight modifiers*:

$$\underline{\kappa}_t^m \triangleq \left( (\eta_t^m)^{-1} \mathbb{1}_{\{V_T^{\boldsymbol{\mu}^b}(\boldsymbol{\pi}) \geq 0\}} + \eta_t^m \mathbb{1}_{\{V_T^{\boldsymbol{\mu}^b}(\boldsymbol{\pi}) < 0\}} \right) \quad (26)$$

and

$$\bar{\kappa}_t^m \triangleq \left( (\eta_t^m)^{-1} \mathbb{1}_{\{V_T^{\boldsymbol{\mu}^b}(\boldsymbol{\pi}) < 0\}} + \eta_t^m \mathbb{1}_{\{V_T^{\boldsymbol{\mu}^b}(\boldsymbol{\pi}) \geq 0\}} \right), \quad (27)$$

and we make use of the BUC results presented in Proposition 2.<sup>28</sup>

**Proposition 4** (Weight-Adjusted Bellman Equation Under BUC). *Suppose  $\boldsymbol{\mu}^b$  satisfies both the BUC condition (25) and positivity. For any policy  $\boldsymbol{\mu}^e$ , define the upper bound and lower bound value functions via the modified weight-adjusted Bellman equations:*

$$\bar{V}_{T-t+1}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}_t) \triangleq \mathbb{E}^m \left[ \bar{\kappa}_t^m \frac{\mu_t^e(A_t | \mathbf{\Pi}_t^m)}{\mu_t^b(A_t | \mathbf{\Pi}_t^m)} \left[ G_t + \beta \bar{V}_{T-t}^{m, \boldsymbol{\mu}^e} \left( T(\mathbf{\Pi}_t^m, A_t, O_t, m) \right) \right] \middle| \mathbf{\Pi}_t^m = \boldsymbol{\pi}_t \right] \quad (28)$$

and

$$\underline{V}_{T-t+1}^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}_t) \triangleq \mathbb{E}^m \left[ \underline{\kappa}_t^m \frac{\mu_t^e(A_t | \mathbf{\Pi}_t^m)}{\mu_t^b(A_t | \mathbf{\Pi}_t^m)} \left[ G_t + \beta \underline{V}_{T-t}^{m, \boldsymbol{\mu}^e} \left( T(\mathbf{\Pi}_t^m, A_t, O_t, m) \right) \right] \middle| \mathbf{\Pi}_t^m = \boldsymbol{\pi}_t \right], \quad (29)$$

along with  $\bar{V}_0^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) \triangleq 0$  and  $\underline{V}_0^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) \triangleq 0$ . Then:

i. For any function  $\phi$  defined on  $\Delta_{\mathcal{G}}$  and for all  $t \in \mathcal{T}$  and  $m \in \mathcal{M}$ , we have

$$\mathbb{E}^m \left[ \bar{\kappa}_t^m \frac{\mu_t^e(A_t | \mathbf{\Pi}_t^m)}{\mu_t^b(A_t | \mathbf{\Pi}_t^m)} \left[ G_t + \beta \bar{V}_{T-t}^{m, \boldsymbol{\mu}^e} \left( T(\mathbf{\Pi}_t^m, A_t, O_t, m) \right) - \bar{V}_{T-t+1}^{m, \boldsymbol{\mu}^e}(\mathbf{\Pi}_t^m) \right] \phi(\mathbf{\Pi}_t^m) \right] = 0 \quad (30)$$

and

$$\mathbb{E}^m \left[ \underline{\kappa}_t^m \frac{\mu_t^e(A_t | \mathbf{\Pi}_t^m)}{\mu_t^b(A_t | \mathbf{\Pi}_t^m)} \left[ G_t + \beta \underline{V}_{T-t}^{m, \boldsymbol{\mu}^e} \left( T(\mathbf{\Pi}_t^m, A_t, O_t, m) \right) - \underline{V}_{T-t+1}^{m, \boldsymbol{\mu}^e}(\mathbf{\Pi}_t^m) \right] \phi(\mathbf{\Pi}_t^m) \right] = 0. \quad (31)$$

ii. For all  $m \in \mathcal{M}$  and  $\boldsymbol{\pi} \in \Delta_{\mathcal{G}}$ , we have  $\underline{V}_T^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi}) \leq \bar{V}_T^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi})$ .

iii. For any  $\alpha \in \mathcal{F}$ , there exists  $\tilde{\alpha} \in \mathcal{F}$  such that  $V_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) \triangleq MEU_{\alpha}[V_T^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi})] = f(\tilde{\alpha}, \boldsymbol{\pi})$ , where  $f(\tilde{\alpha}, \boldsymbol{\pi}) \triangleq \tilde{\alpha} MEU_{\alpha}[\underline{V}_T^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi})] + (1 - \tilde{\alpha}) MEU_{\alpha}[\bar{V}_T^{m, \boldsymbol{\mu}^e}(\boldsymbol{\pi})]$ . Hence,  $\Gamma_T(\boldsymbol{\mu}^e) \triangleq \int f(\tilde{\alpha}, \boldsymbol{\pi}) dF(\boldsymbol{\pi})$ , where  $\Gamma_T(\boldsymbol{\mu}^e) \triangleq \int V_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) dF(\boldsymbol{\pi})$ .

When the behavior policy  $\boldsymbol{\mu}^b$  satisfies sequential ignorability, we have  $\underline{\kappa}_t^m = \bar{\kappa}_t^m = 1$ , and hence, the results

boil down to those in Proposition 3. Proposition 4, however, generalizes Proposition 3 by highlighting the role of weight modifiers  $\underline{\kappa}_t^m$  and  $\bar{\kappa}_t^m$  in analyzing scenarios in which  $\boldsymbol{\mu}^b$  violates sequential ignorability but satisfies it to *some extent*. In particular, after incorporating these modifiers, one can make use of the same procedures as in the previous section using Proposition 4. That is, the sample-average version of (30) and (31) together with regularized learning can be used to learn the upper bound and lower bound value functions in a parametric way, respectively. Once these functions are learned, part (ii) of Proposition 4 guarantees that they can be used to bound the actual value function under each model. More specifically, part (iii) of Proposition 4 states that  $f(\tilde{\alpha}, \boldsymbol{\pi})$  and  $\int f(\tilde{\alpha}, \boldsymbol{\pi}) dF(\boldsymbol{\pi})$  can be used to estimate the APOMDP value function and the overall performance, respectively. It should be noted that (1) the function  $f(\cdot, \cdot)$  is calculable using only the observed part of the data. This resolves the issue that the outcome of interest under the evaluation policy as well as the time-varying confounders needed to estimate it are unobservable. (2) A similar procedure to that discussed after Proposition 2 can be used to tune parameter  $\tilde{\alpha}$ . Specifically, as discussed there, noting that the parameters  $(\eta_t^m)_{t \in \mathcal{T}, m \in \mathcal{M}}$  can be viewed as *design sensitivity parameters*, one can choose them and approximate  $\tilde{\alpha}$  so as to obtain an approximate unbiased  $MEU_{\alpha}$  estimator for  $\Gamma_T(\boldsymbol{\mu}^e)$  with any desired approximation error  $\epsilon > 0$ .

Finally, the results allow us to extend DAV-Learning and SAV-Learning and to introduce their BUC counterparts. These extensions, which we term DAV-Learning-BUC and SAV-Learning-BUC, are presented in Algorithms 3 and 4, respectively, in Online Appendix C. The main difference between these extensions and their original version is that they benefit from weight modifiers (26) and (27) to first obtain estimators for the upper bound and lower bound value functions (as opposed to the main value function itself).

## 7. Performance Analyses: Numerical Results

To gain further insights into the performance of our proposed algorithms, we now perform two sets of numerical experiments. The first is a case study of a medical decision-making problem faced by physicians at our partner hospital and involves using a clinical data set of patients with a kidney transplant operation. In the second set, we make use of synthetic data, in which we simulate patient trajectories under different models while controlling the true data-generating model.

### 7.1. Case Study: NODAT

In this section, we apply our proposed algorithms on a clinical data set that contains over 63,000 data points pertaining to 407 patients who had a kidney transplant

operation during a seven-year period at our partner hospital. Details about the data set can be found in the author’s previous publications (Bolori et al. 2015, 2020; Munshi et al. 2020a, 2021).

Patients who undergo transplantation often face a significant risk of organ rejection. To mitigate this risk, physicians typically use an intensive amount of an immunosuppressive drug (e.g., tacrolimus). Immunosuppressive drugs, however, have a well-established effect known as the diabetogenic effect, and thus, they can elevate the risk of *New Onset Diabetes After Transplantation* (NODAT). NODAT refers to incidence of diabetes in a patient with no history of diabetes prior to transplantation (see, e.g., Chakkerla et al. 2009; Bolori et al. 2015, 2020; and the references therein). To control the risk of NODAT, physicians have to decide whether to put the patient on insulin.<sup>29</sup>

Table 1 describes the observed patient covariates (observations) and their levels. As the table shows, some of these observations are time-varying. Furthermore, most of them are dichotomized to high- versus low-level values. However, the medical tests used to measure the blood glucose (*Fasting Plasma Glucose* (FPG) and *Hemoglobin A1c* (HbA1c)) and the lowest concentration of tacrolimus in the patient’s body—a quantity known as *trough level* or  $C_0$ —have three levels. These levels are defined based on both the medical literature and the practice at our partner hospital. Tables 2 and 3 show the patients’ latent states and physicians’ actions/prescriptions during each visit post-transplant, respectively. Latent states described in Table 2 are summary variables that represent the main condition of the patient in terms of decision-making related to the use of an immunosuppressive drug (e.g., tacrolimus) and insulin

therapy (i.e., the actions in Table 3). These patient summary variables are, however, hidden to physicians because physicians can only rely on medical tests, which have a wide range of false-positive and false-negative errors. In particular, blood glucose levels are measured by two medical tests, FPG and HbA1c, which are subject to false-positive and false-negative errors. Similarly, the concentration of immunosuppressive drugs is measured through tests such as *Abbott Architect* and *Magnetic Immunoassay*, which are error prone.

**7.1.1. Data Preprocessing Steps.** Our data set includes information related to patients’ follow-up visits during months 1, 4, and 12 post-transplantation. However, for the goals of this study, we make use of the same data preprocessing steps as those in Bolori et al. (2020). In particular, we use imputation to replace missing values (see also Munshi et al. 2021) and also make use of cubic spline interpolation to create a test bed with clinical history of patients for months 1–12 after transplant. That is, for the purpose of this study, we consider monthly visits that occur for a year post-transplant. Thus, we let  $T \triangleq 12$  and  $\mathcal{T} \triangleq \{1, 2, \dots, 12\}$ . The imputed data include the 13 variables listed in Table 1 for each of the 407 patients and every month during a year of follow-up post-transplant (a total of  $13 \times 407 \times 12 = 63,492$  data points).

**7.1.2. Behavior Policy.** We estimate the behavior policy based on the actions we observe in our data. These actions are mainly based on the clinical protocols followed at our partner hospital. A detailed summary of the main immunosuppression protocol can be found in Munshi et al. (2021), which includes induction therapy with either rabbit antithymocyte, immunoglobulin, or basiliximab as well as

**Table 1.** Observed Covariates (Observations)

Variate no.	Risk factor (abbreviation)	Unit	Low level	Midlevel	High level	Time varying
1	Glucose test <sup>a</sup> (FPG, HbA1c)	mg/dL, %	Healthy	Prediabetic	Diabetic	Yes
2	Trough-level test <sup>b</sup> ( $C_0$ )	mg/dL	[4, 8)	[8, 10)	[10, 14]	Yes
3	Age	Years	<50	—	≥50	No
4	Gender	—	Female	—	Male	No
5	Race	—	White	—	Non-White	No
6	Diabetes history (Diab Hist)	—	No	—	Yes	No
7	Body mass index (BMI)	kg/m <sup>2</sup>	<30 (nonobese)	—	≥30 (obese)	Yes
8	Blood pressure (BP)	—	Normal <sup>c</sup>	—	Hypertension	Yes
9	Total cholesterol (Chol)	mg/dL	<200	—	≥200	Yes
10	High-density lipoprotein (HDL)	mg/dL	≥40	—	<40	Yes
11	Low-density lipoprotein (LDL)	mg/dL	<130	—	≥130	Yes
12	Triglyceride (TG)	mg/dL	<150	—	≥150	Yes
13	Uric acid (UA)	mg/dL	<7.3	—	≥7.3	Yes

*Notes.* All variables with three levels are coded as 1, 2, 3 (low, mid, high, respectively). All variables with two levels are coded as 1, 2 (low, high, respectively).

<sup>a</sup>A patient with  $FPG \geq 126$  ( $100 \leq FPG < 126$ ) mg/dL or  $HbA1c \geq 6.5\%$  ( $5.7 \leq HbA1c < 6.5\%$ ) is labeled as diabetic (prediabetic), and a patient with  $FPG < 100$  mg/dL or  $HbA1c < 5.7\%$  is labeled as healthy (see, e.g., ADA 2012).

<sup>b</sup> $C_0 \in [4, 8)$ ,  $[8, 10)$ ,  $[10, 14]$  mg/dL is labeled as “low,” “medium,” and “high,” respectively (see, e.g., Bolori et al. 2020).

<sup>c</sup>Normal blood pressure is defined as systolic (diastolic) blood pressure less than 120 (80) mmHg (see, e.g., Whelton et al. 2017).

**Table 2.** Latent Health States

State	Transplant condition (tacrolimus $C_0$ )	Diabetes condition
1	Low	Diabetes (type II)
2	Medium	
3	High	
4	Low	Prediabetes
5	Medium	
6	High	
7	Low	Healthy
8	Medium	
9	High	

a tapering course of glucocorticoids. However, here our focus is on the use of tacrolimus, and we observe that patients are often put on high-dose (i.e., aggressive) tacrolimus during the first months post-transplant, and in later months, depending on the observations made about the patients, they might be transferred to a low (i.e., nonaggressive) dose. This is consistent with the fact that patients in most medical practices are consistently kept on high levels of tacrolimus in the early stages post-transplant (see, e.g., Ghisdal et al. 2012, Boloori et al. 2020). Furthermore, with respect to the use of insulin, patients are primarily put on insulin when their HbA1c and FPG tests indicate that they are not diabetic free (see definitions of prediabetic and diabetic in Table 1). Using the observed actions in our data set as well as the estimated belief vectors  $\{\pi_t^m\}_{t \in \mathcal{T}}$  for each patient (for further details, see Section 7.1.4), we next estimate  $\mu^b(A_t | \Pi_t^m)$  by training a multiclass multiple logistic regression classifier. This classifier is endowed with an  $\ell_2$ -norm penalty, which is tuned to ensure that each action is selected with an estimated probability of 0.05 or higher across all observations (see, e.g., Murphy et al. 2016).

**7.1.3. Immediate Gain Variable.** To calculate the immediate gains, we use a similar approach to our previous work (see, e.g., Boloori et al. 2020). In particular, we make use of *QoL scores*, which take values in  $[0, 1]$ . This allows us to differentiate between the quality of life of being in a diabetic, prediabetic, or healthy state and also having different concentrations of the immunosuppressive drug in the body, which are, in turn, associated with differing risks of organ rejection. Table 4 shows the yearly-based QoL scores associated with each state,<sup>30</sup> which are divided by 12 to represent the fact that patients’ visits are monthly.<sup>31</sup>

**Table 3.** Actions

Action	Prescription (tacrolimus dose)	Prescription (insulin use)
1	Low (nonaggressive)	No
2	High (aggressive)	No
3	Low (nonaggressive)	Yes
4	High (aggressive)	Yes

**Table 4.** Immediate Gain Values

State	Transplant condition (tacrolimus $C_0$ )	Diabetes condition	Immediate gain value <sup>a</sup>
1	Low	Diabetes (type II)	0.68/12
2	Medium		0.72/12
3	High		0.76/12
4	Low	Prediabetes	0.82/12
5	Medium		0.87/12
6	High		0.89/12
7	Low	Healthy	0.90/12
8	Medium		0.92/12
9	High		0.95/12

<sup>a</sup>Immediate gains are average values approximated based on *QoL* scores reported in other studies and include combined disutility of (a) being in a diabetic state and (b) having high risk of organ rejection. Yearly-based values are divided by 12 to represent monthly measures.

**7.1.4. Other Details.** The belief state space in our setting,  $\Delta_{\mathcal{S}}$ , is an 8-simplex because there are nine latent states (Table 2). The vector of basis functions  $\mathbf{b}(\pi)$  maps this 8-simplex to  $\mathbb{R}^{13}$ , which allows us to include enough cut points (while making sure that the value function is piecewise linear and continuous). Thus, both the belief space and the parameter space in our setting are continuous and relatively high dimensional. To perform our analyses, we use a discount factor of  $\beta = 0.95$ . We also tune a penalty parameter  $\theta_t = \theta$ . To create the set of models  $\mathcal{M}$ , we make use of the algorithm in Table 3 of our earlier work (Boloori et al. 2020). Specifically, first the Baum–Welch algorithm is used to obtain point estimations for state transition and observation probability matrices. Next, an entropy ball is constructed (using the Kullback–Leibler divergence criterion) around these point estimate matrices. For tractability, we set  $|\mathcal{M}| = 4$  in this case study. However, our framework is general and can be used for any number of estimated models. In Section 7.2, for example, we change our assumption on the number of models and consider  $|\mathcal{M}| = 10$  different models. Our framework is also not restricted to any specific way of estimating the underlying models. For example, in Section 7.2, we make use of a different way of constructing the set  $\mathcal{M}$ .<sup>32</sup> Finally, we consider the distribution  $F(\pi)$  to be uniform. That is, we use a uniform prior belief at time 0 and implement the Bayesian belief updating operator (see Equation 9) to create a sequence of belief vectors  $\{\pi_t^m\}_{t \in \mathcal{T}}$  for each patient under each model  $m \in \mathcal{M}$  (see, e.g., steps 1–5 in Algorithms 1 and 2 and in steps 1–5 Algorithms 3 and 4 in Online Appendix C).

**7.1.5. Results.** The performances of the three treatment regimes (DAV-Learning, SAV-Learning, and observed) are compared in Table 5. Average and standard deviations in these tables are calculated using Monte Carlo replications.<sup>33</sup> We focus on the performance of DAV-Learning

**Table 5.** Estimated Total Discounted Gain Under Observed and Proposed Regimes (Case Study with  $\beta = 0.95$ )

Pessimism level ( $\alpha$ )	Observed regime <sup>a</sup>	DAV-Learning <sup>a</sup>	SAV-Learning <sup>a</sup>
0.00	1.472 (1.455, 1.489)	<b>2.085</b> (2.061, 2.108)	1.949 (1.770, 2.128)
0.25	1.468 (1.456, 1.480)	<b>1.939</b> (1.920, 1.958)	1.888 (1.566, 2.210)
0.50	1.464 (1.457, 1.471)	<b>1.794</b> (1.779, 1.808)	1.786 (1.534, 2.039)
0.75	1.460 (1.458, 1.462)	1.648 (1.638, 1.658)	<b>1.682</b> (1.658, 1.706)
1.00	1.455 (1.452, 1.458)	<b>1.609</b> (1.560, 1.657)	1.606 (1.585, 1.627)

Note. For all values, only the first three decimal places are shown.

<sup>a</sup>Values in parentheses represent 95% confidence intervals. Values in bold represent the best performance.

and SAV-Learning as opposed to their BUC extensions (DAV-Learning-BUC, SAV-Learning-BUC) because as discussed in Remark 2, the results of Proposition 3 hold in this application. In the next section, we run experiments using DAV-Learning and SAV-Learning as well as DAV-Learning-BUC and SAV-Learning-BUC.

As can be seen from the results in Table 5, DAV-Learning outperforms SAV-Learning in terms of the mean performance for most values of the pessimism level,  $\alpha$ . As both Table 5 and Figure 2 show, however, both DAV-Learning and SAV-Learning approaches significantly outperform the observed regime. In particular, as Figure 2 shows, the improvements over the observed regime when using DAV-Learning and SAV-Learning are in the ranges (10%, 42%) and (10%, 32%), respectively, depending the value of  $\alpha$ . Of note, these ranges also imply that the mean performance of the SAV-Learning regime is much more robust to the value of  $\alpha$  than that of DAV-Learning. This is because of the fact that SAV-Learning uses a “safe estimation” of the underlying parameter of the value function (see, e.g., step 12 of Algorithm 2). This allows SAV-Learning to guard against ambiguity up front (i.e., in parameter estimation) in contrast to DAV-Learning, which combines policy values at the end. Thus, a decisions maker

who uses SAV-Learning does not need to be that concerned about the value of  $\alpha$  he or she uses (or try to tune it).

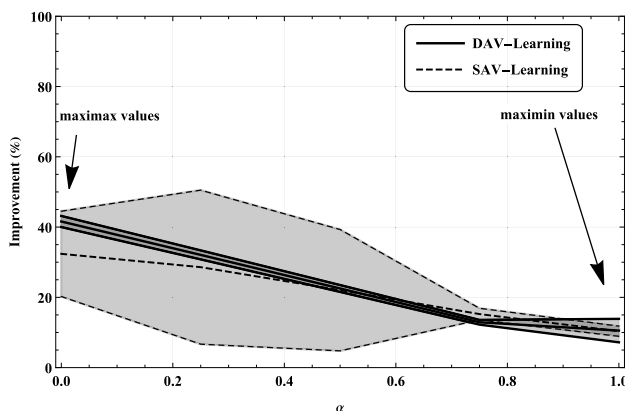
Finally, as can be seen from both Table 5 and Figure 2, the performance of DAV-Learning and SAV-Learning regimes degrades as the pessimism level  $\alpha$  increases. This is fully expected because as we move from a maximax view to a maximin one, DAV-Learning and SAV-Learning tend to put more weight on the worst-case scenario, and hence, perform more *conservatively*. More conservativeness, however, does not necessarily mean more *robustness* to model ambiguity. We further investigate this issue in Section 7.3, and we generate important insights into the values of  $\alpha$  that can provide the highest level of robustness to model ambiguity.

## 7.2. Synthetic Data Analyses

We now use similar assumptions to those described in the case study, but instead of using actual patient trajectories, we simulate random patient trajectories for 100 patients with 10 follow-up periods and use ( $|\mathcal{M}| = 10$ ) different models. These yield randomly generated belief data of the form  $(\pi_i^m)_{i \in \mathcal{I}}$  under each  $m \in \mathcal{M}$ . We keep the other assumptions (e.g., the action space, the number of hidden states, the parameter space, basis functions, etc.) the same as those in the previous section.

We assume patient trajectories are such that for each  $m \in \mathcal{M}$ , the belief vector  $(\pi_i^m)_{i \in \mathcal{I}}$  is generated via a Dirichlet distribution with the vector of parameters  $(p_i^m)_{i \in \{1, 2, \dots, 9\}}$ . All of these models are misspecified, and hence, for each model, we randomly draw each  $p_i^m$  from a Uniform(0, 1) distribution. We assume the true model is such that all  $p_i$ -values are equal to 0.5. Furthermore, we specify the behavior policy as follows. For actions  $a = 1, 2, 3$ , we set  $\mu^b(A = a | \Pi = \pi) = \frac{\exp(\pi \cdot q_a)}{1 + \sum_{a=1}^3 \exp(\pi \cdot q_a)}$ , and for action  $a = 4$ , we set  $\mu^b(A = a | \Pi = \pi) = \frac{1}{1 + \sum_{a=1}^3 \exp(\pi \cdot q_a)}$ , where  $q_1, q_2$ , and  $q_3$  are nine-dimensional predefined vectors. To perform our analyses, we choose each  $q_a$  ( $a = 1, 2, 3$ ) as a vector with all elements equal to 0.1, except the  $a$ th element, which is set to  $-1$ .

Table 6 and Figure EC.1 (Online Appendix D) present our results using the DAV-Learning and SAV-Learning approaches under the same immediate gain values

**Figure 2.** Percentage Improvement over the Observed Regime (Case Study with  $\beta = 0.95$ )

Note. Gray areas represent error bands, with the curve at the center of each error band representing the mean value.

**Table 6.** Estimated Total Discounted Gain Under Observed and Proposed Regimes (Synthetic Data Analyses with  $\beta = 0.95$ )

Pessimism level ( $\alpha$ )	Observed regime <sup>a</sup>	DAV-Learning <sup>a</sup>	SAV-Learning <sup>a</sup>
0.00	1.441 (1.440, 1.442)	<b>1.973</b> (1.969, 1.977)	1.442 (1.441, 1.442)
0.25	1.415 (1.415, 1.416)	<b>1.815</b> (1.811, 1.818)	1.434 (1.433, 1.434)
0.50	1.389 (1.389, 1.390)	<b>1.656</b> (1.654, 1.659)	1.428 (1.428, 1.429)
0.75	1.364 (1.364, 1.364)	<b>1.498</b> (1.496, 1.499)	1.434 (1.434, 1.434)
1.00	1.338 (1.338, 1.339)	1.348 (1.348, 1.348)	<b>1.444</b> (1.444, 1.444)

Note. For all values, only the first three decimal places are shown.

<sup>a</sup>Values in parentheses represent 95% confidence intervals. Values in bold represent the best performance.

as those in the case study (see Table 4). Similar results for the BUC versions of these approaches, DAV-Learning-BUC and SAV-Learning-BUC, are presented in Table 7 and Figure EC.2 (Online Appendix D). Similar to the case study, we observe from these results that all of our proposed algorithms outperform the observed regime. Furthermore, the percentage improvement of DAV-Learning and SAV-Learning over the observed regime ranges in (1%, 37%) and (1%, 8%), respectively, depending on the value of  $\alpha$ . These improvement ranges for DAV-Learning-BUC and SAV-Learning-BUC are (0.7%, 4.9%) and (0.07%, 7.9%), respectively. In addition, similar to our observation in the case study, DAV-Learning outperforms SAV-Learning for most values of  $\alpha$  (Figure EC.1 in Online Appendix D), and a similar result can be seen for the BUC regimes (Figure EC.2 in Online Appendix D). Furthermore, similar to the case study, the performance of SAV-Learning is much more robust to the value of  $\alpha$  compared with DAV-Learning, but in the BUC regimes, DAV-Learning-BUC shows relatively more robustness than SAV-Learning-BUC. In the next section, we further investigate the robustness of our proposed approaches to model ambiguity, and generate insights into the best value of  $\alpha$  that a decision-maker can use to achieve the highest level of robustness.

### 7.3. Robustness to Model Ambiguity

We now compare our proposed approaches in terms of their percentage *gain loss* (also known as *regret*). That is, we first consider an *oracle* who knows both the true

data-generating model and the optimal policy under it, and then, we compare the performance of a decision-maker who is blind to the true data-generating model (is facing model ambiguity) but uses either DAV-Learning or SAV-Learning (or the BUC versions of DAV-Learning or SAV-Learning). How much robustness to model ambiguity using these proposed approaches provide? What is the maximum gain loss of these approaches? For what value of  $\alpha$  is the gain loss minimized? Importantly, in order to minimize the gain loss, should the decision maker use an extreme value of  $\alpha$  (e.g.,  $\alpha = 0, 1$ ) or a middle-level value (e.g.,  $\alpha = 0.5$ )? Additionally, does the answer depend on which learning approach is used?

To answer these questions, we make use of a similar setup to the one discussed in Section 7.2. The results are shown in Figure EC.3 (Online Appendix D), which depicts the percentage gain loss of DAV-Learning and SAV-Learning compared with the imaginary oracle. Similar results for the BUC version of these approaches (DAV-Learning-BUC or SAV-Learning-BUC) are provided in Figure EC.4 (Online Appendix D). From these figures, we make three main observations. (1) Gain loss has a U-shaped curve as  $\alpha$  varies. Importantly, the minimum loss for all four approaches (DAV-Learning, SAV-Learning, DAV-Learning, SAV-Learning) is obtained at a midvalue of  $\alpha$  (approximately  $\alpha = 0.25$ ), which implies that using extreme cases of  $\alpha = 0.0$  (a maximax view) or  $\alpha = 1.0$  (a maximin view) does not provide the highest level of robustness to model ambiguity. That is, neither the maximax view nor the

**Table 7.** Estimated Total Discounted Gain Under Observed and Proposed BUC Regimes (Synthetic Data Analyses with  $\beta = 0.95$ )

Pessimism level ( $\alpha$ )	Observed regime <sup>a</sup>	DAV-Learning-BUC <sup>a,b</sup>	SAV-Learning-BUC <sup>a,b</sup>
0.00	1.441 (1.440, 1.442)	<b>1.511</b> (1.506, 1.517)	1.448 (1.442, 1.447)
0.25	1.415 (1.415, 1.416)	<b>1.468</b> (1.464, 1.472)	1.429 (1.428, 1.429)
0.50	1.389 (1.389, 1.390)	<b>1.425</b> (1.422, 1.428)	1.419 (1.418, 1.420)
0.75	1.364 (1.364, 1.364)	1.382 (1.380, 1.384)	<b>1.418</b> (1.416, 1.420)
1.00	1.338 (1.338, 1.339)	1.344 (1.344, 1.345)	<b>1.444</b> (1.441, 1.447)

Note. For all values, only the first three decimal places are shown.

<sup>a</sup>Values in parentheses represent 95% confidence intervals. Values in bold represent the best performance.

<sup>b</sup>The algorithm is run by assuming  $\eta_t^m = 1.02$  for all  $t \in \mathcal{T}$  and  $m \in \mathcal{M}$ . Results are based on  $\epsilon$ -approximations for a small  $\epsilon$  (see the discussion in Section 6).

maximin view is *robustness maximizing*. (2) The gain loss under SAV-Learning (SAV-Learning-BUC) is more robust to the changes in the value of  $\alpha$  compared with DAV-Learning (DAV-Learning-BUC). (3) All four proposed approaches are able to strongly shield against model ambiguity, regardless of the value of  $\alpha$  used. Specifically, the gain loss under these approaches (compared with the imaginary oracle) is very low (below 0.6%). This implies that a decision-maker who is facing model ambiguity can use these approaches and obtain policies that have similar performance to the very best policy that could be used if the true data generating was known (i.e., if there was no ambiguity regarding the underlying causal model).<sup>34</sup>

**Remark 4** (Dimensionality and Computations). The proposed algorithms DAV-Learning and SAV-Learning as well as their BUC extensions, DAV-Learning-BUC and SAV-Learning-BUC, do not require their main parameter spaces to have low dimensionality. That is, they can safely be used in high dimensions, especially because they make use of regularization to avoid overfitting in high-dimensional settings. In particular, both our theoretical and numerical performance results (Sections 5 and 7) indicate that these algorithms have suitable convergence results and are relatively tractable for use in real-world applications. For example, although the results in Section 5 indicate that only some typical “regularity” conditions are needed to ensure that they have suitable asymptotic convergence behavior, numerical experiments in Section 7 suggest that they are indeed tractable in real-world applications. However, it should be noted that our numerical experiments are motivated by a specific medical decision-making application, in which the belief state space,  $\Delta_{\mathcal{G}}$ , is an 8-simplex (because there are nine latent states), the vector of basis functions  $\mathbf{b}(\boldsymbol{\pi})$  maps this 8-simplex to  $\mathbb{R}^{13}$ , and the main parameter needed to estimate the value function has a reasonable dimensionality (belongs to  $\mathbb{R}^{13}$ ). Furthermore, the set of models and the policies needed in this application are not extremely large. If in an application, the cardinality of these underlying spaces is significantly higher than those in our experiments, the proposed algorithms may lose computational tractability. In such scenarios, further care (e.g., discretization, approximation, etc.) is needed to speed up these algorithms. Finally, it should be noted that the proposed algorithms are suitable for scenarios in which the observed history provides at least some information about latent confounders. In APOMDPs, some models might be naturally more informative than others; see Definition 2 and Lemma 2 of Saghafian (2018) for the notion of model informativeness in APOMDPs, which is based on the Blackwell–Sherman–Stein sufficiency theorem. However, if none of the models are informative, then the Bayesian operator that is used in these

algorithms might fail to update the belief distributions under all models, and hence, the value function cannot be learned from data for different values of  $\boldsymbol{\pi}$ . Intuitively, if the observed data do not provide any information about the dynamic latent confounders, then one should not hope for using observed variables to effectively adjust for the effect of dynamic unobserved confounders.

## 8. Conclusion

We propose a mathematical framework as well as learning algorithms for finding an effective dynamic treatment regime under model ambiguity. Incorporating model ambiguity a priori in the analyses not only provides robustness to inevitable misspecifications (e.g., caused by hidden confounders with unknown dynamics and/or impact on the observed variables) but more broadly can bridge the gap between two philosophical views of causal inference: model based and model free.

Our work also tries to close the gap between RL techniques and dynamic causal inference methods. Specifically, as is common, we view the problem of finding an effective treatment regime as an “off-policy” RL problem. However, unlike the existing work, we allow the learning to occur across a “cloud” of potential data-generating models. This is specifically useful when data are observational, the behavior policy is unknown, and the existence of time-varying unmeasured confounders (which are themselves affected by previous actions) makes the task of learning the causal impact of an evaluation policy challenging.

Unlike the available RL techniques or the methods related to causal inference in dynamic settings, our work also allows for a two-way personalization; the obtained treatment policies are not only personalized based on the subject’s variables (e.g., a patient’s covariates) but also based on the ambiguity attitude and preferences of the decision-maker (e.g., the physician). Given the importance of this two-way personalization in a variety of applications (e.g., medical decision making or public policy), we hope that future research can develop further data-driven methods to learn policies that are personalized in both ways.

We also hope that the future research can test and implement our proposed learning algorithms in a variety of other applications. In this study, we investigate the performance of these learning algorithms in three ways. First, we analytically establish their asymptotic behavior, including (weak) consistency and asymptotic normality. Second, we examine them in a case study using clinical data related to NODAT patients. Third, we make use of simulation experiments (synthetic data), in which we control the true data-generating model and compare the performance of our proposed methods with that of an imaginary oracle who knows both the

true data-generating model and the optimal policy under that model. All these investigations reveal promising results. However, further research is needed to more broadly investigate the performance of our proposed methods in other applications and domains. With the increasing availability of sensor-based devices that are connected via the Internet of Things and benefit from data fusion (see, e.g., Saghafian et al. 2022), future research can also investigate augmenting our approaches to work with data obtained from multiple connected streams. Finally, future research can examine the interpretability of the policies that are obtained via DAV-Learning, SAV-Learning, DAV-Learning-BUC, and SAV-Learning-BUC, and it can propose adjustments (if needed) to ensure that they can be effectively used in practice.

### Acknowledgments

The author is grateful to Susan Murphy (Harvard), Richard Zeckhauser (Harvard), and Guido Imbens (Stanford) for their valuable suggestions and comments.

### Endnotes

<sup>1</sup> This assumption has also appeared in the literature under other names, such as “sequential randomization” (Tsiatis et al. 2019) and “sequential backdoor criterion” (Pearl and Robins 1995).

<sup>2</sup> For example, this can be done under an assumed model for the dynamics of unobserved confounders (e.g., how they are affected by actions taken) and their relationship to observed values (e.g., how unobserved time-varying confounders affect the actions under which data are generated).

<sup>3</sup> This view of data-driven decision-maker under ambiguity has also been shown to be useful in various applications, including in designing and optimizing queueing systems under model ambiguity (Bren and Saghafian 2019) and medical decision-maker (Boloori et al. 2020).

<sup>4</sup> This view is also aligned with that of Arrow (1951), who stated: “There are two types of uncertainty: one as to the hypothesis, which is expressed by saying that the hypothesis is known to belong to a certain class or model, and one as to the future events or observations given the hypothesis, which is expressed by a probability distribution.”

<sup>5</sup> For studies in this literature that consider other measure instead of the expected value of outcomes, we refer to Linn et al. (2017) and Wang et al. (2018) (quantile performance) as well as Leqi and Kennedy (2021) (median performance). These studies, however, do not consider model ambiguity, existence of unobserved confounders, or other challenges we aim to address. Although by using the  $\alpha$ -MEU criterion, we primarily generalize the expected value of outcomes, it should be noted that our results can also be used to study generalizations of other measures, such as the quantile or median measures.

<sup>6</sup> Several studies consider finding treatment regimes that allow shared decision-making between physicians and patients when there are multiple risky (probabilistic)—as opposed to ambiguous (nonprobabilistic)—outcomes. Some available methods include set-valued treatment regimes (Laber et al. 2014, Lizotte and Laber 2016), inverse-preference elicitation (Lizotte et al. 2012), constrained estimation (Linn et al. 2015), and use of item response theory (Butler et al. 2018).

<sup>7</sup> See also Zhang and Bareinboim (2019) for more discussions related to finding the optimal treatment regime under model ambiguity without a Markovian structure.

<sup>8</sup> Because APOMDPs generalize POMDPs, our results can also be viewed as generalizations of those in the literature that use a POMDP setting to perform off-policy evaluation (see, e.g., Thomas and Brunskill 2016, Tennenholtz et al. 2020, Xu et al. 2020, Bennett and Kallus 2021, Hu and Wager 2021, and the references therein).

<sup>9</sup> The fact that the structural properties of the value function in APOMDPs are known (see Saghafian 2018) is a main reason we make use of V-Learning as opposed to other RL methods (e.g., Q-Learning or A-Learning). Furthermore, as we will see, a data transformation approach allows for using a weight-adjusted version of the Bellman equation, thereby directly estimating the value function from observed data.

<sup>10</sup> We do not assume that time points are evenly distributed or homogenous across patient trajectories. Importantly, in some applications, the treatment times are random. For simplicity, we assume that treatment times are fixed. However, extending our results to scenarios with random treatment times is relatively straightforward.

<sup>11</sup> It should be noted that  $S_t$ , in general, depends on the history up to time  $t$ . Thus,  $G_t \triangleq g(S_t, A_t)$  also depends on the history. However, this dependence is only through  $S_t$ , which as noted earlier, can always be achieved with appropriate definition of summary variables  $(S_t)_{t \in \mathcal{T}}$  (see, e.g., Xu et al. 2020).

<sup>12</sup> Although we focus on discounted sum of immediate gains, we note that many of our results readily extend to the average overall gains  $\bar{\Gamma}(\lambda) \triangleq \frac{1}{T} \sum_{t \in \mathcal{T}} G_t^\lambda$  and in particular, to its long-run counterpart  $\liminf_{T \rightarrow \infty} \bar{\Gamma}(\lambda)$ . This is because under some mild conditions,  $\lim_{T \rightarrow \infty} \bar{\Gamma}(\lambda) = \lim_{T \rightarrow \infty} \lim_{\beta \rightarrow 1} \frac{\Gamma_T(\lambda)}{1-\beta}$ .

<sup>13</sup> This assumption links the counterfactual data with the factual one (Robins 1997) and can be violated if treatment of a subject impacts another subject’s variables (e.g., vaccinating a group of individuals may decrease exposure of others to a disease).

<sup>14</sup> We defer discussions on how the space of models can be constructed to the numerical experiments section (see, e.g., the discussion in Section 7.1.4). As described there, there are various ways of constructing the set  $\mathcal{M}$ , and our analysis does not rely on any specific method or assumption in this regard.

<sup>15</sup> See also the *sequential backdoor* criterion (Pearl and Robins 1995).

<sup>16</sup> For related studies on doubly robust estimators, we refer to Jiang and Li (2016), Thomas and Brunskill (2016), Kallus and Uehara (2020), Athey and Wager (2021), Bang and Robins (2021), and the references therein.

<sup>17</sup> For other approaches in modeling confounders as hidden states, see, for example, Xu et al. (2020), Bennett et al. (2021), and the references therein.

<sup>18</sup> For typical POMDPs and APOMDPs, it is known that the belief distribution over latent states can serve as a sufficient statistics (see, e.g., Saghafian 2018, Saghafian and Rasouli 2019, Boloori et al. 2020, and the references therein). In Remark 2 and Section 6, we further discuss handling cases where this might not hold. We also refer interested readers to assumption 1 in Tennenholtz et al. (2020), which establishes the existence of a sufficient statistic as one sufficient condition for unbiasedness of importance sampling in POMDPs, but highlight that our focus is on APOMDPs as opposed to POMDPs.

<sup>19</sup> For a random variable  $X$  with  $n$  observed values denoted by  $x_1, x_2, \dots, x_n$ ,  $\mathbb{E}^\mathbb{P}[X] \triangleq n^{-1} \sum_{i=1}^n x_i$ . Similarly, for a function  $f$ ,  $\mathbb{E}^\mathbb{P}[f(X)] \triangleq n^{-1} \sum_{i=1}^n f(x_i)$ .

<sup>20</sup> This is the case in how the value function in traditional POMDPs is often constructed; see, for example, Equation (8).

<sup>21</sup> Allowing the dimensionality of the parameter space,  $d_t$ , to depend on  $t$  can enable us to increase flexibility as  $t$  grows (e.g., by introducing more knots). The special case where  $d_t$  does not depend on  $t$  is still useful in some settings, including those where the goal is to

learn the long-run impact of a policy (see, e.g., Algorithms 1 and 2 in Section 4.2).

<sup>22</sup> In our case study, simulations experiments, and theoretical results, we make use of the squared Euclidean norm as the penalty function, and hence, assume  $\mathcal{P}(\psi_t) = \psi_t^T \psi_t$ .

<sup>23</sup> We use the “max” operator instead of “sup” because in most real-world applications,  $\Upsilon$  is first identified by a set of domain experts and is such that maximum is obtained. We later make this assumption more implicit (see, e.g., Condition 4 in Section 5). Furthermore, in various practical applications,  $\Upsilon$  is often restricted to the set of policies that satisfy specific attributes, such as fairness or interpretability.

<sup>24</sup> We assume  $\mathcal{M}$  is such that  $\inf_{m \in \mathcal{M}} \|\psi_t^{m, \mu^c}\|$  and  $\sup_{m \in \mathcal{M}} \|\psi_t^{m, \mu^c}\|$  are both finite and that  $\underline{m}$  and  $\overline{m}$  are both in  $\mathcal{M}$ .

<sup>25</sup> For general results related to the asymptotic properties of V-Learning algorithms when all variables are observable and there is no model ambiguity, we refer interested readers to Lockett et al. (2020).

<sup>26</sup> For the definition of the bracketing integral,  $J_{[]}(\infty, \Upsilon, L_\rho(P^m))$ , see, for example, Kosorok (2008).

<sup>27</sup> For the definition of an absolutely regular stationary process and its  $\beta$ -mixing coefficients, see, for example, Dedecker and Louhichi (2002), Kosorok (2008), and the references therein.

<sup>28</sup> These weight modifiers may depend on  $\pi$  in general (based on (26) and (27)). Such dependency is suppressed here for the ease of notation.

<sup>29</sup> Of note, similar to this study, Boloori et al. (2020) also address simultaneous management of immunosuppressive drugs and insulin for NODAT patients. However, the study of Boloori et al. (2020) is not concerned with the main aspects of this work. Namely, it does not deal with (a) causal inference or (b) reinforcement learning. The main ideas we use in this work are also not used in Boloori et al. (2020). For example, we make use of *importance sampling* ideas along with *weight-adjusted* versions of the Bellman equation, but in Boloori et al. (2020), the approach is vastly different: an APOMDP model is directly fitted to the data, and the optimal policy of this APOMDP is established using its Bellman equation (without weight adjustment) based on the theoretical results known for APOMDPs (Saghafian 2018). We believe that the approach used both in Boloori et al. (2020) and in this study are novel. However, they are not directly comparable in a ceteris paribus manner because of the mentioned differences.

<sup>30</sup> In addition to immediate gains, our framework allows for including lump-sum gains (i.e., gains at the end of the horizon to reflect the quality of life associated with the remaining years). For the purposes of this study, however, we simply set  $V_0(\pi) \triangleq 0$ .

<sup>31</sup> Given  $\pi_t^m$  in each period  $t$  under each model  $m$ , the obtained immediate gain in each period  $t$  under each model  $m$  is considered to be the weighted average of immediate gain values shown in Table 4, where weights are given by  $\pi_t^m$  (see also Remark 2).

<sup>32</sup> It should be also noted that any continuous set of models can be approximated via finite sets with any required precision. That is, even if  $\mathcal{M}$  is not finite, one can always consider a finite set  $\mathcal{M}$  as a close approximation to the continuous one.

<sup>33</sup> The number of these replications is chosen so that the confidence intervals are tight enough while maintaining reasonable computational times.

<sup>34</sup> These findings hold in the context of our numerical experiments. However, we avoid making general conclusions because doing so will require a more extensive set of experiments. Our results, however, provide a proof of concept that can be further explored by future research.

## References

ADA (2012) Standards of medical care in diabetes. *Diabetes Care* 35:S11–S63.

- Ahn D, Choi S, Gale D, Kariv S (2014) Estimating ambiguity aversion in a portfolio choice experiment. *Quant. Econom.* 5(2):195–223.
- Angrist JD, Imbens GW, Rubin DB (1996) Identification of causal effects using instrumental variables. *J. Amer. Statist. Assoc.* 91:434–471.
- Arrow KJ (1951) Alternative approaches to the theory of choice in risk-taking situations. *Econometrica* 19(4):404–437.
- Arrow KJ, Hurwicz L (1977) An optimality criterion for decision making under ignorance. Arrow KJ, Hurwicz L, eds. *Studies in Resource Allocation Processes* (Cambridge University Press, Cambridge, UK), 461–472.
- Athey S, Wager S (2021) Policy learning with observational data. *Econometrica* 89(1):133–161.
- Bang H, Robins JM (2021) Doubly robust estimation in missing data and causal inference models. *Biometrics* 61(4):962–972.
- Bennett A, Kallus N (2021) Proximal reinforcement learning: Efficient off-policy evaluation in partially observed Markov decision processes. Preprint, submitted October 28, <https://doi.org/10.48550/arXiv.2110.15332>.
- Bennett A, Kallus N, Li L, Mousavi A (2021) Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *Proc. 24th Internat. Conf. Artificial Intelligence Statist.* (PMLR, New York), 1999–2007.
- Bhidé AV (2000) *The Origin and Evolution of New Business* (Oxford University Press, Oxford, UK).
- Boloori A, Saghafian S, Chakkerla HA, Cook CB (2015) Characterization of remitting and relapsing hyperglycemia in post-renal-transplant recipients. *PLoS One* 10(11):1–16.
- Boloori A, Saghafian S, Chakkerla HA, Cook CB (2020) Data-driven management of post-transplant medications: An ambiguous partially observable Markov decision process approach. *Manufacturing Service Oper. Management* 22(5):1066–1087.
- Box G (1979) Robustness in the strategy of scientific model building. Launer R, Wilkinson G, eds. *Robustness in Statistics* (Academic Press, New York), 201–236.
- Bren A, Saghafian S (2019) Data-driven percentile optimization for Multi-Class queueing systems with model ambiguity: Theory and application. *INFORMS J. Optim.* 1(4):267–287.
- Butler EL, Laber EB, Davis SM, Kosorok MR (2018) Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics* 74(1):18–26.
- Chakkerla HA, Weil EJ, Castro J, Heilman RL, Reddy KS, Mazur MJ, Hamawi K, et al. (2009) Hyperglycemia during the immediate period after kidney transplantation. *Clinical J. Amer. Soc. Nephrol.* 4:853–859.
- Chakraborty B, Murphy SA (2014) Dynamic treatment regimes. *Annual Rev. Statist. Appl.* 1(1):447–464.
- Dedecker J, Louhichi S (2002) Maximal inequalities and empirical central limit theorems. Mikosch T, Sørensen M, eds. *Empirical Process Techniques for Dependent Data* (Birkhäuser, Boston), 137–159.
- Frank RG, Zeckhauser RJ (2007) Custom-made vs. ready-to-wear treatments: Behavioral propensities in physicians' choices. *J. Health Econom.* 26(6):1101–1127.
- Ghiradato P, Maccheroni F, Marinacci M (2004) Differentiating ambiguity and ambiguity attitude. *J. Econom. Theory* 118:133–173.
- Ghisdal L, Van Laecke S, Abramowicz MJ, Vanholder R, Abramowicz D (2012) New-onset diabetes after renal transplantation risk assessment and management. *Diabetes Care* 35(1):181–188.
- Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054.
- Heath C, Tversky A (1991) Preference and belief: Ambiguity and competence in choice under uncertainty. *J. Risk Uncertainty* 4(1):5–28.
- Hu Y, Wager S (2021) Off-policy evaluation in partially observed Markov decision processes. Preprint, submitted October 24, <https://arxiv.org/abs/2110.12343>.
- Hurwicz L (1951a) Optimality criteria for decision making under ignorance. Cowles Commission Discussion Paper: Statistics No. 370, Cowles Commission.

- Hurwicz L (1951b) Some specification problems and applications to econometric models. *Econometrica* 19:343–344.
- Jiang N, Li L (2016) Doubly robust off-policy value evaluation for reinforcement learning. *Proc. 33rd Internat. Conf. Machine Learn.* (JMLR: W&CP), 652–661.
- Kallus N, Uehara M (2020) Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Machine Learn. Res.* 21:1–63.
- Kallus N, Zhou A (2020) Confounding-robust policy evaluation in infinite-horizon reinforcement learning. Preprint, submitted February 11, <https://arxiv.org/abs/2002.04518>.
- Kallus N, Zhou A (2021) Minimax-optimal policy learning under unobserved confounding. *Management Sci.* 67(5):2870–2890.
- Kosorok MR (2008) *Introduction to Empirical Processes and Semiparametric Inference* (Springer, New York).
- Kosorok MR, Laber EB (2019) Precision medicine. *Annual Rev. Statist. Appl.* 6(263–286):1243–1254.
- Laber EB, Lizotte DJ, Ferguson B (2014) Set-valued dynamic treatment regimes for competing outcomes. *Biometrics* 70(1):53–61.
- Leqi L, Kennedy EH (2021) Median optimal treatment regimes. Preprint, submitted March 2, <https://arxiv.org/abs/2103.01802>.
- Linn KA, Laber EB, Stefanski LA (2015) Estimation of dynamic treatment regimes for complex outcomes: Balancing benefits and risks. Kosorok MR, Moodie EEM, eds. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine* (SIAM, Philadelphia), 249–262.
- Linn KA, Laber EB, Stefanski LA (2017) Interactive Q-learning for quantiles. *J. Amer. Statist. Assoc.* 112(518):638–649.
- Lizotte DJ, Laber EB (2016) Multi-objective Markov decision processes for data-driven decision support. *J. Machine Learn. Res.* 17(1):7378–7405.
- Lizotte DJ, Bowling M, Murphy SA (2012) Linear fitted-q iteration with multiple reward functions. *J. Machine Learn. Res.* 13(1): 3253–3295.
- Luckett DJ, Laber EB, Kahkoska AR, Maahs DM, Mayer-Davis E, Kosorok MR (2020) Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* 115(530): 692–706.
- Magnani A, Boyd SP (2009) Convex piecewise-linear fitting. *Optim. Engng.* 10:1–17.
- Manski CF (2007) *Identification for Prediction and Decision* (Harvard University Press, Cambridge, MA).
- Manski CF (2021) Econometrics for decision making: Building foundations sketched by Haavelmo and Wald. *Econometrica* 89(6): 2827–2853.
- Marinacci M (2002) Probabilistic sophistication and multiple priors. *Econometrica* 70(2):755–764.
- Munshi VN, Saghafian S, Cook CB, Aradhyula S, Chakkerla HA (2021) Use of imputation and decision modeling to improve diagnosis and management of patients at risk for newonset diabetes after transplantation. *Ann. Transplantation* 26:1–9.
- Munshi VN, Saghafian S, Cook CB, Werner KT, Chakkerla HA (2020a) Comparison of post-transplantation diabetes mellitus incidence and risk factors between kidney and liver transplantation patients. *PLoS One* 15(1):1–12.
- Munshi VN, Saghafian S, Cook CB, Steidley D, Hardaway B, Chakkerla HA (2020b) Incidence, risk factors, and trends for post-heart transplantation diabetes mellitus. *Amer. J. Cardiology* 125(3):436–440.
- Murphy SA (2003) Optimal dynamic treatment regimes. *J. Roy. Statist. Soc. Ser. B Statist. Methodology* 65(2):331–355.
- Murphy SA (2005) An experimental design for the development of adaptive treatment strategies. *Statist. Medicine* 24(10):1455–1481.
- Murphy SA, van der Laan MJ, Robins JM, CPPRG (2001) Marginal mean models for dynamic regimes. *J. Amer. Statist. Assoc.* 96(456): 1410–1423.
- Murphy SA, Deng Y, Laber EB, Maei HR, Sutton RS, Witkiewitz K (2016) A batch, off-policy, actor-critic algorithm for optimizing the average reward. Preprint, submitted July 18, <https://arxiv.org/abs/1607.05047>.
- Namkoong H, Keramati R, Yadlowsky S, Brunskill E (2020) Off-policy policy evaluation for sequential decisions under unobserved confounding. Preprint, submitted March 12, <https://arxiv.org/abs/2003.05623>.
- Nie X, Brunskill E, Wager S (2021) Learning when-to-treat policies. *J. Amer. Statist. Assoc.* 116(533):392–409.
- Pearl J (2009) *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, UK).
- Pearl J, Robins J (1995) Probabilistic evaluation of sequential plans from causal models with hidden variables. Besnard P, Hanks S, eds. *Uncertainty in Artificial Intelligence 11* (Morgan Kaufmann, San Francisco), 444–453.
- Precup D, Sutton RS, Singh S (2000) Eligibility traces for off-policy policy evaluation. *Proc. 17th Internat. Conf. Machine Learn.*, 759–766.
- Robins J (1986) A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* 7(9–12): 1393–1512.
- Robins J (1997) Causal inference from complex longitudinal data. Berkane M, ed. *Latent Variable Modeling and Applications to Causality* (Springer, New York), 69–117.
- Robins J (2004) Optimal structural nested models for optimal sequential decisions. *Proc. Second Seattle Sympos. Biostatistics* (Springer, New York), 189–326.
- Robins J, Hernán MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–560.
- Rosenbaum PR (2002) *Observational Studies* (Springer, New York).
- Rosenbaum PR (2010) *Design of Observational Studies* (Springer, New York).
- Rubin DB (1986) Comment: Which ifs have causal answers. *J. Amer. Statist. Assoc.* 81:961–962.
- Saghafian S (2018) Ambiguous partially observable Markov decision processes: Structural results and applications. *J. Econom. Theory* 178:1–35.
- Saghafian S, Murphy SA (2021) Innovative healthcare delivery: The scientific and regulatory challenges in designing mHealth interventions. NAM Perspectives. Commentary. Report, National Academy of Medicine, Washington, DC.
- Saghafian S, Rasouli M (2019) Robust partially observable Markov decision processes. Working paper, Harvard University, Cambridge, MA.
- Saghafian S, Tomlin BT (2016) The newsvendor under demand ambiguity: Combining data with moment and tail information. *Oper. Res.* 64(1):167–185.
- Saghafian S, Tomlin B, Biller S (2022) The Internet of things and information fusion: Who talks to who? *Manufacturing Service Oper. Management* 24(1):333–351.
- Savage L (1951) The theory of statistical decision. *J. Amer. Statist. Assoc.* 46:55–67.
- Smallwood R, Sondik EJ (1973) The optimal control of partially observable Markov processes over a finite horizon. *Oper. Res.* 21(5):1071–1088.
- Stoy J (2011) Statistical decisions under ambiguity. *Theory Decision* 70(2):129–148.
- Tennenholtz G, Shalit U, Mannor Sh (2020) Off-policy evaluation in partially observable environments. *Proc. Conf. AAAI Artificial Intelligence* 34:10276–10283.
- Thomas PS, Brunskill E (2016) Data-efficient off-policy policy evaluation for reinforcement learning. *Proc. 33rd Internat. Conf. Machine Learn.*, 2139–2148.
- Tsiatis AA, Davidian M, Holloway ST, Laber EB, Kosorok MR (2019) *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine* (Chapman and Hall/CRC, Boca Raton, FL).

- Wald A (1939) Contribution to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.* 10:299–326.
- Wald A (1945) Statistical decision functions which minimize the maximum risk. *Ann. Math.* 46:265–280.
- Wald A (1950) *Statistical Decision Functions* (Wiley, New York).
- Wang L, Zhou Y, Song R, Sherwood B (2018) Quantile-optimal treatment regimes. *J. Amer. Statist. Assoc.* 113(523):1243–1254.
- Watson J, Holmes C (2016) Approximate models and robust decisions. *Statist. Sci.* 31:465–489.
- Whelton PK, Carey RM, Aronow WS, Casey DE Jr, Collins KJ, Himmel-farb CD, DePalma SM, et al. (2017) 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: A report of the American College of Cardiology/American Heart Association Task Force on clinical practice guidelines. *J. Amer. College Cardiology* 71(19):e127–e248.
- Xu Z, Laber E, Staicu AM, Severus E (2020) Latent-state models for precision medicine. Preprint, submitted May 26, <https://arxiv.org/abs/2005.13001>.
- Zhang J, Bareinboim E (2019) Near-optimal reinforcement learning in dynamic treatment regimes. *Adv. Neural Inform. Processing Systems*, vol. 32 (NeurIPS).
- Zhang Y, Laber EB, Davidian M, Tsiatis AA (2018) Interpretable dynamic treatment regimes. *J. Amer. Statist. Assoc.* 113(524): 1541–1549.
- Zhao YQ, Zeng D, Laber EB, Kosorok MR (2015) New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* 110(510):583–598.