The Best Decisions Are Not the Best Advice: Making Adherence-Aware Recommendations

Julien Grand-Clément
Information Systems and Operations Management Department, Ecole des hautes etudes commerciales de Paris, grand-clement@hec.fr

Jean Pauphilet
Management Science and Operations, London Business School, jpauphilet@london.edu

Many high-stake decisions follow an expert-in-loop structure in that a human operator receives recommendations from an algorithm but is the ultimate decision maker. Hence, the algorithm's recommendation may differ from the actual decision implemented in practice. However, most algorithmic recommendations are obtained by solving an optimization problem that assumes recommendations will be perfectly implemented. We propose an adherence-aware optimization framework to capture the dichotomy between the recommended and the implemented policy and analyze the impact of partial adherence on the optimal recommendation. Our framework provides useful tools to analyze the structure and to compute optimal recommendation policies that are naturally immune against such human deviations, and are guaranteed to improve upon the baseline policy.

Key words: Expert-in-the-loop systems; Prescriptive analytics; Recommender systems; Discretion; Markov Decision Processes

History: .

1. Introduction

While some decisions can be automated and made directly by algorithms based on artificial intelligence (AI), many high-stake decisions follow an expert-in-loop structure in that an expert decision maker (e.g., a doctor) receives information, predictions, or even recommendations, and decides which course of action to follow. Consequently, the human decision maker (DM) does not systematically implement what the algorithm recommended. In other words, they may have a discretionary power to override/reject the recommendations from the algorithm, hence impacting the potential benefits from the AI tool. For instance, in a field experiment, Kesavan and Kushwaha (2020)

observed that merchants overrode the recommendations from a data-driven decision tool 71.24% of the time, resulting in a 5.77% reduction in profitability.

To understand this phenomenon and its ultimate impact on the quality of the decision being made, a growing body of literature has investigated the mechanisms driving non- or partial adherence of humans to algorithmic recommendations. In this work, we ask a complementary question: Given the fact that the decision maker will partially implement recommendations made by an algorithm, should we adjust these recommendations in the first place and how? In other words, we investigate the impact of partial adherence on algorithm design and decision recommendation. Our main contributions are as follows.

A new model of partial adherence. We consider a model of sequential decision-making based on Markov decision processes (MDPs) and assume that the decision maker currently follows a *baseline* policy π_{base} (or state of practice) and is provided with a *recommendation* policy π_{alg} by an algorithm. We propose a framework, namely adherence-aware MDP, to compute recommendations that are immune against human deviations. Our framework is behavioral in that it models the human switching behavior between their baseline policy and the algorithmic recommendations, but without specifying *why* these deviations are undertaken by the DM. Despite its simplicity, we show that our model is consistent with five different models for the DM's adherence decision, including random or adversarial adherence decisions. Furthermore, we provide examples where the co-existence of the human DM and the algorithmic recommendations performs either strictly worse or strictly better than any of the two policies alone, hence illustrating the ability of our model to capture the rich range of situations observed in practice. In particular, we show that (even rare) human deviations from algorithmic recommendations can lead to arbitrarily poor performance compared with both the expected performance of the algorithm and that of the current state of practice. In other words, we show that deploying a recommendation engine that was designed assuming its recommendations will be final decisions can have a dramatic impact on the effective performance. This set of negative results underscores the importance of accounting for the current baseline and the partial adherence phenomenon when building recommendation systems.

A tractable, structured, and flexible model. We study the appealing structural and computational properties of our adherence-aware MDP framework. In particular, we show that an optimal recommendation policy may be chosen stationary and deterministic, which is important from an implementation standpoint, and that it may be computed efficiently by a reduction to a classical MDP problem. We also show several structural properties, such as piecewise constant optimal recommendation policy and monotonicity of the optimal return (both as regards the adherence level). We identify classes of MDPs for which the decision maker may overlook the issue of partial adherence at some states (i.e., where the partial adherence phenomenon has no impact on the algorithmic recommendation to be made). We finally present extensions of our framework, including models where the adherence levels are state-dependent, action-dependent, uncertain, or where the baseline policy is not entirely known.

Numerical study. We evaluate the practical impact of our model on a series of numerical experiments. Our simulations highlight the importance of accounting for the potential non-adherence of the decision maker, showing empirically that severe performance deteriorations can happen when partial adherence is overlooked in the search for an optimal policy. The magnitude of this performance deterioration depends both on the current baseline policy and on the level of adherence of the decision maker. Consequently, in addition to classical sensitivity and robustness analyses used in the literature, we encourage practitioners to conduct a systematic *adherence-robustness* analysis of their algorithms to assess their effective performance prior to deployment.

The rest of the paper is organized as follows: We present related work from the operations literature in Section 2. Section 3 introduces our framework for sequential decision-making under partial adherence, discusses its connection with various models for the DM's adherence decision, and provides examples of situations where the co-existence of human and algorithmic decisions leads to improved or, on the contrary, impaired system performance. In Section 4, we present algorithms to compute optimal recommendation policies, and we analyze their structural properties and their sensitivity to the adherence level. We illustrate the practical impact of imperfect adherence and the value of our framework on numerical experiments in Section 5. Finally, we discuss extensions of our framework in Section 6.

2. Literature review

Our paper contributes to the rich literature of behavioral operations that studies the partial adherence of decision makers to machine recommendations. This phenomenon is also referred to in the literature as *discretion*, *overriding*, or *deviation*.

Many field studies have documented this phenomenon in a wide range of tasks and industries such as demand forecasting (Fildes et al. 2009, Kremer et al. 2011, Kesavan and Kushwaha 2020), warehouse operations (Sun et al. 2022), medical treatment adherence (Lin et al. 2021), or task sequencing (Ibanez et al. 2018). Actually, partial adherence also occurs when the recommendation does not come from a machine. In the context of chronic diseases, for instance, the World Health Organization (WHO) defines *adherence* as "the extent to which a person's behavior-taking medication, following a diet, and/or executing lifestyle changes corresponds with agreed recommendations from a health-care provider" (Sabaté 2003). The WHO notes that adherence of the patients to therapy for chronic illnesses is as low as 50 % in the long-term, and that this partial adherence leads to suboptimal clinical outcomes. To anticipate its potential impact on operational performance, it is important to understand the drivers of partial adherence, such as information asymmetry or algorithmic aversion.

In the context of operations, assuming that humans have more and better information than the machine, deviations due to information asymmetry can be beneficial to effective performance. In an inventory management setting, Van Donselaar et al. (2010) conclude that providing store manager discretion may result in higher profits due to their superior information. In a field experiment with an automotive replacement parts retailer, Kesavan and Kushwaha (2020) evaluate that merchants overriding demand forecasts increases (resp. decreases) profitability for growth- (resp. decline-) stage products, suggesting that the information advantage of merchants increases when the machine has limited access to historical data on the product. However, on average, they observe a negative effect of human overriding power. Similarly, Fildes et al. (2009) document the heterogeneous impact of human adjustment on prediction accuracy, depending on the company but also

the magnitude and direction of the adjustment. In another context, Sun et al. (2022) study the box size recommendation algorithm of Alibaba. Since the algorithm ignores the foldability and compressibility of the items, they observe that warehouse workers are able to pack some orders in smaller boxes than the ones recommended.

Partial adherence can also result from multiple conflicting objectives that are weighted differently by the human and the algorithm. In Alibaba's warehouses for instance, Sun et al. (2022) hypothesize that workers switching to larger boxes might do so to save packing effort at the expense of time and cost. In a healthcare setting, Ibanez et al. (2018) observe that doctors tend to re-prioritize tasks so as to group similar tasks together and reduce mental switching costs, but that such prioritization may reduce long-term productivity.

Another reason that could explain why humans fail to follow machine recommendation is algorithm aversion, as first documented by Dietvorst et al. (2015). Algorithm aversion refers to a general preference to rely on humans instead of algorithms. This general preference could be due to an inflated confidence in human performance. In a lab experiment, for instance, Logg et al. (2019) observed that subjects (and in particular experts) were more prone to follow their own judgment over an algorithm's advice, or advice provided by another human. Alternatively, Dietvorst et al. (2018) hypothesize that decision makers seek control over the output. In an empirical study, they successfully reduced algorithm aversion by offering decision makers some control over the machine's output. Lin et al. (2021) propose and empirically evaluate algorithm use determinants in algorithm aversion.

In an effort to propose alternative explanations to algorithm aversion, de Véricourt and Gurkan (2023) develop a theoretical framework to study the evolution of the decision maker's belief about the performance of a machine and her overruling decisions over time. In their setting, decisions and recommendations are binary (to act or not to act, e.g., collect a biopsy or not) and the decision maker only collects performance data when choosing to act. Because of this verification bias, de Véricourt and Gurkan (2023) identify situations under which a (rational) decision maker

fails to learn the true performance of the machine, and indefinitely overrules its recommendation with some non-zero probability.

Understanding the drivers of partial adherence is useful to propose solutions and incorporate behavioral aspects into the algorithmic recommendations. In a pricing setting, for example, Caro and de Tejada Cuenca (2023) observe adherence patterns that are consistent with the fact that inventory and sales are more salient to managers and conduct two interventions aimed at increasing the salience of revenues. A growing literature has studied *features* of the recommendation system or the recommended policy that could increase adoption, such as partial control over the output (see discussion above and Dietvorst et al. 2018), simplicity (Bastani et al. 2021), or interpretability (see, e.g., Kallus 2017, Bravo and Shaposhnik 2020, Ciocan and Mišić 2022, Jacq et al. 2022). The underlying intuition is that policies that have simple structural forms are more likely to be adopted because of legal requirements for a 'right to explanation' (Goodman and Flaxman 2017) and because decision makers and stake-holders value policy they can understand and audit (Bertsimas et al. 2013, 2022). Assuming that humans are more likely to adhere to recommendations that constitute small changes to their current practice, Bastani et al. (2021) propose a reinforcement learning approach to compute optimal 'tips', i.e., small changes in the current practice, and validate their approach in a controlled experiment. In an attempt to increase interpretability of reinforcement learning policies, Jacq et al. (2022) propose the lazy-MDP framework to learn and recommend *when to act* (i.e., in what states of the system), on the top of the decisions. Meresht et al. (2020) propose to learn when to switch control between machines and human decision makers. Nonetheless, these works assume that the simplicity or interpretability of the recommendation will not only increase adherence, but will lead to perfect adherence. In this paper, we complement this literature by challenging this assumption and investigating the impact of partial adherence directly on the actions to be recommended. We develop a framework to incorporate the potential departure of the human decision maker *within the search for a good recommendation policy.* Our goal resembles that of robust optimization under implementation errors where there is a similar

discrepancy between the computed solution and the implemented one as in Bertsimas et al. (2010), Men et al. (2014), except that their error model is purely adversarial and their decision problem static, and that our model accounts for the current baseline practices.

In a similar vein, Sun et al. (2022) reduce non-adherence in Alibaba's warehouses by 19.3% and packing time by 4.5%, by modifying the box size recommendations for the "at-risk" orders (defined as having $> 50\%$ chance of being overruled). In this paper, we have a similar objective of adjusting the recommendation of the algorithm to the expected adherence level. However, instead of an ad-hoc adjustment, we propose to account for the adherence level directly in the optimization problem which the recommendation is a solution of. Furthermore, our objective is not to increase adherence per se but to adjust the algorithm's recommendation to the adherence level, so as to increase the performance of the human-in-the-loop system.

3. Modeling partial adherence in a decision framework

In this section we formally introduce our model of decision under partial adherence.

We consider a human decision maker (DM) which repeatedly interacts with an environment. The goal of the DM is to maximize a cumulative expected return, which captures both the instantaneous reward and the long-run objective. A policy of the DM is a map from the set of possible states of the environment to the set of actions. We assume that we have access to a *baseline* policy, called π_{base}, which models the historical decisions of the DM. In a healthcare setting, for example, the DM is a medical practitioner, observes the health condition of a patient at each time period, and chooses a treatment to maximize the chances of survival, e.g., intravenous fluids and vasopressors for hospital patients with sepsis (Komorowski et al. 2018), proactive transfers to the intensive care units for patients in the emergency room (**?**), or drug treatment decisions for heart disease in patients with type 2 diabetes (Steimle and Denton 2017). The baseline policy π_{base} captures the current standard of care.

Classical methods from the operations management literature design models and algorithms to compute an alternative *recommendation policy* π_{alg} that leads to improved performance compared

with the baseline. The underlying assumption is that the DM, convinced by the value of the algorithmic approach, will systematically follow π_{alg} and not revert to π_{base}. However, in many practical problems, π_{alg} is only a *recommendation*. The practitioner does not commit to implementing it. She has some discretionary power and the resulting policy is likely to be neither π_{base} nor π_{alg}, but a mixture of the two. The main objective and contribution of our paper is to incorporate this partial adherence phenomenon within the optimization problem that defines π_{alg}, i.e., adjust the recommended policy to the adherence level.

3.1. Preliminaries on Markov decision process

Formally, we adopt the framework of Markov Decision Processes (MDPs; Puterman 2014). The system or environment is described via a set of possible states $\mathcal{S}$. At every decision period, the DM is at a given state $s \in \mathcal{S}$, chooses an action $a \in \mathcal{A}$, transitions to the next state $s' \in \mathcal{S}$ with a probability $P_{sas'} \in [0,1]$ and obtains a reward $r_{sas'} \in \mathbb{R}$. The future rewards are discounted by a factor $\lambda \in (0,1)$ and we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets. An MDP instance $\mathcal{M}$ consists of a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, \boldsymbol{r}, \boldsymbol{p}_0, \lambda)$, with $\boldsymbol{r} = (r_{sas'})_{s,a,s'} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$ and $\boldsymbol{P} = (P_{sas'})_{sas'} \in (\Delta(\mathcal{S}))^{\mathcal{S}\times\mathcal{A}}$, and $\boldsymbol{p}_0 \in \Delta(\mathcal{S})$ is an initial probability distribution over the set of states $\mathcal{S}$. Here, we denote $\Delta(\mathcal{S})$ the simplex over $\mathcal{S}$, defined as

$$\Delta(\mathcal{S}) = \left\{ \boldsymbol{p} \in \mathbb{R}^{\mathcal{S}} \mid p_s \geq 0, \forall\, s \in \mathcal{S}, \sum_{s\in\mathcal{S}} p_s = 1 \right\}.$$

A policy π maps, for each period $t \in \mathbb{N}$, the state-action history $(s_0, a_0, s_1, a_1, ..., s_t)$ to a probability distribution over the set of actions $\mathcal{A}$. A policy π is Markovian if it only depends of the current state s_t, and stationary if it is Markovian and it does not depend on time. Therefore, a stationary policy is simply a map $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We call $\Pi = (\Delta(\mathcal{A}))^{\mathcal{S}}$ the set of stationary policies, Π_{M} the set of Markovian policies, and Π_{H} the set of all policies (possibly history-dependent). In an MDP, the goal of the DM is to compute a policy π to maximize the return $R(\pi)$, defined as

$$R(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t s_{t+1}}\right], \tag{3.1}$$

with s_t the state visited at time period t, a_t the action chosen with probability π_{sa}, and the expectation is as regards with the distribution defined by the policy π on the set of infinite-horizon trajectories. The return $R(\cdot)$ is sometimes called *expected reward*, and we use the term *return* to distinguish it from the instantaneous reward r_{sa}. The *value function* $\boldsymbol{v}^\pi \in \mathbb{R}^{\mathcal{S}}$ of a policy $\pi \in \Pi_{\mathsf{H}}$ represents the return obtained starting from any state: $v_s^\pi = \mathbb{E}^\pi \left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t s_{t+1}} \mid s_0 = s \right], \forall\, s \in \mathcal{S}$. Note that in all generality, the return function $\pi \mapsto R(\pi)$ is neither convex nor concave on Π. An optimal policy can be chosen stationary and deterministic and can be computed efficiently (see Puterman 2014, chapter 6). We will say that a policy π' is an ϵ-*optimal policy* if its return is within $\epsilon > 0$ of the optimal return: $R(\pi') + \epsilon \geq \max\{R(\pi) \mid \pi \in \Pi\}$.

Remark 3.1 (Finite-horizon setting). In this paper, we only consider MDPs with infinite horizon. It is straightforward to extend our framework and results to the case of finite-horizon MDPs by adding an absorbing state with instantaneous reward 0 after the last period.

3.2. Adherence-aware MDP

We now incorporate the phenomenon of partial adherence into an MDP framework. Let $\mathcal{M}$ be an MDP instance, π_{base} a baseline policy, and π_{alg} a recommendation policy. We assume that π_{base} belongs to the set Π of stationary policies. To capture the fact that the DM does not systematically implement π_{alg}, let us introduce a parameter $\theta \in [0,1]$, which we call the *adherence level*. Intuitively, the adherence-level θ quantifies the compliance of the decision maker to follow the recommendation policy π_{alg} instead of the baseline policy π_{base}. Therefore, the policy effectively implemented by the DM depends on π_{alg}, π_{base}, and θ. In particular, we consider an effective policy of the form:

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta) = \theta \pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}. \tag{3.2}$$

According to this model, when $\theta = 0$, the DM always follows the baseline policy π_{base}, and when $\theta = 1$, the DM always follows the recommendation policy π_{alg}. When $\theta \in (0,1)$, the DM follows an effective policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)$, which is a mixture of π_{alg} and π_{base}. Consequently, the effective return for the DM is $R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$, with $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta) = \theta \pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}$. For a fixed adherence level

θ, our objective is to compute an optimal recommendation policy such that the effective return $\pi_{\mathsf{alg}} \mapsto R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$ is maximized, i.e., our goal is to solve the following decision problem, called *Adherence-aware MDP* (AdaMDP):

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)). \tag{AdaMDP}$$

When the supremum in the above optimization program is attained, we write $\pi^\star_{\mathsf{alg}}(\theta)$ for an optimal recommendation policy and we write $\pi^\star_{\mathsf{eff}}(\theta)$ for the resulting optimal effective policy, i.e., $\pi^\star_{\mathsf{eff}}(\theta) = \pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta), \theta)$. For simplicity, we assume for now that θ is the same for all states $s \in \mathcal{S}$, an assumption we will challenge in Section 6. We first note that an optimal policy $\pi^\star_{\mathsf{alg}}(\theta)$ for AdaMDP can be chosen stationary and deterministic, two properties that are appealing from an implementation standpoint.

PROPOSITION 3.1. *The supremum in* AdaMDP *is attained at an optimal recommendation policy* $\pi^\star_{\mathsf{alg}}(\theta)$ *that can be chosen stationary and deterministic:*

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)) = \max_{\pi_{\mathsf{alg}} \in \Pi} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)).$$

The proof of Proposition 3.1 uses some more advanced results that we will introduce in Section 4.2. We present the detailed proof in Appendix F.

REMARK 3.2. Interestingly, a similar type of mixture policies have been studied in the online learning literature, yet with a different motivation. To address the exploration-exploitation trade-off, many policies obtained via reinforcement learning are implemented together with an ad-hoc exploration mechanism. Instead, Shani et al. (2019) propose to compute "exploration-conscious" policies that are designed for a particular exploration policy (e.g., choosing actions uniformly at random) and exploration rate, which play a similar role as π_{base} and $1 - \theta$ in our framework. However, they view the exploration policy and exploration rate as additional parameters one can tune to mitigate the exploration-exploitation tradeoff, while we consider π_{base} and θ as uncontrolled inputs (arising from potential human deviations) and study their impact on actual performance.

3.3. Discussion: Mechanisms for partial adherence and effective policy

Our adherence-aware MDP framework posits that the effective policy can be simply expressed as a convex combination of the algorithmic and the baseline policies, as presented in (3.2). In this section, we further justify the practical relevance of our framework by discussing how different models for the DM's adherence decision connects with our framework.

To model the DM's decision to adhere, we introduce a variable $u_{s,t} \in [0,1]$ indicating, in state s, at time t, whether she follows the recommended policy π_{alg} (the case $u_{s,t} = 1$) or whether she follows π_{base} (the case $u_{s,t} = 0$). We call $u_{s,t}$ the *adherence decision* at state s and period t, and we write $u := (u_{s,t})_{s\in\mathcal{S},t\in\mathbb{N}}$. With this notation, the effective policy at state s at time t is given as

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)_{s,t} = u_{s,t}\pi_{\mathsf{alg}_{s,t}} + (1-u_{s,t})\pi_{\mathsf{base}_{s,t}}, \tag{3.3}$$

and specifying an adherence mechanism is equivalent to specifying how the DM chooses u.

Random model. For example, the DM could sample $u_{s,t}$ following *any distribution* with support included in $[0,1]$ and with a mean θ. For instance, in the case of a Bernoulli distribution with parameter θ, at each time period, the decision maker follows π_{alg} with probability θ and π_{base} with probability $1-\theta$. In practice, this random model of adherence decisions can be interpreted as being agnostic to the reasons for partial adherence. Whatever the cause (e.g., algorithm aversion, information asymmetry), they are inaccessible to the algorithm, hence are perceived by the algorithm as random deviations from the recommended policy. In other words, this model mimics the observed behavior of DM but does not capture from first principles why she sometimes decides to deviate from the recommendations. For example, in a stylized setting with a rational DM trying to learn whether a machine is more accurate than her, de Véricourt and Gurkan (2023) identify regimes where the DM's belief oscillates permanently, hence justifying models like this one, where the DM's adherence decisions $u_{s,t}$ and $u_{s,t'}$ may be different for $t \neq t'$, even though the state is the same. In the next theorem, we show that this model with random adherence decision u is exactly equivalent to AdaMDP.

THEOREM 3.1. *Consider the following model of* **random** *adherence decisions, where each* $(u_{s,t})_{s\in\mathcal{S}}$ *is sampled from a distribution with mean* $(\theta,...,\theta)\in[0,1]^{\mathcal{S}}$*, independently across* $t\in\mathbb{N}$*. Then*

$$\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}}\mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},u))\right]=\max_{\pi_{\mathsf{alg}}\in\Pi}R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\theta))$$

and an optimal recommendation may be chosen stationary and deterministic in the left-hand side of the above equation.

We present a detailed proof in Appendix A. Note that under the assumption of Theorem 3.1, the random variables $u_{s,t}$ and $u_{s',t}$ may be dependent for $s\neq s'$. In fact, the proof relies on showing that $\mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},u))\right]=R\left(\mathbb{E}_u\left[\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},u)\right]\right)$, despite the return $R(\cdot)$ being non-linear. This follows from the properties that $u_{s,t}$ and $u_{s',t'}$ are independent across pairs $(s,t),(s',t')$ such that $t\neq t'$. Noting that $\mathbb{E}_u\left[\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},u)\right]=\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\theta)$ concludes the proof.

Adversarial model. Alternatively, as discussed in the literature review in Section 2, partial adherence can be driven by information asymmetry or conflicting objectives between the algorithm and the DM. In other words, the decision maker could choose to follow the recommendation policy π_{alg} or the baseline policy π_{base} according to a different MDP instance $\mathcal{M}'$ than the MDP instance $\mathcal{M}$ that parametrized the algorithm. Adopting a conservative view, one can assume the DM picks each $u_{s,t}\in[\theta,1]$ *adversarially* in a set $B\subseteq[\theta,1]^{\mathcal{S}\times\mathbb{N}}$:

$$\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}}\min_{u\in B}R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},u)\right). \tag{3.4}$$

Without any restrictions, i.e., in the case $B=[\theta,1]^{\mathcal{S}\times\mathbb{N}}$, the DM could decide to follow the algorithm in state s at time t and, when visiting the same state s at a later stage, decide to override it. Hence, we can enrich the set B with several consistency constraints to model more realistic situations. In some settings, for instance, it might be more realistic to assume a *time-invariant adversarial* model, i.e., to assume that the DM's adherence behavior depends on the state but is consistent over time. For example, one could assume that she chooses an adherence decision $u_s\in[\theta,1]$ adversarially for each state s and adopts this policy throughout, i.e., $u_{s,t}=u_s,\forall\, t\in\mathbb{N}$. Note that the time-invariant

adversarial model assumes that the decision maker has some discretionary power at the beginning but commits to one policy for the rest of the trajectory, which can be seen as contradictory. Another realistic model consists of *state-invariant* adherence decisions, i.e., $u_{s,t} = u_t \in [0,1]$ across all pairs $(s,t) \in \mathcal{S} \times \mathbb{N}$. A fourth model could assume that the adherence decisions are *time- and state-invariant*, i.e., that $u_{s,t} = u \in [0,1]$ across all pairs $(s,t) \in \mathcal{S} \times \mathbb{N}$. Fortunately, as stated (informally) in Theorem 3.2, studying our effective policy (3.2) is equivalent to studying any of these three adherence mechanisms:

THEOREM 3.2. *(Informal statement) An optimal algorithmic recommendation $\pi^\star_{\mathsf{alg}}(\theta)$, solution to* AdaMDP*, is an optimal solution of the decision problem* (3.4)*, whenever the adherence decision u is chosen according to one of the following adversarial models: for all $(s,t) \in \mathcal{S} \times \mathbb{N}$,*

- **(Unconstrained Adversarial)** *$u_{s,t}$ chosen independently and adversarially in $[\theta, 1]$.*
- **(Time-invariant Adversarial)** *$u_{s,t} = u_s$ with u_s chosen independently and adversarially in $[\theta, 1]$.*
- **(State-invariant Adversarial)** *$u_{s,t} = u_t$ with u_t chosen independently and adversarially in $[\theta, 1]$.*
- **(Time- and State-invariant Adversarial)** *$u_{s,t} = u$ with u chosen adversarially in $[\theta, 1]$.*

Additionally, strong duality holds for these models of adversarial adherence decisions.

We defer a formal statement and proof of Theorem 3.2 to Appendix B. Theorem 3.2 shows that AdaMDP can be interpreted as the robust counterpart of the aforementioned adversarial models, and perhaps surprisingly, that these robust models yield the same worst-case return, and from the proof of Theorem 3.2, the same optimal policy as well. The strong duality results show that the case where π_{alg} is chosen *before* the adherence decisions $\boldsymbol{u}$ and the case where π_{alg} is chosen *after* the adherence decisions $\boldsymbol{u}$ are equivalent. We should emphasize, however, that Theorem 3.2 only claims an equivalence in terms of *optimal* effective return. For a given (sub-optimal) policy, its effective return under each model (AdaMDP or one of the adversarial models) can differ.

REMARK 3.3. The proof of Theorem 3.2 shows that for these adversarial models, a worst-case $u_{s,t}$ can be chosen as $u_{s,t} = \theta, \forall (s,t) \in \mathcal{S} \times \mathbb{N}$. Therefore, when $\theta = 0$, we recover the fact that the agent never follows the algorithmic recommendation π_{alg}.

Overall, Theorems 3.1 and 3.2 show that our simple proposal for adherence-aware MDPs subsumes a collection of DM-level models of partial adherence, hence justifying our subsequent analysis of the effective policy (3.2) and the optimal recommendation problem (AdaMDP).

We summarize the equivalences obtained in this section in Table 1. For the adversarial model, *time-invariance* and *state-invariance* are described in Theorem 3.2. For the random model of adherence decisions, *time-invariance* corresponds to a model where there exist two periods $t \neq t'$ for which the random variables $u_{s,t}$ and $u_{s',t'}$ are dependent for some states $s, s' \in \mathcal{S}$, and *state-invariance* corresponds to the case where there exist $s \neq s'$ and $t \in \mathbb{N}$ for which $u_{s,t}$ and $u_{s',t}$ are dependent random variables. The assumption in Theorem 3.1 corresponds to random models that are *not* time-invariant. We provide more discussion on these time-invariant and state-invariant random models at the end of Appendix A.

Constraints		Model of adherence decisions	
Time-invariance	State-invariance	Random	Adversarial
×	×	AdaMDP	AdaMDP
×	✓	AdaMDP	AdaMDP
✓	×	unknown	AdaMDP
✓	✓	unknown	AdaMDP

Table 1 Summary of the adherence decision models considered in this paper and their relations with AdaMDP.

Cardinality-constrained model. Under an adversarial lens, one could model the DM's unwillingness to implement a large number of changes to her current practice by, e.g., imposing a limit on the number of states where she adheres. For example, let us assume that adherence decisions

are time-invariant and let us model the DM's adherence problem as that of finding up to k states where she follows the algorithmic recommendation, with $k \in \mathbb{N}$:

$$\min_{u \in \{0,1\}^{\mathcal{S}}, \sum_{s \in \mathcal{S}} u_s \leq k} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)). \qquad \text{(Constrained-AdaMDP)}$$

The evaluation problem above (let alone the problem of then optimizing for π_{alg}) is hard, as we characterize in the following result:

THEOREM 3.3. Constrained-AdaMDP *is APX-hard, i.e., there exists a constant $\alpha > 0$, for which it is NP-hard to approximate* Constrained-AdaMDP *within a factor smaller than $1 + \alpha$.*

Our proof of Theorem 3.3 is based on a reduction from the constrained assortment optimization under the Markov Chain-based choice model (Désir et al. 2020) and we provide the details in Appendix C. This shows that adding a simple cardinality constraint to AdaMDP makes the decision problem intractable. For the sake of completeness, and since Constrained-AdaMDP may be of independent interest, we provide a mixed-integer optimization formulation for solving Constrained-AdaMDP in Appendix D.

3.4. Examples of competition/complementarity between the human and the algorithm

Before turning to a more formal analysis of our framework, we demonstrate the implications of the effective policy (3.2) on a simple MDP instance, to provide some intuition on the interactions at play between π_{alg} and π_{base} as well as illustrate the rich range of situations that can arise in our framework. Indeed, we provide an example where the co-existence of the algorithmic and baseline policies can lead to arbitrarily bad performance and another example where, on the contrary, they complement each other.

We consider the MDP instance from Figure 1. There are 5 states, the rewards are independent from the chosen action and only depend on the current state. We assume that the transitions are deterministic and are represented with dashed arcs in Figure 1a, along with the rewards above the states. The actions consist in choosing the possible next states. The MDP starts in State 1,

and State 4 and State 5 are absorbing. The MDP instance is parametrized by $\epsilon \in \{-1, 1\}$, which impacts the reward of State 5.

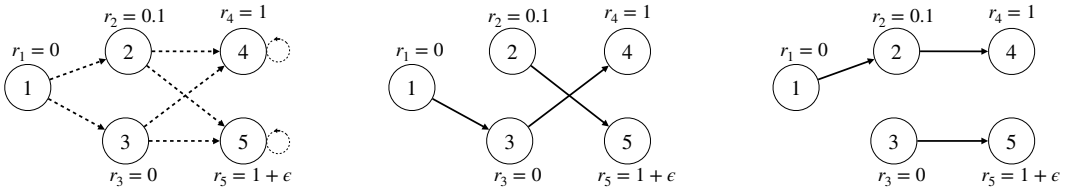
The current policy π_{base} is represented in Figure 1b. Observe that π_{base} prescribes to transition from State 2 to State 5 but that, according to π_{base}, State 2 should not be visited in the first place. For example, in a healthcare setting, State 2 could correspond to a newly introduced treatment, which the practitioner is not used to prescribing. The expected return of π_{base} is

$$R(\pi_{\mathsf{base}}) = \frac{\lambda^2}{1-\lambda},$$

where $\lambda \in (0, 1)$ is the discount factor. Note that, by definition of the effective policy π_{eff}, for any $\theta \in [0, 1]$, $\pi_{\mathsf{base}} = \pi_{\mathsf{eff}}(\pi_{\mathsf{base}}, \theta)$. In other words, for any adherence level $\theta \in [0, 1]$, recommending π_{base} leads exactly to the implementation of π_{base}. We further consider that the algorithm prescribes the policy π_{alg} represented in Figure 1c, whose expected return is

$$R(\pi_{\mathsf{alg}}) = 0.1\lambda + \frac{\lambda^2}{1-\lambda} > R(\pi_{\mathsf{base}}).$$

Detailed computations of policy returns reported in this section are presented in Appendix E.



(a) MDP instance. (b) Baseline policy π_{base}. (c) Representation of π_{alg}.

Figure 1 Details on the transitions and rewards of our MDP instance.

Case 1: partial adherence hurts. We first assume that $\epsilon = -1$. In this case, it is easy to verify that π_{alg} is optimal under perfect adherence ($\theta = 1$). If adherence is not perfect, however, continuing to recommend π_{alg} can lead to sub-optimal performance. Indeed, π_{base} chooses suboptimal actions in State 2, which π_{alg} recommends to visit (unlike π_{base}). So, the mixture policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)$ can lead to

worse performance than either π_{alg} or π_{base}. Formally, the return of the effective policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)$ is equal to

$$R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) = R(\pi_{\mathsf{base}}) + 2\theta \frac{\lambda^2}{1-\lambda}\left(\theta - \tilde{\theta}\right)$$

with $\tilde{\theta} := 1 - 0.1\dfrac{1-\lambda}{2\lambda} \leq 1$. If $\tilde{\theta} \leq 0$, the behavior of the effective return function is intuitive: In this case, we observe that $\theta \mapsto R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right)$ is increasing. In particular, $R\left(\pi_{\mathsf{alg}}\right) = R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, 1)\right) \geq R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right)$, i.e., partial adherence degrades the effective return obtained by recommending π_{alg} compared with the perfect adherence case. Furthermore, $R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) \geq R\left(\pi_{\mathsf{base}}\right)$, i.e., recommending π_{alg} improves over the current standard of practice, π_{base}.

However, the analytic expression above reveals surprising behaviors when $\tilde{\theta} > 0$. In this case, the function $\theta \mapsto R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right)$ is non-monotone (see Figure 2a, obtained with $\lambda = 0.5$, hence $\tilde{\theta} = 0.95$): It decreases on $[0, \tilde{\theta}/2]$ and increases on $[\tilde{\theta}/2, 1]$. Since the effective policy is a convex combination of π_{alg} and π_{base}, it is intuitive to believe that its performance will be bounded above and below by $R(\pi_{\mathsf{alg}})$ and $R(\pi_{\mathsf{base}})$ respectively. This example disproves this intuition. In particular, we have $R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \tilde{\theta})\right) < R\left(\pi_{\mathsf{base}}\right)$. In other words, overlooking the adherence level θ and recommending the same policy π_{alg} may lead to lower return than the baseline policy itself! Actually, as we formally prove in the next section, this sub-optimality gap can be made arbitrarily large.

Finally, via backward induction, we can find an optimal recommendation policy $\pi^{\star}_{\mathsf{alg}}(\theta)$ for any value of $\theta \in [0, 1]$. In particular, we find an optimal recommendation policy of the following form (see derivations in Appendix E): $\pi^{\star}_{\mathsf{alg}}(\theta) = \pi^{\star}$ if $\theta > \max(0, \bar{\theta})$ for $\pi^{\star}$ that chooses $1 \to 2, 2 \to 4, 3 \to 4$ and $\bar{\theta} = 1 - 0.1(1-\lambda)/\lambda$; and $\pi^{\star}_{\mathsf{alg}}(\theta) = \pi_{\mathsf{base}}$ if $\theta \leq \max(0, \bar{\theta})$. Note that by varying λ, the breakpoint $\max(0, \bar{\theta})$ can be made arbitrarily close to 1. In the following section, we show that, for any MDP instance, the optimal recommendation policy $\pi^{\star}_{\mathsf{alg}}(\theta)$ enjoys such piecewise constant structure.

Case 2: partial adherence helps (complementarity). We now consider the case where $\epsilon = 1$ so that neither π_{alg} nor π_{base} are optimal and there is room for improvement. Actually, we show in this example that partial adherence improves upon *both* policies, illustrating *complementarity benefits*

between the human DM and the algorithm. We now compute the expected return of the effective policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta) = \theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}$. In particular, we obtain that

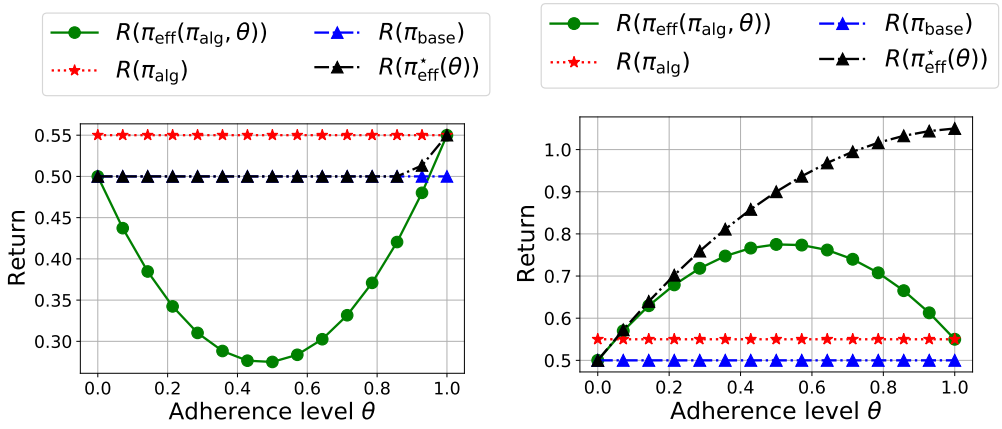
$$R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) = R(\pi_{\mathsf{alg}}) + 2R(\pi_{\mathsf{base}})(1-\theta)(\theta - (1-\tilde{\theta})),$$

with $\tilde{\theta}$ previously defined. Thus, if $1-\tilde{\theta} < 1$, we observe that $R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) > \max\left\{R(\pi_{\mathsf{alg}}), R(\pi_{\mathsf{base}})\right\}$ for any $\theta \in (1-\tilde{\theta}, 1)$. In other words, there exists a regime where the partial implementation of π_{alg} leads to greater performance than π_{alg} or π_{base} alone.

These examples show that, despite its simple form, the class of effective policies defined in (3.2) can capture many realistic situations where the co-existence of the algorithm and the DM hurts or benefits the overall system performance. Because our objective is prescriptive and we are interested in informing the design of the algorithmic recommendations π_{alg}, we assume in the rest of the paper that recommendations are optimal for the true MDP parameter $\boldsymbol{r}, \boldsymbol{P}, \lambda$ and the adherence level θ, i.e., where $\pi_{\mathsf{alg}} = \pi^{\star}_{\mathsf{alg}}(\theta)$ with $\pi^{\star}_{\mathsf{alg}}(\theta)$ an optimal solution to the optimization problem (AdaMDP). This corresponds to the case where there is no model misspecification, and where θ is known. In particular, under this assumption, algorithmic recommendations that ignore the issue of partial adherence correspond to $\pi_{\mathsf{alg}} = \pi^{\star}_{\mathsf{alg}}(1)$, and Case 1 in this section shows that $R(\pi^{\star}_{\mathsf{eff}}(\theta))$ may be much greater than $R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$. Given an estimate of the adherence level θ, our objective is thus to compute an optimal recommendation $\pi^{\star}_{\mathsf{alg}}(\theta)$ as a solution of an optimization problem, enabling us to prove important structural properties and tractability results in the next sections. We should emphasize that diverting from the assumption that the algorithmic recommendation is the solution of an optimization model leaves open the question of how to define (and compute) the algorithmic recommendation in practice.

Remark 3.4. In our MDP instance for the second case (complementarity), neither π_{alg} nor π_{base} are optimal. Indeed, by definition, if π_{alg} or π_{base} is an optimal policy for the nominal MDP, then it is impossible that $R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) > \max\left\{R(\pi_{\mathsf{alg}}), R(\pi_{\mathsf{base}})\right\}$, i.e., complementarity cannot occur. More complex models of partial adherence could lead to interesting human-machine complementarity, for instance in the case where both the algorithm and the human only have access to *partial*

information on the state or action sets or have different objectives. Our agnostic model may adequately complement these cases where more is known (or assumed) about the rational behind partial adherence. Because decision models are necessarily a simplification of real-life decisions, integrating more complex behavioural models behind partial adherence is an important direction for future work.



(a) Co-existence hurts ($\epsilon = -1$) (b) Co-existence helps ($\epsilon = 1$)

Figure 2 **Illustrating the impact of the partial adherence phenomenon (hence the coexistence of a baseline and algorithmic policy) in the MDP instance from Figure 1a. We choose $\lambda = 0.5$ in our simulations.**

4. Analyzing adherence-aware MDPs

We now theoretically analyze the class of adherence-aware MDPs we introduced in the previous section. As a motivation, we first provide negative results showing the worst-case performance deterioration that can be experienced by overlooking the partial adherence phenomenon, i.e., by recommending $\pi^{\star}_{\text{alg}}(1)$ instead of $\pi^{\star}_{\text{alg}}(\theta)$. We then show how to compute optimal adherence-aware recommendations efficiently and investigate how they depend structurally on θ.

4.1. Worst-case analysis of the performance of $\pi^{\star}_{\text{alg}}(1)$

As the example in Section 3.4 shows, an optimal recommendation policy $\pi^{\star}_{\text{alg}}(\theta)$ may be different from an optimal nominal policy $\pi^{\star}_{\text{alg}}(1)$, which itself can lead to worse performance than the baseline policy π_{base} alone. We now formalize these observations.

First, we analyze the performance of $\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)$ for $\pi^\star_{\mathsf{alg}}(1)$ an optimal nominal policy and show that recommending $\pi^\star_{\mathsf{alg}}(1)$ (i.e., ignoring the partial adherence effect) can lead to arbitrarily worse returns than the baseline policy.

PROPOSITION 4.1. *For any scalar $M \geq 0$, for any adherence level $\theta \in (0,1)$, there exists an MDP instance $\mathcal{M}$ such that $R(\pi_{\mathsf{base}}) \geq M + R\left(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)\right)$, where $\pi^\star_{\mathsf{alg}}(1)$ is an optimal policy for the nominal MDP instance $\mathcal{M}$.*

Proof of Proposition 4.1 Fix $M \geq 0$ and $\theta \in (0,1)$ and consider the MDP instance of Section 3.4 with $\epsilon = -1$, with $\pi^\star_{\mathsf{alg}}(1)$ as in Figure 1c. In the limit where $\lambda \to 1$, we have $R\left(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)\right) - R(\pi_{\mathsf{base}}) \sim 2\theta\frac{\lambda^2}{1-\lambda}(\theta - \tilde{\theta}) \to -\infty$ since $\theta < 1$. Hence, we can have $R\left(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)\right) - R(\pi_{\mathsf{base}}) \leq -M$ for λ close to 1. □

Proposition 4.1 generalizes the observation that $\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)$ can lead to *arbitrarily worse* performance than the current baseline policy itself (e.g., the current state of practice). As elicited in the example from Section 3.4, this phenomenon happens when the baseline policy π_{base} chooses sub-optimal actions in some states. As a result, the effective policy $\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)$ can also end up in these bad states that are overlooked by $\pi^\star_{\mathsf{alg}}(1)$, which assumes that the actions are always chosen from $\pi^\star_{\mathsf{alg}}(1)$. Consequently, for any value of $\theta \in (0,1)$, the policy $\pi^\star_{\mathsf{alg}}(1)$ can be arbitrarily sub-optimal.

COROLLARY 4.1. *For any scalar $M \geq 0$, for any adherence level $\theta \in (0,1)$, there exists an MDP instance $\mathcal{M}$ such that $R(\pi^\star_{\mathsf{eff}}(\theta)) \geq M + R\left(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1),\theta)\right)$.*

Proof of Corollary 4.1 The result follows from Proposition 4.1 since $R(\pi_{\mathsf{base}}) = R(\pi_{\mathsf{eff}}(\pi_{\mathsf{base}},\theta)) \leq R(\pi^\star_{\mathsf{eff}}(\theta))$. □

While Proposition 4.1 and Corollary 4.1 show that ignoring the adherence level θ can lead to arbitrarily large losses in performance, there are worst-case statements where, for each value of $\theta \in [0,1)$, a particular MDP instance $\mathcal{M}$ is constructed. In practice, one might be interested in a single MDP instance and the impact of varying $\theta \in [0,1]$ on this instance in particular, which is the focus of the rest of this section.

4.2. Solving adherence-aware MDPs

We now show how to efficiently compute an optimal policy $\pi^{\star}_{\mathsf{eff}}(\theta)$ for adherence-aware MDPs. Note that when $\theta = 1$, the DM is simply solving a classical MDP problem, which can be done efficiently with various algorithms such as value iteration, policy iteration, and linear programming (see chapter 6 in Puterman 2014). Additionally, for the classical MDP problem, it is well-known that an optimal policy can be chosen stationary and deterministic without loss of optimality, which greatly simplifies implementation and interpretation of such policies in practice. We show that the same holds for the adherence-aware MDP problem in the next proposition.

PROPOSITION 4.2. *There exists a unique vector $\boldsymbol{v}^{\infty} \in \mathbb{R}^{\mathcal{S}}$ defined as*

$$v_s^{\infty} = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \theta \cdot \sum_{a \in \mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^{\top} \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}^{\infty}\right) + (1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},sa} \boldsymbol{P}_{sa}^{\top} \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}^{\infty}\right), \forall\, s \in \mathcal{S}, \tag{4.1}$$

and an optimal recommendation policy $\pi^{\star}_{\mathsf{alg}}(\theta)$ can be computed as a stationary deterministic policy attaining the $\arg\max$ *of Equation* (4.1) *for each $s \in \mathcal{S}$.*

The proof of Proposition 4.2 is akin to our proof of Proposition 3.1, presented in Appendix F, and we omit it for conciseness. We note that we can rewrite Equation (4.1) as

$$v_s^{\infty} = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \pi_{sa} \left(r'_{sa} + \lambda \boldsymbol{P}'^{\top}_{sa} \boldsymbol{v}^{\infty}\right), \forall\, s \in \mathcal{S}, \tag{4.2}$$

with $\boldsymbol{P}' \in \left(\Delta(\mathcal{S})\right)^{\mathcal{S}\times\mathcal{A}}, \boldsymbol{r}' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ defined as

$$\begin{aligned} \boldsymbol{P}'_{sa} &:= \theta \cdot \boldsymbol{P}_{sa} + (1-\theta) \cdot \sum_{a' \in \mathcal{A}} \pi_{\mathsf{base},sa'} \boldsymbol{P}_{sa'}, \\ r'_{sa} &:= \theta \cdot \boldsymbol{P}_{sa}^{\top} \boldsymbol{r}_{sa} + (1-\theta) \cdot \sum_{a' \in \mathcal{A}} \pi_{\mathsf{base},sa'} \boldsymbol{P}_{sa'}^{\top} \boldsymbol{r}_{sa'}, \end{aligned} \tag{4.3}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This shows that for any $\theta \in [0,1]$, an optimal recommendation $\pi^{\star}_{\mathsf{alg}}(\theta)$ can be viewed as the optimal policy for another MDP instance $\mathcal{M}' = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}', \boldsymbol{r}', \boldsymbol{p}_0, \lambda)$, where the new transition probabilities $\boldsymbol{P}'$ and the new rewards $\boldsymbol{r}'$ are defined as (4.3), and, interestingly, where the instantaneous rewards only depend on the current state-action pair (s,a) but not on the subsequent state s'. In the context of "exploration-conscious" reinforcement learning and in the simpler case

where $r_{sas'} = r_{sa}, \forall\,(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ in the MDP instance $\mathcal{M}$, Shani et al. (2019) refer to the MDP instance $\mathcal{M}'$ as the *surrogate MDP*. This shows that we can efficiently compute an optimal recommendation policy by computing an optimal policy of the surrogate MDP. Note that even though $\pi^\star_{\mathsf{alg}}(\theta)$ can be chosen deterministic since it is an optimal policy to the surrogate MDPs, the effective policy $\pi^\star_{\mathsf{eff}}(\theta)$ may be randomized, since by definition $\pi^\star_{\mathsf{eff}}(\theta) = \theta\pi^\star_{\mathsf{alg}}(\theta) + (1-\theta)\pi_{\mathsf{base}}$.

For the sake of completeness, we now describe two efficient methods to compute $\boldsymbol{v}^\infty$.

Iterative method: value iteration. Let us define the operator $f : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ as

$$f_s(\boldsymbol{v}) = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \theta \cdot \sum_{a\in\mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}) + (1-\theta) \sum_{a\in\mathcal{A}} \pi_{\mathsf{base},sa} \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}), \forall\, s \in \mathcal{S}. \tag{4.4}$$

Note that when $\theta = 1$, this is the classical Bellman operator. The operator f is a contraction for ℓ_∞: for any $\boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^{\mathcal{S}}$, we have $\|f(\boldsymbol{v}) - f(\boldsymbol{w})\|_\infty \leq \lambda \|\boldsymbol{v} - \boldsymbol{w}\|_\infty$. Therefore, as for classical MDPs, the fixed-point $\boldsymbol{v}^\infty$ can be computed efficiently via value iteration (VI): $\boldsymbol{v}^0 = \boldsymbol{0}, \boldsymbol{v}^{t+1} = f(\boldsymbol{v}^t), \forall\, t \in \mathbb{N}$. To obtain an ϵ-optimal recommendation policy, we can stop as soon as $\|\boldsymbol{v}^t - f(\boldsymbol{v}^t)\|_\infty \leq \epsilon(1-\lambda)(2\lambda)^{-1}$, which is satisfied after $O(\log(\epsilon^{-1}))$ iterations (Puterman 2014, theorem 6.3.3).

Linear programming formulation. The optimal value function $\boldsymbol{v}^\infty \in \mathbb{R}^{\mathcal{S}}$ can also be computed with linear programming (Puterman 2014, section 6.9). In particular, $\boldsymbol{v}^\infty$ is the unique solution to the optimization problem $\min\left\{\sum_{s\in\mathcal{S}} v_s \mid v_s \geq f_s(\boldsymbol{v}), \forall\, s \in \mathcal{S}\right\}$, which can reformulated in the following linear program with $|\mathcal{S}|$ decision variables and $|\mathcal{S}| \times |\mathcal{A}|$ linear constraints:

$$\min\{\boldsymbol{p}_0^\top \boldsymbol{v} \mid v_s \geq \theta \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}) + (1-\theta) \sum_{a'\in\mathcal{A}} \pi_{\mathsf{base},sa'} \boldsymbol{P}_{sa'}^\top (\boldsymbol{r}_{sa'} + \lambda \boldsymbol{v}), \forall\,(s,a) \in \mathcal{S}\times\mathcal{A}\}.$$

4.3. Structure and sensitivity of $\pi^\star_{\mathsf{alg}}(\theta)$ with respect to the adherence level

We now investigate how the optimal recommendation $\pi^\star_{\mathsf{alg}}(\theta)$ and its performance $R(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta), \theta))$ depend on the adherence level θ.

First, the example from Section 3.4 illustrates that the mapping $\theta \mapsto R(\pi_{\mathsf{eff}}(\pi, \theta))$, for a fixed policy π, is not necessarily monotone. Still, we can recover monotonicity when considering $\pi^\star_{\mathsf{alg}}(\theta)$ instead, as shown in the next proposition.

PROPOSITION 4.3. *For any MDP instance $\mathcal{M}$, the map $\theta \mapsto R(\pi^\star_{\mathsf{eff}}(\theta))$ is non-decreasing on $[0,1]$.*

Proof. This is straightforward from the equivalence of AdaMDP and the models of adversarial adherence decisions from Theorem 3.2. We provide a simple, more direct proof below. Let $\theta_1, \theta_2 \in [0,1]$ with $\theta_1 \leq \theta_2$. We will show that $R(\pi^\star_{\mathsf{eff}}(\theta_1)) \leq R(\pi^\star_{\mathsf{eff}}(\theta_2))$. Following the definition of $\pi^\star_{\mathsf{eff}}(\theta_1)$, we have $\pi^\star_{\mathsf{eff}}(\theta_1) = \theta_1 \pi^\star_{\mathsf{alg}}(\theta_1) + (1-\theta_1)\pi_{\mathsf{base}}$. We can rewrite this as

$$\pi^\star_{\mathsf{eff}}(\theta_1) = \theta_2 \left(\frac{\theta_1}{\theta_2} \pi^\star_{\mathsf{alg}}(\theta_1) + \frac{\theta_2 - \theta_1}{\theta_2} \pi_{\mathsf{base}} \right) + (1-\theta_2)\pi_{\mathsf{base}},$$

and $\hat{\pi} := \frac{\theta_1}{\theta_2} \pi^\star_{\mathsf{alg}}(\theta_1) + \frac{\theta_2 - \theta_1}{\theta_2} \pi_{\mathsf{base}}$ is a policy since $0 \leq \theta_1 \leq \theta_2 \leq 1$. Overall, we conclude that $R(\pi^\star_{\mathsf{eff}}(\theta_1)) = R(\pi_{\mathsf{eff}}(\hat{\pi}, \theta_2)) \leq R(\pi^\star_{\mathsf{eff}}(\theta_2))$, by optimality of $\pi^\star_{\mathsf{alg}}(\theta_2)$. □

Proposition 4.3 shows that as the DM deviates more and more from the recommendation policy (i.e., as θ decreases), the optimal effective return decreases. Note that this result holds because we consider $\pi^\star_{\mathsf{alg}}(\theta)$, in other words because we adjust our recommended policy as the adherence level varies. Since $\pi^\star_{\mathsf{eff}}(0) = \pi_{\mathsf{base}}$, Proposition 4.3 also implies that $R(\pi^\star_{\mathsf{eff}}(\theta)) \geq R(\pi_{\mathsf{base}})$: recommending $\pi^\star_{\mathsf{eff}}(\theta)$ can only improve performance compared with the current baseline, which may not be the case when recommending $\pi^\star_{\mathsf{alg}}(1)$, as highlighted in Proposition 4.1. Overall, Proposition 4.3 also suggests that it is always beneficial to try to increase the compliance of the decision maker (i.e., increase the value of θ), as this leads to more returns for the optimal effective policy $\pi^\star_{\mathsf{eff}}(\theta)$.

Actually, we now show that the optimal recommendation $\pi^\star_{\mathsf{alg}}(\theta)$ does not vary continuously in θ but rather enjoys a piecewise constant structure:

PROPOSITION 4.4. *For any MDP instance $\mathcal{M}$:*

1. *There exists $\bar{\theta} \in [0,1)$, such that $\pi^\star_{\mathsf{alg}}(\theta) = \pi^\star_{\mathsf{alg}}(1)$ for any $\theta \in [\bar{\theta}, 1]$.*
2. *There exists $n \in \mathbb{N}$ and $0 = \theta_1 < \theta_2 < \cdots < \theta_n = 1$ such that, for any $i \in \{1, ..., n-1\}$, $\pi^\star_{\mathsf{alg}}(\theta)$ can be chosen constant over the interval $[\theta_i, \theta_{i+1}]$.*
3. *If $\pi_{\mathsf{base}} = \pi^\star_{\mathsf{alg}}(\underline{\theta})$ for some $\underline{\theta} \in [0,1]$, then $\pi^\star_{\mathsf{alg}}(\theta) = \pi_{\mathsf{base}}$ for any $\theta \in [0, \underline{\theta}]$.*

Combined with the fact that $\pi^\star_{\mathsf{alg}}(1)$ is an optimal recommendation for $\theta = 1$, Statement 1 shows that, when the adherence level is sufficiently close to 1, we can overlook the issue of partial adherence

and output the same recommendation as when $\theta = 1$, which reduces to the classical MDP model. More generally, Statement 2 in Proposition 4.4 shows that, in general, $\pi^\star_{\mathsf{alg}}(\theta)$ has a piecewise constant structure. The piecewise constant structure of $\pi^\star_{\mathsf{alg}}(\theta)$ combined with the fact that π_{base} is an optimal recommendation for $\theta = 0$ also ensures that π_{base} is an optimal recommendation in a neighborhood of 0. Statement 3 generalizes this observation and states that if the baseline policy is an optimal recommendation policy for an adherence level $\underline{\theta}$, then it is optimal for any lower adherence level. A trivial example is the case where $\underline{\theta} = 1$, i.e., when π_{base} is optimal in the classical MDP model, then we should systematically recommend the baseline. To motivate our study, we implicitly assumed that $R(\pi_{\mathsf{base}}) < R(\pi^\star_{\mathsf{alg}}(1))$, i.e., that the baseline policy could be improved.

Lastly, we uncover two conditions on the MDP instance under which the partial adherence phenomenon can be ignored by the decision-maker. We start with a simple example where the optimal recommendation $\pi^\star_{\mathsf{alg}}(\theta)$ does not depend on θ and π_{base}. We observe that when the transitions $\boldsymbol{P}_{sa} \in \Delta(\mathcal{S})$ do not depend on the action but only on the current state: $\boldsymbol{P}_{sa} = \boldsymbol{P}_s \in \Delta(\mathcal{S})$ and when $r_{sas'} = r_{sa}$ for all $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, then the optimality equation (4.1) becomes

$$v^\infty_s = \theta \cdot \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \left\{ \boldsymbol{\pi}_s^\top \boldsymbol{r}_s \right\} + \theta \cdot \lambda \boldsymbol{P}_s^\top \boldsymbol{v}^\infty + (1-\theta) \cdot \boldsymbol{\pi}_{\mathsf{base},s}^\top \boldsymbol{r}_s + (1-\theta) \cdot \lambda \boldsymbol{P}_s^\top \boldsymbol{v}^\infty, \forall\, s \in \mathcal{S},$$

and we can choose an optimal recommendation policy $\pi^\star_{\mathsf{alg}}(\theta)$ that is *independent* from θ and π_{base}. In other words, partial adherence only impacts the effective return but *it does not change the optimal recommendation*. This special case occurs, for example, when the DM faces a sequence of independent single-stage decision problems (e.g., patients arriving independently to be treated) where each decision provides an immediate reward but does not impact the next decision problem, see de Véricourt and Gurkan (2023) for a detailed study of this case in a learning setting.

We now describe a condition under which the decision-maker may ignore partial adherence at a given state. Inspecting the surrogate MDP defined in Equation (4.3), we note that the new pair of rewards and transitions $(\boldsymbol{r}', \boldsymbol{P}')$ is a convex combination of the nominal parameters $(\boldsymbol{r}, \boldsymbol{P})$ and the rewards and transitions induced by π_{base}. Therefore, if π_{base} chooses an optimal action at a state $\bar{s} \in \mathcal{S}$, we may expect that the algorithmic recommendation coincides with π_{base} at $\bar{s}$. We show that this intuition is true in the next proposition.

PROPOSITION 4.5. *Let $\bar{s} \in \mathcal{S}$ such that $v_{\bar{s}}^{\pi^\star_{\mathsf{alg}}(1)} = v_{\bar{s}}^{\pi_{\mathsf{base}}}$. Then for any $\theta \in [0,1]$, we have $v_{\bar{s}}^{\pi^\star_{\mathsf{eff}}(\theta)} = v_{\bar{s}}^{\pi_{\mathsf{base}}}$ and we can choose $\pi^\star_{\mathsf{alg}}(\theta)_{\bar{s}} = \pi_{\mathsf{base},\bar{s}}$.*

We provide the proof of Proposition 4.5 in Appendix H. Proposition 4.5 shows that if the baseline policy obtains the optimal nominal value at a given state $\bar{s} \in \mathcal{S}$, then the decision-maker can guarantee this same value at $\bar{s}$ for any value of the adherence level $\theta \in [0,1]$ by recommending the same action as the baseline policy. We conclude this section by noting that obtaining a meaningful bound on the suboptimality of a policy π_{alg} against $\pi^\star_{\mathsf{alg}}(\theta)$ for a given value $\theta \in [0,1]$ of the adherence level is an interesting direction for future work. We derive a bound in Appendix I, noting that it may be hard to interpret, due to the piece-wise constant structure of the optimal recommendation policies (Proposition 4.4).

5. Numerical experiments

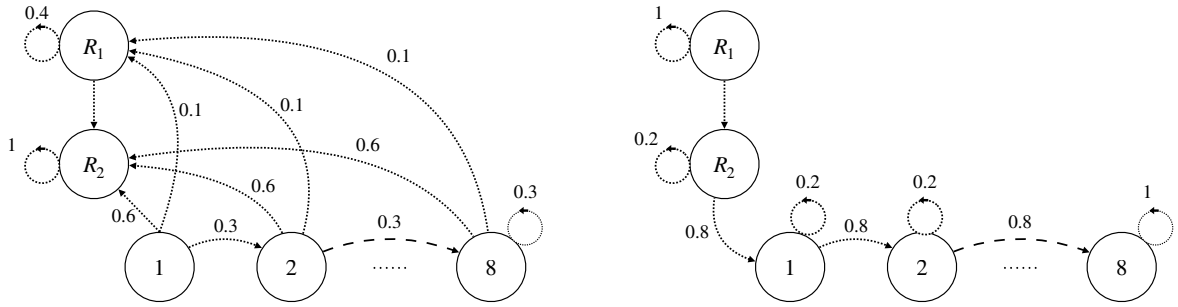
In this section, we numerically study the impact of the adherence level and of the baseline policy on two decision-making examples, in machine replacement and healthcare respectively, that have been studied in the MDP literature. We solve all the decision problems using the value iteration algorithm presented in Section 4.2. Among others, these numerical results illustrate the importance of taking into account the current state of practice and the adherence level when designing algorithmic recommendations. In particular, the adherence-aware optimization framework we develop in this paper provides simple tools to evaluate the robustness of a policy with respect to the adherence level and to obtain improved solutions in situations where the performance is the most impacted.

5.1. Machine replacement problem

We start with the a machine replacement problem introduced in Delage and Mannor (2010) and studied in Wiesemann et al. (2013), **?**.

MDP instance. We represent the machine replacement MDP in Figure 3. The set of states is $\{1,2,3,4,5,6,7,8,R_1,R_2\}$ and the set of actions is $\{\mathsf{repair}, \mathsf{wait}\}$. Each state models the condition of the same machine. In State 8 the machine is broken, while State R_1 and State R_2 model some ongoing reparations. State R_1 is a normal repair while State R_2 is a long repair. We use the same

rewards and transitions as in Delage and Mannor (2010). In particular, there is a reward of 0 in State 8, a reward of 18 in State R_1, a reward of 10 in State R_2, and a reward of 20 in the remaining states. We set a discount factor of $\lambda = 0.99$ and the DM starts in State 1.



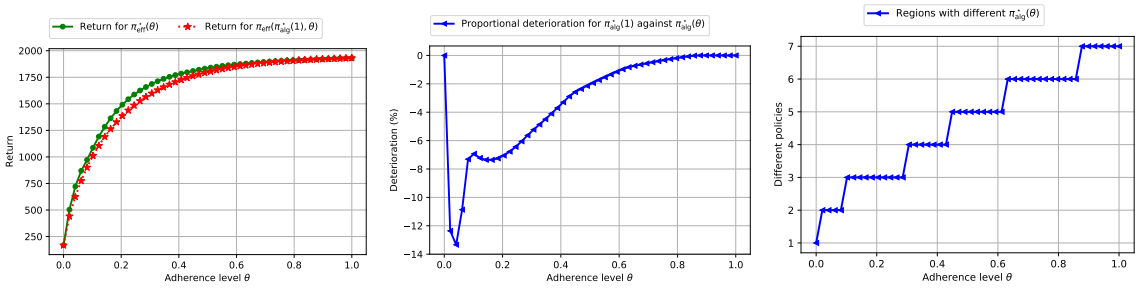
(a) Transition probabilities for action repair.

(b) Transition probabilities for action wait.

Figure 3 **Transition probabilities for the machine replacement MDP. There is a reward of** 18 **in state** $R1$**, of** 10 **in state** $R2$ **and of** 0 **in state** 8**. All others states have a reward of** 20**.**

Numerical results. Assuming $\theta = 1$, an optimal policy $\pi^\star_{\mathsf{alg}}(1)$ is to choose action wait in States $1, 2, 3, 4, R_2$ and action repair in States $5, 6, 7, 8, R_1$. We now compare the effective return of $\pi^\star_{\mathsf{alg}}(1)$ with that of the best recommendation $\pi^\star_{\mathsf{alg}}(\theta)$, for varying values of the adherence level θ. We first consider the case where π_{base} chooses to always wait instead of repairing the machine. We present the results of our empirical study in Figure 4. In Figure 4a, we report the effective return of both policies, namely $R(\pi^\star_{\mathsf{eff}}(\theta))$ and $R(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1), \theta))$, for varying $\theta \in [0, 1]$. We also compute the proportional deterioration in performance, $\left(R(\pi^\star_{\mathsf{eff}}(\theta)) - R(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1), \theta))\right) / R(\pi^\star_{\mathsf{eff}}(\theta))$ in Figure 4b. As expected from Proposition 4.4, when θ is sufficiently close to 1 (here, for $\theta \geq 0.88$), we have $\pi^\star_{\mathsf{eff}}(\theta) = \pi^\star_{\mathsf{eff}}(1)$ and there is no deterioration in performance. However, as the value of θ decreases towards 0, overlooking the adherence level and recommending $\pi^\star_{\mathsf{alg}}(1)$ can lead to as much as 13.34% proportional deterioration compared with the optimal return $R(\pi^\star_{\mathsf{eff}}(\theta))$. We also note in Figure 4b that small changes in θ can lead to very severe deterioration, for instance in the region $\theta \in [0, 0.20]$, i.e., for very low adherence from the human decision maker. The different regions over which the

optimal decision $\theta \mapsto \pi^{\star}_{\mathsf{alg}}(\theta)$ is constant are shown in Figure 4c, which highlights that the optimal recommendation policy may change many times as the adherence level decreases.



(a) Returns for recommending $\pi^{\star}_{\mathsf{alg}}(\theta)$ and $\pi^{\star}_{\mathsf{alg}}(1)$.

(b) Proportional deterioration for $\pi^{\star}_{\mathsf{alg}}(1)$ against $\pi^{\star}_{\mathsf{alg}}(\theta)$.

(c) Subregions with constant recommendation policies.

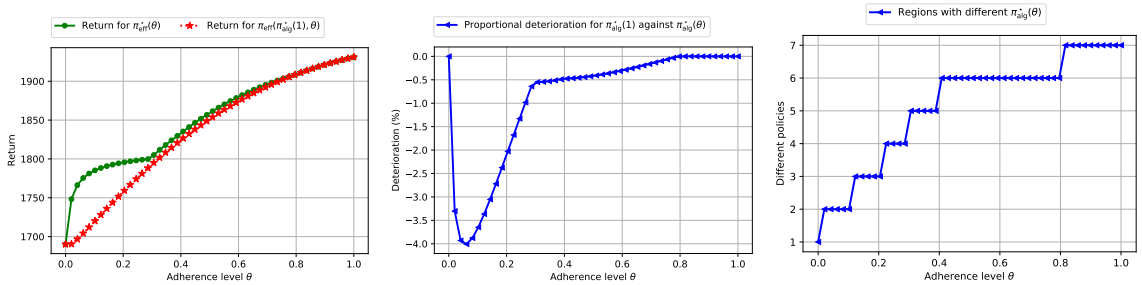
Figure 4 Numerical results for the machine replacement MDP with π_{base} always choosing action wait.

We also study the impact of the adherence level when π_{base} is the policy that avoids being trapped in the "bad" states (States $8, R_1, R_2$). In particular, let us consider a policy π_{base} that always waits when the machine is not broken (State 1 to State 7) or in the normal repair state (State R_2), but chooses to repair in State 8 and in the long repair state (State R_1). The numerical results are presented in Figure 5. In this case, we see that the performance of $\pi^{\star}_{\mathsf{alg}}(1)$ are robust for $\theta \geq 0.35$, with a proportional deterioration of only 0.5% compared to the return of the optimal recommendation policy $\pi^{\star}_{\mathsf{alg}}(\theta)$ (Figure 5b). However, for $\theta \leq 0.35$, there is a significant drop in performance, leading to a 4.01% reduction in effective return.

5.2. Stylized healthcare decision problem

We consider an MDP instance inspired from sequential decision-making in healthcare. In particular, we approximate the evolution of the patient's health dynamics using a Markov chain, using a simplification of the models in Goh et al. (2018) and **?**.

MDP instance. The dynamics of the MDP is represented in Figure 6. There are 5 states representing the severity of the health condition of the patient, and an absorbing *mortality* state m. State 1 represents a healthy condition for the patient while State 5 is more likely to lead to mortality.



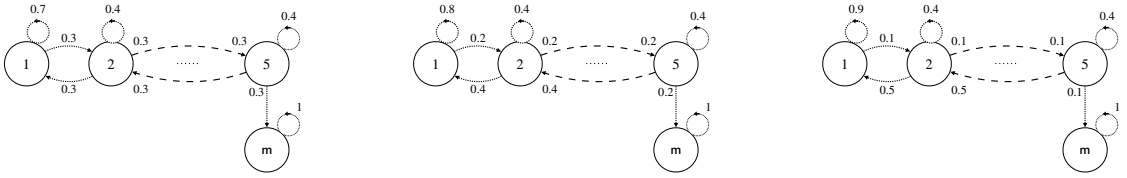
(a) Returns for recommending $\pi^\star_{\text{alg}}(\theta)$ and $\pi^\star_{\text{alg}}(1)$.

(b) Proportional deterioration for $\pi^\star_{\text{alg}}(1)$ against $\pi^\star_{\text{alg}}(\theta)$.

(c) Subregions with constant recommendation policies.

Figure 5 **Numerical results for the machine replacement MDP with π_{base} repairing in the absorbing states $8, R_1$ and waiting in the other states.**

There are three actions {low, medium, high}, corresponding to prescription of a given drug dosage at every state. In any given state (except mortality), there is a reward of 20 for choosing action low, a reward of 15 for choosing action medium, and a reward of 10 for choosing action high. There is a reward of 0 in the mortality state m. The goal of the decision maker is to choose a policy to keep the patient alive (by avoiding the mortality state m) while minimizing the invasiveness of the treatment. We choose a discount factor of $\lambda = 0.99$ and the patient starts in State 1.



(a) Transition probabilities for action low.

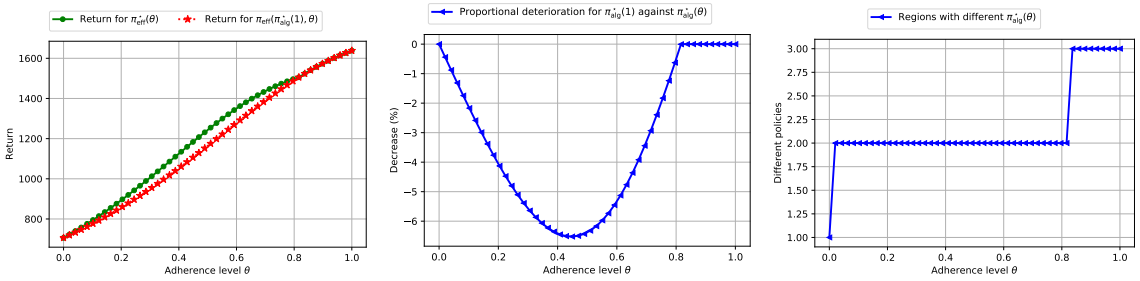
(b) Transition probabilities for action medium.

(c) Transition probabilities for action high.

Figure 6 **Transition probabilities for the healthcare MDP instance.**

Numerical results. An optimal policy $\pi^\star_{\text{alg}}(1)$ is to choose action low in States $1, 2$, and to choose action high in States $3, 4, 5$. We now test the robustness of $\pi^\star_{\text{alg}}(1)$ to partial adherence of the patient. In particular, we consider three different baseline policies π_{base}. In Figure 7, Figure 8 and

Figure 9, we consider baseline policies π_{base} that always chooses action low, medium or high in every health states, respectively. Our simulations highlights the sensitivity of the effective performance of $\pi^{\star}_{\mathsf{alg}}(1)$, with respect to both the baseline policy and the adherence level. In particular, while $\pi^{\star}_{\mathsf{alg}}(1)$ may loose up to 6.52% of the optimal effective return when the baseline policy always chooses low dosage (Figure 7b), it only loses a maximum of 0.97% of the optimal effective return when the baseline policy always chooses medium dosage (Figure 8b), and loses close to 0% of the optimal effective return when the baseline policy always chooses high dosage (Figure 9b). In addition, we observe that the range of the θ-values for which $\pi^{\star}_{\mathsf{alg}}(1)$ is optimal differs greatly from one baseline policy to another (Figures 7c-8c-9c): when π_{base} always chooses low dosage, $\pi^{\star}_{\mathsf{alg}}(1)$ is optimal for $\theta \geq 0.82$, whereas when π_{base} always chooses medium dosage, $\pi^{\star}_{\mathsf{alg}}(1)$ is optimal for $\theta \geq 0.51$, and when π_{base} always chooses high dosage, $\pi^{\star}_{\mathsf{alg}}(1)$ is optimal for $\theta \geq 0.20$.

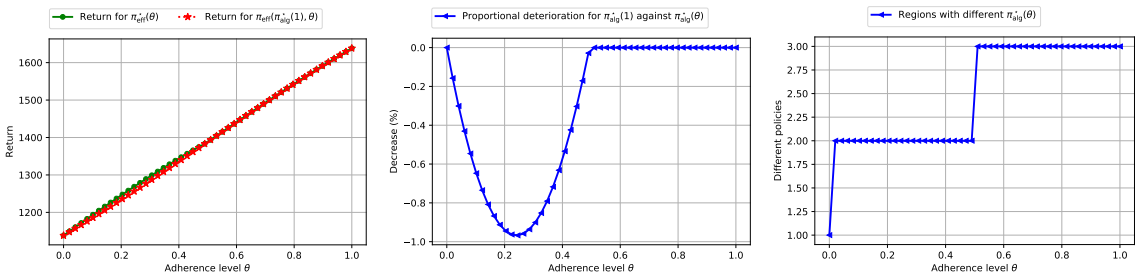


(a) Returns for recommending $\pi^{\star}_{\mathsf{alg}}(\theta)$ and $\pi^{\star}_{\mathsf{alg}}(1)$.

(b) Proportional deterioration for $\pi^{\star}_{\mathsf{alg}}(1)$ against $\pi^{\star}_{\mathsf{alg}}(\theta)$.

(c) Subregions with constant recommendation policies.

Figure 7 Numerical results for the healthcare MDP with π_{base} choosing action low in all states.

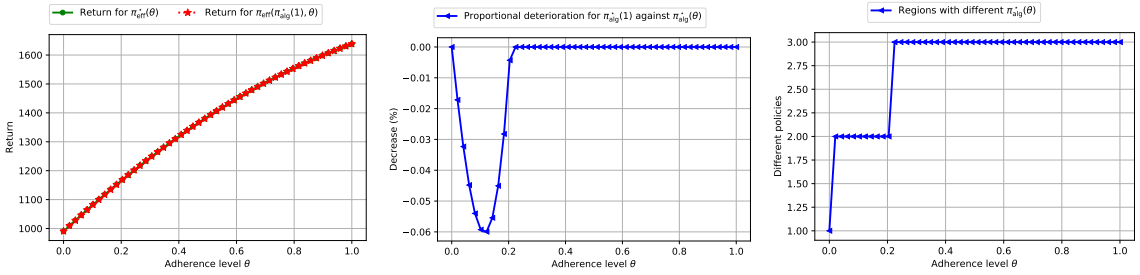


(a) Returns for recommending $\pi^\star_{\text{alg}}(\theta)$ and $\pi^\star_{\text{alg}}(1)$.

(b) Proportional deterioration for $\pi^\star_{\text{alg}}(1)$ against $\pi^\star_{\text{alg}}(\theta)$.

(c) Subregions with constant recommendation policies.

Figure 8 **Numerical results for the healthcare MDP with π_{base} choosing action** medium **in all states.**



(a) Returns for recommending $\pi^\star_{\text{alg}}(\theta)$ and $\pi^\star_{\text{alg}}(1)$.

(b) Proportional deterioration for $\pi^\star_{\text{alg}}(1)$ against $\pi^\star_{\text{alg}}(\theta)$.

(c) Subregions with constant recommendation policies.

Figure 9 **Numerical results for the healthcare MDP with π_{base} choosing action** high **in all states.**

6. Extensions and discussion

Finally, we discuss additional properties and potential extensions of our adherence-aware decision framework.

6.1. Heterogeneous adherence levels across states

We have restricted our previous analysis to the case of a *homogeneous* adherence level $\theta \in [0, 1]$, common to all states $s \in \mathcal{S}$. However, in practice, it is possible that the adherence level differs across states. For instance, in a healthcare setting, practitioners may be more prone to overlook the algorithms' recommendations when the patient is in a critical health condition because any error may have life-threatening consequences. To model this practical consideration, we can extend our model to *heterogeneous adherence levels*, $\theta_s \in [0, 1]$ for each state $s \in \mathcal{S}$. In this model, at every

decision period $t \in \mathbb{N}$ and visited state s_t, the decision maker decides to follow the recommendation policy π_{alg} (with probability θ_s) or the baseline policy π_{base} (with probability $1-\theta_s$). The effective policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})$ is now defined as

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})_s = \theta_s \boldsymbol{\pi}_{\mathsf{alg},s} + (1-\theta_s)\boldsymbol{\pi}_{\mathsf{base},s}, \forall\, s \in \mathcal{S}. \tag{6.1}$$

All the structural results from Section 4.1 would generalize to this simple extension. In particular, Proposition 4.3 still holds provided the non-decreasing property of $\theta \mapsto R(\pi^\star_{\mathsf{eff}}(\theta))$ is replaced with an order-preserving property:

$$\theta_s \leq \theta'_s, \forall\, s \in \mathcal{S} \Rightarrow R(\pi^\star_{\mathsf{eff}}(\boldsymbol{\theta})) \leq R(\pi^\star_{\mathsf{eff}}(\boldsymbol{\theta}')).$$

Importantly, we can still efficiently find an optimal recommendation policy $\pi^\star_{\mathsf{alg}}(\boldsymbol{\theta})$ for any adherence level $\boldsymbol{\theta} \in [0,1]^{\mathcal{S}}$, by adapting the value iteration and the linear programming formulation to the map $f_{\boldsymbol{\theta}} : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$, defined as

$$f_{\boldsymbol{\theta},s}(\boldsymbol{v}) = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \theta_s \cdot \sum_{a \in \mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}\right) + (1-\theta_s) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},sa} \boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}\right), \forall\, s \in \mathcal{S}.$$

6.2. Heterogeneous adherence levels across states and actions

Furthermore, it is plausible in practice that recommendations that are close to the baseline actions are more likely to be followed than drastically different ones, e.g., in a healthcare setting where the actions correspond to drug dosages. To model this situation, we can extend our framework further to involve an adherence level that depends on each state $s \in \mathcal{S}$ and each action in $a \in \mathcal{A}$. Formally, we could study policies of the form

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})_{sa} = \theta_{sa} \pi_{\mathsf{alg},sa} + (1-\theta_{sa})\pi_{\mathsf{base},sa}, \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}.$$

However, for every state $s \in \mathcal{S}$, we need $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})_s \in \Delta(\mathcal{A})$, which imposes some non-trivial restrictions on the values of θ_{sa} (which would depend on the probability of playing each action according to π_{alg} and π_{base}).

To circumvent this issue, we propose an alternative model where $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})_s \in \Delta(\mathcal{A})$ by design. For the sake of simplicity, in this section, we assume that π_{base} is a deterministic stationary policy: for each state $s \in \mathcal{S}$ we write $\pi_{\mathsf{base}}(s) \in \mathcal{A}$ for the action chosen by the policy π_{base}. At a state $s \in \mathcal{S}$, a recommended action a is sampled from the probability distribution $\boldsymbol{\pi}_{\mathsf{alg},s} \in \Delta(\mathcal{A})$. Then with probability $\theta_{sa} \in [0,1]$ the DM follows the recommendation (action a), otherwise the action selected by the DM is $\pi_{\mathsf{base}}(s)$. With this model, the effective policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})$ for some $(\theta_{sa})_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in [0,1]^{\mathcal{S}\times\mathcal{A}}$ is such that

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})_{sa} = \begin{cases} \pi_{\mathsf{alg},sa}\theta_{sa} & \text{if } a \neq \pi_{\mathsf{base}}(s), \\ 1 - \sum_{a' \in \mathcal{A}\setminus\{\pi_{\mathsf{base}}(s)\}} \pi_{\mathsf{alg},sa'}\theta_{sa'} & \text{if } a = \pi_{\mathsf{base}}(s). \end{cases}$$

Note that the expression for the case $a = \pi_{\mathsf{base}}(s)$ simply follows from

$$1 - \sum_{a' \in \mathcal{A}\setminus\{\pi_{\mathsf{base}}(s)\}} \pi_{\mathsf{alg},sa'}\theta_{sa'} = \pi_{\mathsf{alg},s\pi_{\mathsf{base}}(s)} + \sum_{a' \in \mathcal{A}\setminus\{\pi_{\mathsf{base}}(s)\}} \pi_{\mathsf{alg},sa'}(1-\theta_{sa'}), \tag{6.2}$$

i.e., action $\pi_{\mathsf{base}}(s)$ is chosen either because it has been sampled following $\boldsymbol{\pi}_{\mathsf{alg},s}$ or because another action a' was sampled but the decision maker chose to follow π_{base}, which happens with probability $1-\theta_{sa'}$. We can now write the value function of a policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \boldsymbol{\theta})$. For any $s \in \mathcal{S}$, we obtain, using (6.2):

$$v_s^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\boldsymbol{\theta})} = \sum_{a\in\mathcal{A}} \pi_{sa}\left(\theta_{sa}\boldsymbol{P}_{sa}^\top\left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\boldsymbol{\theta})}\right) + (1-\theta_{sa})\boldsymbol{P}_{s\pi_{\mathsf{base}}(s)}^\top\left(\boldsymbol{r}_{s\pi_{\mathsf{base}}(s)} + \lambda \boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\boldsymbol{\theta})}\right)\right).$$

Overall, we have obtained that the value function $\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\boldsymbol{\theta})}$ satisfies

$$v_s^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\boldsymbol{\theta})} = \sum_{a\in\mathcal{A}} \pi_{sa}\left(r'_{sa} + \lambda \boldsymbol{P}'^\top_{sa} \boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\boldsymbol{\theta})}\right), \forall\, s \in \mathcal{S}$$

with $\boldsymbol{P}' \in (\Delta(\mathcal{S}))^{\mathcal{S}\times\mathcal{A}}, \boldsymbol{r}' \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ the transition probabilities and the instantaneous rewards of another surrogate MDP $\mathcal{M}'$ with transitions and rewards defined as $\boldsymbol{P}'_{sa} := \theta_{sa}\cdot\boldsymbol{P}_{sa} + (1-\theta_{sa})\cdot\boldsymbol{P}_{s\pi_{\mathsf{base}}(s)}, r'_{sa} := \theta_{sa}\cdot\boldsymbol{P}_{sa}^\top\boldsymbol{r}_{sa} + (1-\theta_{sa})\cdot\boldsymbol{P}_{s\pi_{\mathsf{base}}(a)}^\top\boldsymbol{r}_{s\pi_{\mathsf{base}}(a)}$, for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. This shows that for this model of state-action-dependent adherence level, we can efficiently find an optimal recommendation policy by computing an optimal (nominal) policy for the surrogate MDP $\mathcal{M}'$.

6.3. Uncertain adherence level

In our framework, we have assumed that the adherence level $\theta \in [0,1]$ was known and used as an input to design the recommendation policy π_{alg}. This assumption is likely violated in practice, where θ is not perfectly known. Instead, we can assume that the true adherence level θ is uncertain but belongs to an interval $[\underline{\theta}, \bar{\theta}]$. Under this assumption, we take a *robust optimization* approach (Bertsimas and Sim 2004, Ben-Tal et al. 2009) and model the uncertainty in the value of θ as an adversarial choice from the set $[\underline{\theta}, \bar{\theta}]$ of all possible realizations. The goal is to compute an optimal *robust* recommendation policy, that optimizes the worst-case objective over all plausible values of the adherence levels:

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{\theta \in [\underline{\theta}, \bar{\theta}]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)). \tag{6.3}$$

The optimization problem (6.3) is reminiscent to *robust MDPs*, which consider the case where the rewards and/or the transition probabilities are unknown (Iyengar 2005, Wiesemann et al. 2013), but in our setting the same adherence level θ has an impact on the transition probabilities out of every states $s \in \mathcal{S}$ in the surrogate MDP, which contradicts the classical rectangularity assumption for robust MDPs. However, thanks to the structural properties highlighted in Section 4.1, the optimization problem (6.3) can be solved as efficiently as AdaMDP, the adherence-aware decision-making problem with known adherence level θ. Crucially, an optimal recommendation policy can still be chosen stationary (i.e., in the set Π) instead of history-dependent (i.e., in the set Π_{H}), and deterministic. Formally, we have the following theorem (proof detailed in Appendix J):

THEOREM 6.1. *An optimal robust recommendation policy in* (6.3) *may be chosen stationary:*

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{\theta \in [\underline{\theta}, \bar{\theta}]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)) = \max_{\pi_{\mathsf{alg}} \in \Pi} \min_{\theta \in [\underline{\theta}, \bar{\theta}]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)).$$

Additionally, the pair $\left(\pi^{\star}_{\mathsf{alg}}(\underline{\theta}), \underline{\theta}\right)$ *with* $\pi^{\star}_{\mathsf{alg}}(\underline{\theta})$ *a deterministic policy is an optimal solution to* (6.3).

Theorem 6.1 is remarkable in that it shows that the same value of θ (in particular, the most pessimistic value $\underline{\theta}$) is attaining the worst-case return for all policies. In practice, it reduces the problem of estimating the true adherence level to the (admittedly easier) task of obtaining a valid

lower bound only. Furthermore, Theorem 6.1 also has significant computational impact since it shows that solving (6.3) can be done by applying the same algorithms as the one described in Section 4.2 with $\theta = \underline{\theta}$. The resulting recommendation will also be a deterministic policy, which is desirable in practice. The proof is very similar to the case of time- and state-invariant adversarial adherence decision in Theorem 3.2 and we present it in Appendix J.

6.4. Uncertain baseline policy

Similarly, the baseline policy π_{base} is currently a known input to our adherence-aware MDP framework. However, in practice it is possible that the algorithm only has access to an estimation $\hat{\pi}_{\mathsf{base}}$ of the baseline policy, learned from a finite dataset, and that the true baseline policy differs from $\hat{\pi}_{\mathsf{base}}$. We consider a robust approach where the recommendation policy optimizes over the worst-case baseline policy $\pi_{\mathsf{base}} \in \Gamma$, where the set $\Gamma \subseteq (\Delta(\mathcal{A}))^{\mathcal{S}}$ represents feasible baseline policies that are close to the estimation $\hat{\pi}_{\mathsf{base}}$, i.e., we consider

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{\pi_{\mathsf{base}} \in \Gamma} R(\theta \pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}). \tag{6.4}$$

The following theorem shows that (6.4) is still a tractable optimization problem under some mild assumption on Γ. We provide the detailed proof in Appendix K.

THEOREM 6.2. *Assume that the set of feasible baseline policies* Γ *satisfies the following* rectangularity assumption*:* $\Gamma = \times_{s \in \mathcal{A}} \Gamma_s$ *where* $\Gamma_s \subseteq \Delta(\mathcal{A})$ *is a convex, compact set for each* $s \in \mathcal{S}$*. Then an optimal solution to* (6.4) *exists and can be chosen stationary. Additionally, if the set* Γ *is a polytope or defined with conic constraints, then an optimal solution to* (6.4) *can be computed efficiently.*

Our proof is based on showing that the optimization problem (6.4) can be reformulated as an s-rectangular robust MDP (Wiesemann et al. 2013) with uncertain pair $(\boldsymbol{r}, \boldsymbol{P})$ of instantaneous rewards and transition probabilities. This follows from the interpretation of AdaMDP as solving a surrogate MDP, where the rewards and transitions, defined in (4.3), are dependent on π_{base}.

6.5. Varying adherence level

The adherence level θ may also vary over time. As the DM observes the recommendation made by the algorithm over time, her trust in the recommendation, hence her adherence, may increase (or decrease).

One could endogeneize these dynamics by making θ explicitly dependent on the recommended policy π_{alg}. However, the works of Boyacı et al. (2023), de Véricourt and Gurkan (2023) highlight how complex these dynamics can be, even for highly stylized decision problems, because of cognitive limitations and asymmetric performance evaluation. Therefore, we conjecture that such game-theoretic approaches (where π_{alg} and θ are updated at each step) would be intractable for the type of complex multi-stage decision problems we consider in this paper. Furthermore, as discussed in Section 2, many mechanisms could explain partial adherence. Consequently, any method that restricts the reasons for non-adherence (e.g., information asymmetry, algorithm aversion, cognitive limitations) and derives update rules for the adherence level θ based on these mechanisms could suffer from model misspecification.

Alternatively, one could capture the dynamic nature of θ by estimating it from past observations in an online fashion. At a high-level, the optimization problem to which $\pi^{\star}_{\mathsf{alg}}(\theta)$ is a solution resembles that of an MDP whose transition probabilities depend on θ (and π_{base}). Hence, a varying adherence level would lead to non-stationary transition probabilities. In the multi-armed bandit literature, two types of assumptions are used to address non-stationarity. Garivier and Moulines (2011) introduced a piecewise stationary assumption, where the parameters are constant over certain time periods and change at unknown time steps. Alternatively, Besbes et al. (2014, 2015) considered a slowly varying setting where the absolute difference between parameters at two consecutive time-steps are bounded (by a so-called variation budget). Although originally derived for multi-armed bandit problems, both these frameworks have been extended and used to solve non-stationary MDPs (or non-stationary reinforcement learning problems) as well. We refer to Auer et al. (2008) and Cheung et al. (2023) for an analysis of non-stationary MDPs under the piecewise

stationary and slowly varying assumptions respectively. Beyond the technical difficulties addressed by the aforementioned works, learning θ from past historical data also suffers from a censorship issue: if both π_{alg} and π_{base} recommend the same action at a given state s_t, then it is impossible to distinguish adherence from non-adherence.

We see our model based on partial adherence in *offline* sequential decision-making as a first step towards a better understanding of the phenomena arising in expert-in-loop systems and a better design of algorithmic recommendations. The online extension of our framework, where the adherence level (and potentially the baseline policy π_{base}) needs to be continuously learned from past observations constitutes an interesting future direction, as well as the case where the real MDP parameters $(\boldsymbol{r}, \boldsymbol{P})$ themselves are only partially known to the human agent and the algorithm and must be learned over time.

Acknowledgements

We would like to thank the three anonymous reviewers and the associate editor for their insightful comments that lead to a more complete version of the paper.

References

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*, 2021.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*, volume 28. Princeton university press, 2009.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

Dimitris Bertsimas, Omid Nohadani, and Kwong Meng Teo. Nonconvex robust optimization for problems with constraints. *INFORMS Journal on Computing*, 22(1):44–58, 2010.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. Fairness, efficiency, and flexibility in organ allocation for kidney transplantation. *Operations Research*, 61(1):73–87, 2013.

Dimitris Bertsimas, Jean Pauphilet, Jennifer Stevens, and Manu Tandon. Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 24(6): 2809–2824, 2022.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, 27, 2014.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Jose Blanchet, Guillermo Gallego, and Vineet Goyal. A markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905, 2016.

Tamer Boyacı, Caner Canyakmaz, and Francis de Véricourt. Human and machine: The impact of machine input on decision-making under cognitive limitations. *Management Science*, 2023.

Fernanda Bravo and Yaron Shaposhnik. Mining optimal policies: A pattern recognition approach to model analysis. *INFORMS Journal on Optimization*, 2(3):145–166, 2020.

Felipe Caro and Anna Sáez de Tejada Cuenca. Believing in analytics: Managers' adherence to price recommendations from a DSS. *Manufacturing & Service Operations Management*, 2023.

Wang Chi Cheung, David Simchi-Levi, and Ruihao Zhu. Non-stationary reinforcement learning: The blessing of (more) optimism. *Management Science*, 2023.

Dragos Florin Ciocan and Velibor V Mišić. Interpretable optimal stopping. *Management Science*, 68(3): 1616–1638, 2022.

Francis de Véricourt and Huseyin Gurkan. Is your machine better than you? you may never know. *Management Science*, 2023.

Eric Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203 – 213, 2010.

Cyrus Derman. *Finite State Markovian Decision Processes*. Academic Press, Inc., 1970.

Antoine Désir, Vineet Goyal, Danny Segev, and Chun Ye. Constrained assortment optimization under the markov chain–based choice model. *Management Science*, 66(2):698–721, 2020.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170, 2018.

Eugene A Feinberg and Adam Shwartz. *Handbook of Markov Decision Processes: Methods and Applications*, volume 40. Springer Science & Business Media, 2012.

Robert Fildes, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, 25(1):3–23, 2009.

Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

Joel Goh, Mohsen Bayati, Stefanos A Zenios, Sundeep Singh, and David Moore. Data uncertainty in Markov chains: Application to cost-effectiveness analyses of medical innovations. *Operations Research*, 66(3): 697–715, 2018.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017.

Vineet Goyal and Julien Grand-Clément. Robust Markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 2022.

Julien Grand-Clément, Carri W Chan, Vineet Goyal, and Gabriel Escobar. Robust policies for proactive ICU transfers. *arXiv preprint arXiv:2002.06247*, 2020.

Maria R Ibanez, Jonathan R Clark, Robert S Huckman, and Bradley R Staats. Discretionary task ordering: Queue management in radiological services. *Management Science*, 64(9):4389–4407, 2018.

Garud N Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

Alexis Jacq, Johan Ferret, Olivier Pietquin, and Matthieu Geist. Lazy-MDPs: Towards interpretable reinforcement learning by learning when to act. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 669–677, 2022.

Nathan Kallus. Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, pages 1789–1798. PMLR, 2017.

Saravanan Kesavan and Tarun Kushwaha. Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science*, 66(11):5182–5190, 2020.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

Mirko Kremer, Brent Moritz, and Enno Siemsen. Demand forecasting behavior: System neglect and change detection. *Management Science*, 57(10):1827–1843, 2011.

Wilson Lin, Song-Hee Kim, and Jordan Tong. Does algorithm aversion exist in the field? An empirical analysis of algorithm use determinants in diabetes self-management. *SSRN (July 23, 2021)*, 2021.

Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

Han Men, Robert M Freund, Ngoc C Nguyen, Joel Saa-Seoane, and Jaime Peraire. Fabrication-adaptive optimization with an application to photonic crystal design. *Operations Research*, 62(2):418–434, 2014.

Vahid Balazadeh Meresht, Abir De, Adish Singla, and Manuel Gomez-Rodriguez. Learning to switch between machines and humans. *arXiv preprint arXiv:2002.04258*, 2020.

Martin L Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 2014.

Eduardo Sabaté. *Adherence to Long-term Therapies: Evidence for Action*. World Health Organization, 2003.

Lior Shani, Yonathan Efroni, and Shie Mannor. Exploration conscious reinforcement learning revisited. In *International Conference on Machine Learning*, pages 5680–5689. PMLR, 2019.

Lauren N Steimle and Brian T Denton. Markov decision processes for screening and treatment of chronic diseases. In *Markov Decision Processes in Practice*, pages 189–222. Springer, 2017.

Jiankun Sun, Dennis J Zhang, Haoyuan Hu, and Jan A Van Mieghem. Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 68(2):846–865, 2022.

Karel H Van Donselaar, Vishal Gaur, Tom Van Woensel, Rob ACM Broekmeulen, and Jan C Fransoo. Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, 56(5):766–784, 2010.

Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

Appendix A: Proof of Theorem 3.1

Proof of Theorem 3.1. Let us assume that the random variables $(u_{s,t})_{s,t}$ are such that $u_{s,t}$ an $u_{s',t'}$ are independent for any $t \neq t'$, and that for any $t \in \mathbb{N}$, $\mathbb{E}_u\left[(u_{s,t})_{s\in\mathcal{S}}\right] = (\theta, ..., \theta) \in [0,1]^{\mathcal{S}}$. We prove that $\max_{\pi_{\mathsf{alg}}\in\Pi} R\left(\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}\right) = \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right]$. The proof proceeds in three steps.

Step 1. We first show that we can restrict ourselves to Markovian policies: $\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{M}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right]$. Note that for some fixed values of $u \in [0,1]^{\mathcal{S}\times\mathbb{N}}$, the map $\pi_{\mathsf{alg}} \mapsto R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$ is a function of the values of $\mathbb{P}^{\pi_{\mathsf{alg}}}(s_t = s, a_t = a)$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $t \in \mathbb{N}$. Following Puterman (corollary 5.5.2, 2014), for any history-dependent policy $\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}$, there exists a Markovian policy $\pi'_{\mathsf{alg}} \in \Pi_{\mathsf{M}}$, potentially randomized, such that for any pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ and any time $t \in \mathbb{N}$, we have $\mathbb{P}^{\pi_{\mathsf{alg}}}(s_t = s, a_t = a) = \mathbb{P}^{\pi'_{\mathsf{alg}}}(s_t = s, a_t = a)$. Therefore, for any history-dependent policy $\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}$, we can find a Markovian policy π'_{alg} such that $\mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi'_{\mathsf{alg}}, u))\right]$. From this we conclude that $\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{M}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right]$.

Step 2. We now show that for any $\pi_{\mathsf{alg}} \in \Pi_{\mathsf{M}}$, we have $\mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$. Indeed, let us define, for $\pi \in (\Delta(\mathcal{A}))^{\mathcal{S}}$, the transition matrix $\boldsymbol{P}^{\pi} \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ as $P^{\pi}_{ss'} = \sum_{a\in\mathcal{A}} \pi_{sa} P_{sas'}, \forall\, (s,s') \in \mathcal{S}\times\mathcal{S}$ and $\boldsymbol{r}_{\pi} \in \mathbb{R}^{\mathcal{S}}, r_{\pi,s} = \sum_{a\in\mathcal{A}} \pi_{sa} \boldsymbol{P}^{\top}_{sa} \boldsymbol{r}_{sa}$. Note that $\boldsymbol{P}^{\pi}$ and $\boldsymbol{r}_{\pi}$ depend linearly on π. Then by definition we have, for a Markovian policy $\pi = (\pi_t)_{t\in\mathbb{N}}$ with $\pi_t \in (\Delta(\mathcal{A}))^{\mathcal{S}}$ for $t \in \mathbb{N}$,

$$R(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t s_{t+1}}\right] = \boldsymbol{p}_0^{\top}\left(\sum_{t=0}^{+\infty} \lambda^t \prod_{t'=0}^{t-1} \boldsymbol{P}^{\pi_{t'}} \boldsymbol{r}_{\pi_t}\right).$$

We have

$$\begin{aligned}
\mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] &= \boldsymbol{p}_0^{\top}\left(\mathbb{E}_u\left[\sum_{t=0}^{+\infty} \lambda^t \prod_{t'=0}^{t-1} \boldsymbol{P}^{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_{t'}} \boldsymbol{r}_{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_t}\right]\right) \\
&= \boldsymbol{p}_0^{\top}\left(\sum_{t=0}^{+\infty} \lambda^t \mathbb{E}_u\left[\prod_{t'=0}^{t-1} \boldsymbol{P}^{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_{t'}} \boldsymbol{r}_{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_t}\right]\right) \qquad \text{(A.1)} \\
&= \boldsymbol{p}_0^{\top}\left(\sum_{t=0}^{+\infty} \lambda^t \prod_{t'=0}^{t-1} \mathbb{E}_u\left[\boldsymbol{P}^{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_{t'}}\right] \mathbb{E}_u\left[\boldsymbol{r}_{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_t}\right]\right) \qquad \text{(A.2)} \\
&= \boldsymbol{p}_0^{\top}\left(\sum_{t=0}^{+\infty} \lambda^t \prod_{t'=0}^{t-1} \boldsymbol{P}^{\mathbb{E}_u\left[(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_{t'}\right]} \boldsymbol{r}_{\mathbb{E}_u\left[(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))_t\right]}\right) \qquad \text{(A.3)} \\
&= \boldsymbol{p}_0^{\top}\left(\sum_{t=0}^{+\infty} \lambda^t \prod_{t'=0}^{t-1} \boldsymbol{P}^{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))} \boldsymbol{r}_{(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))}\right) \qquad \text{(A.4)} \\
&= R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) \\
&= R\left(\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}\right) \qquad \text{(A.5)}
\end{aligned}$$

where (A.1) follows from the dominated convergence theorem, (A.2) follows from the adherence decisions being independent across time, (A.3) follows from linearity of the expectation and the definition of $\boldsymbol{P}^{\pi}$ and $\boldsymbol{r}_{\pi}$, (A.4) follows from $\mathbb{E}_u\left[\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)\right] = \pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)$, and finally (A.5) follows from the definition of $R\left(\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}\right)$.

Step 3. In Step 1, we have shown $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = \sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{M}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right]$. In Step 2, we have shown $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{M}}} \mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = \sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{M}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$. Proposition 3.1 shows that $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{M}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)) = \max_{\pi_{\mathsf{alg}} \in \Pi} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$, which concludes our proof. □

Other random models of adherence decisions. We briefly discuss here the viability of *Time-invariant random* adherence decision models, where there are some correlations across the adherence decisions across times. One possible *time-invariant* random models corresponds to

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)_{s,t} = u_s \pi_{\mathsf{alg}}{}_{s,t} + (1 - u_s)\pi_{\mathsf{base}}{}_{s,t}$$

with u_s sampled following a distribution with mean θ independently across all $s \in \mathcal{S}$. Another random model of adherence decisions corresponds to *Time- and State-invariant random* adherence decisions, where

$$\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)_{s,t} = u \pi_{\mathsf{alg}}{}_{s,t} + (1 - u)\pi_{\mathsf{base}}{}_{s,t},$$

with u sampled following a distribution with mean θ and support in $[0, 1]$. These models appear harder to analyze than the random models of deviation from Theorem 3.1, where the fact that the decisions are *independent* over time plays a crucial role in our proof. We simply note that an interesting property arises when the adherence decisions $u_{s,t}$ is common across all states and times: $u_{s,t} = u, \forall\, (s,t) \in \mathcal{S} \times \mathbb{N}$, and chosen at random following a distribution *supported in* $\{0, 1\}$, with mean $\theta \in [0, 1]$. In this case, the decision maker chooses either to follow π_{alg} (in every state and at every period) with probability θ, or to follow π_{base} with probability $1 - \theta$. This situation may occur in the case where the decision maker is reluctant to changing policy along a trajectory and is constrained to follow the same policy at all states, e.g. because of concerns about the consistency of the resulting effective policy. Consequently, we have $\mathbb{E}_u\left[R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))\right] = \theta R(\pi_{\mathsf{alg}}) + (1 - \theta) R(\pi_{\mathsf{base}})$, so that an optimal recommendation policy $\pi^{\star}_{\mathsf{alg}}(\theta)$ may be chosen *independent* of the true value of the adherence level θ and it is equal to the optimal nominal policy for the MDP instance $\mathcal{M}$.

Appendix B: Proof of Theorem 3.2

In this section we study the adversarial models of adherence decisions and their equivalence with AdaMDP. For concision, we denote

$$B_\infty := [\theta, 1]^{\mathcal{S}\times\mathbb{N}}, \qquad \text{(Unconstrained Adversarial)}$$

$$B_1 := \{u \in [\theta, 1]^{\mathcal{S}\times\mathbb{N}} \mid u_{s,t} = u_{s,t'}, \forall s \in \mathcal{S}, \forall\, t, t' \in \mathbb{N}\}, \qquad \text{(Time-invariant Adversarial)}$$

$$B_2 := \{u \in [\theta, 1]^{\mathcal{S}\times\mathbb{N}} \mid u_{s,t} = u_{s',t}, \forall s, s' \in \mathcal{S}, \forall\, t \in \mathbb{N}\}, \qquad \text{(State-invariant Adversarial)}$$

$$B_3 := \{u \in [\theta, 1]^{\mathcal{S}\times\mathbb{N}} \mid u_{s,t} = u_{s',t'}, \forall s, s' \in \mathcal{S}, \forall\, t, t' \in \mathbb{N}\} \qquad \text{(Time- and State-invariant Adversarial)}$$

We will prove the following theorems, showing the connection between adversarial models of adherence decisions and AdaMDP. We then turn to showing strong duality in Appendix B.3.

THEOREM B.1 **(Unconstrained Adversarial)**. *For a given adherence level $\theta \in [0, 1]$, we have the following equality:*

$$\max_{\pi_{\mathsf{alg}} \in \Pi} R\left(\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}\right) = \sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)).$$

Additionally, there exists an optimal stationary deterministic policy that is a solution to the right-hand side optimization problem above.

THEOREM B.2 **(Other Adversarial models)**. *Let $B \subset [\theta, 1]^{\mathcal{S}\times\mathbb{N}}$ be either B_1 (***Time-invariant Adversarial***), B_2 (***State-invariant Adversarial***), or B_3 (***Time- and State-invariant Adversarial***). For a given adherence level $\theta \in [0, 1]$, we have the following equality:*

$$\max_{\pi_{\mathsf{alg}} \in \Pi} R\left(\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}\right) = \sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{u \in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)).$$

Additionally, there exists an optimal stationary deterministic policy that is a solution to the right-hand side optimization problem above.

The proofs of Theorem B.1 and Theorem B.2 proceed in several steps.

- First, we show that the optimization problem with adversarial adherence decisions in B_∞:

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) \tag{B.1}$$

admits a robust Bellman equation in Proposition B.1.

- In Corollary B.1, we then show that this robust Bellman equation can be interpreted as the Bellman equation of an alternate MDP, which shows the equivalence between (B.1) and AdaMDP as in

$$\max_{\pi_{\mathsf{alg}} \in \Pi} R\left(\theta \pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}\right), \tag{B.2}$$

hence concluding the proof of Theorem B.1.

- The proof for Theorem B.2 follows from $B_1 \subset B_\infty, B_2 \subset B_\infty, B_3 \subset B_\infty$, and from the worst-case $u^\infty \in B_\infty$ for the **Unconstrained Adversarial** model being $u^\infty_{s,t} = \theta, \forall\, (s,t) \in \mathcal{S} \times \mathbb{N}$, which is feasible in B_1, B_2 and B_3.

B.1. Proof of Theorem B.1 (Unconstrained Adversarial)

For the sake of conciseness, for a given stationary policy π and $\boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}$, we define $T^\pi(\boldsymbol{v}) \in \mathbb{R}^{\mathcal{S}}$ as $T^\pi_s(\boldsymbol{v}) = \sum_{a \in \mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}\right), \forall\, s \in \mathcal{S}$. Note that for each $s \in \mathcal{S}$, the scalar $T^\pi_s(\boldsymbol{v})$ only depends on $\boldsymbol{\pi}_s \in \Delta(\mathcal{A})$ and not on $\boldsymbol{\pi}_{s'}$ for $s' \neq s$. The next proposition shows that (B.1) admis a robust Bellman equation.

Proposition B.1. *Let $\boldsymbol{v}^\infty \in \mathbb{R}^{\mathcal{S}}$ satisfying*

$$v^\infty_s = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \min_{u \in [\theta, 1]} u \cdot T^\pi_s(\boldsymbol{v}^\infty) + (1-u) \cdot T^{\pi_{\mathsf{base}}}_s(\boldsymbol{v}^\infty), \forall\, s \in \mathcal{S}. \tag{B.3}$$

Additionally, let π^∞ be a stationary policy attaining the maximum in the right-hand side in (B.3) *for each $s \in \mathcal{S}$. Then π^∞ can be chosen deterministic, and π^∞ is an optimal solution to* (B.1).

Proof. We first note that the vector $\boldsymbol{v}^\infty$ is well defined and is unique because the following map $f : \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ is a contraction for the ℓ_∞-norm:

$$f : \boldsymbol{v} \mapsto \left(\max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \min_{u \in [\theta, 1]} u \cdot T^\pi_s(\boldsymbol{v}) + (1-u) \cdot T^{\pi_{\mathsf{base}}}_s(\boldsymbol{v}) \right).$$

Since f is a contraction, there exists a unique vector $\boldsymbol{v}^\infty \in \mathbb{R}^{\mathcal{S}}$ such that $\boldsymbol{v}^\infty$ is a fixed-point of f, i.e., such that $f(\boldsymbol{v}^\infty) = \boldsymbol{v}^\infty$. Let us define π^∞ as the policy attaining the arg max in (B.3) and $u^\star_s \in [\theta, 1]$ attaining its worst-case on each state $s \in \mathcal{S}$. We define $u^\infty \in [\theta, 1]^{\mathcal{S} \times \mathbb{N}}$ with $u^\infty_{s,t} = u^\star_s, \forall\, (s,t) \in \mathcal{S} \times \mathbb{N}$. We will show that (π^∞, u^∞) is an optimal solution to $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$. To show this, we will show that π^∞ is an ϵ-optimal policy to $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$, for any $\epsilon > 0$.

- Let $\epsilon > 0$. Recall that the infinite-horizon return of a policy π is defined as $R(\pi) = \mathbb{E}^\pi \left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t s_{t+1}} \right]$. Let $T \in \mathbb{N}$. For any policy π, we define $R_T(\pi)$, *the truncated return with terminal reward* $\boldsymbol{v}^\infty$, as

$$R_T(\pi) = \mathbb{E}^\pi \left[\sum_{t=0}^{T-1} \lambda^t r_{s_t a_t s_{t+1}} + \lambda^T v^\infty_{s_T} \right]. \tag{B.4}$$

Since $\mathcal{S}, \mathcal{A}$ are finite sets, the rewards $r_{s,a,s'}$ are bounded. Therefore, for any $\epsilon > 0$, there exists a corresponding T such that $|R_T(\pi) - R(\pi)| \leq \epsilon$ for any policy $\pi \in \Pi$. For instance we can take T such that $\lambda^T \left(\frac{\max_{s,a} |r_{s,a}|}{1-\lambda} + \|\boldsymbol{v}^\infty\|_\infty \right) < \epsilon$.

- We can define the *worst-case truncated return with terminal reward* $\boldsymbol{v}^\infty$ as

$$\min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi, u)) = \min_{u \in B_\infty} \mathbb{E}^{\pi_{\mathsf{eff}}(\pi,u)} \left[\sum_{t=0}^{T-1} \lambda^t r_{s_t a_t s_{t+1}} + \lambda^T v^\infty_{s_T} \right]. \tag{B.5}$$

Note that the worst-case return and the worst-case truncated return of π^∞ coincide:

$$\min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi^\infty, u)) = \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty, u)).$$

This is by definition of π^∞ and $\boldsymbol{v}^\infty$ as the fixed-point of f, i.e., as the continuation values of π^∞.

- We claim that for any value of $T \in \mathbb{N}$, the decisions $\pi_{\mathsf{eff}}(\pi^\infty, u^\infty), ..., \pi_{\mathsf{eff}}(\pi^\infty, u^\infty)$ (repeated T times) are optimal for (B.5). Indeed, the terminal rewards are given by $(v^\infty_s)_{s \in \mathcal{S}}$, and f is the Bellman operator that relates the worst-case values at period $t \in \{1, ..., T\}$ to the worst-case values at period $t-1$. Since $\boldsymbol{v}^\infty = f(\boldsymbol{v}^\infty)$ and π^∞_s, u^∞ is a solution to $f_s(\boldsymbol{v}^\infty)$ as a saddle-point program for each $s \in \mathcal{S}$, we conclude that repeating $\pi_{\mathsf{eff}}(\pi^\infty, u^\infty)$ T-times optimizes the worst-case truncated return.

- Overall, we have shown the following inequalities. First, we have shown that the worst-case return and the worst-case truncated return of π^∞ coincides: $\min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi^\infty, u)) = \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty, u))$ and we have shown that π^∞ is optimal for the worst-case truncated return: $\min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi^\infty, u)) \geq \min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi, u)), \forall\, \pi \in \Pi_{\mathsf{H}}$. Let $\pi^\star$ an optimal policy for the worst-case adherence model (B.1). Then we have

$$\min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty, u)) = \min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi^\infty, u)) \geq \min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi^\star, u)) \geq \min_{u \in B_\infty} R_T(\pi_{\mathsf{eff}}(\pi^\star, u)) - \epsilon.$$

where the last inequality follows from the worst-case truncated return approximating the worst-case return up to ϵ. This shows that for any $\epsilon > 0$, we have $\min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty, u)) \geq \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\star, u)) - \epsilon$, from which we conclude that $\min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty, u)) \geq \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\star, u))$. Since we have chosen $\pi^\star$ as an optimal policy for (B.1), we can conclude that π^∞ is an optimal policy for (B.1). This concludes our proof of Proposition B.1. □

We now have the following corollary, which shows the equivalence (at optimality) between the model with worst-case time-varying adherence and our model of adherence-aware MDP (B.2).

COROLLARY B.1. *For $\boldsymbol{v}^\infty$ defined as in* (B.3)*, we have*

$$v_s^\infty = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \theta \cdot T_s^{\pi}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty), \forall\, s \in \mathcal{S}, \tag{B.6}$$

i.e., the worst-case deviation at optimality in (B.1) *is attained at $u_{s,t}^\infty = \theta$ for all $(s,t) \in \mathcal{S} \times \mathbb{N}$. In particular, we have the following equality:*

$$\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u^\infty)) = \min_{u \in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty, u)). \tag{B.7}$$

Proof. Because the inner objective function is linear in u, we have

$$\begin{aligned}\max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \min_{u \in [\theta,1]} u \cdot T_s^{\pi}(\boldsymbol{v}^\infty) + (1-u) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) &= \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \min_{u \in \{\theta,1\}} u \cdot T_s^{\pi}(\boldsymbol{v}^\infty) + (1-u) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) \\ &= \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \min\{\theta \cdot T_s^{\pi}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty), T_s^{\pi}(\boldsymbol{v}^\infty)\}.\end{aligned}$$

Therefore, we want to prove that the minimum $\min\{\theta \cdot T_s^{\pi^\infty}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty), T_s^{\pi}(\boldsymbol{v}^\infty)\}$ is always attained at $\theta \cdot T_s^{\pi^\infty}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty)$, i.e., we want to show that $\theta \cdot T_s^{\pi^\infty}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) \leq T_s^{\pi^\infty}(\boldsymbol{v}^\infty)$. Note that by choosing $\pi = \pi_{\mathsf{base}}$ in the max-min program (4.1), we always have

$$v_s^\infty \geq T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty), \forall\, s \in \mathcal{S}. \tag{B.8}$$

Now if for some $s \in \mathcal{S}$ we have $\theta \cdot T_s^{\pi^\infty}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) > T_s^{\pi^\infty}(\boldsymbol{v}^\infty)$, then $T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) > T_s^{\pi^\infty}(\boldsymbol{v}^\infty)$. But since $T_s^{\pi^\infty}(\boldsymbol{v}^\infty) = v_s^\infty$, we would obtain $T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) > v_s^\infty$, which is a contradiction with (B.8). Therefore, we always have $\theta \cdot T_s^{\pi^\infty}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty) \leq T_s^{\pi^\infty}(\boldsymbol{v}^\infty)$, which shows that $v_s^\infty = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \theta \cdot T_s^{\pi}(\boldsymbol{v}^\infty) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^\infty), \forall\, s \in \mathcal{S}$. Therefore, the worst-case adherence decisions u^∞ for π^∞ can be chosen as $u_{s,t}^\infty = \theta$ for any pair $(s,t) \in \mathcal{S} \times \mathbb{N}$, which concludes the proof of Corollary B.1. This also shows that we can choose an optimal policy for (B.1) as a stationary deterministic policy, because π^∞ attains the right-hand side in (B.6), which maximizes a linear form over the simplex $\Delta(\mathcal{A})$. □

We can now interpret the equation (B.6) as the Bellman equation of a decision maker which chooses π_{alg} and where the effective policy is $\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}$. This is because for any $\boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}, s \in \mathcal{S}$ and $\pi_{\mathsf{alg}} \in \Pi$, we have

$$\theta \cdot T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}) + (1-\theta) \cdot T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}) = T_s^{\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}}(\boldsymbol{v})$$

which shows that (B.1) is equal to (B.2) and that the sets of optimal policies of (B.1) and (B.2) share a common stationary deterministic policy π^∞, attaining the arg max in (B.6).

B.2. Proof of Theorem B.2 (other adversarial models)

We now study the other models of adversarial adherence decisions presented in Theorem 3.2. We provide the proof of Theorem B.2 below.

Proof of Theorem B.2. Let $B \subset [\theta, 1]^{\mathcal{S}\times\mathbb{N}}$ be either B_1 (**Time-invariant Adversarial**), B_2 (**State-invariant Adversarial**), or B_3 (**Time- and State-invariant Adversarial**).

Let $u^\infty \in [\theta, 1]^{\mathcal{S}\times\mathbb{N}}$ be defined as $u^\infty_{s,t} = \theta, \forall\,(s,t) \in \mathcal{S}\times\mathbb{N}$, and π^∞ be defined as in Proposition B.1. We have

$$\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) \leq \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u^\infty)) = \min_{u\in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty_{\mathsf{alg}}, u)) \leq \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi^\infty_{\mathsf{alg}}, u))$$

where the first inequality comes from $u^\infty \in B$, the equality comes from (B.7), and the second inequality comes from $B \subset B_\infty$. This shows that π^∞ is an optimal policy in $\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$ and the two saddle-point formulations are equal: $\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) = \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B_\infty} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$. We have proved in Theorem B.1 that the right-hand side in the previous equation is equal to $\max_{\pi_{\mathsf{alg}}\in\Pi} R(\theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}})$, which concludes the proof of Theorem B.2. □

B.3. Proof of strong duality for the adversarial models

We now turn to proving that strong duality holds for all the adversarial models considered in Theorem 3.2. In particular, we show the following theorem.

THEOREM B.3. *Let $B \subset [\theta, 1]^{\mathcal{S}\times\mathbb{N}}$ be either B_∞ (**Unconstrained Aversarial**), B_1 (**Time-invariant Adversarial**), B_2 (**State-invariant Adversarial**), or B_3 (**Time- and State-invariant Adversarial**). Then*

$$\sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) = \min_{u\in B} \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)).$$

Proof. Let $B \in \{B_\infty, B_1, B_2, B_3\}$. Since weak duality always holds, we only have to prove that

$$\min_{u\in B} \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) \leq \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)).$$

We have

$$\min_{u\in B} \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) \leq \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u^\infty)) = \min_{u\in B_\infty} R(\pi_{\mathsf{eff}}(\pi^\infty_{\mathsf{alg}}, u)) \leq \sup_{\pi_{\mathsf{alg}}\in\Pi_{\mathsf{H}}} \min_{u\in B} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$$

where the first inequality comes from $u^\infty \in B$, the equality comes from (B.7), and the second inequality comes from $B \subset B_\infty$ and from maximizing over Π_{H}. Therefore strong duality holds. □

Appendix C: Proof of Theorem 3.3

We will show that the constrained assortment optimization with a Markov chain-based choice model (Blanchet et al. 2016, Désir et al. 2020) can be reduced to Constrained-AdaMDP. We first introduce the Markov chain-based choice model below. We follow the lines of Désir et al. (2020) here.

Markov chain model. Let $n \in \mathbb{N}$. The set $\mathcal{N} = \{1, ..., n\}$ represents n items. The no-purchase option is represented by 0 and we write $\mathcal{N}_+ = \mathcal{N} \cup \{0\}$. There are scalars $\nu_i \geq 0$ which represents the initial arrival probabilities for every state $i \in \mathcal{N}_+$ and some transition probabilities $\rho_{ij} \in [0,1]$ for all $(i,j) \in \mathcal{N}_+^2$. The return for each item i is written as $\xi_i \geq 0$. The goal of the decision maker is to choose a subset of items $\mathcal{I} \subseteq \mathcal{N}$ to display to the customers, in order to maximize its expected return $R_{\mathsf{MC}}(\mathcal{I})$, computed as follows:

- For any state i in the chosen set of items $\mathcal{I}$, the state i is absorbing.
- A customer arrives in state $i \in \mathcal{N}_+$ with an initial probability ν_i. If the state i is non-absorbing, the customer transitions to a different state $j \in \mathcal{N}_+, j \neq i$ with probability ρ_{ij}.
- The process continues until an absorbing state is reached (either in $\mathcal{I}$ or in $\{0\}$).
- Let $\gamma(i, \mathcal{I})$ be the probability that item $i \in \mathcal{I}$ is chosen by the customer when the assortment $\mathcal{I} \subset \mathcal{N}$ is offered. Note that $\gamma(i, \mathcal{I})$ is equal to the probability that the customer reaches state i before any other absorbing states. Then the return $R_{\mathsf{MC}}(\mathcal{I})$ associated with a chosen subset $\mathcal{I}$ is

$$R_{\mathsf{MC}}(\mathcal{I}) = \sum_{i \in \mathcal{I}} \gamma(i, \mathcal{I}) \xi_i. \tag{C.1}$$

The *cardinality assortment* (Card-Assort) problem is the following optimization problem, which solves for an optimal assortment $\mathcal{I}$ with constraints on the number of items selected for display:

$$\max \{R_{\mathsf{MC}}(\mathcal{I}) \mid \mathcal{I} \subset \mathcal{N}, |\mathcal{I}| \leq k\}. \tag{Card-Assort}$$

The authors in Désir et al. (2020) prove the following hardness result for Card-Assort, even under some conditions on the parameters ρ, ξ and ν.

THEOREM C.1 **(Reformulation of Theorem 5, Désir et al. (2020))**. Card-Assort *is APX-hard, even when* $\rho_{i0} = 1/4$ *and* $\xi_i = 1, \nu_i = 1/n$ *for all* $i \in \mathcal{N}$.

Our proof of Theorem 3.3 follows from Theorem C.1. On the one hand, we can interpret Constrained-AdaMDP as an optimization problem where there are two Markov chains, one induced by π_{alg} and one by π_{base}, and where the variable u_s decides to follow the Markov chain induced by π_{alg} or the Markov chain

induced by π_{base} at each state $s \in \mathcal{S}$. On the other hand, Card-Assort can be interpreted as follows: given a Markov chain following a transition matrix ρ, the decision maker chooses a subset $\mathcal{I}$ (with $|\mathcal{I}| \leq k$) for which the states in $\mathcal{I}$ become absorbing. Based on this interpretation of Constrained-AdaMDP and Card-Assort, we can reformulate any instance of Card-Assort as an instance of Constrained-AdaMDP in a straightforward manner. The only technical difficulty is that $R_{\mathsf{MC}}(\mathcal{I})$ a priori does not involve a discount factor, whereas we have defined the objective function in Constrained-AdaMDP based on the discounted return (3.1), which depends on a discount factor λ. We show how to circumvent this issue in our proof below. In particular, we prove Theorem 3.3 in two steps.

Step 1: reformulating the objective in Card-Assort. We consider the instance of Card-Assort from the proof of Theorem 5 in Désir et al. (2020). In this instance, $\rho_{i0} = 1/4$ in all state $i \in \mathcal{N}$, and for each item $i \in \mathcal{N}$ there is a subset $\mathcal{N}_i \subset \mathcal{N}\{i\}$ such that $|\mathcal{N}_i| = 3$ and $\rho_{ij} = 1/4, \forall\, j \in \mathcal{N}_i$. Additionally, the return ξ_i is equal to 1 for each item $i \in \mathcal{N}$. We first note that $R_{\mathsf{MC}}(\mathcal{I})$, defined as $R_{\mathsf{MC}}(\mathcal{I}) = \sum_{i \in \mathcal{I}} \gamma(i, \mathcal{I}) \xi_i$, is equal to $\mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \tilde{r}_{s_t a_t s_{t+1}}\right]$ for a certain MDP instance $\tilde{\mathcal{M}}$ and a certain policy π that represents the subset $\mathcal{I}$ of chosen items. In particular, let us consider the following MDP instance $\tilde{\mathcal{M}}$ with states $\tilde{\mathcal{S}} = \mathcal{N}_+$, actions $\{a_0, a_1\}$ which represent choosing or not an item to display, and where the instantaneous rewards $\tilde{\boldsymbol{r}}$ and the transition probabilities $\tilde{\boldsymbol{P}}$ are defined as follows:

- All the instantaneous rewards are equal to 0, except $r_{sa_1 s} = 1/4, \forall\, s \in \mathcal{N}$.
- $\tilde{P}_{sa_0 s'} = 1/4$ if $s' \in \mathcal{N}_s \cup \{0\}, \forall\, s \in \mathcal{N}$,
- $\tilde{P}_{sa_1 s} = 3/4, \tilde{P}_{sa_1 0} = 1/4, \forall\, s \in \mathcal{N}$,
- $\tilde{P}_{0a_0 0} = \tilde{P}_{0a_1 0} = 1$.

A stationary deterministic policy π is a map $\mathcal{N}_+ \to \{a_0, a_1\}$, and we can construct a policy π representing an assortment $\mathcal{I}$ as follows: π chooses action a_1 in state s if and only if $s \in \mathcal{I}$; otherwise, π chooses action a_0. Finally, the initial probability distribution is just $p_{0,i} = \nu_i = 1/n, \forall\, i \in \mathcal{N}$. Let us give some intuition on the values of $\tilde{\boldsymbol{r}}$ and $\tilde{\boldsymbol{P}}$ given above. In a state where action a_0 is chosen (which corresponds to a state not included the subset $\mathcal{I}$), the Markov chain induced by $\tilde{\boldsymbol{P}}$ evolves exactly as for the kernel ρ defined above and no instantaneous reward is obtained. When action a_1 is chosen at a state s, i.e. when s is chosen to be included in $\mathcal{I}$, the decision maker transitions to the absorbing state 0 after a number of period that follows a geometric distribution with parameter $\lambda = 3/4$, earning an instantaneous reward of $1 - \lambda = 1/4$ while remaining in state s and then an instantaneous reward of 0 while in state 0. Overall, we have shown that Card-Assort can

be reformulated as optimizing an undiscounted return, in contrast to the discounted returns considered in this paper. However, the Markov chain $\tilde{\mathcal{M}}$ has a very particular structure: from any state $s \in \mathcal{N}$, there is a probability $1/4$ to reach the absorbing state 0, and no reward is obtained when transitioning to state 0. We now show in the next step that this can be interpreted as computing a discounted return with a discount factor of $\lambda = 1 - 1/4 = 3/4$.

Step 2: from undiscounted objective to discounted objective. In this section we show that we can reformulate the undiscounted objective function $\mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \tilde{r}_{s_t a_t s_{t+1}}\right]$ as a discounted objective $\mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t s_{t+1}}\right]$ for a certain discount factor λ and a certain Markov chain $\mathcal{M}$. We follow the lines of Section 5.3 in Puterman (2014), which shows that the discount factor $\lambda \in [0,1)$ can be interpreted as a termination probability. This idea dates back to Derman (1970). More precisely, we have the following proposition, which is a reformulation of the results in Section 5.3 in Puterman (2014).

PROPOSITION C.1. *Let $\mathcal{M}$ be any MDP instance and π be a policy. Then the discounted return $R(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t}\right]$ is equal to the following undiscounted return $\tilde{R}(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \tilde{r}_{s_t a_t s_{t+1}}\right]$, where $\tilde{\boldsymbol{r}} \in \mathbb{R}^{\mathcal{S}_+ \times \mathcal{A}}, \tilde{\boldsymbol{P}} \in \mathbb{R}^{\mathcal{S}_+ \times \mathcal{A} \times \mathcal{S}_+}$ are the instantaneous rewards and the transition probabilities for an MDP instance $\tilde{\mathcal{M}}$ defined over an augmented state space $\mathcal{S}_+ = \mathcal{S} \cup \{\Delta\}$, defined as follows:*

- $\tilde{r}_{sas'} = r_{sas'}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \forall\, s' \neq \Delta$,
- $\tilde{r}_{sa\Delta} = 0, \forall (s,a) \in \mathcal{S}_+ \times \mathcal{A}$
- $\tilde{P}_{sas'} = \lambda P_{sas'}, \forall (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$,
- $\tilde{P}_{sa\Delta} = 1 - \lambda, \forall\, s \in \mathcal{S}$,
- $\tilde{P}_{\Delta a \Delta} = 1$.

Applying Proposition C.1 to the MDP instance $\tilde{\mathcal{M}}$ defined in the first step of this proof, we find that the objective function $\mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \tilde{r}_{s_t a_t s_{t+1}}\right]$ can be reformulated as the discounted return in the following MDP instance $\mathcal{M}$: the set of states is $\mathcal{N}$, the set of actions is $\{a_0, a_1\}$, the discount factor is $\lambda = 3/4$, the initial probability distribution $\boldsymbol{p}_0$ for the MDP instance $\tilde{\mathcal{M}}$ is equal to the probability distribution $\boldsymbol{\nu}$ for the instance of Card-Assort, and the instantaneous rewards $\boldsymbol{r}$ and the transition probabilities $\boldsymbol{P}$ are defined as follows:

$$r_{sa_0s'} = 0, r_{sa_1s} = 1/4, \forall\, (s,s') \in \mathcal{N} \times \mathcal{N}, P_{sa_0s'} = 1/3, P_{sa_1s} = 1, \forall s' \in \mathcal{N}_s, \forall\, s \in \mathcal{N}. \tag{C.2}$$

Overall, we have shown the following proposition.

PROPOSITION C.2. *Let us consider the instance of* Card-Assort *from the proof of Theorem 5 in Désir et al. (2020). Then* $R_{\mathsf{MC}}(\mathcal{I})$ *can be reformulated as* $R(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{+\infty} \lambda^t r_{s_t a_t}\right]$ *for the MDP instance described in* (C.2) *with* $\pi_s = a_1$ *if and only if* $s \in \mathcal{I}$ *for* $s \in \mathcal{N}$ *and* $\lambda = 3/4$.

Let $\mathcal{I} \subset \mathcal{N}$ a subset of displayed items and let π be the policy representing $\mathcal{I}$ in the MDP $\mathcal{M}$. Then $\pi_s = \pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u) = u_s \pi_{\mathsf{alg}\,s} + (1-u_s)\pi_{\mathsf{base}\,s}$, with π_{base} the policy that chooses a_0 in all states, π_{alg} the policy that chooses a_1 in all states, and $u \in \{0,1\}^{\mathcal{S}}$. The cardinality constraint $|\mathcal{I}| \leq k$ can be directly rewritten $\sum_{s\in\mathcal{S}} u_s \leq k$, which concludes our proof.

Appendix D: Mixed-integer optimization formulation for Constrained-AdaMDP

In this section we provide a mixed-integer optimization formulation for Constrained-AdaMDP. We start with the following lemma, which is a direct consequence of a classical contraction lemma, see for instance Lemma 2 in **?** or Lemma 3.1 in **?**.

LEMMA D.1. *Let* $\pi \in \Pi$. *Then* $R(\pi) = \min\{\boldsymbol{p}_0^\top \boldsymbol{v} \mid v_s \geq \sum_{a\in\mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}), \forall\, s \in \mathcal{S}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}\}$.

Based on Lemma D.1, for a fixed $\pi_{\mathsf{alg}} \in \Pi$ we can reformulate Constrained-AdaMDP as follows:

$$\min\left\{\boldsymbol{p}_0^\top \boldsymbol{v} \mid v_s \geq \sum_{a\in\mathcal{A}} u_s \pi_{\mathsf{alg},sa} \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}) + (1-u_s)\pi_{\mathsf{base},sa} \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}), \forall\, s \in \mathcal{S}, \sum_{s\in\mathcal{S}} u_s \leq k, \boldsymbol{u} \in \{0,1\}^{\mathcal{S}}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}\right\}. \tag{D.1}$$

In the optimization program above, the terms $u_s \times \left(\sum_{a\in\mathcal{A}} (\pi_{\mathsf{alg},sa} - \pi_{\mathsf{base},sa}) \boldsymbol{P}_{sa}^\top \boldsymbol{v}\right)$ are bilinear in the variables $(u_s, \boldsymbol{v}) \in \{0,1\} \times \mathbb{R}^{\mathcal{S}}$ for any $s \in \mathcal{S}$. However, $u_s \in \{0,1\}, v_s \in [0, r_\infty/(1-\lambda)], \sum_{s'\in\mathcal{S}} P_{sas'} = 1$ for any $s, a \in \mathcal{S}$, so we can use classical reformulation techniques to linearize the bilinear constraints. In particular, it is well-known that we can linearize the term $x \times y$ with $x \in \{0,1\}, y \in [L,U]$ with $L \leq U$ by introducing an auxiliary variable $z \in \mathbb{R}$ such that $z \geq Lx, z \leq Ux, z \geq y - (1-x)\max\{|L|,|U|\}, z \leq y + (1-x)\max\{|L|,|U|\}$. Applying this method with $x = u_s \in \{0,1\}$ and $y = \sum_{a\in\mathcal{A}} (\pi_{\mathsf{alg},sa} - \pi_{\mathsf{base},sa}) \boldsymbol{P}_{sa}^\top \boldsymbol{v} \in [-\frac{r_\infty}{1-\lambda}, 2\frac{r_\infty}{1-\lambda}]$ to linearize the bilinear terms appearing in the reformulation (D.1), we obtain a mixed-integer program for Constrained-AdaMDP, only involving a linear objective and linear constraints over continuous and binary variables:

$$\begin{aligned}
&\min \boldsymbol{p}_0^\top \boldsymbol{v} \\
&\quad v_s \geq \lambda z_s + \sum_{a\in\mathcal{A}} \left(u_s \pi_{\mathsf{alg},sa} + (1-u_s)\pi_{\mathsf{base},sa}\right) \boldsymbol{P}_{sa}^\top \boldsymbol{r}_{sa} + \lambda \pi_{\mathsf{base},sa} \boldsymbol{P}_{sa}^\top \boldsymbol{v}, \forall\, s \in \mathcal{S}, \\
&\quad -2\frac{r_\infty}{1-\lambda}(1-u_s) \leq z_s - \sum_{a\in\mathcal{A}} (\pi_{\mathsf{alg},sa} - \pi_{\mathsf{base},sa}) \boldsymbol{P}_{sa}^\top \boldsymbol{v} \leq 2\frac{r_\infty}{1-\lambda}(1-u_s), \forall\, s \in \mathcal{S},
\end{aligned}$$

$$-\frac{r_\infty}{1-\lambda}u_s \le z_s \le 2\frac{r_\infty}{1-\lambda}u_s, \forall\, s \in \mathcal{S},$$

$$\sum_{s\in\mathcal{S}} u_s \le k,$$

$$\boldsymbol{u} \in \{0,1\}^{\mathcal{S}}, \boldsymbol{v} \in \mathbb{R}^{\mathcal{S}}, \boldsymbol{z} \in \mathbb{R}^{\mathcal{S}}.$$

Appendix E: Detailed computation for Section 3.4

We detail the computation of the simple MDP instance we presented in Section 3.4.

Considering the policies π_{base} and π_{alg} represented in Figures 1b and 1c respectively, we compute the return of the effective policy $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta) = \theta\pi_{\mathsf{alg}} + (1-\theta)\pi_{\mathsf{base}}$. For concision, let us denote $r := r_2 = 0.1$. Then, for any ϵ, we have

$$\begin{aligned}
R\left(\pi_{\mathsf{base}}\right) &= \frac{\lambda^2}{1-\lambda},\\
R\left(\pi_{\mathsf{alg}}\right) &= r\lambda + \frac{\lambda^2}{1-\lambda},\\
R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) &= \theta\cdot\left(r\lambda + \theta\cdot\frac{\lambda^2}{1-\lambda} + (1-\theta)\cdot(1+\epsilon)\frac{\lambda^2}{1-\lambda}\right) + (1-\theta)\cdot\left(0 + \theta\cdot(1+\epsilon)\frac{\lambda^2}{1-\lambda} + (1-\theta)\cdot\frac{\lambda^2}{1-\lambda}\right)\\
&= \theta\lambda r + \theta R(\pi_{\mathsf{base}}) + \theta(1-\theta)\epsilon R(\pi_{\mathsf{base}}) + (1-\theta)R(\pi_{\mathsf{base}}) + \theta(1-\theta)\epsilon R(\pi_{\mathsf{base}})\\
&= [\lambda r + R(\pi_{\mathsf{base}})] + (\theta-1)\lambda r + 2\theta(1-\theta)\epsilon R(\pi_{\mathsf{base}})\\
&= R(\pi_{\mathsf{alg}}) + [R(\pi_{\mathsf{alg}}) - R(\pi_{\mathsf{base}})](\theta-1) + 2\epsilon R(\pi_{\mathsf{base}})\theta(1-\theta).
\end{aligned}$$

In other words, $R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right)$ is a second-order polynomial in θ that equals $R(\pi_{\mathsf{alg}})$ (resp. $R(\pi_{\mathsf{base}})$) for $\theta = 1$ (resp. $\theta = 0$). Since we would like to compare $R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right)$ with both $R(\pi_{\mathsf{alg}})$ and $R(\pi_{\mathsf{base}})$, we provide two convenient reformulations:

$$\begin{aligned}
R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) &= R(\pi_{\mathsf{base}}) + \theta\left([R(\pi_{\mathsf{alg}}) - R(\pi_{\mathsf{base}})] + 2\epsilon R(\pi_{\mathsf{base}})(1-\theta)\right)\\
&= R(\pi_{\mathsf{alg}}) + (1-\theta)\left(2\epsilon R(\pi_{\mathsf{base}})\theta - [R(\pi_{\mathsf{alg}}) - R(\pi_{\mathsf{base}})]\right),
\end{aligned}$$

and define $\tilde{\theta} - 1 := \dfrac{R(\pi_{\mathsf{alg}}) - R(\pi_{\mathsf{base}})}{2\epsilon R(\pi_{\mathsf{base}})} = \dfrac{\lambda r}{2\epsilon R(\pi_{\mathsf{base}})}$.

Case 1: partial adherence hurts ($\epsilon = -1$). When $\epsilon = -1 < 0$, the reward of State 5 is strictly less than 1 so π_{alg} is optimal and any deviation from π_{alg} can only deteriorate performance. In this case, we have

$$R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) = R(\pi_{\mathsf{base}}) + 2\epsilon R(\pi_{\mathsf{base}})\theta\left(\tilde{\theta} - 1 + (1-\theta)\right) = R(\pi_{\mathsf{base}}) + 2R(\pi_{\mathsf{base}})\theta\left(\theta - \tilde{\theta}\right),$$

as announced in Section 3.4.

Furthermore, for any value of $\theta \in [0,1]$, we can find an optimal recommendation policy $\pi^\star_{\mathsf{alg}}(\theta)$ via backward induction. Let us write v_i for the value derived by the DM starting from state $i \in \{1,2,3,4,5\}$. Clearly, $v_5 = 0, v_4 = 1/(1-\lambda)$. In State 3, $v_5 < v_4$ so the best action is to choose to go to State 4. Since π_{base} also chooses to go to State 4, we obtain $v_3 = \lambda/(1-\lambda)$. In State 2, the best action is to choose to State 4. Since we follow the recommendation policy with probability θ and the baseline policy with probability $1-\theta$, we have $v_2 = 0.1 + \theta\lambda \cdot v_4 + (1-\theta)\lambda \cdot v_5 = 0.1 + \theta\frac{\lambda}{1-\lambda}$. Finally, for the best action in State 1, we have to choose between going to State 2 and going to State 3. If we recommend going to state $i \in \{2,3\}$, then the value derived by the DM from state 1 will be $v_1 = \theta\lambda \cdot v_i + (1-\theta)\lambda v_3$. Hence, the optimal recommendation depends on the comparison between v_2 and v_3. If $v_2 > v_3$, we should recommend to go to State 2 from State 1 and $v_1 = \theta\lambda \cdot v_2 + (1-\theta)\lambda v_3$. Otherwise, we recommend State 3 and $v_1 = v_3$. Observe that $v_2 \geq v_3 \iff 0.1 + \theta\frac{\lambda}{1-\lambda} \geq \frac{\lambda}{1-\lambda} \iff \theta \geq 1 - 0.1 \cdot \frac{1-\lambda}{\lambda} =: \bar{\theta}$. All in all, we have the following two cases,

- If $\theta \leq \bar{\theta}$, $v_2 \leq v_3$ and an optimal recommendation policy $\pi^\star_{\mathsf{alg}}(\theta)$ should recommend the following transitions: $\pi^\star_{\mathsf{alg}}(\theta)$ is $1 \to 3$, $3 \to 4$. Observe that, as long as $\pi^\star_{\mathsf{alg}}(\theta)$ recommends $1 \to 3$, $\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta), \theta)$ will never visit State 2. As a result, π_{base} is also an optimal recommendation policy in this case, despite the fact that it prescribes a sub-optimal action at State 2.
- If $\theta \geq \bar{\theta}$, $v_2 \geq v_3$ and an optimal recommendation policy $\pi^\star_{\mathsf{alg}}(\theta)$ should recommend the following transitions: $1 \to 2$, $2 \to 4$, $3 \to 4$. Unlike in the previous case, $\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta), \theta)$ will visit State 3 even if $\pi^\star_{\mathsf{alg}}(\theta)$ does not recommend $1 \to 3$. Consequently, an optimal recommendation recommends State 4 when at State 3.

Case 2: partial adherence helps ($\epsilon = 1$). When $\epsilon = 1$, the human DM is currently taking the optimal decision when visiting State 2 (but never visits State 2 in the first place), while the algorithm plays optimally in State 3 but never visits it. In this case, a mixture of π_{alg} and π_{base} can lead to greater performance than any of the two policy alone. Formally, in this case,

$$R\left(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)\right) = R(\pi_{\mathsf{alg}}) + 2R(\pi_{\mathsf{base}})(1-\theta)\left(\theta - [\tilde{\theta} - 1])\right),$$

which is increasing over the interval $[\tilde{\theta} - 1, 1]$.

Appendix F: Proof of Proposition 3.1

Proof of Proposition 3.1. Our proof is very similar to the proof of Theorem B.1 and we only give a sketch here. The gist of the proof is to show that AdaMDP admits a Bellman equation, i.e., to show that the stationary deterministic policy π^∞ as defined in Equation (4.1) is an optimal recommendation policy.

We first note that the vector $\boldsymbol{v}^\infty \in \mathbb{R}^{\mathcal{S}}$, defined as the unique solution to Equation (4.1), does exist since the following map is a contraction for $\|\cdot\|_\infty$:

$$\boldsymbol{v} \mapsto \left(\max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \theta \cdot \sum_{a \in \mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v} \right) + (1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},sa} \boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v} \right) \right)_{s \in \mathcal{S}}.$$

This is a straightforward consequence of $\boldsymbol{\pi}_s \in \Delta(\mathcal{A}), \boldsymbol{\pi}_{\mathsf{base},s} \in \Delta(\mathcal{A})$ and $\boldsymbol{P}_{sa} \in \Delta(\mathcal{S})$ for each pair $(s,a) \in \mathcal{S} \times \mathcal{A}$.

We can then show that π^∞ is ϵ-optimal in AdaMDP, for any value of $\epsilon > 0$. This is done in the same way as for our proof of Theorem B.1, where we build a time T such that the instantaneous rewards obtained after more than $T+1$ periods only accounts for at most ϵ in the (untruncated) return $R(\cdot)$. We then use a truncated return $R_T(\cdot)$ with continuation value $\boldsymbol{v}^\infty \in \mathbb{R}^{\mathcal{S}}$ for $\boldsymbol{v}^\infty$ as defined in Equation (4.1). The policy π^∞ is an optimal recommendation policy for $R_T(\cdot)$, so that it is ϵ-optimal for the untruncated return $R(\cdot)$. Since this holds for all $\epsilon > 0$, we conclude that π^∞ as defined in Proposition 4.2 is an optimal recommendation policy. Since π^∞ is stationary and deterministic, this concludes the proof of Proposition 3.1. □

Appendix G: Proof of Proposition 4.4

Proof. 1. The proof uses similar ideas as the proof of *Blackwell optimality* for classical MDPs (Feinberg and Shwartz 2012) and robust MDPs (**?**), which studies the sensitivity of optimal policies as regards the values of the discount factor $\lambda \in [0,1)$. In particular, we recall the following lemma from Puterman (2014).

LEMMA G.1 **(Lemma 10.1.2, Puterman (2014))**. *Let $\phi : \mathcal{I} \to \mathbb{R}$ be a rational function on an interval $\mathcal{I} \subset \mathbb{R}$, that is, ϕ is the ratio of two polynomials and the denominator does not have any zeros in the interval $\mathcal{I}$. Then either $\phi(\theta) = 0$ for all $\theta \in \mathcal{I}$, either $\phi(\theta) = 0$ for finitely many values of $\theta \in \mathcal{I}$.*

We now proceed with the proof of this statement. From Proposition 3.1, for any value of $\theta \in [0,1]$, an optimal recommendation policy $\pi^\star_{\mathsf{alg}}(\theta)$ can be chosen stationary and deterministic. Since there are finitely many deterministic policies, the map $\theta \mapsto \pi^\star_{\mathsf{alg}}(\theta), [0,1] \to \Pi$ takes a finite number of values. This shows that at least one deterministic policy is visited infinitely often as $\theta \to 1$. In particular, let $(\theta_n)_{n \geq 1} \in [0,1]^{\mathbb{N}}$ such that $\theta_n \to 1$ and the same deterministic recommendation policy $\hat{\pi}_{\mathsf{alg}}$ is optimal for this sequence of adherence levels: $R(\pi_{\mathsf{eff}}(\hat{\pi}_{\mathsf{alg}}, \theta_n)) \geq R(\pi_{\mathsf{eff}}(\pi, \theta_n)), \forall\, \pi \in \Pi, \forall\, n \in \mathbb{N}$. In particular, for each deterministic recommendation policy π, let us write $\phi_\pi : [0,1] \to \mathbb{R}$ for the map $\phi_\pi(\theta) = R(\pi_{\mathsf{eff}}(\hat{\pi}_{\mathsf{alg}}, \theta)) - R(\pi_{\mathsf{eff}}(\pi, \theta))$. We know that $\phi_\pi(\theta_n) \geq 0, \forall\, n \geq 1$. We want to show that this inequality $\phi_\pi(\theta) \geq 0$ holds for all values of θ sufficiently close to 1. From Lemma 10.1.3 in Puterman (2014), we know that ϕ_π is a rational function. From Lemma G.1, we know that either ϕ_π is identically equal to 0, or it is equal to 0 for finitely many values of θ in the

interval $[0,1]$. If ϕ_π is identically equal to 0, then indeed $\phi_\pi(\theta) \geq 0$ holds for all $\theta \in [0,1]$. Otherwise, ϕ_π is equal to 0 only for a finite number of values in $[0,1]$, which implies that ϕ_π can change sign only finitely many times. Since $\phi_\pi(\theta_n) \geq 0, \forall\, n \geq 1$, for $\theta_n \to 1$, there exists a threshold $\theta_\pi \in [0,1]$, such that $\phi_\pi(\theta) \geq 0$, for all $\theta \in [\theta_\pi, 1]$. If we take $\hat{\theta} \geq \theta_\pi$ for any deterministic policy π, we find that for all $\theta \in [\hat{\theta}, 1]$, we have $R(\pi_{\mathsf{eff}}(\hat{\pi}_{\mathsf{alg}}, \theta)) \geq R(\pi_{\mathsf{eff}}(\pi, \theta)), \forall\, \pi \in \Pi$. This concludes the proof.

2. We can extend the proof of the previous statement to any $\theta \in (0,1)$. In particular, for any $\theta \in (0,1)$, there exists an open interval $\mathcal{I}_\theta \subset (0,1)$ containing θ such that the optimal recommendation policy is constant on $\mathcal{I}_\theta^+ = \mathcal{I}_\theta \bigcap [\theta, 1]$ and constant on $\mathcal{I}_\theta^- = \mathcal{I}_\theta \bigcap [0, \theta]$. If the optimal recommendation policies on $\mathcal{I}_\theta^-$ and $\mathcal{I}_\theta^+$ are different, then they are still both optimal at θ. We can construct two additional intervals, $\mathcal{I}_0 = [0, \theta')$ and $\mathcal{I}_1 = (\theta'', 1]$, on which the optimal recommendation policy is constant. Note that $\mathcal{I}_0, \mathcal{I}_1$ are open sets for the subspace topology of $[0,1]$ induced by the classical topology of $\mathbb{R}$. Overall, we obtain a covering of the compact set $[0,1]$ with open sets $\{\mathcal{I}_\theta \mid \theta \in [0,1]\}$: $\bigcup_{\theta \in [0,1]} \mathcal{I}_\theta = [0,1]$. From the Heine-Lebesgue covering theorem, there exists a finite number of adherence levels $\theta'_1, ..., \theta'_m \in [0,1]$ such that $\bigcup_{i=1}^m \mathcal{I}_{\theta'_i} = [0,1]$. Finally, from the set $\{\theta'_i \mid i = 1, ..., m\}$, for some $p \in \mathbb{N}$ we can construct $\theta_1 = 0 < \theta_2 < ... < \theta_p = 1$ such that the optimal recommendation policy is constant on each of the interval $[\theta_i, \theta_{i+1}]$ for each $i \in \{1, ..., p\}$, with multiple optimal policies at the breakpoints θ_i for $i \in \{1, ..., p-1\}$.

3. We show this statement for $\underline{\theta} = 1$. Let $\theta \in [0,1]$ and assume that $\pi_{\mathsf{base}} = \pi^\star_{\mathsf{alg}}(1)$. Note that

$$\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1), \theta) = \theta \pi^\star_{\mathsf{alg}}(1) + (1-\theta)\pi_{\mathsf{base}} = \theta \pi^\star_{\mathsf{alg}}(1) + (1-\theta)\pi^\star_{\mathsf{alg}}(1) = \pi^\star_{\mathsf{alg}}(1).$$

By definition, $R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}(\theta), \theta)) \geq R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)), \forall\, \pi_{\mathsf{alg}} \in \Pi$. But we have shown in Proposition 4.3 that $R(\pi^\star_{\mathsf{eff}}(\theta)) \leq R(\pi^\star_{\mathsf{eff}}(1))$ and $R(\pi^\star_{\mathsf{eff}}(1)) = R(\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(1), \theta))$. We conclude that $\pi^\star_{\mathsf{eff}}(\theta) = \pi^\star_{\mathsf{eff}}(1)$ and that $\pi^\star_{\mathsf{alg}}(\theta) = \pi_{\mathsf{alg}}(1)$. The proof of this statement for $\underline{\theta} < 1$ is similar and we omit it for conciseness.

□

Appendix H: Proof of Proposition 4.5

Proof of Proposition 4.5. Let $\bar{s} \in \mathcal{S}$ such that $v_{\bar{s}}^{\pi_{\mathsf{base}}} = v_{\bar{s}}^{\pi^\star_{\mathsf{alg}}(1)}$. We first prove that $v_{\bar{s}}^{\pi^\star_{\mathsf{eff}}(\theta)} = v_{\bar{s}}^{\pi_{\mathsf{base}}}$ for any $\theta \in [0,1]$. We can adapt the proof of Proposition 4.3 to show that for any $s \in \mathcal{S}$, the map $\theta \mapsto v_s^{\pi^\star_{\mathsf{alg}}(\theta)}$ is non-decreasing. Since we can choose $\pi^\star_{\mathsf{alg}}(0) = \pi_{\mathsf{base}}$, this shows that

$$v_s^{\pi_{\mathsf{base}}} \leq v_s^{\pi^\star_{\mathsf{eff}}(\theta)} \leq v_s^{\pi^\star_{\mathsf{alg}}(1)}, \forall\, s \in \mathcal{S}, \forall\, \theta \in [0,1]. \tag{H.1}$$

Combining (H.1) with $v_{\bar{s}}^{\pi_{\mathsf{base}}} = v_{\bar{s}}^{\pi^\star_{\mathsf{alg}}(1)}$, we obtain that $v_{\bar{s}}^{\pi^\star_{\mathsf{eff}}(\theta)} = v_{\bar{s}}^{\pi_{\mathsf{base}}}$ for any $\theta \in [0,1]$. Now let $\theta \in [0,1]$. We show that we can choose $\pi^\star_{\mathsf{alg}}(\theta)_{\bar{s}} = \pi_{\mathsf{base},\bar{s}}$. Recall that there exists a unique vector $\boldsymbol{v}_\theta^\infty$ such that

$$v_{\theta,\bar{s}}^\infty = \max_{\boldsymbol{\pi}_{\bar{s}} \in \Delta(\mathcal{A})} \theta \cdot \sum_{a \in \mathcal{A}} \pi_{\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) + (1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) \tag{H.2}$$

and $\boldsymbol{v}_\theta^\infty = \boldsymbol{v}^{\pi^\star_{\mathsf{eff}}(\theta)}$. To show $\pi^\star_{\mathsf{alg}}(\theta)_{\bar{s}} = \pi_{\mathsf{base},\bar{s}}$, we need to show that

$$\pi_{\mathsf{base},\bar{s}} \in \arg\max_{\boldsymbol{\pi}_{\bar{s}} \in \Delta(\mathcal{A})} \theta \cdot \sum_{a \in \mathcal{A}} \pi_{\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) + (1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right). \tag{H.3}$$

Since $v_{\theta,s}^\infty \geq v_s^{\pi_{\mathsf{base}}}, \forall\, s \in \mathcal{S}$, we have

$$(1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) \geq (1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}^{\pi_{\mathsf{base}}}\right) = (1-\theta) v_{\bar{s}}^{\pi_{\mathsf{base}}} \tag{H.4}$$

where the equality follows from the fixed-point equation for the value function of a policy. Similarly, we obtain

$$\max_{\boldsymbol{\pi}_{\bar{s}} \in \Delta(\mathcal{A})} \theta \cdot \sum_{a \in \mathcal{A}} \pi_{\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) \geq \theta \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) \geq \theta v_{\bar{s}}^{\pi_{\mathsf{base}}}.$$

Overall, we obtain that

$$\theta \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) + (1-\theta) \cdot \sum_{a \in \mathcal{A}} \pi_{\mathsf{base},\bar{s}a} \boldsymbol{P}_{\bar{s}a}^\top \left(\boldsymbol{r}_{\bar{s}a} + \lambda \boldsymbol{v}_\theta^\infty\right) \geq v_{\bar{s}}^{\pi_{\mathsf{base}}}.$$

But $v_{\bar{s}}^{\pi_{\mathsf{base}}} = v_{\bar{s}}^\infty$, and $v_{\bar{s}}^\infty$ satisfies Equation (H.2). Therefore, (H.3) holds, and we can choose $\pi^\star_{\mathsf{alg}}(\theta)_{\bar{s}} = \pi_{\mathsf{base},\bar{s}}$.

□

Appendix I: Bounding the suboptimality of a recommendation policy

In this section we show the following proposition, which provides a bound between $\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta),\theta)}$, the optimal value functions at a given adherence level θ, and $\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\theta)}$, the value function of π_{alg}.

PROPOSITION I.1. *Let $\pi_{\mathsf{alg}} \in \Pi$ be a recommendation policy and $\theta \in [0,1]$. Then we have*

$$\left\|\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta),\theta)} - \boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\theta)}\right\|_\infty \leq \frac{\theta}{1-\lambda} \cdot \max_{s \in \mathcal{S}} \left\|\boldsymbol{\pi}^\star_{\mathsf{alg}}(\theta)_s - \boldsymbol{\pi}_{\mathsf{alg},s}\right\|_1 \cdot \left\|\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta),\theta)}\right\|_\infty. \tag{I.1}$$

Proposition I.1 bounds the difference between the value function of the optimal recommendation, $\pi^\star_{\mathsf{alg}}(\theta)$ and that of any policy π_{alg}. We delay the proof of Proposition I.1 below and we first analyze the bound. Several comments are in order:

- Our bound is parametrized by the ℓ_1-norm between the distribution induced by $\pi^\star_{\mathsf{alg}}(\theta)$ and π_{alg} at each state $s \in \mathcal{S}$; note that if both policies are deterministic, we always have $\|\boldsymbol{\pi}^\star_{\mathsf{alg}}(\theta)_s - \boldsymbol{\pi}_{\mathsf{alg},s}\|_1 \in \{0,2\}$. This term also reflects the piecewise constant structure of optimal recommendation policies as regards the adherence level, see Proposition 4.4.

- Our bound reflects the fact that when $\theta = 0$, we have $\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta) = \pi_{\mathsf{base}}$ for any π_{alg}, so that in this case all recommendation policies have the same performances.
- The multiplicative factor in $1/(1-\lambda)$ is common for bounds on the difference between two value functions. However, the multiplicative factor $\|\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta),\theta)}\|_\infty$ may also be of order $1/(1-\lambda)$, in which case our bound (I.1) is not tight.

Proof of Proposition I.1. Recall the notation $T^\pi(\boldsymbol{v}) \in \mathbb{R}^{\mathcal{S}}$, defined as $T^\pi_s(\boldsymbol{v}) = \sum_{a\in\mathcal{A}} \pi_{sa} \boldsymbol{P}_{sa}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}), \forall\, s \in \mathcal{S}$. For the sake of clarity, in this proof we use the notation $\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta), \theta) = \pi^\infty_{\mathsf{eff}}, \pi^\star_{\mathsf{alg}}(\theta) = \pi^\infty_{\mathsf{alg}}, \pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta) = \pi_{\mathsf{eff}}$. We want to obtain a bound on $\|\boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi^\star_{\mathsf{alg}}(\theta),\theta)} - \boldsymbol{v}^{\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}},\theta)}\|_\infty = \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}} - \boldsymbol{v}^{\pi_{\mathsf{eff}}}\|_\infty$. By definition, we have the following two fixed-point equations: for all $s \in \mathcal{S}$,

$$v_s^{\pi^\infty_{\mathsf{eff}}} = \theta T_s^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) + (1-\theta) T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}), v_s^{\pi_{\mathsf{eff}}} = \theta T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}) + (1-\theta) T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}).$$

Therefore, for all $s \in \mathcal{S}$, $v_s^{\pi^\infty_{\mathsf{eff}}} - v_s^{\pi_{\mathsf{eff}}} = \theta \left(T_s^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}) \right) + (1-\theta) \left(T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}) \right)$. Since $\boldsymbol{v} \mapsto T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v})$ is a contraction, we have, for all $s \in \mathcal{S}$,

$$|T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{base}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}})| \leq \lambda \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}} - \boldsymbol{v}^{\pi_{\mathsf{eff}}}\|_\infty. \tag{I.2}$$

We now consider the term $T_s^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}})$. We have, for all $s \in \mathcal{S}$,

$$T_s^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}) = T_s^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) + T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}).$$

Note that $\boldsymbol{v} \mapsto T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v})$ is a contraction, therefore we have, for all $s \in \mathcal{S}$,

$$|T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}})| \leq \lambda \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}} - \boldsymbol{v}^{\pi_{\mathsf{eff}}}\|_\infty. \tag{I.3}$$

Combining (I.2) and (I.3), we obtain that, for all $s \in \mathcal{S}$,

$$v_s^{\pi^\infty_{\mathsf{eff}}} - v_s^{\pi_{\mathsf{eff}}} \leq \theta \cdot \lambda \cdot \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}} - \boldsymbol{v}^{\pi_{\mathsf{eff}}}\|_\infty + \theta \cdot \|T^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}})\|_\infty + (1-\theta) \cdot \lambda \cdot \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}} - \boldsymbol{v}^{\pi_{\mathsf{eff}}}\|_\infty.$$

This shows that $\|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}} - \boldsymbol{v}^{\pi_{\mathsf{eff}}}\|_\infty \leq \frac{\theta}{1-\lambda} \cdot \|T^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}})\|_\infty$.

There remains to bound $\|T^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}})\|_\infty$. We have, for all $s \in \mathcal{S}$,

$$T_s^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) - T_s^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}) = \sum_{a\in\mathcal{A}} \left(\pi^\infty_{\mathsf{alg},sa} - \pi_{\mathsf{alg},sa}\right) \boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}\right) \leq \|\boldsymbol{\pi}^\infty_{\mathsf{alg},s} - \boldsymbol{\pi}_{\mathsf{alg},s}\|_1 \| \left(\boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}\right)\right)_{a\in\mathcal{A}} \|_\infty.$$

Now note that from Corollary B.1, we have $\| \left(\boldsymbol{P}_{sa}^\top \left(\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}\right)\right)_{a\in\mathcal{A}} \|_\infty \leq \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}\|_\infty$. This shows that $\|T^{\pi^\infty_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}}) - T^{\pi_{\mathsf{alg}}}(\boldsymbol{v}^{\pi_{\mathsf{eff}}})\|_\infty \leq \left(\max_{s\in\mathcal{S}} \|\boldsymbol{\pi}^\infty_{\mathsf{alg},s} - \boldsymbol{\pi}_{\mathsf{alg},s}\|_1\right) \cdot \|\boldsymbol{v}^{\pi^\infty_{\mathsf{eff}}}\|_\infty$. Combining this with (I.2) and (I.3) concludes the proof of Proposition I.1. □

Appendix J: Proof of Theorem 6.1

Proof. We want to show that $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{\theta \in [\underline{\theta}, \bar{\theta}]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta)) = \max_{\pi_{\mathsf{alg}} \in \Pi} \min_{\theta \in [\underline{\theta}, \bar{\theta}]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, \theta))$ and that $\left(\pi^\star_{\mathsf{alg}}(\underline{\theta}), \underline{\theta}\right)$ is an optimal solution to the optimization problem in the above equation. Recall that in Theorem B.2 in Appendix B we have studied the **Time- and State-invariant Adversarial** model, for which we have shown that $\sup_{\pi_{\mathsf{alg}} \in \Pi_{\mathsf{H}}} \min_{u \in [\theta, 1]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u)) = \max_{\pi_{\mathsf{alg}} \in \Pi} \min_{u \in [\theta, 1]} R(\pi_{\mathsf{eff}}(\pi_{\mathsf{alg}}, u))$ and that $\left(\pi^\star_{\mathsf{alg}}(\theta), \theta\right)$ is an optimal solution to the optimization problem above. Therefore, we can prove Theorem 6.1 by applying the exact same proof as for Theorem B.2 but replacing the interval $[\theta, 1]$ by the interval $[\underline{\theta}, \bar{\theta}]$. We omit the proof for the sake of conciseness.

Appendix K: Proof of Theorem 6.2

We start from the surrogate MDP $\mathcal{M}'$ defined in Section 4.2, where the transition probabilities $\boldsymbol{P}' \in (\Delta(\mathcal{S}))^{\mathcal{S} \times \mathcal{A}}$ and the instantaneous rewards $\boldsymbol{r}' \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}$ are defined in (4.3). In particular, recall that the Bellman equation of the surrogate MDP (4.2) is $v_s^\infty = \max_{\boldsymbol{\pi}_s \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \pi_{sa} \left(r'_{sa} + \lambda \boldsymbol{P}'^{\top}_{sa} \boldsymbol{v}^\infty\right), \forall\, s \in \mathcal{S}$. From this, we see that the optimization problem (6.4) is an s-rectangular robust Markov decision process (Iyengar 2005, Wiesemann et al. 2013), where the set $\mathcal{U}$ of admissible pairs of instantaneous rewards and transition probabilities $(\boldsymbol{r}', \boldsymbol{P}')$ is described as

$$
\begin{aligned}
\mathcal{U} = \{(\boldsymbol{r}', \boldsymbol{P}') \mid \boldsymbol{r}' \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}, \boldsymbol{P}' \in (\Delta(\mathcal{S}))^{\mathcal{S} \times \mathcal{A}}, \pi_{\mathsf{base}} \in \Gamma,& \\
\boldsymbol{P}'_{sa} := \theta \cdot \boldsymbol{P}_{sa} + (1-\theta) \cdot \sum_{a' \in \mathcal{A}} \pi_{\mathsf{base},sa'} \boldsymbol{P}_{sa'}, \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}& \\
r'_{sa} := \theta \cdot \boldsymbol{P}_{sa}^\top \boldsymbol{r}_{sa} + (1-\theta) \cdot \sum_{a' \in \mathcal{A}} \pi_{\mathsf{base},sa'} \boldsymbol{P}_{sa'}^\top \boldsymbol{r}_{sa'}, \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}\}.&
\end{aligned}
$$

By construction, the set $\mathcal{U}$ is s-rectangular, i.e., it can be written $\mathcal{U} = \times_{s \in \mathcal{S}} \mathcal{U}_s$ with $\mathcal{U}_s \subset \Delta(\mathcal{S})^{\mathcal{A}}$ convex for each $s \in \mathcal{S}$. From Wiesemann et al. (2013), we conclude that an optimal policy for the decision problem (6.4) can be chosen stationary. Additionally, an optimal policy can be computed efficiently when the sets Γ_s are described by affine and conic constraints, see corollary 3 in Wiesemann et al. (2013) for a more precise complexity statement.