Submitted to
manuscript

Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach

Soroush Saghafian*
Harvard Kennedy School, Harvard University, Cambridge, MA

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, since 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 and faced a medical condition known as New Onset Diabetes After Transplantation (NODAT), we extend DTRs to a new class termed Ambiguous Dynamic Treatment Regimes (ADTRs), in which the casual impact of treatment regimes is evaluated based on a "cloud" of potential causal models. We then connect ADTRs to Ambiguous Partially Observable Mark 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 efficiently 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 to an imaginary oracle who knows both the true causal model (of the data generating process) and the optimal regime under that model.

Key words: Observational Data; Dynamic Treatment Regimes; Unobserved Confounders; APOMDPs, Reinforcement Learning
History: Version: December 8, 2021

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).

* The author is grateful to Susan Murphy (Harvard University) and Richard Zeckhauser (Harvard University) for their valuable suggestions and comments.

Dynamic Treatment Regimes (DTRs) have been widely studied for this goal, enabling finding effective alternative policies from observational data (Robins 1986, 1997, Murphy et al. 2001, Murphy 2003, Robins 2004, Zhao et al. 2015, Zhang et al. 2018, Wang et al. 2018, Tsiatis et al. 2019, Kosorok and Laber 2019, Luckett et al. 2020, Nie et al. 2021, Leqi and Kennedy 2021). A DTR is, in essence, a set of rules that prescribe 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 is *observational*. Notably, one needs to assume *sequential ignorability*[1] (Robins 1986, 1997, Murphy et al. 2001, Murphy 2003, Robins 2004), meaning that the data is 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 Micro-Randomized Trials (MRTs) in some mobile health studies where the goals 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 typically 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.[2] Assuming such a model can allow estimating a distribution for potential trajectories under any alternative decision-making policy (i.e., treatment regime), which is central to estimating its effect. However, since time-varying confounders are often unobserved, estimating and assuming any such model is subject to significant misspecifications (a.k.a. model ambiguity). We address this challenge by extending the analyses of DTRs to a new class termed *Ambiguous DTRs (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.

[1] 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).

[2] 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 is generated).

Specifically, we allow for non-probabilistic ambiguity (a.k.a. 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 is generated (see, e.g., Saghafian (2018) and Chapter 11 of Manski (2007) 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). 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).[3]

In extending DRTs to ADTRs, we are particularly motivated by our various collaborations with our partner hospital, the Mayo Clinic. In various studies (see, e.g., Boloori 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., Boloori et al. 2015, 2020). Due to 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., Boloori et al. 2020), and the existence of unobserved confounders that are time-varying disallow 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 Mark Decision Processes* (APOMDPs) proposed by Saghafian (2018), and develop Reinforcement Learning (RL) algorithms that allow learning an optimal treatment regime from the observed data in efficient ways.

The utility function we use is based on a generalization of the traditional *maximin expected utility (MEU)* theory (a.k.a. Wald's or robust optimization criterion). The MEU theory assumes that outcomes should by maximizing with respect to the worst possible member of the ambiguity

[3] 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.".

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, does not allow for representing meaningful human choices such as those of ambiguity seeking individuals established in some behavioral studies (see, e.g., Bhidé 2000, Heath and Tversky 1991, Ahn et al. 2014). This was also recognized in the seminal work of Savage (1951) who wrote that this criterion is "ultrapessimisitic" and "can lead to absurd conclusion[s]". The generalization we use is known as *α-maximin expected utility (α-MEU)*, which allows for both optimistic and pessimistic views of the world (Arrow and Hurwicz 1977, Hurwicz 1951a,b, Ghiradato et al. 2004, Saghafian 2018). Unlike studies that use the MEU criterion, using the α-MEU criterion avoids overly conservative decisions by allowing for a controllable *pessimism level* (denoted by the parameter α) that can take values in [0, 1].

Within the utility theory literature, early studies (see, e.g., Arrow and Hurwicz 1977, Hurwicz 1951a,b) provided four axioms that a choice operator must salsify. 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., Ghiradato et al. 2004, Marinacci 2002) further axiomatized preferences under the α-MEU criterion and also highlighted another importance of using the α-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 α-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[4]—but also allows 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 are personalized based on both the subject's and the decision-maker's characteristics. This is important in various applications 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/her preferences in providing the best course of treatment.

We start our analyses by showing how a generalization of importance sampling methods (a.k.a. inverse-probability-weighting) widely used in the literature (see, e.g., Robins et al. 2000, Precup 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., sat-

[4] For studies in this literature that consider other measure instead of the expected value of outcomes, we refer to Wang et al. (2018) (quantile performance) and 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. While by using the α-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.

isfy the Markov property).[5] Specifically, we start by generalizing importance sampling methods by allowing sampling across a *cloud* of potential data generating models (a.k.a. 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 POMDPs by allowing model ambiguity. Connecting ADTRs to APOMDPs, thus, allows considering time-varying unobserved confounders as dynamic latent states while allowing ambiguity regarding the true (data generating) causal model.[6]

We then make use of known structural results for APODMPs (e.g., piecewise linearity and continuity of the value function) established in the literature (Saghafian 2018), and develop two RL approaches that can efficiently 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., Luckett 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).

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 part-

[5] See also Zhang and Bareinboim (2019) for more discussions related to fining the optimal treatment regime under model ambiguity without a Markovian structure.

[6] Since APOMDPs generalize POMPDs, our results can also be viewed as generalizations of those in the literature that use a POMDP setting to perform policy evaluation (see, e.g., Tennenholtz et al. 2020, Xu et al. 2020, and the references therein).

ner 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 α) than that of `DAV-Learning`, and hence, a decisions-maker who uses `SAV-Learning` does not need to be worried about the value of α s/he 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 to 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 α. 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 to such an imaginary decision-maker has a U-shape curve in the pessimism level: the minimum loss for both `DAV-Learning` and `SAV-Learning` are obtained at a mid-value of α. 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 α as a tuning parameter in our proposed approaches, one can efficiently 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, while 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 $\mathscr{T} \triangleq \{1,2,3,\cdots,T\}$, the notations $(X_t)_{t\in\mathscr{T}}$ and $\mathscr{T}_{\leq t}$ are used to represent the vector $(X_1,X_2,\cdots,X_T)$ and the set $\mathscr{T}\setminus\{t+1,t+2,\cdots,T\}$, respectively. All vectors are consider to be in the column format (e.g., $(X_t)_{t\in\mathscr{T}}$ is $|\mathscr{T}|\times 1$). For any finite set $\Xi\subset\mathbb{R}$, we let Δ_Ξ denote the probability simplex induced by Ξ. The notations $\xrightarrow{p}$ and $\xrightarrow{d}$ denote convergence in probability and distribution, respectively. The set $\mathscr{I}$ 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 $(O_t,A_t)_{t\in\mathscr{T}}$. For a realized trajectory, $(o_t,a_t)_{t\in\mathscr{T}}$, $o_t\in\mathscr{O}$ 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 $t\in\mathscr{T}$, and $A_t\in\mathscr{A}$ denotes the action/treatment assigned at time $t\in\mathscr{T}$, where $\mathscr{T}\triangleq\{1,2,3,\cdots,T\}$ is the set of time periods (e.g., patients' visits/follow-ups).[7] 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 (non-aggressive) or high-dose (aggressive) tacrolimus prescriptions as well as information on whether insulin has been used (see Table 3), Finally, $\mathscr{T}\triangleq\{1,2,3,\cdots,12\}$, since 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 $\mathscr{S}$ 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 $t\in\mathscr{T}$ by

[7] 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 treatment times are fixed. However, extending our results to scenarios with random treatment times is relatively straightforward.

$\mathbf{H}_t^o \triangleq (O_1, A_1, O_2, A_2, \cdots, O_t)$ and let $\mathscr{H}_t^o$ be the support of $\mathbf{H}_t^o$. Similarly, we denote the (partially) unobservable history up to each time $t \in \mathscr{T}$ by $\mathbf{H}_t^u \triangleq (S_1, O_1, A_1, S_2, O_2, A_2, \cdots, S_t, O_t)$ and let $\mathscr{H}_t^u$ 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 $(a_t)_{t\in\mathscr{T}_{\leq t-1}} \triangleq (a_1, a_2, \cdots, a_{t-1})$.

We assume the latent state summaries $(S_t)_{t\in\mathscr{T}}$ 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 $(S_t)_{t\in\mathscr{T}}$ (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. The set of all possible immediate gains can be denoted by $\mathscr{G} \triangleq \{\mathbf{g}^a \in \mathbb{R}^{|\mathscr{S}|} : \forall a \in \mathscr{A}\}$, where $\mathbf{g}^a \triangleq (g(s,a))_{s\in\mathscr{S}}$.

A treatment regime (hereafter also "policy" for simplicity) $\boldsymbol{\lambda} \triangleq (\lambda_t)_{t\in\mathscr{T}}$ 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, $\Delta_{\mathscr{A}}$. 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

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

where $\beta \in \mathscr{I} \setminus \{1\}$ is a discount factor. Similarly, the long-run impact of $\boldsymbol{\lambda}$ can be analyzed using $\Gamma_\infty(\boldsymbol{\lambda}) \triangleq \lim_{T\to\infty} \Gamma_T(\boldsymbol{\lambda})$.[8] Here, we shall note that $G_t^{\boldsymbol{\lambda}}$, and hence $\Gamma_T(\boldsymbol{\lambda})$, should be viewed with a potential outcomes lens (for more discussions, see, e.g., Robins 1986, Rubin 1986, Angrist et al. 1996, Robins 1997, Murphy et al. 2001); equivalently, in the language of do calculus, $G_t^{\boldsymbol{\lambda}}$ 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*[9], 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

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

[9] 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).

observed part of the data, but with different implications about the distribution of $\Gamma_T(\boldsymbol{\lambda})$. We let $\mathscr{M}$ denote the set of all such models (a.k.a. an ambiguity set). Each given model $m \in \mathscr{M}$ implies a distribution for $\Gamma_T(\boldsymbol{\lambda})$, which we denote by $f_m \in \mathscr{F}$, where $\mathscr{F} \triangleq \{f_m : m \in \mathscr{M}\}$.

Finally, since the distribution of $\Gamma_T(\boldsymbol{\lambda})$ varies across the models in $\mathscr{M}$, we define a utility function that allows us to compare the performance of different policies. To this end, we make use of the α-MEU utility, which is suitable for decision-making under ambiguity (see, e.g., Ghiradato et al. 2004, Marinacci 2002, 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

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

as the utility of $Y(\boldsymbol{\lambda})$, where α 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 $|\mathscr{M}| = 1$, $MEU_\alpha[Y(\boldsymbol{\lambda})]$ 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 "α-MEU identifiable," if $\left|MEU_\alpha[Y(\boldsymbol{\lambda})]\right| < \infty$ and $MEU_\alpha[Y(\boldsymbol{\lambda})]$ can be identified given $\mathscr{M}$. Since 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 Λ be the set of all α-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

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

To perform our analyses, it is useful to differentiate between the policy under which the data has 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 $\boldsymbol{\lambda}^b \triangleq (\lambda_t^b)_{t \in \mathscr{T}}$ is a vector of time-dependent mappings $\lambda_t^b : \mathscr{H}^u \to \Delta_{\mathscr{A}}$ whereas the evaluation policy denoted by $\boldsymbol{\lambda}^e \triangleq (\lambda_t^e)_{t \in \mathscr{T}}$ is a vector of time-dependent mappings $\lambda_t^e : \mathscr{H}^o \to \Delta_{\mathscr{A}}$. An important difference between the evaluation and the behavior policies relates to a condition known as *sequential ignitability*[10] (see, e.g., Robins 1986, 1997, Murphy et al. 2001, Murphy 2003, Robins 2004), which we define next.

[10] See also the *sequential backdoor* criterion (Pearl and Robins 1995).

DEFINITION 2 (**Sequential Ignorability**). For any policy $\boldsymbol{\lambda} \triangleq (\lambda_t)_{t\in\mathscr{T}}$, let $\mathbf{H}_t^{o,m}(\boldsymbol{\lambda}) \triangleq (O_1^m, A_1^m, O_2^m, A_2^m, \cdots, O_t^m)$ denote the observable history up to time $t \in \mathscr{T}$, generated under $\boldsymbol{\lambda}$ and model $m \in \mathscr{M}$. We say that $\boldsymbol{\lambda}$ satisfies sequential ignorability under model $m \in \mathscr{M}$, if for all $t \in \mathscr{T}$, the action generated by λ_t is independent of $G_t^m, O_{t+1}^m, G_{t+1}^m, O_{t+2}^m, \cdots, G_T^m$ conditional on $\mathbf{H}_t^{o,m}(\boldsymbol{\lambda})$.

Both by definition and naturally, any evaluation policy $\boldsymbol{\lambda}^e \triangleq (\lambda_t^e)_{t\in\mathscr{T}}$ (where $\lambda_t^e : \mathscr{H}^o \to \Delta_{\mathscr{A}}$) satisfies sequential ignorability under any model $m \in \mathscr{M}$. In contrast, a behavior policy $\boldsymbol{\lambda}^b \triangleq (\lambda_t^b)_{t\in\mathscr{T}}$ (where $\lambda_t^b : \mathscr{H}^u \to \Delta_{\mathscr{A}}$) may or may not satisfy this condition, since it might depend on unobservable confounders (variables in $(S_t)_{t\in\mathscr{T}}$ that affect both the gain and the actions selected by $\boldsymbol{\lambda}^b$). In fully randomized experiments (e.g., Micor Randomized Trials), the behavior policy may satisfy sequential ignorability. However, when the data is observational, it is often impossible to test whether the behavior policy satisfies this assumption, and in addition, it is often highly likely that this assumption does not hold.

2.1. Analyzing ADTRs via Generalized Sequential Importance Sampling (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)*. While allowing for model ambiguity, GSIS assigns weights under each model and sequentially adjusts the trajectory probabilities that occur under a given evaluation policy compared to those observed in the data set. Of note, we use GSIS in this section to study ADTRs that do not satisfy any Markovian (a.k.a. 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.

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 actions A_t is chosen under $\boldsymbol{\lambda}^e$ when the observed history is $\mathbf{H}_t^o(\boldsymbol{\lambda}^e)$. Furthermore, while the behavior policy is not known (e.g., due to 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 alow us to define importance sampling weights

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

Proposition 1 establishes that, under some conditions, the optimal policy for an ADTR governed by a set of models $\mathscr{M}$ can be found via GSIS. The proof follows by showing that GSIS provides a *distributionally robust* way of estimating the outcome variable of interest under the evaluation policy (see the appendix for further details). This result, in turn, is built by understanding how

the impact of an evaluation policy can first be analyzed for any given (a) sequence of actions, and (b) model $m \in \mathscr{M}$ under which the data might be generated (Lemma 1 below).

LEMMA 1 **(Evaluation Using Fixed Sequences of Actions)**. *Suppose that an evaluation policy* $\boldsymbol{\lambda}^e \triangleq (\lambda_t^e)_{t\in\mathscr{T}}$ *satisfies sequential ignorability under a given model* $m \in \mathscr{M}$*, and let the outcome of interest be* $Y(\boldsymbol{\lambda}^e) \triangleq \Gamma_T(\boldsymbol{\lambda}^e)$*. Defining* $\boldsymbol{\tau}_T \triangleq (a_t)_{t\in\mathscr{T}}$ *and* $\boldsymbol{\tau}_{t-1} \triangleq (a_t)_{t\in\mathscr{T}_{\leq t-1}}$*, we have:*

$$\mathbb{E}^{f_m}\big[Y(\boldsymbol{\lambda}^e)\big] = \sum_{\boldsymbol{\tau}_T} \mathbb{E}^m\big[Y(\boldsymbol{\tau}_T) \prod_{t\in\mathscr{T}} \lambda_t^e(a_t|\mathbf{H}_t^o(\boldsymbol{\tau}_{t-1}))\big], \tag{5}$$

where $Y(\boldsymbol{\tau}_T)$ *and* $\mathbf{H}_t^o(\boldsymbol{\tau}_{t-1})$ *denote* $Y(\boldsymbol{\lambda}^e)$ *and* $\mathbf{H}_t^o(\boldsymbol{\lambda}^e)$ *when the actions taken are given by* $\boldsymbol{\tau}_T$ *and* $\boldsymbol{\tau}_{t-1}$*, respectively,*

As mentioned earlier, any evaluation policy satisfies sequential ignorability both naturally and by definition. Thus, the condition in Lemma 1 is not restrictive. Notably, however, this lemma allows us to establish an MEU_α–unbiased estimator of $\Gamma_T(\boldsymbol{\lambda}^e)$ in Proposition 1, where the notion of MEU_α–unbiased estimation is defined below.

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

To establish an MEU_α–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 below).

DEFINITION 4 (**Almost Sure Overlap**). We say that the evaluation and the behavior policy almost surely overlap if, and only 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 $t \in \mathscr{T}$ and $a \in \mathscr{A}$.

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 $t \in \mathscr{T}$ (except perhaps on histories that might happen with probability zero).

PROPOSITION 1 **(Generalized Sequential Importance Sampling (GSIS))**. *Suppose that the evaluation and behavior policies (a) satisfy sequential ignorability under all models* $m \in \mathscr{M}$*, and (b) almost surely overlap. Then, for any* $\alpha \in \mathscr{I}$*, we have*

$$MEU_\alpha\big[\Gamma_T(\boldsymbol{\lambda}^e)\big] = MEU_\alpha\big[\Gamma_T(\boldsymbol{\lambda}^b) \prod_{t\in\mathscr{T}} w_t(\boldsymbol{\lambda}^e)\big], \tag{6}$$

and hence, $\hat{\Gamma}_T(\boldsymbol{\lambda}^e) \triangleq \Gamma_T(\boldsymbol{\lambda}^b) \prod_{t\in\mathscr{T}} w_t(\boldsymbol{\lambda}^e)$ *is an* MEU_α*–unbiased estimator of* $\Gamma_T(\boldsymbol{\lambda}^e)$*.*

Of note, Proposition 1 also provides a partial way of characterizing the optimal evaluation policy, since it provides a way of estimating $MEU_\alpha\big[\Gamma_T(\boldsymbol{\lambda}^e)\big]$ under any given evaluation policy. That is, using this proposition and optimizing $MEU_\alpha\big[\Gamma_T(\boldsymbol{\lambda}^e)\big]$ over a given set of policies (which for use in practice might be restricted to those satisfying desired attributes such as fairness, interpretability, etc.) 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 $\mathscr{M}$). 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 to 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 $|\mathscr{T}|=1$, the treatment variable is binary $|\mathscr{A}|=2$, and there is no model ambiguity $|\mathscr{M}|=1$, to ADTRs in which these restrictions are all relaxed. Two notable challenges in doing so are: (1) since future actions depend on the history, a confounding decision/treatment in any period can make future decisions confounding as well; (2) since 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 (BUC)**). We say that the behavioral policy satisfies Bounded Unobserved Confounding (BUC) under a model $m\in\mathscr{M}$ if, and only if, there exist constants $\eta_t^m\in[1,\infty)$ such that

$$(\eta_t^m)^{-1}\leq\frac{\ell(a_t,a_t',\mathbf{H}_t^{o,m},\mathbf{S}_t^m=\mathbf{s})}{\ell(a_t,a_t',\mathbf{H}_t^{o,m},\mathbf{S}_t^m=\mathbf{s}')}\leq\eta_t^m \tag{7}$$

a.s. over observable history $\mathbf{H}_t^{o,m}$, for all $t\in\mathscr{T}$, $a,a'\in\mathscr{A}$, $\mathbf{s},\mathbf{s}'\in\mathscr{S}^{(T)}$, where $\mathbf{S}_t^m\triangleq(S_t^m)_{t\in\mathscr{T}_{\leq t}}$ 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 above definition bounds the impact of the vector of the unobservable confounder variables, $\mathbf{S}_t^m$, in each period. In Lemma EC.1 (Appendix B) we show that this definition results in

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

over $\mathbf{H}_t^{o,m}$ and $\mathbf{H}_t^{u,m} = (\mathbf{H}_t^{o,m}, \mathbf{S}_t^m)$ for all $t \in \mathscr{T}$ and $a \in \mathscr{A}$. 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).

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

(i) Under each model $m \in \mathscr{M}$ we have:

$$\mathbb{E}^m\big[\Gamma_T(\boldsymbol{\lambda}^b)\prod_{t\in\mathscr{T}}\underline{w}_t(\boldsymbol{\lambda}^e)\big] \le \mathbb{E}^m\big[\Gamma_T(\boldsymbol{\lambda}^e)\big] \le \mathbb{E}^m\big[\Gamma_T(\boldsymbol{\lambda}^b)\prod_{t\in\mathscr{T}}\overline{w}_t(\boldsymbol{\lambda}^e)\big],$$

where

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

and

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

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

Similar to Proposition 1, part (ii) of Proposition 2 provides a way of fining the optimal evaluation policy, since 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 α-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. But to use Proposition 2 part (ii), one needs to tune the parameter $\tilde{\alpha}$. Since f is a decreasing function and $\tilde{\alpha} \in \mathscr{I}$, tuning $\tilde{\alpha}$ can be done in an 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 $f(\tilde{\alpha}) \ge MEU_\alpha\big[\Gamma_T(\boldsymbol{\lambda}^b)\big]$, implying that one can start tuning $\tilde{\alpha}$ using the threshold value $\tilde{\alpha}^* \triangleq \min\Big\{f^{(-1)}\Big(MEU_\alpha\big[\Gamma_T(\boldsymbol{\lambda}^b)\big]\Big), 1\Big\}$ and only consider values of $\tilde{\alpha}$ that are in $[0, \tilde{\alpha}^*]$. More importantly, it should be noted that the parameters $(\eta_t^m)_{t\in\mathscr{T},m\in\mathscr{M}}$ are design

sensitivity parameters. Specifically, for any $\epsilon > 0$, they can be chosen so that $f(0) - f(\tilde{\alpha}^*) < \epsilon$. Since $f(0) - f(\tilde{\alpha}^*) \geq 0$, this allows one to use any $\tilde{\alpha}$ in $[0, \tilde{\alpha}^*]$ and obtain an *approximate* unbiased MEU_α estimator for $\Gamma_T(\boldsymbol{\lambda}^e)$ with a guaranteed approximation error of ϵ.

Finally, we note that one can extend Propositions 1 and 2 to provide doubly robust estimators[11] to account for the fact that, under each given model $m \in \mathscr{M}$, the variance of an importance sampling based 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_α 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 find that our proposed learning methods provide strong robustness to model ambiguity (see, e.g., Section 6.3).

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 efficient RL methods.[12]

An APOMDP 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 (see, e.g., Xu et al. 2020). Specifically, while the observed history $\mathbf{H}_t^o$ grows over time, we can assume that there are summary functions $\nu_t : \mathscr{H}_t^o \to \Delta_{\mathscr{S}}$ such that $\boldsymbol{\pi}_t \triangleq \nu_t(\mathbf{h}_t^o)$ (a belief distribution over the latent states) is a sufficient statistics.[13] Using the belief distribution $\boldsymbol{\pi}_t$, we can work with transformed policies: we can consider $\boldsymbol{\mu}^e \triangleq \big(\mu_t^e(\boldsymbol{\pi}_t)\big)_{t\in\mathscr{T}}$ and $\boldsymbol{\mu}^b \triangleq \big(\mu_t^b(\boldsymbol{\pi}_t)\big)_{t\in\mathscr{T}}$ as the evaluation and behavior policies, respectively, where $\mu_t^e, \mu_t^b : \Delta_{\mathscr{S}} \to \Delta_{\mathscr{A}}$. 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.

[11] For related studies on doubly robust estimators, we refer to Bang and Robins (2021), Jiang and Li (2016), Thomas and Brunskill (2016), Kallus and Uehara (2020), Athey and Wager (2021), and the references therein.

[12] For other approaches in modeling confounders as hidden states see, e.g., Bennett et al. (2021), Xu et al. (2020), and the references therein.

[13] For POMDPs and APOMDPs, it is known that the belief distribution over latent states is a sufficient statistics (see, e.g., Saghafian 2018, Boloori et al. 2020, Saghafian and Rasouli 2019, and the references therein).

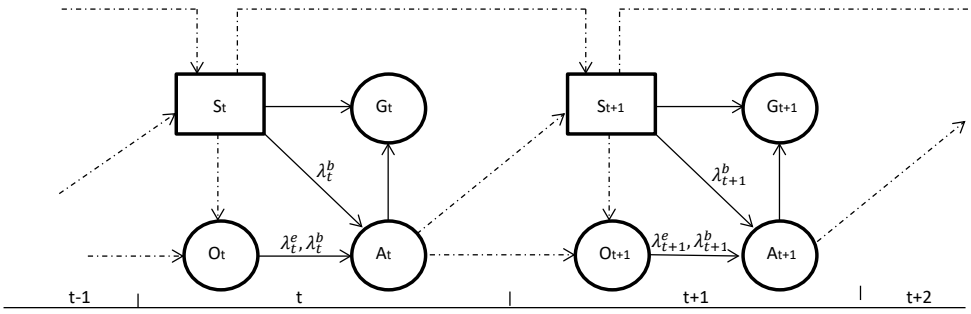

Figure 1 DAG representation of APOMDPs. *Circles:* observable variables; *Rectangles:* unobservable variables; *Solid arrows:* unambiguous causal mechanisms; *Dashed arrows:* ambiguous causal mechanisms.

In what follows, we first define APOMDPs and then develop two RL algorithms that enable finding the optimal policy by efficiently 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 (α, β, $\mathscr{S}$, $\mathscr{O}$, $\mathscr{A}$, $\mathscr{G}$, $\mathscr{P}$, $\mathscr{Q}$). The notation used in the first part of this tuple is as introduced earlier. $\mathscr{P}$ and $\mathscr{Q}$ 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 $\mathscr{P} \times \mathscr{Q}$ using $\mathscr{M}$ so that each $m \in \mathscr{M}$ represents a specific (unambiguous) POMDP model. In particular, associated with each $m \in \mathscr{M}$ is a set of the form $P_m \times Q_m$ with $P_m \in \mathscr{P}$ and $Q_m \in \mathscr{Q}$ denoting the set of state and observation transition probabilities under model m, respectively (Saghafian 2018). In this setting, (a) $P_m \triangleq \{P_m^a : a \in \mathscr{A}\}$, where for each $a \in \mathscr{A}$, $P_m^a \triangleq [p_{ij}^a(m)]_{i,j\in\mathscr{S}}$ is an $|\mathscr{S}| \times |\mathscr{S}|$ 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) $Q_m \triangleq \{Q_m^a : a \in \mathscr{A}\}$, where for each $a \in \mathscr{A}$, $Q_m^a \triangleq [q_{jo}^a(m)]_{j\in\mathscr{S},o\in\mathscr{O}}$ is an $|\mathscr{S}| \times |\mathscr{O}|$ 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 $\mathscr{M}$ was a singleton with its only member being m, the optimal gain and policy for any $t \in \mathscr{T}$ and $\boldsymbol{\pi} \in \Delta_{\mathscr{S}}$ could be obtained by a traditional POMDP Bellman equation (along with the terminal condition $V_0^m(\boldsymbol{\pi}) \triangleq 0$):

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

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 $T : \Delta_{\mathscr{S}} \times \mathscr{A} \times \mathscr{O} \times \mathscr{M} \to \Delta_{\mathscr{S}}$ is defined by the Bayes' rule (in the

matrix form):

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

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

Unlike POMDPs, in AMPOMDs $\mathscr{M}$ is not a singleton. However, it is shown in Saghafian (2018) that the APOMDP value function, a model independent function which 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 β on a complete metric space (under some mild conditions), which in turn allows 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 and efficient 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 $|\mathscr{M}| = 1$), and then 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, $\boldsymbol{\mu}^b$, satisfies *positivity* defined below.

Definition 6 (Positivity). We say that a policy $\boldsymbol{\mu} \triangleq (\mu_t)_{t\in\mathscr{T}}$ satisfies positivity if, and only if, there exists a constant $c_0 > 0$ such that $\mu_t(a_t|\boldsymbol{\pi}_t) \geq c_o$ for all $t \in \mathscr{T}$, $\boldsymbol{\pi}_t \in \Delta_{\mathscr{S}}$, and $a_t \in \mathscr{A}$

Positivity implies that all actions have a positive chance of being selected (appear in the observed data) regardless of the belief. The behavior policy, $\boldsymbol{\mu}^b$, automatically satisfies positivity when the data is 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 $\boldsymbol{\pi}_t$) 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 Luckett et al. (2020) for related results in settings with fully observable states).

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

$$V_{T-t+1}^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi}_t) = \mathbb{E}^m\left[\frac{\mu_t^e(A_t|\boldsymbol{\Pi}_t^m)}{\mu_t^b(A_t|\boldsymbol{\Pi}_t^m)}\Big[G_t + \beta\, V_{T-t}^{m,\boldsymbol{\mu}^e}(T(\boldsymbol{\Pi}_t^m, A_t, O_t, m))\Big]\Big|\boldsymbol{\Pi}_t^m = \boldsymbol{\pi}_t\right], \tag{10}$$

for all $t \in \mathscr{T}$ *and* $\boldsymbol{\pi}_t \in \Delta_{\mathscr{S}}$*, where* $V_0^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi}) \triangleq 0$*. Therefore, for any function* ϕ *defined on* $\Delta_{\mathscr{S}}$*, and for all* $t \in \mathscr{T}$*, we have:*

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

The importance of Proposition 3 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):

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

where $\mathbb{E}^{\mathbb{P}}$ denotes average with respect to the empirical probability measure.[14] It is important to note that while we are using sample-average in (12), the result still depends on the assumed m, because while the sequence $\{(A_t, O_t)\}_{t\in\mathscr{T}}$ is observable to us, to form the sequence $\{\boldsymbol{\Pi}_t^m\}_{t\in\mathscr{T}}$, we need to have an assumed model. That is, due to the existence of unobserved variables, the empirical measure alone is *insufficient* for our goal.

Remark 1 (**Efficient Approximation**). In establishing the properties of our approaches (e.g., consistency, asymptotic normality, etc.) 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 a data-efficient 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 POMDPs and APOMDP settings, we only assume *piecewise linearity* and *continuity*, but do not impose any assumption on convexity. This, in turn, helps us in another way: while piecewise linear and continuous functions can be efficiently 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).

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

Let $\mathscr{V}$ denote the set of real-valued piecewise linear and continuous bounded functions defined on $\Delta_{\mathscr{S}}$, and assume $V_t^{m,\boldsymbol{\mu}^e} \in \mathscr{V}$. To learn $V_t^{m,\boldsymbol{\mu}^e} \in \mathscr{V}$ using (12), we consider the parametric version of the value function: $V_t^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi};\boldsymbol{\psi}_t) \triangleq \big(\mathbf{b}(\boldsymbol{\pi})\big)'\boldsymbol{\psi}_t$, where $\mathbf{b}(\boldsymbol{\pi}) \triangleq \big(\mathbf{b}_1(\boldsymbol{\pi}), \mathbf{b}_2(\boldsymbol{\pi}), \cdots, \mathbf{b}_{d_t}(\boldsymbol{\pi})\big)'$ is a predefined *basis function* that allows us to ensure that the learned function is in $\mathscr{V}$, and $\boldsymbol{\psi}_t \in \boldsymbol{\Psi}_t \subseteq \mathbb{R}^{d_t}$ is the parameter.[15] This also enables us to set $\phi(\boldsymbol{\pi}) \triangleq \mathbf{b}(\boldsymbol{\pi})$ in (12), since $\mathbf{b}(\boldsymbol{\pi})$ can be thought of as the gradient of $V_t^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi};\boldsymbol{\psi}_t)$ with respect to its parameter, which only depends on $\boldsymbol{\pi}$ (and not the parameter) and is almost everywhere defined.

Furthermore, since $\boldsymbol{\psi}_t$ 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 backwards iteratively, having an estimation of $T-t$ periods to go value function in hand ($\hat{V}_{T-t}^{m,\boldsymbol{\mu}^e}$), we define

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

We then obtain the estimate

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

where $\boldsymbol{\Omega}$ is an arbitrary positive definite matrix, $\mathcal{P}(\cdot)$ is a penalty function, and θ_t is a tuning parameter.[16] Consequently, we plug in $\hat{\boldsymbol{\psi}}_t^{\boldsymbol{\mu}^e}$ in $V_{T-t+1}^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi}_t;\boldsymbol{\psi}_t)$ and thereby obtain an estimate for the value function $V_{T-t+1}^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi}_t)$, and move to the next period (backward). This procedure, under a given model $m \in \mathscr{M}$, yields an estimator for the gain under $\boldsymbol{\mu}^e$. That is, $\hat{\Gamma}_T^m(\boldsymbol{\mu}^e) \triangleq \int \hat{V}_T^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi})\,dF(\boldsymbol{\pi})$ can be used as an estimator for $\Gamma_T^m(\boldsymbol{\mu}^e) = \int V_T^{m,\boldsymbol{\mu}^e}(\boldsymbol{\pi})\,dF(\boldsymbol{\pi})$, where $dF(\boldsymbol{\pi})$ is a given distribution on (starting) belief values. Since we have an estimator for the gain under any policy $\boldsymbol{\mu}^e$, we can obtain $\hat{\boldsymbol{\mu}}^{e*} \triangleq \text{argmax}_{\boldsymbol{\mu}^e \in \Upsilon}\,\hat{\Gamma}_T^m(\boldsymbol{\mu}^e)$ as an estimate for the optimal policy under model m, where Υ is a given set of policies.[17] 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 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 $t \to \infty$ (see, e.g., Proposition 1 of Saghafian 2018). Therefore, in (12) we can

[15] Allowing the dimensionality of the parameter space, d_t, to depend on t can enable us 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 the next sections).

[16] In our case study, simulations experiments, and theoretical results, we make use of the squared Euclidean norm as the penalty function, and hence, assume $\mathcal{P}(\boldsymbol{\psi}_t) = \boldsymbol{\psi}_t'\boldsymbol{\psi}_t$.

[17] We use the "max" operator instead of "sup," because in most real-world applications, Υ 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 (C4) in Section 5. Furthermore, in various practical applications, Υ is often restricted to the set of policies that satisfy specific attributes such as fairness or interpretability.

replace both $V_{T-t}^{m,\boldsymbol{\mu}^e}(\cdot)$ and $V_{T-t+1}^{m,\boldsymbol{\mu}^e}(\cdot)$ with the same function. This removes the need for recursive calculations and allows us to follow a one-shot data-efficient 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 $|\mathscr{M}|=1$ 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 $\mathscr{M}$. 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 upfront that takes into account ambiguity and removes the need to obtain a value function separately for each POMDP model in $\mathscr{M}$.

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

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

Next, to estimate the optimal policy, we note that for any policy $\boldsymbol{\mu}^e$, the estimator of the gain is $\hat{\Gamma}_T(\boldsymbol{\mu}^e)=\int \hat{V}_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi})\,dF(\boldsymbol{\pi})$, 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 $\hat{\boldsymbol{\mu}}^{e*}\triangleq \operatorname{argmax}_{\boldsymbol{\mu}^e\in\Upsilon}\hat{\Gamma}_T(\boldsymbol{\mu}^e)$. 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 includes 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 $n\to\infty$ (see Section 5). Our estimation equations for the infinite-horizon gain are

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

and

$$\hat{\boldsymbol{\psi}}_n^{m,\boldsymbol{\mu}^e} = \operatorname*{argmin}_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\left\{\big(\varphi_n^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi})\big)'\,\boldsymbol{\Omega}\,\varphi_n^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi})+\theta_n\mathcal{P}(\boldsymbol{\psi})\right\}, \tag{17}$$

Algorithm 1: `DAV-Learning`

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

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

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 $m \in \mathscr{M}$ 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 $\hat{\Gamma}_\infty(\boldsymbol{\mu}^e) \triangleq \alpha \inf_{m\in\mathscr{M}} \hat{\Gamma}_\infty^m(\boldsymbol{\mu}^e) + (1-\alpha)\sup_{m\in\mathscr{M}} \hat{\Gamma}_\infty^m(\boldsymbol{\mu}^e)$, which provides an estimation for the overall gain $\Gamma_\infty(\boldsymbol{\mu}^e) \triangleq \alpha \inf_{m\in\mathscr{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e) + (1-\alpha)\sup_{m\in\mathscr{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)$.

Finally, the estimated optimal policy and its infinite-horizon value for the APOMDP are obtained as $\hat{\boldsymbol{\mu}}^{e*} \triangleq \operatorname{argmax}_{\boldsymbol{\mu}^e\in\Upsilon} \hat{\Gamma}_\infty(\boldsymbol{\mu}^e)$ 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 $\hat{\boldsymbol{\mu}}^{e*,m} \triangleq \operatorname{argmax}_{\boldsymbol{\mu}^e\in\Upsilon} \hat{\Gamma}_\infty^m(\boldsymbol{\mu}^e)$, 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 $\Gamma_\infty^m(\boldsymbol{\mu}^{e*,m}) \triangleq \max_{\boldsymbol{\mu}^e\in\Upsilon} \Gamma_\infty^m(\boldsymbol{\mu}^{e*,m})$.

4.2.2. Safe Augmented V-Learning (`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" is overcome at the end. In contrast, in `SAV-Learning`, this curse is overcome upfront 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 $V_t^{\boldsymbol{\mu}^e} \in \mathscr{V}$, and parameterize it via $V_t^{\boldsymbol{\mu}^e}(\boldsymbol{\pi};\boldsymbol{\psi}_t) \triangleq \big(\mathbf{b}(\boldsymbol{\pi})\big)' \boldsymbol{\psi}_t$. We then estimate its parameter as

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

where $\alpha \in \mathscr{I}$ can be viewed as a tuning parameter, $\underline{m} \triangleq \text{arginf}_{m\in\mathscr{M}} ||\hat{\boldsymbol{\psi}}_t^{m,\boldsymbol{\mu}^e}||$, $\overline{m} \triangleq \text{argsup}_{m\in\mathscr{M}} ||\hat{\boldsymbol{\psi}}_t^{m,\boldsymbol{\mu}^e}||$,[18] and

$$\hat{\boldsymbol{\psi}}_t^{m,\boldsymbol{\mu}^e} = \arg\min_{\boldsymbol{\psi}_t\in\boldsymbol{\Psi}_t} \Big\{ \big(\varphi^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi}_t)\big)' \boldsymbol{\Omega}\, \varphi^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi}_t) + \theta_t \mathcal{P}(\boldsymbol{\psi}_t) \Big\}, \tag{19}$$

where $\varphi^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi}_t)$ is defined in (13). Consequently, we plug $\hat{\boldsymbol{\psi}}_t^{\boldsymbol{\mu}^e}$ obtained in (18) in $V_{T-t+1}^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}_t;\boldsymbol{\psi}_t)$, which yields an estimate for the APOMDP value function $V_{T-t+1}^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}_t)$, and move to the next period (backwards) 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: $\hat{\boldsymbol{\mu}}^{e*} \triangleq \text{argmax}_{\boldsymbol{\mu}^e\in\Upsilon} \hat{\Gamma}_T(\boldsymbol{\mu}^e) = \text{argmax}_{\boldsymbol{\mu}^e\in\Upsilon} \int \hat{V}_T^{\boldsymbol{\mu}^e}(\boldsymbol{\pi})\, dF(\boldsymbol{\pi})$. 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 since 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.

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).[19]

The main results of this section are as follows. Under some conditions discussed below, we first establish weak consistency and asymptotic normality of the estimators under any policy $\boldsymbol{\mu}^e \in \Upsilon$

[18] We assume $\mathscr{M}$ is such that $\inf_{m\in\mathscr{M}} ||\boldsymbol{\psi}_t^{m,\boldsymbol{\mu}^e}||$ and $\sup_{m\in\mathscr{M}} ||\boldsymbol{\psi}_t^{m,\boldsymbol{\mu}^e}||$ are both finite, and $\underline{m}$ and $\overline{m}$ are both in $\mathscr{M}$.

[19] 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 Luckett et al. (2020).

Algorithm 2: SAV-Learning

for *each observed trajectory and model* $m \in \mathscr{M}$ **do**
  Initialize $\boldsymbol{\pi}_0^m$ using a random draw from $F(\boldsymbol{\pi})$;
  set t=1;
  while $t+1 \in \mathscr{T}$ **do**
    $\boldsymbol{\pi}_{t+1}^m \leftarrow T(\boldsymbol{\pi}_t^m, a_t, o_t, m)$;
for *any given* $\boldsymbol{\mu}^e \in \Upsilon$ *and* $m \in \mathscr{M}$ **do**
  $\varphi_n^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi}) \leftarrow \mathbb{E}^{\mathbb{P}}\left[\sum_{t\in\mathscr{T}}\left[\frac{\mu^e(A_t|\boldsymbol{\Pi}_t^m)}{\mu^b(A_t|\boldsymbol{\Pi}_t^m)}\left[G_t + \beta V_\infty^{m,\boldsymbol{\mu}^e}(T(\boldsymbol{\Pi}_t^m, A_t, O_t, m)) - V_\infty^{m,\boldsymbol{\mu}^e}(\boldsymbol{\Pi}_t^m)\right]\mathbf{b}(\boldsymbol{\Pi}_t^m)\right]\right]$;
  $\hat{\boldsymbol{\psi}}_n^{m,\boldsymbol{\mu}^e} \leftarrow \text{argmin}_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\left\{\left(\varphi_n^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi})\right)' \boldsymbol{\Omega}\, \varphi_n^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi}) + \theta_n \mathcal{P}(\boldsymbol{\psi})\right\}$;
for *any given* $\boldsymbol{\mu}^e \in \Upsilon$ **do**
  $\underline{m} \leftarrow \text{arginf}_{m\in\mathscr{M}} ||\hat{\boldsymbol{\psi}}_n^{m,\boldsymbol{\mu}^e}||$;
  $\overline{m} \leftarrow \text{argsup}_{m\in\mathscr{M}} ||\hat{\boldsymbol{\psi}}_n^{m,\boldsymbol{\mu}^e}||$;
  $\hat{\boldsymbol{\psi}}_n^{\boldsymbol{\mu}^e} \leftarrow \alpha\, \hat{\boldsymbol{\psi}}_n^{\underline{m},\boldsymbol{\mu}^e} + (1-\alpha)\, \hat{\boldsymbol{\psi}}_n^{\overline{m},\boldsymbol{\mu}^e}$;
  $\hat{V}_\infty^{\boldsymbol{\mu}^e}(\boldsymbol{\pi}) \leftarrow \left(\mathbf{b}(\boldsymbol{\pi})\right)' \hat{\boldsymbol{\psi}}_n^{\boldsymbol{\mu}^e}$;
  $\hat{\Gamma}_\infty(\boldsymbol{\mu}^e) \leftarrow \int \hat{V}_\infty^{\boldsymbol{\mu}^e}(\boldsymbol{\pi})\, dF(\boldsymbol{\pi})$;
$\hat{\boldsymbol{\mu}}^{e*} \leftarrow \text{argmax}_{\boldsymbol{\mu}^e\in\Upsilon} \hat{\Gamma}_\infty(\boldsymbol{\mu}^e)$;
$\hat{\Gamma}_\infty(\hat{\boldsymbol{\mu}}^{e*}) \leftarrow \max_{\boldsymbol{\mu}^e\in\Upsilon} \hat{\Gamma}_\infty(\boldsymbol{\mu}^e)$;

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

(C1) For every $\boldsymbol{\mu}^e \in \Upsilon$ and $m \in \mathscr{M}$ there exist a unique solution to $\varphi^{m,\boldsymbol{\mu}^e}(\boldsymbol{\psi}) = 0$ denoted by $\boldsymbol{\psi}_\diamond^{m,\boldsymbol{\mu}^e} \in \boldsymbol{\Psi} \subseteq \mathbb{R}^d$, 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 $\boldsymbol{\Psi}$, and $\boldsymbol{\Psi}$ is compact subset of $\mathbb{R}^d$.

(C2) There exists a $2 < \rho < \infty$ such that for all $m \in \mathscr{M}$:

  (C2a) The class of policies (Υ) 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 $\{(\boldsymbol{\Pi}_t^m, A_t)\}_{t\geq 1}$.[20]

[20] For the definition of the bracketing integral, $J_{[]}(\infty, \Upsilon, L_\rho(P^m))$, see, e.g., Kosorok (2008).

(C2b) The sequence $\{(\mathbf{\Pi}_t^m, A_t)\}_{t\geq 1}$ is an absolutely regular stationary process with its β-mixing coefficients $\zeta^m(t)$ satisfying $\sum_{t=1}^{\infty} k^{2/(\rho-2)}\zeta^m(t) < \infty$.[21]

(C3) There exists a constant $c_1 > 0$ such that for all $m \in \mathscr{M}$, $t \in \mathscr{T}$, $\boldsymbol{\mu}^e \in \Upsilon$, and $\mathbf{c} \in \mathbb{R}^d$:

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

(C4) $\boldsymbol{\mu}^{e*}$ is a unique and well septated maximizer of $\Gamma_\infty(\boldsymbol{\mu}^e)$ and $\boldsymbol{\mu}^{e*}$ is in the interior Υ.

(C5) For every $\boldsymbol{\mu}^e \in \Upsilon$: $|\inf_{m\in\mathscr{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)| < \infty$, $|\sup_{m\in\mathscr{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)| < \infty$, and $\mathscr{M}$ contains both $\operatorname{arginf}_{m\in\mathscr{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)$ and $\operatorname{argsup}_{m\in\mathscr{M}} \Gamma_\infty^m(\boldsymbol{\mu}^e)$.

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 6.1) and in our simulation experiments (Section 6.2). (C1) 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." (C2a) is a regularity condition on the policy space, and requires that the set of policies under consideration satisfy 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). (C2b) is a regulatory condition on the space of trajectories and allows viewing their formation as a suitable stationary process. The β-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. (C3) ensures that the matrix $\mathbf{C}^m(\boldsymbol{\mu}^e)$ defined in Theorem 1 below is positive-definite, and hence, invertible. One can empirically check whether (C3) holds by creating certain matrixes using data and testing whether they are positive-definite. (C4) 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. (C5) is a regularity condition on the space of models, $\mathscr{M}$, which holds in most real-wrold 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 (C1)-(C3). The proof is based on some additional results provided in Appendix B (see Lemmas EC.2 and EC.3), which establish Donsker properties and asymptotic normality in $\ell^\infty(\Upsilon)$ for the underlying absolutely regular stationary process in our setting .

[21] For the definition of an absolutely regular stationary process and its β-mixing coefficients, see, e.g., Dedecker and Louhichi (2002), Kosorok (2008), and the references therein.

THEOREM 1 **(Asymptotic Behavior: Fixed Policy and its Value)**. *Suppose (C1)-(C3) hold and the behavior policy satisfies positivity. Then under* `DAV-Learning` *(Algorithm 1), for any* $\boldsymbol{\mu}^e \in \Upsilon$ *and* $m \in \mathscr{M}$*, we have:*

(i) $\hat{\boldsymbol{\psi}}_n^{m,\boldsymbol{\mu}^e} \xrightarrow{p} \boldsymbol{\psi}_\diamond^{m,\boldsymbol{\mu}^e}$.

(ii) $\sqrt{n}\left[\hat{\boldsymbol{\psi}}_n^{m,\boldsymbol{\mu}^e} - \boldsymbol{\psi}_\diamond^{m,\boldsymbol{\mu}^e}\right] \xrightarrow{d} \mathbb{G}(\boldsymbol{\mu})$ *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*

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

where

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

$$\tilde{\mathbf{C}}^m(\boldsymbol{\mu}^e, \tilde{\boldsymbol{\mu}}^e) \triangleq \mathbb{E}^m\left[\frac{\mu^e(A_t|\boldsymbol{\Pi}_t^m)\tilde{\mu}^e(A_t|\boldsymbol{\Pi}_t^m)}{\mu^b(A_t|\boldsymbol{\Pi}_t^m)\,\mu^b(A_t|\boldsymbol{\Pi}_t^m)}\,\boldsymbol{\vartheta}(\boldsymbol{\Pi}_t^m, \boldsymbol{\psi}_\diamond^{\boldsymbol{\mu}^e})\,\boldsymbol{\vartheta}(\boldsymbol{\Pi}_t^m, \boldsymbol{\psi}_\diamond^{\tilde{\boldsymbol{\mu}}^e})\,\mathbf{b}(\boldsymbol{\Pi}_t^m)\Big(\mathbf{b}(\boldsymbol{\Pi}_t^m)\Big)'\right], \tag{23}$$

and

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

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

(iv) $\hat{\Gamma}_\infty(\boldsymbol{\mu}^e) \xrightarrow{p} \Gamma_\infty(\boldsymbol{\mu}^e)$ *assuming (C5) 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 Appendix B (see Lemma EC.4), which in turn relies on results from the M-estimation theory.

THEOREM 2 **(Asymptotic Behavior: Optimal Policy and its Value)**. *Suppose (C1)-(C5) hold and the behavior policy satisfies positivity. Then, considering a metric space* (Υ, d_Υ)*, under* `DAV-Learning` *(Algorithm 1) we have:*

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

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

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

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

6. Performance Analyses: Numerical Results

To gain further insights into the performance of our purposed 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.

6.1. Case Study: New Onset Diabetes After Transplantation (NODAT)

In this section, we apply our proposed algorithms (`DAV-Learning` and `SAV-Learning`) on a clinical data set that contains over 63,000 data points pertaining 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 (Boloori et al. 2015, 2020, Munshi et al. 2020b, 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, 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., Chakkera et al. 2009, Boloori et al. 2015, 2020, and the references therein). To control the risk of NODAT, physicians have to decide whether or not to put the patient on insulin.

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 (FPG and Hb1Ac) 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 describe the main condition of the patient in terms of decision-making related to 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, since 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 *Fasting Plasma Glucose* (FPG) and *Hemoglobin A1c* (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.

Data Pre-processing 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 (Boloori 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 to 12 after

Table 1 Observed Covariates (Observations)

Var. No.	Risk Factor (Abbr.)	Unit	Low Level	Mid Level	High Level	Time-Varying
1	Glucose test† (FPG, HbA1c)	mg/dL, %	Healthy	Pre-Diabetic	Diabetic	Yes
2	Trough level test‡ (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/m²	<30 (non-obese)	—	≥30 (obese)	Yes
8	Blood pressure (BP)	—	Normal♯	—	Hypertension	Yes
9	Total cholesterol (Chol)	mg/dL	<200	—	≥200	Yes
10	High-density lipoportein (HDL)	mg/dL	≥40	—	<40	Yes
11	Low-density lipoportein (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

†A patient with FPG≥126 (100 ≤FPG< 126) mg/dL or HbA1c≥6.5% (5.7 ≤HbA1c<6.5%) is labeled as diabetic (pre-diabetic), and a patient with FPG<100 mg/dL or HbA1c<5.7% is labeled as healthy (see, e.g., ADA 2012).
‡$C_0 \in [4,8)$, $[8,10)$, $[10,14]$ mg/dL is label as "low," "medium," and "high," respectively (see, e.g., Boloori et al. 2020).
♯Normal Blood Pressure (BP) is defined as systolic (diastolic) BP less than 120 (80) mmHg (see, e.g., Whelton et al. 2017).
Note: All variables with three levels are coded as 1,2, 3 (low, mid, high). All variables with two levels are coded as 1, 2 (low, high).

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 $\mathscr{T} \triangleq \{1, 2, \cdots, 12\}$. The imputed data includes 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).

Behavior Policy. We estimate the behavior policy based on the actions we observe in our data. These actions are mainly based on the 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 anti-thymocyte, immunoglobulin, or basiliximab, as well as a tapering course of glucocorticoids. However, here our focus is on the use of tacrulimus, and we observe that patients are often put on high (i.e., aggressive) dose tacrolimus during the first months post-transplant, and in later months, depending on the observations made about the patient patients, they might be transferred to a low (i.e., non-aggressive) dose. This is consistent with the fact that patients in most medical practices are consistently kept on high levels of tacrolimus in 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 Hb1Ac and FPG tests indicates that they are not diabetic free (see definitions of pre-diabetic and diabetic in Table 1). Using the observed actions in our data set as well as the estimated belief vectors $\{\boldsymbol{\pi}_t^m\}_{t\in\mathscr{T}}$ for each patient (for further details, see the "Time-Dependent Belief Vectors" paragraph below), we next estimate $\mu^b(A_t|\boldsymbol{\Pi}_t^m)$ by training a multi-class multiple logistic regression classifier. This classifier is endowed with an ℓ_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).

Table 2 Latent Health States

State	Transplant Condition (Tacrolimus C_0)	Diabetes Condition
1	Low	Diabetes (type II)
2	Medium	
3	High	
4	Low	Pre-diabetes
5	Medium	
6	High	
7	Low	Healthy
8	Medium	
9	High	

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 *Quality of Life (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 concentration of the immunosuppressive 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, which are divided by 12 to represent the fact that patients' visits are monthly.[22]

Other Details. The belief state space in our setting, $\Delta_{\mathscr{S}}$, is a 8-simplex, since there are 9 latent states (Table 2). The vector of basis functions $\mathbf{b}(\boldsymbol{\pi})$ 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 $\mathscr{M}$, 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,

[22] The QoL values shown in Table 4 are average values and are approximate values based on various reports in the medical literature (see, e.g., the extended appendix of Boloori et al. 2020, and the references therein). In addition to immediate gains, our framework allows 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 $V_0(\boldsymbol{\pi}) \triangleq 0$.

Table 3 Actions

Action	Prescription (Tacrolimus dose)	Prescription (Insulin use)
1	Low (Non-Aggressive)	No
2	High (Aggressive)	
3	Low (Non-Aggressive)	Yes
4	High (Aggressive)	

Table 4 Immediate Gain Values

State	Transplant Condition (Tacrolimus C_0)	Diabetes Condition	Immediate Gain Value†
1	Low	Diabetes (type II)	0.68/12
2	Medium		0.72/12
3	High		0.76/12
4	Low	Pre-diabetes	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

†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.

an entropy ball is constructed (using the Kullback–Leibler divergence criterion) around these point estimate matrices. For tractability, we set $|\mathscr{M}| = 4$ in this case study. However, our framework is general and can be used for any number of estimated models. In Section 6.2, for example, we change our assumption on the number of models and consider $|\mathscr{M}| = 10$ different models. Our framework is also not restricted to any specific way of estimating the underlying models. For example, in Section 6.2, we make use of a different way of constructing the set $\mathscr{M}$. Finally, we consider the distribution $F(\boldsymbol{\pi})$ to be uniform. That is, we use a uniform prior belief at time zero, and implement the Bayesian belief updating operator (see Eq. 9) to create a sequence of belief vectors $\{\boldsymbol{\pi}_t^m\}_t \in \mathscr{T}$ for each patient under each model $m \in \mathscr{M}$ (see, e.g., steps 1-5 in Algorithms 1 and 2).

Results. The performance 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.[23] 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, α. 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 α. Of note, these ranges also imply that the mean performance of the `SAV-Learning` regime is much more robust to the value of α than that of `DAV-Learning`. This is due to 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

[23] The number of these replications is chosen so that the confidence intervals are tight enough, while maintaining reasonable computational times.

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

Pessimism Level (α)	Observed Regime†	DAV-Learning†	SAV-Learning†
0.00	1.472 (1.455, 1.489)	**2.085** (2.061, 2.108)	1.949 (1.770, 2.128)
0.25	1.468 (1.456, 1.480)	**1.939** (1.920, 1.958)	1.888 (1.566, 2.210)
0.50	1.464 (1.457, 1.471)	**1.794** (1.779, 1.808)	1.786 (1.534, 2.039)
0.75	1.460 (1.458, 1.462)	1.648 (1.638, 1.658)	**1.682** (1.658, 1.706)
1.00	1.455 (1.452, 1.458)	**1.609** (1.560, 1.657)	1.606 (1.585, 1.627)

†Values in parenthesis represent 95% confidence intervals. Values in bold font represent the best performance. For all values, only the first three decimal places are shown.

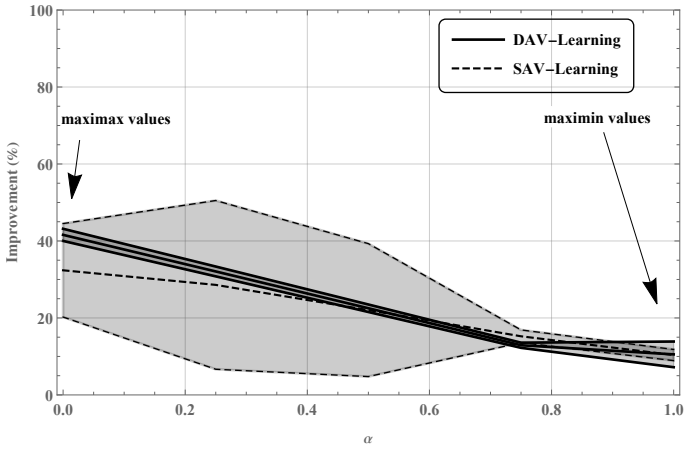

Figure 2 Percentage improvement over the observed regime (case study with $\beta = 0.95$). Gray areas represent error bands with the curve at the center of each error band representing the mean value.

uses `SAV-Learning` does not need to be that concerned about the value of α s/he 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 α increases. This is fully expected, since 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 6.3, and generate important insights into the values of α that can provide the highest level of robustness to model ambiguity.

6.2. Synthetic Data Analyses

We now use similar assumptions to those described in the case study, but instead of using actual patient traceries, simulate random patient trajectories for 100 patients with 10 follow-up periods, and use ($|\mathscr{M}| = 10$) different models. These yield randomly generated belief data of the form $(\boldsymbol{\pi}_t^m)_{t\in\mathscr{T}}$ under each $m \in \mathscr{M}$. 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 $m \in \mathscr{M}$ the belief vector $(\boldsymbol{\pi}_t^m)_{t\in\mathscr{T}}$ is generated via a Dirichlet distribution with the vector of parameters $(p_i^m)_{i\in\{1,2\cdots,9\}}$. All of these

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

Pessimism Level (α)	**Observed Regime**†	`DAV-Learning`†	`SAV-Learning`†
0.00	1.441 (1.440, 1.442)	**1.973** (1.969, 1.977)	1.442 (1.441, 1.442)
0.25	1.415 (1.415, 1.416)	**1.815** (1.811, 1.818)	1.434 (1.433, 1.434)
0.50	1.389 (1.389, 1.390)	**1.656** (1.654, 1.659)	1.428 (1.428, 1.429)
0.75	1.364 (1.364, 1.364)	**1.498** (1.496, 1.499)	1.434 (1.434, 1.434)
1.00	1.338 (1.338, 1.339)	1.348 (1.348, 1.348)	**1.444** (1.444, 1.444)

†Values in parenthesis represent 95% confidence intervals. Values in bold font represent the best performance. For all values, only the first three decimal places are shown.

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

$$\mu^b(A = a|\mathbf{\Pi} = \boldsymbol{\pi}) = \frac{exp(\boldsymbol{\pi}'\,\boldsymbol{\varrho}_a)}{1 + \sum_{a=1}^{3} exp(\boldsymbol{\pi}'\,\boldsymbol{\varrho}_a)}, \tag{25}$$

and for action $a = 4$ we set

$$\mu^b(A = a|\mathbf{\Pi} = \boldsymbol{\pi}) = \frac{1}{1 + \sum_{a=1}^{3} exp(\boldsymbol{\pi}'\,\boldsymbol{\varrho}_a)}. \tag{26}$$

where $\boldsymbol{\varrho}_1$, $\boldsymbol{\varrho}_2$, and $\boldsymbol{\varrho}_3$ are 9-dimensional predefined vectors. To perform our analyses, we choose each $\boldsymbol{\varrho}_a$ $(a = 1, 2, 3)$ as a vector with all elements equal to 0.1, except the a-th element, which is set to -1.

Table 6 and Figure 3 present our results using the same immediate gain values as those in the case study (see Table 4). Similar to the case study, we observer that both `DAV-Learning` and `SAV-Learning` approaches outperform the observed regime. The percentage improvement of `DAV-Learning` an `SAV-Learning` over the observe regime ranges in $(1\%, 37\%)$ and $(1\%, 8\%)$, respectively, depending on the value of α. In addition, similar to our observation in the case study, `DAV-Learning` outperforms `SAV-Learning` for most vales of the pessimism level, α. Another similar observation is that the performance of `SAV-Learning` is much more robust to the value of α compared to `DAV-Learning`. Hence, a decision-maker who uses `SAV-Learning` has the advantage that s/he does not need to be that concerned with the value of α that s/he uses (or with tuning it). In the next section, we further investigate the robustness of our proposed approaches to model ambiguity, and generate insights into the best value of α that a decision-maker can use to achieve the highest level of robustness.

6.3. Robustness to Model Ambiguity

We now compare our proposed approaches in terms of their percentage *gain loss* (a.k.a. *regret*). That is, we first consider an *oracle* who knows both the true data generating model and the optimal policy under it, and then compare the performance of a decision-maker who is blind to the true

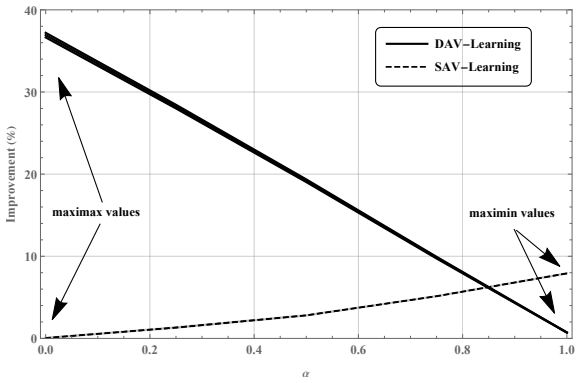

Figure 3 Percentage improvement over the observed regime (synthetic data analyses with $\beta = 0.95$). Error bands for both approaches, and especially for the `SAV-Learning` approach, are very tight (hence, not depicted).

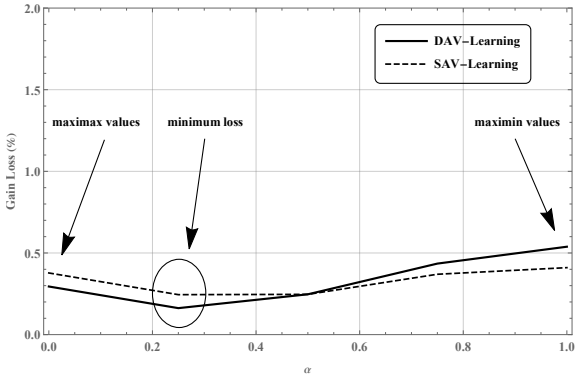

Figure 4 Percentage Gain Loss of `DAV-Learning` and `SAV-Learning` (Synthetic Data Analyses with $\beta = 0.95$). Minimum loss is obtained for a mid level value of the pessimism level ($\alpha = 0.25$).

data generating model (is facing model ambiguity) but uses either `DAV-Learning` or `SAV-Learning`. How much robustness to model ambiguity using the proposed `DAV-Learning` or the `SAV-Learning` approaches provide? What is the maximum gain loss of these approaches? For what value of α the gain loss is minimized? Importantly, in order to minimize the gain loss, should the decision-maker use an extreme value of α (e.g., $\alpha = 0, 1$) or a mid level value (e.g., $\alpha = 0.5$)? And does the answer depend on which of the two learning approaches is used?

To answer these questions, we make use of a similar setup to the one discussed in Section 6.2. The results are shown in Figure 4, which depicts the percentage gain loss of `DAV-Learning` and `SAV-Learning` compared to the imaginary oracle. From this figure, we make three main observations: (1) Gain loss has a U-shape curve as α varies. Importantly, the minimum loss for both `DAV-Learning` and `SAV-Learning` are obtained at a mid value of α (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 maximin view is *robustness-maximizing*. (2) The gain loss under `SAV-Learning` is much more robust to the changes in value of α compared to `DAV-Learning`, which is consistent

with our observations in Sections 6.1 and 6.2 that the performance of the `SAV-Learning` regime is in general more robust to the value of α. (3) Both `DAV-Learning` and `SAV-Learning` are able to strongly shield against model ambiguity, regardless of the value of α used. Specifically, the gain loss under these approaches (compared to 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).

7. 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 is observational, the behavior policy is unknown, and the existence of time-varying unmeasured confounders (which are themselves affected by previous actions) make 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 prosed learning algorithms (`DAV-Learning` and `SAV-Learning`) 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. Finally, future research can also examine the interpretability of the policies that are obtained via `DAV-Learning` and `SAV-Learning`, and propose adjustments (if needed) to ensure that they can be effectively used in practice.

References

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

Ahn, D., S. Choi, D. Gale, S. Kariv. 2014. Estimating ambiguity aversion in a portfolio choice experiment. *Quantitative Economics* **5**(2) 195–223.

Angrist, J.D., G.W. Imbens, D.B. Rubin. 1996. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91** 434—471.

Arrow, K. J., L. Hurwicz. 1977. An optimality criterion for decision making under ignorance. K. J. Arrow, L. Hurwicz, eds., *Studies in Resource Allocation Processes*. Cambridge University Press.

Arrow, K.J. 1951. Alternative approaches to the theory of choice in risk-taking situations. *Econometrica* **19**(4) 404–437.

Athey, S., S. Wager. 2021. Policy learning with observational data. *Econometrica* **89**(1) 133–161.

Bang, H., J.M. Robins. 2021. Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**(4) 962–972.

Bennett, A., N. Kallus, L. Li, A. Mousavi. 2021. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. 1999–2007.

Bhidé, A.V. 2000. *The Origin and Evolution of New Business*. Oxford University Press, Oxford.

Boloori, A., S. Saghafian, H.A. Chakkera, C.B. Cook. 2015. Characterization of remitting and relapsing hyperglycemia in post-renal-transplant recipients. *PLOS ONE* **10**(11) 1–16.

Boloori, A., S. Saghafian, H.A. Chakkera, C.B. Cook. 2020. Data-driven management of post-transplant medications: An ambiguous partially observable Markov decision process approach. *Manufacturing and Service Operations Management* **22**(5) 1066–1087.

Box, G. 1979. Robustness in the strategy of scientific model building. R. Launer, G. Wilkinson, eds., *Robustness in Statistics*. Academic Press, NY, 201–236.

Chakkera, H. A., E. J. Weil, J. Castro, R. L. Heilman, K. S. Reddy, M. J. Mazur, K. Hamawi, D. C. Mulligan, A. A. Moss, K. L. Mekeel, F. G. Cosio, C. B. Cook. 2009. Hyperglycemia during the immediate period after kidney transplantation. *Clinical Journal of the American Society of Nephrology* **4** 853–859.

Chakraborty, B., S.A. Murphy. 2014. Dynamic treatment regimes. *Annual Review of Statistics and Its Application* **1**(1) 447–464.

Dedecker, J., S. Louhichi. 2002. Maximal inequalities and empirical central limit theorems. T. Mikosch, M. Sørensen, eds., *Empirical Process Techniques for Dependent Data*. Birkhäuser, Boston, 137–159.

Ghiradato, P, F Maccheroni, M Marinacci. 2004. Differentiating ambiguity and ambiguity attitude. *Journal of Economic Theory* **118** 133–173.

Ghisdal, L., S. Van Laecke, M.J. Abramowicz, R. Vanholder, D. Abramowicz. 2012. New-onset diabetes after renal transplantation risk assessment and management. *Diabetes Care* **35**(1) 181–188.

Hansen, L.P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* **50**(4) 1029–1054.

Heath, C., A. Tversky. 1991. Preference and belief: ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty* **4**(1) 5–28.

Hurwicz, L. 1951a. Optimality criteria for decision making under ignorance. *Cowles Commission discussion paper: Statistics no. 370* .

Hurwicz, L. 1951b. Some specification problems and applications to econometric models. *Econometrica* **19** 343–344.

Jiang, N, L. Li. 2016. Doubly robust off-policy value evaluation for reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*. 652–661.

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

Kallus, N., A. Zhou. 2020. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *arXiv preprint arXiv:2002.04518* .

Kallus, N., A. Zhou. 2021. Minimax-optimal policy learning under unobserved confounding. *Management Science* **67**(5) 2870–2890.

Kosorok, Michael R, Eric B Laber. 2019. Precision medicine. *Annual Review of Statistics and its Application* **6**(263–286) 1243–1254.

Kosorok, M.R. 2008. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York, NY.

Leqi, L., E.H. Kennedy. 2021. Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802* .

Luckett, D.J., E.B. Laber, A.R. Kahkoska, D.M. Maahs, E. Mayer-Davis, M.R. Kosorok. 2020. Estimating dynamic treatment regimes in mobile health using V-Learning. *Journal of the American Statistical Association* **115**(530) 692–706.

Magnani, A., S.P. Boyd. 2009. Convex piecewise-linear fitting. *Optimization and Engineering* **10** 1–17.

Manski, C.F. 2007. *Identification for Prediction and Decision*. Harvard Univeristy Press, Cambridge, MA.

Manski, C.F. 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, V.N., S. Saghafian, C.B. Cook, S. Aradhyula, H.A. Chakkera. 2021. Use of imputation and decision modeling to improve diagnosis and management of patients at risk for newonset diabetes after transplantation. *Annals of Transplantation* **26** 1–9.

Munshi, V.N., S. Saghafian, C.B. Cook, D. Steidley, B. Hardaway, H.A. Chakkera. 2020a. Incidence, risk factors, and trends for post-heart transplantation diabetes mellitus. *The American Journal of Cardiology* **125**(3) 436–440.

Munshi, V.N., S. Saghafian, C.B. Cook, K.T. Werner, H.A. Chakkera. 2020b. Comparison of post-transplantation diabetes mellitus incidence and risk factors between kidney and liver transplantation patients. *PLOS ONE* **15**(1) 1–12.

Murphy, S.A. 2003. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2) 331–355.

Murphy, S.A. 2005. An experimental design for the development of adaptive treatment strategies. *Satitics in Medicine* **24**(10) 1455–1481.

Murphy, S.A., Y. Deng, E.B. Laber, H.R Maei, R.S. Sutton, K. Witkiewitz. 2016. A batch, off-policy, actor-critic algorithm for optimizing the average reward. *arXiv preprint arXiv:1607.05047* .

Murphy, S.A., M.J. van der Laan, J.M. Robins, CPPRG. 2001. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* **96**(456) 1410–1423.

Namkoong, H., R. Keramati, S. Yadlowsky, E. Brunskill. 2020. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623* .

Nie, X., E. Brunskill, S. Wager. 2021. Learning when-to-treat policies. *Journal of the American Statistical Association* **116**(533) 392–409.

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J., J. Robins. 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables. P. Besnard, S. Hanks, eds., *Uncertainty in Artificial Intelligence 11*. Morgan Kaufmann, San Francisco, 444—453.

Precup, D., R.S. Sutton, S. Singh. 2000. Eligibility traces for off-policy policy evaluation. *Proceedings of the 17th International Conference on Machine Learning*. 759–66.

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. *Mathematical Modeling* **7**(9-12) 1393–1512.

Robins, J. 1997. Causal inference from complex longitudinal data. *Latent variable modeling and applications to causality*. Springer, 69–117.

Robins, J. 2004. Optimal structural nested models for optimal sequential decisions. *Proceedings of the Second Seattle Symposium in Biostatistics*. Springer, 189–326.

Robins, J., M.A. Hernán, B. Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* **11**(5) 550–560.

Rosenbaum, PR. 2002. *Observational Studies*. Springer, New York, NY.

Rosenbaum, PR. 2010. *Design of Observational Studies*. Springer, New York, NY.

Rubin, D.B. 1986. Comment: Which ifs have causal answers. *Journal of the American Statistical Association* **81** 961—962.

Saghafian, S. 2018. Ambiguous partially observable Markov decision processes: Structural results and applications. *Journal of Economic Theory* **178** 1–35.

Saghafian, S., S.A. Murphy. 2021. Innovative health care delivery: The scientific and regulatory challenges in designing mHealth interventions. *NAM Perspectives. Commentary, National Academy of Medicine, Washington, DC.* .

Saghafian, S., M. Rasouli. 2019. Robust partially observable Markov decision processes. *Working Paper, Harvard University* .

Saghafian, S., B.T. Tomlin. 2016. The newsvendor under demand ambiguity: Combining data with moment and tail information. *Operations Research* **64**(1) 167–185.

Savage, L. 1951. The theory of statistical decision. *Journal of the American Statistical Association* **46** 55–67.

Smallwood, R., E.J. Sondik. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations Research* **21**(5) 1071–1088.

Stoy, J. 2011. Statistical decisions under ambiguity. *Theory and Decision* **70**(2) 129–148.

Tennenholtz, G., U. Shalit, Sh. Mannor. 2020. Off-policy evaluation in partially observable environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34. 10276–10283.

Thomas, Ph., E. Brunskill. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. *Proceedings of the 33rd International Conference on Machine Learning*. 2139–2148.

Tsiatis, A.A., M. Davidian, S.T. Holloway, E.B. Laber, M.R. Kosorok. 2019. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Chapman and Hall/CRC, Boca Raton.

Wald, A. 1939. Contribution to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics* **10** 299–326.

Wald, A. 1945. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics* **46** 265–280.

Wald, A. 1950. *Statistical Decision Functions*. Wiley, New York, NY.

Wang, L., Y. Zhou, R. Song, B. Sherwood. 2018. Quantile-optimal treatment regimes. *Journal of the American Statistical Association* **113**(523) 1243–1254.

Watson, J., C. Holmes. 2016. Approximate models and robust decisions. *Statistical Science* **31** 465–489.

Whelton, P.K., R.M. Carey, W.S. Aronow, D.E. Casey Jr, K.J. Collins, et al. 2017. 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. *Journal of the American College of Cardiology* **71**(19) e127–e248.

Xu, Z., E. Laber, A.M. Staicu, E. Severus. 2020. Latent-state models for precision medicine. *arXiv preprint arXiv:2005.13001* .

Zhang, J., E. Bareinboim. 2019. Near-optimal reinforcement learning in dynamic treatment regimes. *Advances in Neural Information Processing Systems*, vol. 32.

Zhang, Y., E.B. Laber, M. Davidian, A.A. Tsiatis. 2018. Interpretable dynamic treatment regimes. *Journal of the American Statistical Association* **113**(524) 1541–1549.

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