arXiv:1906.05988v3 [math.OC] 8 Dec 2020

Distributionally Robust Partially Observable Markov Decision Process with Moment-based Ambiguity

Hideaki Nakao* Ruiwei Jiang† Siqian Shen‡

Abstract

We consider a distributionally robust Partially Observable Markov Decision Process (DR-POMDP), where the distribution of the transition-observation probabilities is unknown at the beginning of each decision period, but their realizations can be inferred using side information at the end of each period after an action being taken. We build an ambiguity set of the joint distribution using bounded moments via conic constraints and seek an optimal policy to maximize the worst-case (minimum) reward for any distribution in the set. We show that the value function of DR-POMDP is piecewise linear convex with respect to the belief state and propose a heuristic search value iteration method for obtaining lower and upper bounds of the value function. We conduct numerical studies and demonstrate the computational performance of our approach via testing instances of a dynamic epidemic control problem. Our results show that DR-POMDP can produce more robust policies under misspecified distributions of transition-observation probabilities as compared to POMDP, but has less costly solutions than robust POMDP. The DR-POMDP policies are also insensitive to varying parameter in the ambiguity set and to noise added to the true transition-observation probability values obtained at the end of each decision period.

Keywords: Partially Observable Markov Decision Process (POMDP), distributionally robust optimization, moment-based ambiguity set, heuristic search value iteration (HSVI), epidemic control

1 Introduction

Partially Observable Markov Decision Processes (POMDPs) are useful for modeling sequential decision making problems, where a decision maker (DM) is only able to obtain partial information about the present state of a system of interest. Similar to the Markov Decision Processes (MDPs), the transition probabilities in between the states of the system depend on the current state and the action chosen by the DM. In addition, POMDPs are accompanied with a set of observation outcomes that are realized probabilistically given the DM's action and the state into which the system has transitioned. Different from MDPs where the DM is able to directly observe the current state of the system, in POMDPs the DM can only view an observation instead of the true state. Applications of POMDPs include clinical decision making, inventory control, machine repair, epidemic intervention and many more Cassandra (1998); Hauskrecht and Fraser (2000); Treharne and Sox (2002).

A general objective in sequential decision making is to devise a policy of taking dynamic actions to maximize (minimize) the expected value of the cumulative reward (cost). In MDPs, the DM

*Department of Industrial and Operations Engineering, University of Michigan at Ann Arbor, USA;

†Department of Industrial and Operations Engineering, University of Michigan at Ann Arbor, USA;

‡Corresponding author; Department of Industrial and Operations Engineering, University of Michigan at Ann Arbor, USA. Email: siqian@umich.edu.

gains a reward (or pays a cost) for each action made on a state of the system. In POMDPs, since the DM has no access to the true state, she is uncertain about the reward (cost) received. Instead, the DM retains her belief of the present state based on past actions and observations, and anticipates an expected value of the reward (or the expected cost) based on the belief. The DM's belief is represented by a probability mass associated with each state of the system, which is a sufficient statistic of the history of past actions and observations (Kumar and Varaiya, 2015, Chapter 6.6). Since a policy is a function of the past actions and observations, this property is useful to compactly represent an increasing sequence of information.

In POMDPs, a critical assumption is that the exact transition and observation probabilities are known to the DM for each action-state combination. In practice, there may exist estimation errors about either the transition or observation probability values, to handle which, Rasouli and Saghafian (2018) builds an uncertainty set of probabilities and develops an exact algorithm for the problem of maximizing the expected reward in the worst-case realization of the unknown probabilities in POMDPs. We will numerically compare actions of robust POMDP (see Osogami (2015)) with decision policies of DR-POMDP and POMDP in Section 6.

In this paper, using bounded moments, we construct an ambiguity set of the unknown joint distribution of the transition-observation probabilities, in which the true joint distribution lies with high probability. We consider a distributionally robust optimization framework of POMDPs (called DR-POMDP) to seek an optimal policy against the worst-case distribution in the ambiguity set, when realizations of the transition and observation probabilities in each decision period are generated from this distribution. Moreover, we allow transition-observation probabilities to vary in different decision periods, and assume that at the end of each period, the DM can gather side information to infer the true values of the transition-observation probabilities realized in that period, even these values were unknown to the DM when decisions were made. Admittedly, it is rather restrictive to have this assumption where the transition-observation probabilities can be observed retrospectively. However, there exist a wide range of applications where the underlying dynamics are understood and can be simulated to produce unknown parameters (i.e., transition-observation probabilities) once values of some exogenous parameters are gained after the decisions are made. For example, Mannor et al. Mannor et al. (2016) justify the electric power system as one case where the system performance can be reliably simulated when environmental factors, such as wind and solar radiation levels, are known. In Section 3, we provide a few examples to further illustrate and justify this assumption and in Section 6, we conduct numerical tests on dynamic epidemic control problem instances, which satisfy the assumption.

In distributionally robust optimization (DRO), we seek solutions to optimize the worst-case objective given by possible distributions contained in an ambiguity set. Compared with robust optimization that accounts for the worst-case objective outcome given by all possible realizations of uncertain parameters in an uncertainty set, optimal solutions to DRO models are less conservative and can be adjusted through the amount of data/information we have. Ref. Delage and Ye (2010) develops a moment-based ambiguity set, considering a set of distributions with an ellipsoidal condition on the mean and a conic constraint on the second-order moment, to derive tractable reformulations of several distributionally robust convex programs. Standardization of ambiguity sets via conic representable sets is proposed by Wiesemann et al. (2014). Ref. Zymler et al. (2013) considers tractable reformulations of DR chance-constrained programs using moment-based ambiguity set. Other types of ambiguity sets used in DRO models bound the ϕ-divergence Ben-Tal et al. (2013); Jiang and Guan (2016) or Wasserstein distance Esfahani and Kuhn (2018); Gao and Kleywegt (2016) in between possible distributions to a nominal distribution. In this paper, we also use a moment-based ambiguity set where the moment information is bounded via conic constraints. We establish the Bellman equation for DR-POMDP and prove the piecewise-linear-convex prop-

erty of the value function, using which we further develop efficient computational algorithms and demonstrate the efficacy of the DR-POMDP model by testing epidemic control problem instances with diverse parameter settings.

The remainder of the paper is organized as follows. In Section 2, we review the most relevant POMDP, robust MDP/POMDP, and DRO literature. In Section 3, we formally present DR-POMDP and provide a few examples to show possible applications. In Section 4, we formulate the Bellman equation and show that the value function is piecewise linear convex under general moment-based ambiguity sets described in Yu and Xu (2016). In Section 5, we develop an approximation algorithm for DR-POMDP based on a distributionally robust variant of the heuristic value search iteration algorithm. In Section 6, we demonstrate the computational results of solving DR-POMDP on randomly generated instances of a dynamic epidemic control problem, and compare it with POMDP and robust POMDP through different out-of-sample tests. Section 7 concludes the paper and presents future research directions.

2 Literature Review

Although strong modeling connections exist in between MDP and POMDP, techniques applied to solve MDP models where the states are discrete, are not directly applicable to solving POMDP since belief states are continuous. Ref. Smallwood and Sondik (1973) shows that the value function of POMDP is piecewise linear convex (PWLC) with respect to the belief state, and derives an exact algorithm to find an optimal policy. The exact algorithm, which keeps a set of vectors for characterizing the value function, is intractable as the search space increases exponentially over periods. Ref. Pineau et al. (2003) proposes a point-based value iteration (PBVI) algorithm by only keeping characterizing vectors for a subset of belief states, and thus maintains a lower bound of the true value function that aims to maximize the reward. The PBVI algorithm is polynomial in the number of states, observations, and actions, and the error induced by taking a subset of belief states is shown to be convergent if the subset is sampled densely in the reachable set of belief states. Ref. Smith and Simmons (2004) develops a heuristic search value iteration (HSVI) algorithm to derive an upper bound of the value function via finding the reachable set through simulation. Ref. Smith and Simmons (2004) shows that HSVI is guaranteed to terminate after the gap between the upper and lower bounds converges within a certain threshold.

The research on robust MDP is motivated by possible estimation errors of transition matrices and how they may have a significant impact to the solution quality (see, e.g., Abbad and Filar (1992); Abbad et al. (1990)). In Wiesemann et al. (2013), the authors show probabilistic guarantees for solutions to robust MDPs by building an uncertainty set using fully observable history. By construction, their robust policy achieves or exceeds its worst-case performance with a certain confidence. Ref. Nilim and El Ghaoui (2005) considers robust control for a finite-state, finite-action MDP, where uncertainty on the transition matrices is described by particular uncertainty sets such as likelihood regions or entropy bounds, and the authors present a robust dynamic programming algorithm for solving the problem. Ref. Iyengar (2005) analyzes a robust formulation for discrete-time dynamic programming where the transition probabilities are uncertain and ambiguously known, and shows that it is equivalent to stochastic zero-sum games with perfect information. Ref. Delage and Mannor (2010) argues that robust MDP models may produce over-conservative solutions, as they do not incorporate the distributional information of uncertain parameters. Then Xu and Mannor (2012) presents a distributionally robust MDP model, where the ambiguity set is characterized by a sequence of nested sets, each having a confidence level to guarantee that the true value is in the set with a certain probability. Ref. Yu and Xu (2016) generalizes the distributionally robust

MDP to include multi-modal distributions and the information of mean and variance. Ref. Yang (2017) proposes a distributionally robust MDP model by building an ambiguity set of distributions on transition probability using a Wasserstein ball centered around a nominal distribution. The use of Wasserstein ball ambiguity set results in a Kantorovich-duality-based convex reformulation for distributionally robust MDP.

Ref. Saghafian (2018) presents a modeling framework of ambiguous POMDP (called APOMDP), which generalizes the robust POMDP in Rasouli and Saghafian (2018). APOMDP optimizes over the α-maxmin expected utility, resulting in a policy that can achieve the intermediate performance of the worst case and the best case in the uncertainty set of parameters. Ref. Saghafian (2018) describes conditions under which the value function of APOMDP is PWLC. Meanwhile, Rasouli and Saghafian (2018) considers a general setting of robust POMDP, where the DM may not be able to obtain the exact transition-observation probabilities even after taking actions at the end of each period. In this case, the sufficient statistic is no longer a single belief state, but a collection of belief states, and the expected reward up to the current period must be taken into account to realize a policy that is robust in terms of the entire cumulative expected reward. The authors also derive an exact algorithm for robust POMDP where the uncertainty set is discrete. Here we note that robust POMDP with a continuous uncertainty set is computationally challenging even in a very simple setting. Moreover, Osogami (2015) formulates a robust counterpart for POMDP, where the transition-observation matrix is assumed to lie in a fixed support within the probability simplex. The realized transition-observation probability values are assumed to be observable to the DM at the end of each decision period, similar to the setting in this paper. While the value function for the standard POMDP can be described by a PWLC function, the value function of the robust POMDP is not necessarily piecewise linear, as there are possibly infinitely many supporting hyperplanes. The authors derive an efficient algorithm based on PBVI to approximate the exact solution, and discusses a method to conduct a robust belief update.

3 Problem Description

Figure 1 depicts the sequence of events that occur during one decision period. In a distributionally robust setting, we consider another agent (the "nature"), who chooses a distribution μ of the transition-observation probabilities from a pre-assumed ambiguity set. The DM expects that the nature may access to the same information as the DM and acts adversarially against the DM's action a taken at the beginning of each period. Therefore, the distribution μ is expected to lead to the worst-case expected reward. Next, the joint transition-observation probability $\boldsymbol{p}$ is realized from the distribution μ. The state makes a transition according to $\boldsymbol{p}$, and the observation outcome z is shown. Finally, the DM obtains the values of z and $\boldsymbol{p}$ at the end of the period.

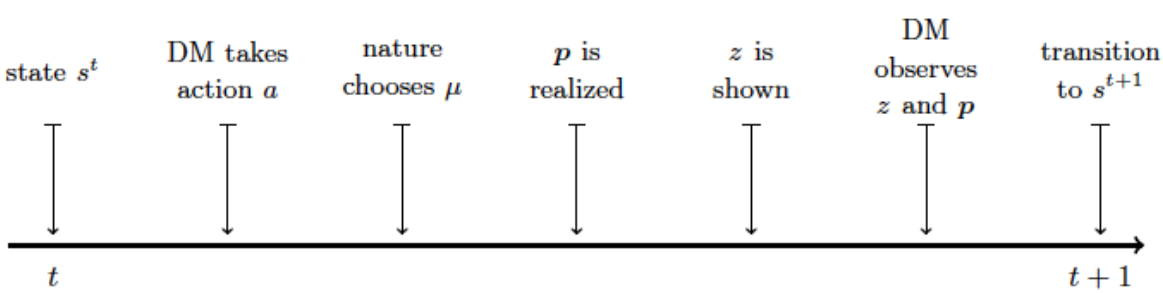


Figure 1: Sequence of events during one decision period in a DR-POMDP

We denote $\mathcal{S}$ as the set of states, $\mathcal{A}$ as the set of actions, and $\mathcal{Z}$ as the set of observation outcomes. For all $(s, s', z, a) \in \mathcal{S}^2 \times \mathcal{Z} \times \mathcal{A}$, we define $p_{as}(s', z) = \Pr(s', z|s, a)$ as the probability of transitioning between (s, s') and observing z, given action a. For $(s, a) \in \mathcal{S} \times \mathcal{A}$, let r_{as} be the reward for taking action a at state s. For all $s \in \mathcal{S},\ a \in \mathcal{A}$, we define a vector of probabilities $\boldsymbol{p}_{as} = (p_{as}(s', z),\ (s', z) \in \mathcal{S} \times \mathcal{Z})^\top$ and assume that the Cartesian product $(\boldsymbol{p}_{as}, r_{as})$ is a member of a set $\mathcal{X}_{as} \subseteq \Delta(\mathcal{S} \times \mathcal{Z}) \times \mathbb{R}$, where $\Delta(\cdot)$ is a probability simplex of set $\cdot$. We denote $\boldsymbol{p}_a = (p_{as}(s', z),\ (s, s', z) \in \mathcal{S}^2 \times \mathcal{Z})^\top$ and $\boldsymbol{r}_a = (r_{as},\ s \in \mathcal{S})^\top$ for all $a \in \mathcal{A}$. We assume that $(\boldsymbol{p}_{as}, r_{as})$ follows a distribution μ_{as}, which is unknown but is included in an ambiguity set $\mathcal{D}_{as} \subseteq \mathcal{P}(\mathcal{X}_{as})$, where $\mathcal{P}(\cdot)$ represents a set of all probability distributions with support $\cdot$. Furthermore, the set of distributions is rectangular with respect to the set of actions $\mathcal{A}$ and the set of states $\mathcal{S}$, i.e., the overall ambiguity set is $\mathcal{D} = \bigotimes_{\substack{a \in \mathcal{A} \\ s \in \mathcal{S}}} D_{as}$. This assumption is analogous to the (s, a)-rectangularity in Wiesemann et al. (2013). The above conditions increase the conservativeness of the model in general. In the online supplement Nakao, Hideaki and Jiang, Ruiwei and Shen, Siqian (2020) A, we discuss a relaxation of the a-rectangularity assumption for DR-POMDP.

Below we describe several examples in which the above settings of DR-POMDP can be justified, and therefore our approach can be applied to optimize corresponding policies. The key is to justify whether the DM can obtain the true value of $\boldsymbol{p}$ using side information at the end of each decision period. In Section 6, we also numerically show that our approach can produce quite stable reward in out-of-sample simulation tests even we add noise to the true $\boldsymbol{p}$-value obtained at the end of each period and thus the assumption is relatively weak.

First, consider dynamic epidemic surveillance and control. During a flu season, the number of weekly visits of patients who show influenza-like illness (ILI) symptoms is reported to the public. The number of ILI patients divided by the total population, called the ILI rate, is frequently used to estimate the prevalence of an epidemic. For example, Rath et al. (2003) studies a two-state MDP model (i.e., epidemic vs. non-epidemic) and shows that the ILI rate follows a Gaussian and an exponential distribution for the epidemic and non-epidemic state, respectively; Le Strat and Carrat (1999) uses ILI rate to predict influenza epidemics through a hidden Markov model. The hidden states correspond to the current epidemic level, which is unobservable to the DM due to incubation period and patient arrival latency. Different epidemic levels also cause different probabilities of the population visiting healthcare providers, which will then be reflected in ILI rate.

Arguably, the transition probabilities and ILI rates are dependent on government control policies, such as restricting travels, stopping mass gatherings, and so on. These decisions often have to be made before knowing the true transition matrix and observation probabilities between ILI rate and the true epidemic state. The DR-POMDP seeks a policy to minimize the worst-case expected cost (e.g., the total infected count, death toll, etc.) and at the end of each decision period, side information such as humidity, antigenic evolution of the virus, and population travels in the past period can be used to infer the true transition and ILI-rate observation probabilities (see, e.g., Du et al., 2017). Note that the side information is not available at the beginning of each decision period when the DM takes an action, but can be collected at the end of each period.

Another example arises in clinical decision-making such as deciding prostate cancer treatment plans Zhang and Denton (2018), where different treatment plans can probabilistically vary cancer conditions (i.e., states) of a patient. The true state of a cancer patient is hard to know but can be inferred probabilistically from belief states. Using DR-POMDP, a doctor's objective is to provide treatment and inspection as needed in order to minimize the maximum expected quality-adjusted life years for each patient under ambiguously known transition-observation probabilities. According to Zhang and Denton (2018), the detection of prostate-specific antigen (PSA), has a varying accuracy rate depending on the patient's condition. After treatment in each period, the

doctor can utilize the PSA information to infer the true transition and observation probabilities happening to the patient and update her belief to make treatment plans for the next period.

One can also consider planning production or maintaining inventory in highly seasonal industries such as agriculture Treharne and Sox (2002), where system states correspond to market trends in each decision period. The trend makes a transition according to a probability mass function that is unknown to the DM and each trend is associated with a certain distribution of demand that the DM aims to satisfy. For a certain product, the market transition probability and the demand distribution are correlated with climate factors, such as temperature and precipitation, which are uncertain to the DM when she makes a production plan and thus using DR-POMDP, the goal is to minimize the maximum demand loss due to distributional ambiguity. After each period, the DM observes the realized temperature and precipitation and also the true demand, to identify the true value of $\boldsymbol{p}$.

4 Optimal Policy for DR-POMDP

We derive an optimal policy for DR-POMDP when the DM can obtain the value of transition-observation probability at the end of each decision period. In Section 4.1, we formulate DR-POMDP as an optimization problem and construct the Bellman equation to derive the optimal policy. In Section 4.2, we show that the value function satisfying the Bellman equation is PWLC. Finally, in Section 4.3, we consider the infinite-horizon case, and demonstrate that the value function converges under the Bellman update operation.

4.1 Distributionally Robust Bellman Equation

We formulate a dynamic game involving two players: The DM selects $a \in \mathcal{A}$ and then the nature selects $\mu_a = \bigotimes_{s\in\mathcal{S}} \mu_{as}$ from the ambiguity set $D_a = \bigotimes_{s\in\mathcal{S}} \mathcal{D}_{as}$ to minimize the expected reward given the DM's action a. Let a^t, $\boldsymbol{p}^t_{a^t}$, z^t be the action, transition-observation probability outcome, and observation during decision period t. We denote $\mathcal{H}^t$ as the set of all possible histories up to period t, and denote $h^t = \left(a^1, \boldsymbol{p}^1_{a^1}, z^1, \ldots, a^{t-1}, \boldsymbol{p}^{t-1}_{a^{t-1}}, z^{t-1}\right)$ as a history in $\mathcal{H}^t$. The DM's objective is to find an optimal policy of selecting an action $a \in \mathcal{A}$ based on the history from $t = 1$ to T, i.e., finding the best policy $\pi = (\pi^1, \ldots, \pi^{T-1})$ with $\pi^t : \mathcal{H}^t \to \mathcal{A}$. We denote the set of all such policies as Π, and define an extended history $\tilde{h}^t = \left(a^1, \boldsymbol{p}^1_{a^1}, z^1, \ldots, a^{t-1}, \boldsymbol{p}^{t-1}_{a^{t-1}}, z^{t-1}, a^t\right) \in \tilde{\mathcal{H}}^t$, on which the nature bases its decision for choosing μ_{a^t}. The nature's objective is to find the best policy (from the nature's perspective) $\gamma = (\gamma^1, \ldots, \gamma^{T-1})$, with $\gamma^t : \tilde{\mathcal{H}}^t \to \mathcal{D}_{a^t}$ to minimize the expected reward. Similarly, we denote the set of all the nature's policies as Γ.

Rasouli and Saghafian Rasouli and Saghafian (2018) point out that the sufficient statistic for robust POMDP is no longer a single belief state, but a set of belief states. Moreover, they discuss that the set of belief states by itself cannot be used to construct an optimal policy since there exists uncertainty for the reward accumulated in the past, associated with each of the belief states. Because of the uncertainty in the expected reward, the DM must consider a belief state that achieves the smallest expected reward both in the past and the future, posing great challenge for optimization. We claim that a similar observation holds true for the distributionally robust case. However, when the DM can obtain the value of transition-observation probability at the end of each decision period, the ambiguity of the belief state, as well as the expected reward diminishes and the single belief state becomes a sufficient statistic for DR-POMDP, which can also be used to characterize the optimal policy.

Let the belief state in period t be $(b^t_s,\ s \in \mathcal{S}) = \boldsymbol{b}^t \in \Delta(\mathcal{S})$. Given action a, transition-observation probability $\boldsymbol{p}_a$, and observation outcome z, the sufficient statistic for the history $h^{t+1} =$

$(h^t, a, \boldsymbol{p}_a, z)$, or the belief state in period $t+1$ is given by

$$\boldsymbol{b}^{t+1} = \boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) = \frac{\sum_{s\in\mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s}{\sum_{s\in\mathcal{S}} \boldsymbol{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} b_s}, \tag{1}$$

where $\boldsymbol{1}$ represents a vector of ones having the length $|\mathcal{S}|$; $\boldsymbol{J}_z \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\times|\mathcal{Z}|)}$ is a matrix of zeros and ones that projects the vector $\boldsymbol{p}_{as}$ to a vector $\boldsymbol{p}_{asz} = (p_{as}(s', z),\ s' \in \mathcal{S})^\top$, whose entries correspond to the outcome z. That is, $\boldsymbol{p}_{asz} = \boldsymbol{J}_z \boldsymbol{p}_{as},\ \forall a,\ s,\ z$. Note that the belief state cannot be updated using (1) and will not be a sufficient statistic of the history of past actions and observations if we do not have the true values of $\boldsymbol{p}_{as}$.

With slight abuse of notation, let π be a policy that maps belief states to the actions, i.e., $\pi^t :\ \Delta(\mathcal{S}) \to \mathcal{A}$ for all $t \in \{1, \ldots, T-1\}$. Similarly, let $\gamma^t :\ \Delta(\mathcal{S})\times\mathcal{A} \to \mathcal{D}_{a^t}$ for all $t \in \{1, \ldots, T-1\}$. Note that the nature's policy is dependent on the belief state since the nature acts adversarial to the DM.

Remark 1 *Note that the deterministic policy is optimal since the nature is able to access to the same information as the DM, plus the action that the DM has performed. This does not hold true when the nature is not able to perfectly access to the DM's immediate action.*

Given the nature's choice of distribution μ_a, the expected value of the instantaneous reward given belief state $\boldsymbol{b}$ and action a is denoted as $\mathbb{E}_{(\boldsymbol{p}_a, \boldsymbol{r}_a)\sim\mu_a}\left[\boldsymbol{b}^\top \boldsymbol{r}_a\right]$, where "$\sim$" expresses the relation between random variables and probability distributions. Let $\beta \in (0, 1]$ be a discount factor. The objective of the DM is to find a policy to maximize the minimum cumulative discounted expected reward given all possible policies (i.e., distributions of transition-observation probabilities) by the nature. That is, DR-POMDP aims to solve

$$\begin{aligned}
\max_{\pi\in\Pi}\min_{\gamma\in\Gamma} \quad & \mathbb{E}\left[\sum_{t=1}^{T-1} \beta^t \boldsymbol{b}^{t\top} \boldsymbol{r}^t_{a^t}\right] && && \text{(2a)}\\
\text{s.t.} \quad & a^t = \pi^t(\boldsymbol{b}^t), && \forall t \in \{1, \ldots, T-1\} && \text{(2b)}\\
& \mu^t_{a^t} = \gamma^t(\boldsymbol{b}^t, a^t), && \forall t \in \{1, \ldots, T-1\} && \text{(2c)}\\
& (\boldsymbol{p}^t_{a^t}, \boldsymbol{r}^t_{a^t}) \sim \mu^t_{a^t}, && \forall t \in \{1, \ldots, T-1\} && \text{(2d)}\\
& (s^{t+1}, z^t) \sim \boldsymbol{p}^t_{a^t s^t}, && \forall t \in \{1, \ldots, T-1\} && \text{(2e)}\\
& \boldsymbol{b}^{t+1} = \boldsymbol{f}(\boldsymbol{b}^t, a^t, \boldsymbol{p}^t_{a^t}, z^t), && \forall t \in \{1, \ldots, T-1\} && \text{(2f)}
\end{aligned}$$

where the terminal reward is zero without loss of generality. The initial belief state is given as $\boldsymbol{b}$. Alternatively, we denote the problem (2) as

$$\max_{\pi\in\Pi}\min_{\gamma\in\Gamma} \mathbb{E}\left[\sum_{t=1}^{T-1} \beta^t \boldsymbol{b}^{t\top} \boldsymbol{r}^t_{a^t} \,\middle|\, \boldsymbol{b}^1 = \boldsymbol{b}\right]. \tag{3}$$

Here we omit all the constraints in (2) for presentation simplicity.

To solve (2), we propose to use dynamic programming, and derive the Bellman equation below.

Proposition 1 *Denote* $\pi^{t:T-1} = (\pi^t, \pi^{t+1}, \ldots, \pi^{T-1})$ *and* $\gamma^{t:T-1} = (\gamma^t, \gamma^{t+1}, \ldots, \gamma^{T-1})$ *as sequences of policies from* t *to* $T-1$. *Let* $\Pi^{t:T-1}$ *and* $\Gamma^{t:T-1}$ *be the sets of all policies* $\pi^{t:T-1}$ *and* $\gamma^{t:T-1}$, *respectively. Consider the value function in period* t *as*

$$V^t(\boldsymbol{b}) = \max_{\pi^{t:T-1}\in\Pi^{t:T-1}} \min_{\gamma^{t:T-1}\in\Gamma^{t:T-1}} \mathbb{E}\left[\sum_{n=t}^{T-1} \beta^{n-t} \boldsymbol{b}^{n\top} \boldsymbol{r}^n_{a^n} \,\middle|\, \boldsymbol{b}^t = \boldsymbol{b}\right]. \tag{4}$$

Then,

$$V^t(\boldsymbol{b}) = \max_{a\in\mathcal{A}} \min_{\mu_a\in\mathcal{D}_a} \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a}\left[\sum_{s\in\mathcal{S}} b_s\left\{r_{as} + \beta\sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1}\left(\boldsymbol{f}\left(\boldsymbol{b}, a, \boldsymbol{p}_a, z\right)\right)\right\}\right]. \tag{5}$$

Proof: We first isolate the term associated with period t inside the expectation of (4) as follows.

$$V^t(\boldsymbol{b}) = \max_{\pi^{t:T-1}\in\Pi^{t:T-1}} \min_{\gamma^{t:T-1}\in\Gamma^{t:T-1}} \mathbb{E}\left[\boldsymbol{b}^{t\top}\boldsymbol{r}^t_{a^t} + \beta\sum_{n=t+1}^{T-1}\beta^{n-(t+1)}\boldsymbol{b}^{n\top}\boldsymbol{r}^n_{a^n}\,\middle|\,\boldsymbol{b}^t = \boldsymbol{b}\right].$$

Given $a^t = \pi^t(\boldsymbol{b}),\ \boldsymbol{p}^t_a = \boldsymbol{p}_{\pi^t(\boldsymbol{b})},\ z^t = z$, the probability of observing z is

$$\sum_{s\in\mathcal{S}} b_s \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{\pi_t(\boldsymbol{b})s}.$$

Thus, we can calculate the expectation conditioned on the values of $a^t,\ \boldsymbol{p}^t_a,\ z^t$ in the value function as:

$$\begin{aligned}
V^t(\boldsymbol{b}) \quad = \quad & \max_{\pi^{t:T-1}\in\Pi^{t:T-1}} \min_{\gamma^{t:T-1}\in\Gamma^{t:T-1}} \mathbb{E}_{(\boldsymbol{p}_{\pi_t(\boldsymbol{b})},\boldsymbol{r}_{\pi_t(\boldsymbol{b})})\sim\mu_{\pi^t(\boldsymbol{b})}}\Bigg[\sum_{s\in\mathcal{S}} b_s r^t_{\pi^t(\boldsymbol{b})s} \\
& + \sum_{z\in\mathcal{Z}}\sum_{s\in\mathcal{S}} b_s \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{\pi_t(\boldsymbol{b})s} \mathbb{E}\left[\sum_{n=t+1}^{T-1}\beta^{n-(t+1)}\boldsymbol{b}^{n\top}\boldsymbol{r}^n_{a^n}\,\middle|\,\boldsymbol{b}^t = \boldsymbol{b}, a^t = \pi^t(\boldsymbol{b}), \boldsymbol{p}^t_a = \boldsymbol{p}_{\pi^t(\boldsymbol{b})}, z^t = z\right]\Bigg\}\Bigg] \\
= \quad & \max_{\pi^{t:T-1}\in\Pi^{t:T-1}} \min_{\gamma^{t:T-1}\in\Gamma^{t:T-1}} \mathbb{E}_{(\boldsymbol{p}_{\pi_t(\boldsymbol{b})},\boldsymbol{r}_{\pi_t(\boldsymbol{b})})\sim\mu_{\pi^t(\boldsymbol{b})}}\Bigg[\sum_{s\in\mathcal{S}} b_s\Bigg\{r^t_{\pi^t(\boldsymbol{b})s} \\
& + \beta\sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{\pi_t(\boldsymbol{b})s} \mathbb{E}\left[\sum_{n=t+1}^{T-1}\beta^{n-(t+1)}\boldsymbol{b}^{n\top}\boldsymbol{r}^n_{a^n}\,\middle|\,\boldsymbol{b}^{t+1} = \boldsymbol{f}\left(\boldsymbol{b}, \pi^t(\boldsymbol{b}), \boldsymbol{p}_{\pi^t(\boldsymbol{b})}, z\right)\right]\Bigg\}\Bigg],
\end{aligned}$$

where the second equality is due to rearranging the terms and the fact that $\boldsymbol{b}$ is an information state. Because policies beyond period t do not affect $(\boldsymbol{p}^t_{a^t}, \boldsymbol{r}^t_{a^t})$, we have

$$\begin{aligned}
V^t(\boldsymbol{b}) = & \max_{a\in\mathcal{A}} \min_{\mu_a\in\mathcal{D}_a} \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a}\Bigg[\sum_{s\in\mathcal{S}} b_s\Bigg\{r_{as} + \beta\sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} \\
& \times \max_{\pi^{t+1:T-1}\in\Pi^{t+1:T-1}} \min_{\gamma^{t+1:T-1}\in\Gamma^{t+1:T-1}} \mathbb{E}\left[\sum_{n=t+1}^{T-1}\beta^{n-(t+1)}\boldsymbol{b}^{n\top}\boldsymbol{r}^n_{a^n}\,\middle|\,\boldsymbol{b}^{t+1} = \boldsymbol{f}\left(\boldsymbol{b}, a, \boldsymbol{p}_a, z\right)\right]\Bigg\}\Bigg] \\
= & \ (5).
\end{aligned}$$

The final equality follows the definition of V^{t+1}. This completes the proof. Following Proposition 1, the policies optimal to (3) can be determined by recursively solving (5) from period T to $t = 1$.

Now define two functions:

$$U^t(\boldsymbol{b}, a, \mu_a) = \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a}\left[\sum_{s\in\mathcal{S}} b_s\left\{r_{as} + \beta\sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1}\left(\boldsymbol{f}\left(\boldsymbol{b}, a, \boldsymbol{p}_a, z\right)\right)\right\}\right], \tag{6}$$

$$Q^t(\boldsymbol{b}, a) = \min_{\mu_a\in\mathcal{D}_a} U^t(\boldsymbol{b}, a, \mu_a). \tag{7}$$

The solution to the Bellman equation provides the optimal action given belief state $\boldsymbol{b}$. That is, an optimal action for the DM in period t is

$$\arg\max_{a\in\mathcal{A}} Q^t(\boldsymbol{b}, a),$$

whereas the optimal distribution chosen by the nature, under belief state $\boldsymbol{b}$ and the DM's action a, is

$$\arg\min_{\mu_a\in\mathcal{D}_a} U^t(\boldsymbol{b}, a, \mu_a).$$

4.2 Properties of Distributionally Robust Bellman Equation (5)

We consider an ambiguity set based on mean absolute deviation of transition-observation probabilities as described below. We refer the readers to the online supplement Nakao, Hideaki and Jiang, Ruiwei and Shen, Siqian (2020) B for a more general ambiguity set that can also involve ambiguity in the reward, and the mean values are on an affine manifold with conic representable support. The same property here holds for DR-POMDP with the general ambiguity set and we omit the details for presentation simplicity.

Suppose that the expected value of the deviation of the transition-observation probability from its mean value $\bar{\boldsymbol{p}}_{as}$ is at most $\boldsymbol{c}_{as}$. Then for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, the unknown distribution μ_{as} satisfies $\mathbb{E}_{\boldsymbol{p}_{as}\sim\mu_{as}}\left[|\boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as}|\right] \leq \boldsymbol{c}_{as}$, which is reformulated as:

$$\begin{aligned} &\mathbb{E}_{(\boldsymbol{p}_{as},\tilde{\boldsymbol{u}}_{as})\sim\tilde{\mu}_{as}}\left[\tilde{\boldsymbol{u}}_{as}\right] = \boldsymbol{c}_{as}, \\ &\tilde{\mu}_{as}\left(\begin{array}{cc} \tilde{\boldsymbol{u}}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as}, & \mathbf{1}^\top \boldsymbol{p}_{as} = 1 \\ \tilde{\boldsymbol{u}}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as}, & \boldsymbol{p}_{as} \geq 0 \end{array}\right) = 1. \end{aligned}$$

Here, $\tilde{\boldsymbol{u}}_{as} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{Z}|}$ denotes a vector of auxiliary variables, and $\tilde{\mu}_{as}$ is a joint distribution of $(\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as})$. This notation is introduced to differentiate from μ_{as}, which represents the true distribution of $\boldsymbol{p}_{as}$. The ambiguity set for distribution $\tilde{\mu}_{as}$ is therefore

$$\tilde{\mathcal{D}}_{as} = \left\{ \tilde{\mu}_{as}\begin{pmatrix} \boldsymbol{p}_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \middle| \begin{array}{l} \mathbb{E}_{(\boldsymbol{p}_{as},\tilde{\boldsymbol{u}}_{as})\sim\tilde{\mu}_{as}}\left[\tilde{\boldsymbol{u}}_{as}\right] = \boldsymbol{c}_{as} \\ \tilde{\mu}_{as}\left(\mathcal{X}_{as}\right) = 1 \end{array} \right\}, \tag{8}$$

while the support $\tilde{\mathcal{X}}_{as}$ for $(\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as})$ is given by

$$\tilde{\mathcal{X}}_{as} = \left\{ \begin{pmatrix} \boldsymbol{p}_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \in \begin{array}{c} \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{Z}|} \\ \mathbb{R}^L \end{array} \middle| \begin{array}{c} \tilde{\boldsymbol{u}}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as} \\ \tilde{\boldsymbol{u}}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as} \\ \mathbf{1}^\top \boldsymbol{p}_{as} = 1 \end{array} \right\}. \tag{9}$$

For ambiguity sets and supports respectively defined in terms of (8) and (9), we show that the value function is convex with respect to the belief state $\boldsymbol{b}$ for each decision period.

Theorem 1 *For all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, let the ambiguity set and support be* (8) *and* (9)*, respectively. For all $t \in \{1, \ldots, T\}$, there exists a set Λ^t of slopes such that the value function can be expressed as follows.*

$$V^t(\boldsymbol{b}) = \max_{\boldsymbol{\alpha}\in\Lambda^t} \boldsymbol{\alpha}^\top \boldsymbol{b}. \tag{10}$$

A detailed proof of Theorem 1 is shown in the online supplement Nakao, Hideaki and Jiang, Ruiwei and Shen, Siqian (2020) C. Following this result, having provided the values of a and $\boldsymbol{\alpha}_{az}$, the inner minimization in (37) can be solved efficiently using linear programming. The issue, however, is that there are possibly infinitely many elements in $\text{Conv}\left(\Lambda^{t+1}\right)$, and even if there are finitely many, the number of supporting hyperplanes $\boldsymbol{\alpha}$ inside Λ^t increases exponentially as the value functions are calculated from period $t = T$ to $t = 1$. We describe in Section 5 a heuristic search value iteration (HSVI) algorithm for efficiently computing optimal policies in DR-POMDP.

4.3 Case of Infinite Horizon

We show that the PWLC property of the value function can be extended to the case with infinite horizon. We prove the result by following the Banach fixed point theorem (see, e.g., Puterman (2014)), and show that by repeatedly updating the value function in (5), it converges to a unique function corresponding to the optimal value V^* of the infinite-horizon DR-POMDP problem.

Theorem 2 *The operator $\mathcal{L}$ defined as*

$$\mathcal{L}V(\boldsymbol{b}) = \max_{a\in\mathcal{A}} \min_{\mu_a\in\widehat{\mathcal{D}}_a} \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a}\left[\sum_{s\in\mathcal{S}} b_s\left(r_{as} + \beta\sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V\left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z)\right)\right)\right] \tag{11}$$

is a contraction for $0 < \beta < 1$.

We refer the readers to a detailed proof provided in C in the online supplement Nakao, Hideaki and Jiang, Ruiwei and Shen, Siqian (2020). Theorem 2 suggests that by employing the exact algorithm discussed in the finite horizon case, starting from any initial value function, the value function V converges to an optimal function V^* with rate β by iteratively performing the Bellman operator $\mathcal{L}$. Therefore, we can use the same solution approach to be discussed in Section 5 for handling both finite-horizon and infinite-horizon cases of DR-POMDP.

5 Solution Method

We present a variant of the HSVI algorithm proposed in Smith and Simmons (2004) (originally for solving POMDP) for efficiently computing upper and lower bounds for DR-POMDP. We maintain a set of finite number of hyperplanes $\Lambda_{\underline{V}}$, where the resulting PWLC function $\underline{V}$ bounds the true value function from below. We also maintain a set of points $\Upsilon_{\overline{V}}$ whose elements are $(\boldsymbol{b}, v)$, which is a combination of a belief $\boldsymbol{b}$ and an upper bound v of the true value function at the belief $\boldsymbol{b}$. Therefore, the resulting PWLC function $\overline{V}$ bounds the value function from above. The upper bound v corresponding to a belief $\boldsymbol{b}$ is obtained through sampling. The sampling follows a greedy strategy to close the gap between the upper bound $\overline{V}$ and the lower bound $\underline{V}$ for the belief points that are reachable from the initial belief.

Algorithm 1 Heuristic Search Value Iteration (HSVI)

1: **Input:** initial belief state $\boldsymbol{b}^0$, tolerance ϵ
2: **Initialize:** $\overline{V}$, $\underline{V}$ (see details in Section 5.1)
3: **while** $\overline{V}(\boldsymbol{b}^0) - \underline{V}(\boldsymbol{b}^0) > \epsilon$ or time limit is reached **do**
4: *DR-BoundExplore*$(\boldsymbol{b}^0, 0)$ (see details in Algorithm 2)
5: **end while**
6: **Output:** $\overline{V}$, $\underline{V}$

Algorithm 1 presents the main algorithmic steps in HSVI, where the details of Step 4 are later provided in Algorithm 2. During Step 4, one sample path of DM, the nature's action and the observation outcomes are greedily selected, and then the bounds are updated using Bellman equations. Figure 2 demonstrates how the lower bound of the value function can be described as the maximum of the lower bounding hyperplanes, and the upper bound can be described as a convex hull of the upper bounding points. Figure 3 illustrates an example of how newly discovered bounding hyperplanes and points can be used to locally update the bounds.

In Section 5.1, we explain how the upper and lower bounds of the value function are initialized (i.e., the details for Step 2), and in Section 5.2, we present an exploration strategy to close the gap to a pre-determined tolerance level. Finally, in Section 5.3, we discuss how the value functions are updated given a belief state $\boldsymbol{b}$.

5.1 Initialization

Recall the ambiguity set and support defined in (8) and (9), respectively. In the initialization step, we compute the lower bound for the true value function by taking the best action for obtaining the

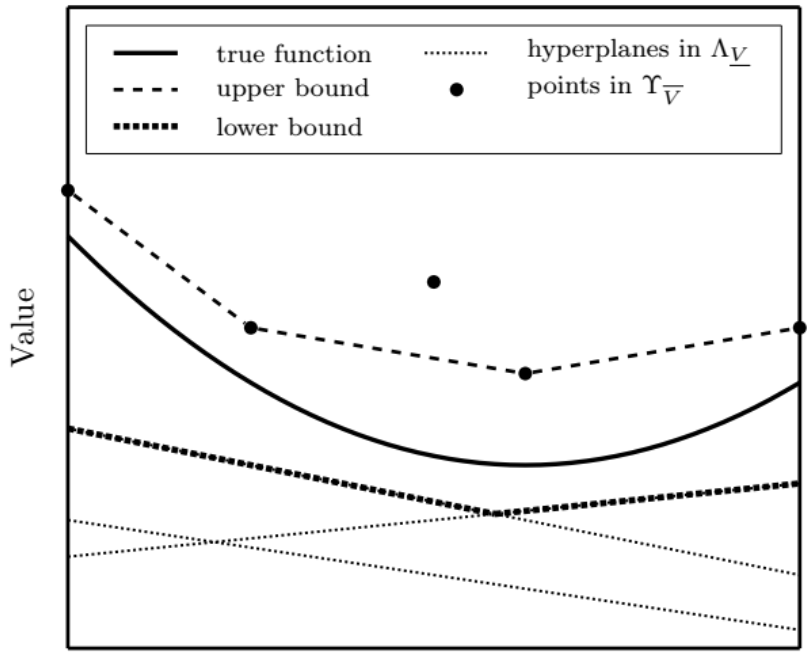


Figure 2: An example of upper- and lower-bounds of a value function

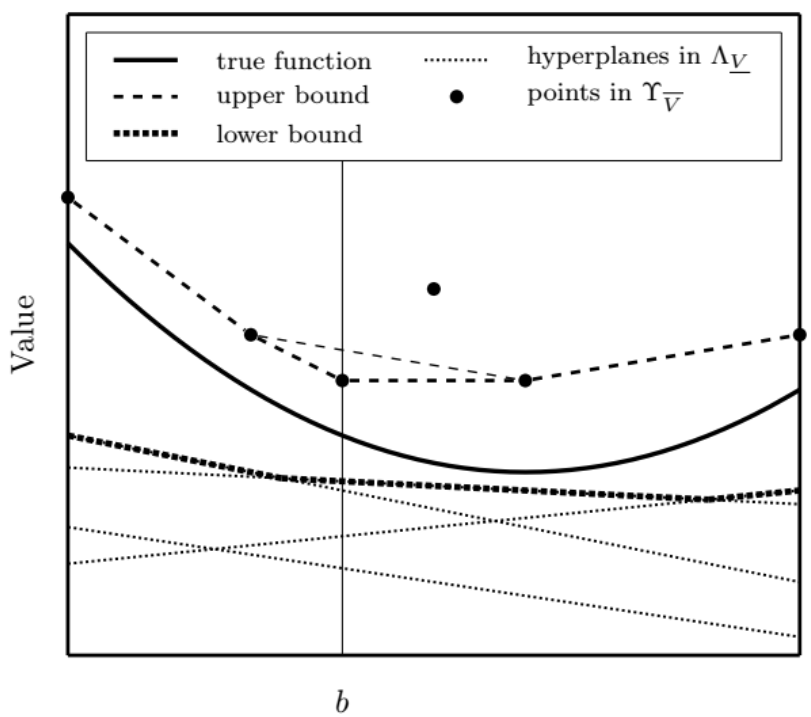


Figure 3: An example of updated upper- and lower-bounds

worst-case expected reward in each decision period. That is, for each action a, we solve

$$\underline{R}_a = \sum_{t=0}^{\infty} \beta^t \min_{s\in\mathcal{S}} \min_{\mu_{as}\in\mathcal{D}_{as}} \mathbb{E}_{(\boldsymbol{p}_{as},r_{as})\sim\mu_{as}}\left[r_{as}\right] = \frac{1}{1-\beta} \min_{s\in\mathcal{S}} \min_{\mu_{as}\in\mathcal{D}_{as}} \mathbb{E}_{(\boldsymbol{p}_{as},r_{as})\sim\mu_{as}}\left[r_{as}\right].$$

In the case of mean absolute deviation based ambiguity set (8), the second minimization is trivial as r_{as} is fixed. The minimum value for all $s \in \mathcal{S}$ is computed by enumeration. We then define an initial lower bounding hyperplane $\alpha'_s = \max_{a\in\mathcal{A}} \underline{R}_a$, $\forall s \in \mathcal{S}$ and set $\Lambda_{\underline{V}} = \{\boldsymbol{\alpha}'\}$, where $\boldsymbol{\alpha}' = (\alpha'_s, s \in \mathcal{S})^\top$.

The upper bound for the true value function is obtained by considering full observability of the system and computing the MDP for the best-case scenario in the ambiguity set. Let $\boldsymbol{V}^{MDP} \in \mathbb{R}^{|\mathcal{S}|}$ be a value function for the distributionally-optimistic MDP. It satisfies

$$V_s^{MDP} = \max_{a\in\mathcal{A}} \max_{\mu_{as}\in\mathcal{D}_{as}} \mathbb{E}_{(\boldsymbol{p}_{as},r_{as})\sim\mu_{as}}\left[r_{as} + \beta \boldsymbol{V}^{MDP\top} \sum_{z\in\mathcal{Z}} \boldsymbol{J}_z \boldsymbol{p}_{as}\right], \qquad \forall s \in \mathcal{S}.$$

To solve this, we take a linear programming approach by formulating

$$\min_{\boldsymbol{V}^{MDP}} \quad \boldsymbol{1}^\top \boldsymbol{V}^{MDP} \tag{12a}$$

$$\text{s.t.} \quad V_s^{MDP} \geq \max_{\mu_{as}\in\mathcal{D}_{as}} \mathbb{E}_{(\boldsymbol{p}_{as},r_{as})\sim\mu_{as}}\left[r_{as} + \beta \boldsymbol{V}^{MDP\top} \sum_{z\in\mathcal{Z}} \boldsymbol{J}_z \boldsymbol{p}_{as}\right], \ \forall a \in \mathcal{A}, s \in \mathcal{S}. \tag{12b}$$

In the case of ambiguity set (8), model (12) becomes

$$\min_{\boldsymbol{\rho},\boldsymbol{\kappa},\boldsymbol{V}^{MDP}} \quad \boldsymbol{1}^\top \boldsymbol{V}^{MDP} \tag{13a}$$

$$\text{s.t.} \quad V_s^{MDP} - \boldsymbol{c}_{as}^\top \boldsymbol{\rho}_{as} - \bar{\boldsymbol{p}}_{as}^\top \boldsymbol{\kappa}_{as}^1 + \bar{\boldsymbol{p}}_{as}^\top \boldsymbol{\kappa}_{as}^2 - \sigma_{as} \geq r_{as}, \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{13b}$$

$$\beta \sum_{z\in\mathcal{Z}} \boldsymbol{J}_z{}^\top \boldsymbol{V}^{MDP} - \boldsymbol{\kappa}_{as}^1 + \boldsymbol{\kappa}_{as}^2 - \boldsymbol{1}\sigma_{as} \leq 0, \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{13c}$$

$$\boldsymbol{\kappa}_{as}^1 + \boldsymbol{\kappa}_{as}^2 - \boldsymbol{\rho}_{as} = 0, \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{13d}$$

$$\boldsymbol{\kappa}_{as}^1, \kappa_{as}^2 \in \mathbb{R}_+^{|\mathcal{S}|\times|\mathcal{A}|}, \sigma_{as} \in \mathbb{R}\ \boldsymbol{\rho}_{as} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}, \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{13e}$$

$$\boldsymbol{V}^{MDP} \in \mathbb{R}^{|\mathcal{S}|}. \tag{13f}$$

After the optimal solution is discovered, we initialize $\Upsilon_{\overline{V}} = \left\{\left(\boldsymbol{e}_s, V_s^{MDP}\right), \ \forall s \in \mathcal{S}\right\}$, where $\boldsymbol{e}_s$ is a column vector with 1 in the element corresponding to s and zero elsewhere. Overall, the initialization step consists of solving a polynomial number of convex optimization problems.

To obtain $\underline{V}(\boldsymbol{b})$, we solve

$$\max\left\{\boldsymbol{\alpha}^\top \boldsymbol{b} \mid \forall \boldsymbol{\alpha} \in \Lambda_{\underline{V}}\right\}$$

by enumerating all the values of $\boldsymbol{\alpha}^\top \boldsymbol{b}$. To obtain $\overline{V}(\boldsymbol{b})$, we consider a convex combination of points $(\boldsymbol{b}^i, v^i) \in \Upsilon_{\overline{V}}$, and find a point $(\boldsymbol{b}, v)$ so that v is the smallest attainable value. That is, we let w^i be a weight corresponding to a point $(\boldsymbol{b}^i, v^i)$ and solve

$$v = \min\left\{\sum_{i\in[|\Upsilon_{\overline{V}}|]} w^i v^i \,\middle|\, \sum_{i\in[|\Upsilon_{\overline{V}}|]} w^i \boldsymbol{b}^i = \boldsymbol{b}, \ \sum_{i\in[|\Upsilon_{\overline{V}}|]} w^i = 1,\ w^i \geq 0,\ \forall i \in [|\Upsilon_{\overline{V}}|]\right\}, \tag{14}$$

where $[N]$ denotes the set $\{1, \ldots, N\}$ for some integer N.

5.2 Forward Exploration Heuristics

The forward heuristics follow from the HSVI algorithm from Smith and Simmons (2004), where the selection of a suboptimal action leads to lowering the upper bound of the value function, eventually being replaced by another action having higher upper bound. Then, the scenario of the observation is chosen such that the expected value of the gap is the highest in the child node. This process is repeated until the discounted value of the gap is smaller than a tolerance. The algorithmic steps described in this section are based on a greedy sampling strategy to close the gap between the upper and lower bounds of the value function. Samples in the simulation are branched by the DM's actions a, the nature's distribution choices μ_a, and their outcomes z and $\boldsymbol{p}_a$.

We consider the following function:

$$U_V(\boldsymbol{b}, a, \mu_a) = \mathbb{E}_{(\boldsymbol{p}_a, \boldsymbol{r}_a)\sim\mu_a}\left[\sum_{s\in\mathcal{S}} b_s \left\{ r_{as} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V\left(\boldsymbol{f}\left(\boldsymbol{b}, a, \boldsymbol{p}_a, z\right)\right)\right\}\right].$$

We can obtain $U_{\overline{V}}$ and $U_{\underline{V}}$ by letting $V = \overline{V}$ and $V = \underline{V}$, respectively.

First, we select the DM and nature's decision pair $(a^*, \mu^*_{a^*})$. The gap between $U_{\overline{V}}$ and $U_{\underline{V}}$ at belief state $\boldsymbol{b}$ is

$$\begin{aligned}
& U_{\overline{V}}(\boldsymbol{b}, a^*, \mu^*_{a^*}) - U_{\underline{V}}(\boldsymbol{b}, a^*, \mu^*_{a^*}) \\
= \;& \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \left(r_{a^*s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \overline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right)\right)\right] \\
& - \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \left(r_{a^*s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \underline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right)\right)\right] \\
= \;& \beta \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu^*_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \left(\overline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right) - \underline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right)\right)\right].
\end{aligned} \tag{15}$$

Here we describe a greedy strategy to select the branches. For a given action a, we define $\mu^*_a = \operatorname{argmin}_{\mu_a\in\tilde{\mathcal{D}}_a} U_{\underline{V}}(\boldsymbol{b}, a, \mu_a)$. Then, we let $a^* = \operatorname{argmax}_{a\in\mathcal{A}} U_{\overline{V}}(\boldsymbol{b}, a, \mu^*_a)$. We therefore have

$$\begin{aligned}
\overline{V}(\boldsymbol{b}) - \underline{V}(\boldsymbol{b}) &= \max_{a\in\mathcal{A}} \min_{\mu_a\in\tilde{\mathcal{D}}_a} U_{\overline{V}}(\boldsymbol{b}, a, \mu_a) - \max_{a\in\mathcal{A}} \min_{\mu_a\in\tilde{\mathcal{D}}_a} U_{\underline{V}}(\boldsymbol{b}, a, \mu_a) \\
&\leq \max_{a\in\mathcal{A}} U_{\overline{V}}(\boldsymbol{b}, a, \mu^*_a) - \max_{a\in\mathcal{A}} U_{\underline{V}}(\boldsymbol{b}, a, \mu^*_a) \\
&\leq U_{\overline{V}}(\boldsymbol{b}, a^*, \mu^*_{a^*}) - U_{\underline{V}}(\boldsymbol{b}, a^*, \mu^*_{a^*}).
\end{aligned} \tag{16}$$

This greedy strategy ensures that a suboptimal decision pair $(a^*, \mu^*_{a^*})$ gets replaced by better ones as updating the value functions reduces the gap.

To achieve the gap ϵ at the initial state $\boldsymbol{b}_0$, the condition for the gap at depth level t starting from the initial one is only $\epsilon\beta^{-t}$, which can readily be seen from (15) and (16). We define the difference of the gap and the required condition as the excess uncertainty, which is

$$\text{excess}(\boldsymbol{b}, t) = \overline{V}\left(\boldsymbol{b}\right) - \underline{V}\left(\boldsymbol{b}\right) - \epsilon\beta^{-t}.$$

Using (16) and applying the identity (15), we have

$$\text{excess}(\boldsymbol{b}, t) \leq \beta \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu^*_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \text{excess}(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z), t+1)\right]. \tag{17}$$

Next, we greedily choose $(z^*, p^*_{a^*})$ so that the quantity associated to the pair in right-hand side of (17) has the maximum expected value, i.e.,

$$(z^*, \boldsymbol{p}^*_{a^*}) \in \underset{z\in\mathcal{Z},\ \boldsymbol{p}_{a^*}\in\mathcal{X}_{a^*}}{\arg\max}\ \mu^*_{a^*}(\boldsymbol{p}_{a^*}) \times \sum_{s\in\mathcal{S}} b_s \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}^*_{a^*s} \text{excess}(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z), t+1). \tag{18}$$

Note that because the worst-case distribution under ambiguity set (8) is a point mass distribution, obtaining $\boldsymbol{p}^*_{a^*}$ is trivial. Algorithm 2 describes the detailed algorithmic steps. In the HSVI approach, Algorithm 2 is called recursively to make decisions on which branch to choose in the next depth level $t+1$. After the simulation is terminated, the updates on the lower and upper bounds are made for the belief states that are discovered through the simulation.

Algorithm 2 DR-BoundExplore$(\boldsymbol{b}, t)$

Input: belief state $\boldsymbol{b}$, depth level t
if $\overline{V}(\boldsymbol{b}) - \underline{V}(\boldsymbol{b}) > \epsilon\beta^{-t}$ **then**
  $(\mu^*_a,\ \forall a \in \mathcal{A}) \leftarrow \text{argmin}_{\mu_a\in\mathcal{D}_a} U_{\underline{V}}(\boldsymbol{b}, a, \mu_a)$
  $a^* \leftarrow \text{argmax}_{a\in\mathcal{A}} U_{\overline{V}}(\boldsymbol{b}, a, \mu^*_a)$
  $z^*, \boldsymbol{p}^*_{a^*} \leftarrow \text{argmax}_{z\in\mathcal{Z},\ \boldsymbol{p}_{a^*}\in\mathcal{X}_{a^*}} \mu^*_{a^*}(\boldsymbol{p}_{a^*}) \times \sum_{s\in\mathcal{S}} b_s \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}^*_{a^*s} \times \text{excess}(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z), t+1)$
  DR-BoundExplore$(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}^*_{a^*}, z^*), t+1)$
  $\Lambda_{\underline{V}} \leftarrow \Lambda_{\underline{V}} \cup$ *DR-backup*$(\boldsymbol{b}, \Lambda_{\underline{V}})$ (see the details in Algorithm 3)
  $\Upsilon_{\overline{V}} \leftarrow \Upsilon_{\overline{V}} \cup$ *DR-update*$(\boldsymbol{b}, \Upsilon_{\overline{V}})$ (see the details in Algorithm 4)
end if

5.3 Local Updates

In this section, we describe the details of *DR-backup* and *DR-update* steps in Algorithm 2. We first illustrate how the lower bound is updated in *DR-backup*. For each $a \in \mathcal{A}$, we solve the two inner maximization problems in (36) provided a and $\boldsymbol{b}$, where we set $\Lambda^{t+1} = \Lambda_{\underline{V}}$. The convex hull of $\Lambda_{\underline{V}}$ is therefore,

$$\text{Conv}\left(\Lambda_{\underline{V}}\right) = \left\{ \sum_{i\in[|\Lambda_{\underline{V}}|]} w^i \boldsymbol{\alpha}^i \,\middle|\, \sum_{i\in[|\Lambda_{\underline{V}}|]} w^i = 1,\ \boldsymbol{\alpha}^i \in \Lambda_{\underline{V}},\ w^i \geq 0,\ i \in [|\Lambda_{\underline{V}}|] \right\}. \tag{19}$$

Thus, we combine the two inner maximization problems in (36) as

$$\max_{\boldsymbol{\rho}_a, \boldsymbol{\kappa}^1_a, \boldsymbol{\kappa}^2_a, \boldsymbol{\sigma}_a} \quad \sum_{s\in\mathcal{S}} \boldsymbol{c}^\top_{as}\boldsymbol{\rho}_{as} + \sum_{s\in\mathcal{S}} b_s r_{as} + \sum_{s\in\mathcal{S}} \left(-\bar{p}^\top_{as}\boldsymbol{\kappa}^1_{as} + \bar{p}^\top_{as}\boldsymbol{\kappa}^2_{as} + \sigma_{as}\right) \tag{20a}$$

$$\text{s.t.} \quad \beta b_s \sum_{z\in\mathcal{Z}} \sum_{i\in[|\Lambda_{\underline{V}}|]} w^i_{az} \boldsymbol{J}_z{}^\top \boldsymbol{\alpha}^i_{az} + \boldsymbol{\kappa}^1_{as} - \boldsymbol{\kappa}^2_{as} - \mathbf{1}\sigma_{as} \geq 0, \qquad \forall s \in \mathcal{S} \tag{20b}$$

$$\sum_{i\in[|\Lambda_{\underline{V}}|]} w^i_{az} = 1, \qquad \forall z \in \mathcal{Z} \tag{20c}$$

$$w^i_{az} \in \mathbb{R}_+, \qquad \forall i \in [|\Lambda_{\underline{V}}|],\ z \in \mathcal{Z} \tag{20d}$$

$$(35c), (35d), (36b).$$

We denote the optimal solutions to (20) using a superscript $\star$, and let the optimal dual solutions associated with constraints (20b) be $\hat{\boldsymbol{p}}^\star_{as}$. For each action $a \in \mathcal{A}$, we can generate a lower bounding hyperplane

$$\boldsymbol{\alpha}' = \left(r_{as} + \beta \sum_{z\in\mathcal{Z}} \boldsymbol{\alpha}^{\star\top}_{az} \boldsymbol{J}_z \hat{\boldsymbol{p}}^\star_{as},\ s \in \mathcal{S} \right)^\top, \tag{21}$$

where $\boldsymbol{\alpha}^{\star}_{az} = \sum_{i\in[N]} w^{i\star}_{az}\boldsymbol{\alpha}^{i}_{az}$. We present the detailed algorithmic steps in Algorithm 3.

Algorithm 3 DR-backup($\boldsymbol{b}, \Lambda_{\underline{V}}$)

Input: belief $\boldsymbol{b}$, lower bounding hyperplanes $\Lambda_{\underline{V}}$
for $\forall a \in \mathcal{A}$ **do**
  solve (20) for action a
  $\mathcal{L}(a) \leftarrow \boldsymbol{\alpha}'$ (calculated using (21))
end for
Output: $\text{argmax}_{\boldsymbol{\alpha}\in\mathcal{L}} \boldsymbol{\alpha}^{\top}\boldsymbol{b}$

Next, we discuss how to update the upper bound and describe the algorithmic steps of *DR-update* in Algorithm 4. Combining (36) and the dual representation of (14), for each $a \in \mathcal{A}$, we solve

$$\max_{\boldsymbol{\rho}_a,\boldsymbol{\kappa}^1_a,\boldsymbol{\kappa}^2_a,\boldsymbol{\sigma}_a} \quad \sum_{s\in\mathcal{S}} \boldsymbol{c}^{\top}_{as}\boldsymbol{\rho}_{as} + \sum_{s\in\mathcal{S}} b_s r_{as} + \sum_{s\in\mathcal{S}} \left(-\bar{p}^{\top}_{as}\boldsymbol{\kappa}^1_{as} + \bar{p}^{\top}_{as}\boldsymbol{\kappa}^2_{as} + \sigma_{as}\right) \tag{22a}$$

$$\text{s.t.} \quad \beta b_s \sum_{z\in\mathcal{Z}} \boldsymbol{J}_z{}^{\top}\varphi_{az} + \beta b_s \sum_{z\in\mathcal{Z}} \psi_{az}\boldsymbol{J}_z{}^{\top}\mathbf{1} + \boldsymbol{\kappa}^1_{as} - \boldsymbol{\kappa}^2_{as} - \mathbf{1}\sigma_{as} \geq 0, \qquad \forall s \in \mathcal{S} \tag{22b}$$

$$\boldsymbol{b}^{i\top}\varphi_{az} + \psi_{az} \leq v_i, \qquad \forall z \in \mathcal{Z}, i \in [|\Upsilon_{\overline{V}}|] \tag{22c}$$

$$\varphi_{az} \in \mathbb{R}^{|\mathcal{S}|},\ \psi_{az} \in \mathbb{R}, \qquad \forall z \in \mathcal{Z}, i \in [|\Upsilon_{\overline{V}}|] \tag{22d}$$

$$(35c), (35d), (36b).$$

Here φ_{az} and ψ_{az} are the dual variables associated with the two sets of constraints, $\sum_{i\in[|\Upsilon_{\overline{V}}|]} w^i\boldsymbol{b}^i = \boldsymbol{b}$, $\sum_{i\in[|\Upsilon_{\overline{V}}|]} w^i = 1$, respectively. The maximum objective value among all $a \in \mathcal{A}$ is added to $\Upsilon_{\overline{V}}$.

Algorithm 4 DR-update($\boldsymbol{b}, \Upsilon_{\overline{V}}$)

Input: belief $\boldsymbol{b}$, upper bounding points $\Upsilon_{\overline{V}}$
for $\forall a \in \mathcal{A}$ **do**
  $\mathcal{Q}(a) \leftarrow$(optimal objective value of (22) for action a)
end for
Output: $(\boldsymbol{b}, \max_{a\in\mathcal{A}}\{\mathcal{Q}(a)\})$

Remark 2 *The complexity of the related algorithm presented in Smith and Simmons (2004) is based on the finiteness of the scenario tree up to a tolerance level ϵ. In the DR-HSVI algorithm, the scenario tree is not finite as the nature is able to choose from a continuous ambiguity set of distributions, and therefore the scenario tree has an infinite number of elements. Later we numerically demonstrate the convergence of the DR-HSVI algorithm in Section 6 for different combinations of parameter choices.*

6 Numerical Studies

We test DR-POMDP policies for dynamic epidemic control (Sections 6.1 and 6.2), and compare the results of a two-state epidemic control problem with the ones given by POMDP and robust POMDP (Section 6.1.1). We vary parameter choices to test the robustness and sensitivity of DR-POMDP policies (i) under various types of ambiguity sets used in the in-sample tests (Sections 6.1.2, 6.1.3) and (ii) given certain noise added to the transition-observation probability value obtained at the end of each decision period in out-of-sample tests (Sections 6.1.4, 6.1.5). In Sections 6.2.1 and 6.2.2, we increase the sizes of the two-state influenza epidemic control instances in Section 6.1,

demonstrate the algorithmic convergence, and present computational time results of using POMDP and DR-POMDP for solving larger-scale epidemic control instances.

6.1 Two-state Influenza Epidemic Control Problem

We study the problem of influenza epidemic control mentioned in Section 3. In the base setting, we consider two states, epidemic (E) and non-epidemic (N), and four actions as $a \in \{$Level 0, Level 1, Level 2, Inspection$\}$. Here Level 0 corresponds to the minimum disease prevention and intervention plan, e.g., doing nothing, while Level 2 corresponds to the most restrictive strategy. The "Inspection" action refers to the same disease-control strategy as the Level 0 action, except that the DM pays extra cost to improve the observation of disease spread to obtain more accurate ILI rate.

For actions $a \in \{0, 1, 2\}$, the transition probability matrix is given by

$$\begin{pmatrix} 0.99 - 0.1a & 0.01 + 0.1a \\ 0.3 - 0.1a & 0.7 + 0.1a \end{pmatrix}. \tag{23}$$

When $a = 0$ (i.e., the DM does nothing), the above transition probabilities follow studies on influenza epidemics (see, e.g., Le Strat and Carrat (1999)). The setting of the matrix (23) indicates that higher-level actions (i.e., more restrictive control strategies) will lead to greater chances that an epidemic state turns into non-epidemic and that a non-epidemic state remains itself. The transition probability for $a =$ 'Inspection' ('I') is the same as the one for $a = 0$. The observation outcome is the ILI rate, calculated as the number of ILI patients per 1000 population. For actions $a \in \{0, 1, 2\}$, we follow Rath et al. (2003) and assume that the ILI rate follows a Gaussian distribution with mean value $\mu_E = 2 - 0.5a$ and variance $\mathrm{Var}_E = 30 - \mu_E^2$ for $s =$ 'Epidemic' ('E'), and with mean $\mu_N = 0.2 - 0.05a$ and variance $\mathrm{Var}_N = 2 - \mu_N^2$ for $s =$ 'Non-epidemic' ('N'). We discretize the observation outcome into five levels as $\{(-\infty, 0], (0, 1/3], (10/3, 20/3], (20/3, 10], (10, \infty)\}$. For $a =$ 'I', the probabilities of observing the five outcomes are $\{0.01, 0.1/3, 0.1/3, 0.1/3, 0.89\}$ when $s =$ 'E', and the ILI rate follows the same distribution as the one of $a = 0$ if $s =$ 'N', to model the situation where more careful inspection action can result in more ILI patients showing up. The rewards for each action-state combination are presented in Table 1, reflecting the negative number of total infections minus the effort paid for different actions in different states.

Table 1: Reward setting for each state-action pair

State/Action	Level 0	Level 1	Level 2	Inspection
Epidemic	-100	-50	-25	-110
Non-epidemic	0	-20	-40	-20

When implementing the HSVI algorithm in Section 5 for solving DR-POMDP, we set the discount factor $\beta = 0.95$ and the gap tolerance $\epsilon = 1.0$. The computation is terminated when the gap between the upper and lower bounds is less than ϵ, at the initial states $b_E^0 = 0.5$, $b_N^0 = 0.5$. We code the algorithm in Python and execute all the tests on a computer with Intel Core i5 CPU running at 2.9 GHz and 8 GB of RAM. We solve all the linear programming models using the Gurobi solver. Note that the complexity of computing the lower bound is linear in the number of elements in $\Lambda_{\underline{V}}$, and the complexity of computing the upper bound is polynomial in the size of set $\Upsilon_{\overline{V}}$ as we need to solve linear programs. Both $|\Lambda_{\underline{V}}|$ and $|\Upsilon_{\overline{V}}|$ increase monotonically, but most elements in the two sets are dominated by others. We follow a heuristic to prune all the dominated elements whenever the number of elements increases by 10%.

6.1.1 Policy Comparison

We compare DR-POMDP policies with the ones by POMDP and robust POMDP via cross testing. We randomly generate ten samples of the transition probability for Level 2 action (i.e., $a = 2$) and epidemic state (i.e., $s =$ 'E'), by keeping all the values the same as the base setting in (23) but letting the probability $p_2(N|E) = 0.99 - 0.1 \times 2 + 0.1 \times x$, where x follows a standard Normal distribution. (We make sure that $0 \leq p_2(N|E) \leq 1$ and re-sample if not.) For all three approaches, the mean value of the ten samples is used as the nominal transition probability. For robust POMDP, the maximum L1 norm from the mean defines an uncertainty set centered around the nominal probability. For DR-POMDP, we use the mean absolute deviation to define the ambiguity set.

Table 2: Estimated median values of the cross-tested rewards

	Nature's policy		
DM's policy	POMDP(std)	DR-POMDP(std)	Robust(std)
POMDP	**−541.22** (1.08)	−609.63 (0.93)	−597.06 (2.19)
DR-POMDP	−559.02 (0.95)	−589.93 (0.92)	**−594.30** (1.31)
Robust	−570.16 (1.44)	**−585.99** (1.22)	−597.75 (1.18)

Table 3: Estimated five-percentile values of the cross-tested rewards

	Nature's policy		
DM's policy	POMDP(std)	DR-POMDP(std)	Robust(std)
POMDP	**−656.99** (2.39)	−696.34 (1.34)	−711.14 (1.43)
DR-POMDP	−669.26 (2.35)	**−677.87** (1.95)	−705.61 (1.60)
Robust	−689.26 (1.78)	−691.77 (2.07)	**−698.93** (2.19)

We implement the DM's optimal polices given by different approaches in out-of-sample environments where the nature follows the settings of POMDP, DR-POMDP, and robust POMDP to realize the transition probabilities in each period. The number of simulated instances is 5000 each. We report the estimated value of the median and the 5-percentile values of the reward in each case in Tables 2 and 3, respectively using Harrell-Davis quantile estimator Harrell and Davis (1982). We also include the standard deviation of the estimator. Note that the 5-percentile of the reward is equivalent to the 95-percentile of the cost, indicating the tail (worse) performance of different policies. Therefore, Tables 2 and 3 indicate that POMDP has the smallest reward when the nature agrees with the DM to pick the nominal transition probabilities at each decision period, but it can lead to much worse reward (both in terms of the mean value and tail performance) if the transition probabilities are realized as the worst-case (in robust POMDP) or from the worst-case distribution (in DR-POMDP). On the other hand, the performance of DR-POMDP solutions is quite stable and robust under all out-of-sample circumstances but the tail performance is worse than the mean results. Lastly, the robust POMDP policy yields worse mean value and tail performance when the true environment is POMDP or DR-POMDP.

6.1.2 Results of Varying Ambiguity Set Sizes

We first only consider an ambiguity in the transition-observation probabilities of Level 0 action and epidemic state. We build the ambiguity set based on the mean absolute deviation such that $\mathbb{E}_{\boldsymbol{p}_{as} \sim \mu_{as}} [|\boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as}|] \leq \boldsymbol{c}_{as}$ for $a = 0$ and $s =$ 'E', where $\bar{\boldsymbol{p}}_{as} \in \Delta(\mathcal{S} \times \mathcal{Z})$ is the mean value of given

probability samples and $\boldsymbol{c}_{as} \in \mathbb{R}^{|\mathcal{S}\times\mathcal{Z}|}$. We let $\boldsymbol{c}_{as}$ be $c \cdot \mathbf{1}$ for some $c \in \mathbb{R}$ and vary the values of c in our tests to vary the size of the ambiguity set.

We vary $c = 0.03, 0.06, 0.09$ for DR-POMDP and also compute the POMDP policy using $\bar{\boldsymbol{p}}_{as}$ as the transition-observation probabilities for all a and s, which corresponds to a special case of DR-POMDP with $c = 0.00$. Figure 4 depicts the upper bound (dashed line) and the lower bound (solid line) of the value functions of POMDP and DR-POMDP, as well as optimal actions corresponding to different beliefs of the epidemic. The region of the belief in red (horizontal shade) corresponds to Level 0 action, blue (dotted shade) to Level 1 action, green (cross shade) to Level 2 action, and white (diagonal shade) to Inspection action. Because the ambiguity is in the transition-observation probabilities related to $a = 0$, in all the subfigures, as compared to POMDP, the DR-POMDP policy relies less on Level 0 action and replaces it with the 'Inspection' action when the belief of epidemic is relatively higher. When the belief increases further, both DR-POMDP and POMDP agree on implementing Level 1 or Level 2 action. As the ambiguity set size increases (i.e., c increases), the DR-POMDP policy becomes more conservative and shifts to the 'Inspection' action earlier, even in relatively low belief of epidemic.

6.1.3 Results of Multiple Ambiguities

Next, we increase the number of action-state pairs that have distributional ambiguity in the transition-observation probabilities. We use $c = 0.05$ for all ambiguity sets and vary the number of action-state pairs among $\{2, 3, 4, 5\}$. In Figure 5a, action-state pairs (Level 0, E) and (Level 0, N) have ambiguous probability distributions and then we add pairs (Level 1, E), (Level 1, N), and (Level 2, E) one by one in the subsequent Figures 5b, 5c, 5d.

We observe that the reward becomes smaller as we increase the number of action-state pairs with distributional ambiguity. This is because the worst-case scenario is considered jointly for all action-state pairs and the DR-POMDP policy aims to achieve a conservative reward outcome. Moreover, the belief range where Level 1 action is taken becomes smaller as we consider the distributional ambiguity in the transition-observation probabilities associated with $a = 1$. The 'Inspection' action also replaces the Level 0 action as we increase the number of ambiguity sources.

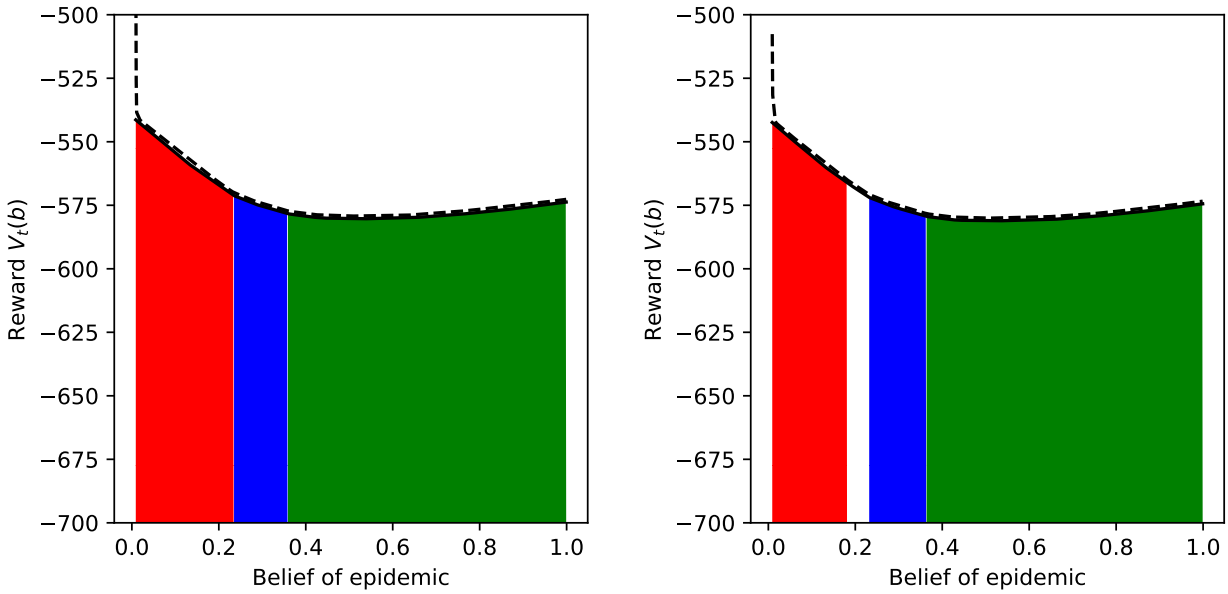
6.1.4 Solution Robustness under Different Ambiguity Sets

We simulate the DR-POMDP policies on instances with an initial state 'E' chosen with probability 50%. We use different sizes of ambiguity sets for the nature to choose the worst-case distributions in the in-sample computation. Specifically, we consider $c = 0.03, 0.06, 0.09$ to compute DR-POMDP policies using the ambiguity setting in Section 6.1.2 and then vary $c' = 0.00, 0.03, 0.06, 0.09$ to change the nature's ambiguity set size for testing each DR-POMDP policy.

Figure 6 presents the statistics of the reward, including mean, standard deviation, 5-percentile and 95-percentile values, by implementing the DR-POMDP policies in in-sample tests when the nature uses different sizes of ambiguity sets to choose the worst-case distribution for the transition-observation probabilities. We observe that DR-POMDP policies are robust and not sensitive to the ambiguity set size change, especially in the mean, worst and best reward values.

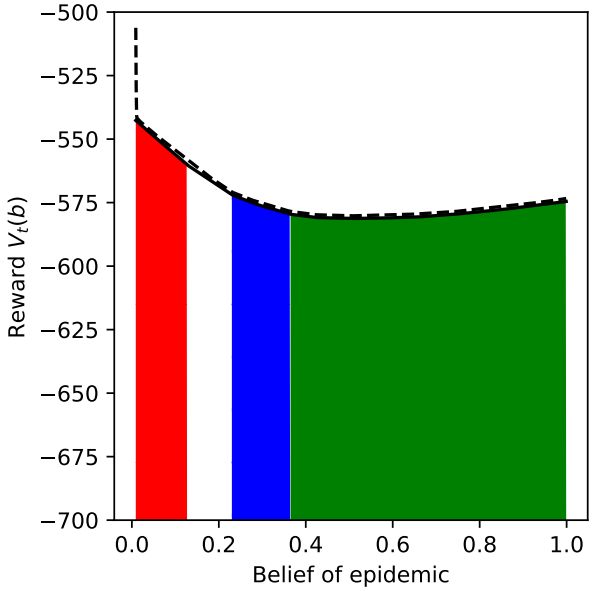
6.1.5 Solution Sensitivity under Noise Added to the Realized Transition-Observation Probabilities

We argue that our assumption about the true transition-observation probabilities being accessible at the end of each decision period is relatively weak, by testing the DR-POMDP policies in out-of-sample scenarios while adding noise to the $\boldsymbol{p}$-value obtained at the end of each period. Specifically,

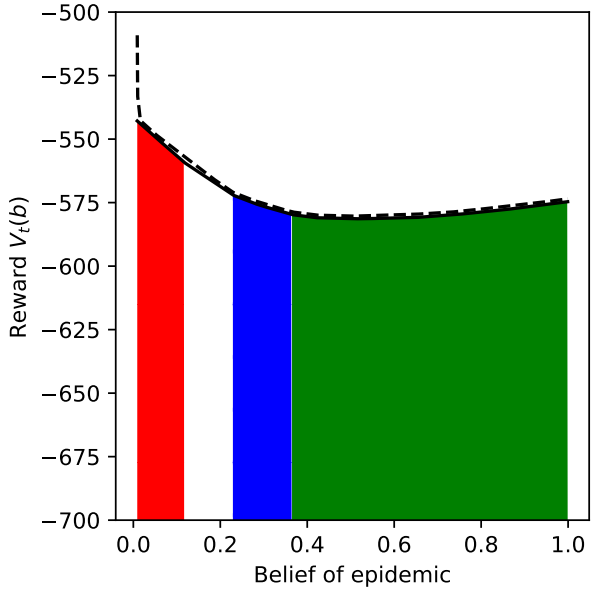


(a) POMDP ($c = 0.00$)

(b) DR-POMDP ($c = 0.03$)

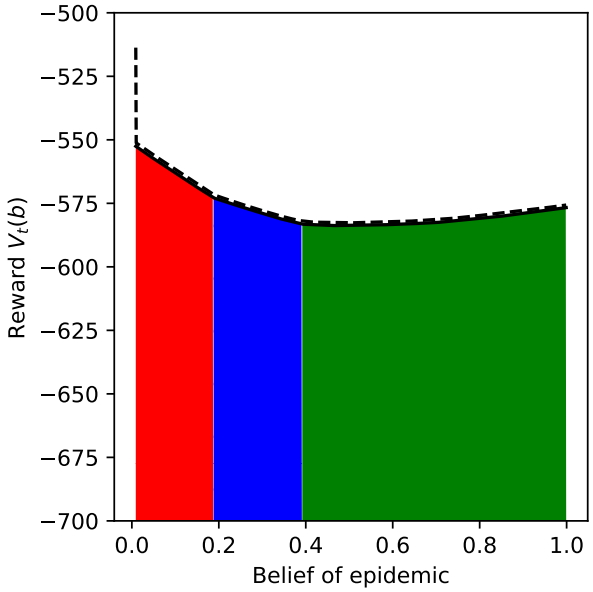


(c) DR-POMDP ($c = 0.06$)

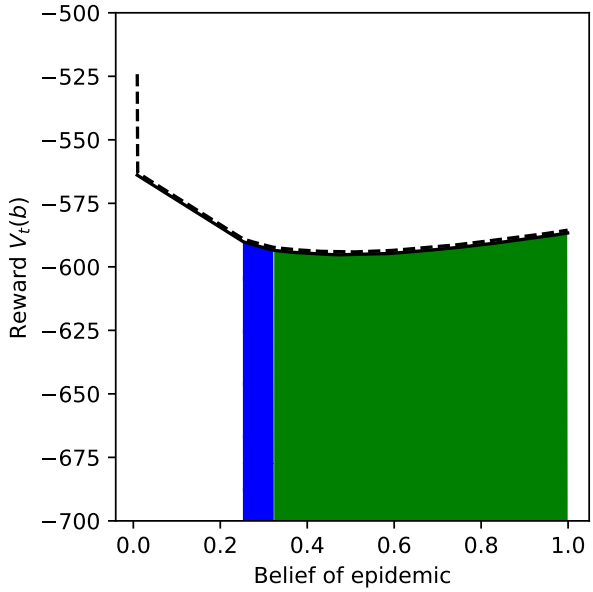


(d) DR-POMDP ($c = 0.09$)

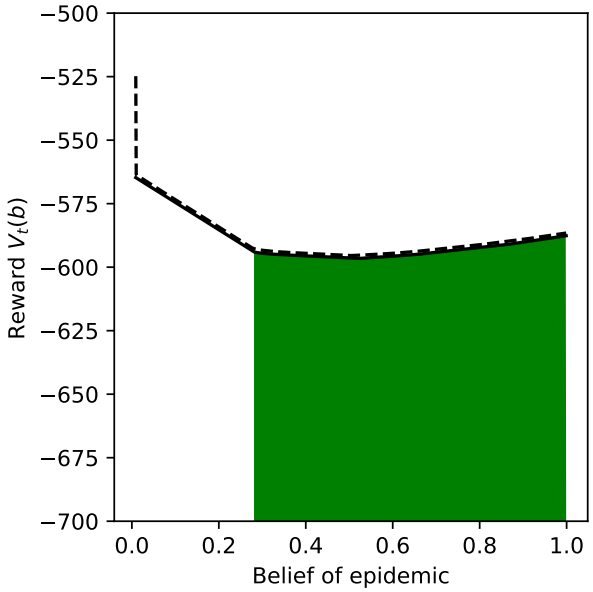
Figure 4: Value functions for different ambiguity-set sizes. Solid line: lower bound, dashed line: upper bound. Corresponding actions: Level 0 – (red, horizontal), Level 1 – (blue, dot), Level 2 – (green, cross), Inspection – (white, diagonal)



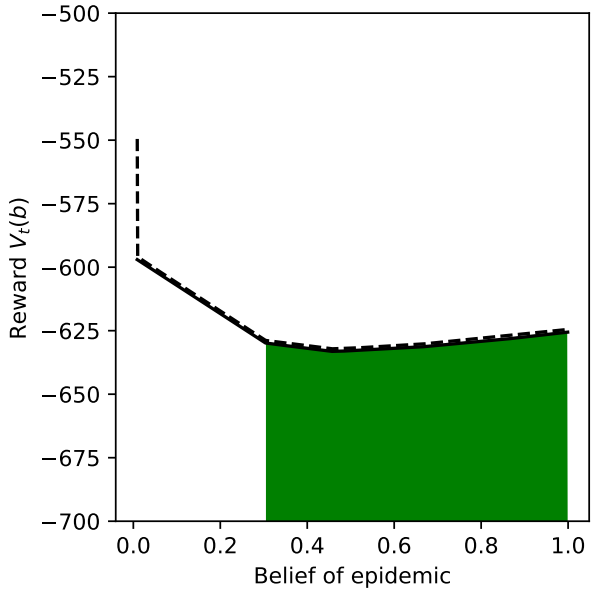
(a) {(Level 0, E), (Level 0, N)}



(b) {(Level 0, E), (Level 0, N), (Level 1, E)}

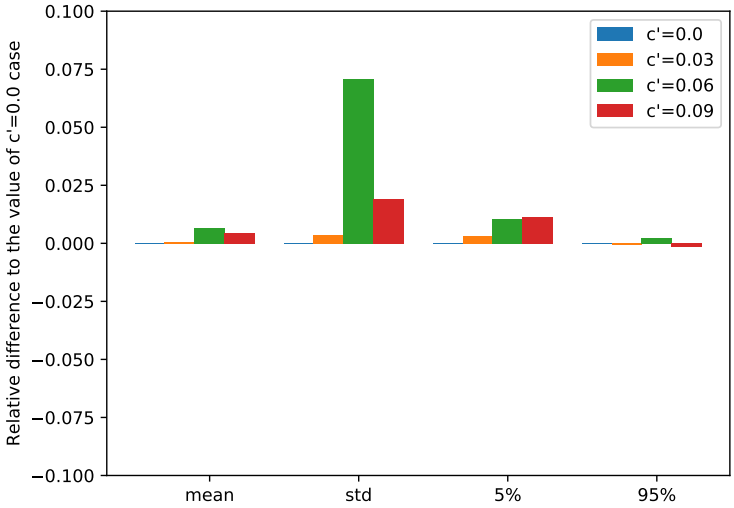


(c) {(Level 0, E), (Level 0, N), (Level 1, E), (Level 1, N)}

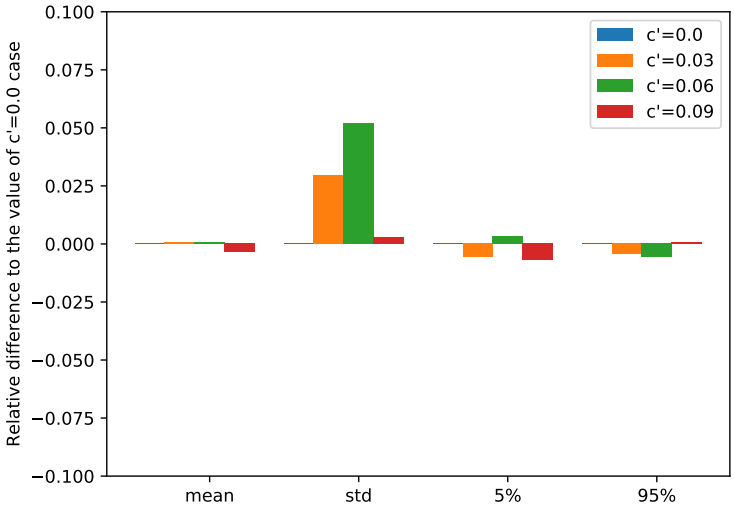


(d) {(Level 0, E), (Level 0, N), (Level 1, E), (Level 1, N), (Level 2, E)}

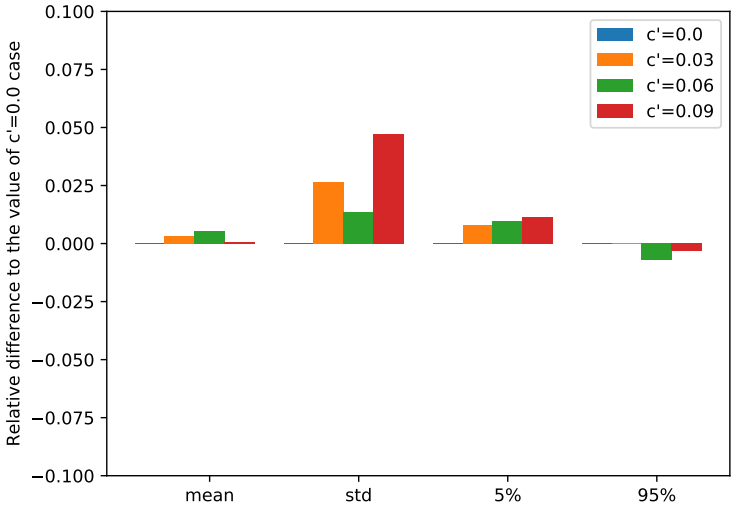
Figure 5: Value functions for increasing number of action-state pairs with distributional ambiguity. Solid line: lower bound, dashed line: upper bound. Corresponding actions: Level 0 – (red, horizontal), Level 1 – (blue, dot), Level 2 – (green, cross), Inspection – (white, diagonal)



(a) DR-POMDP ($c = 0.03$)



(b) DR-POMDP ($c = 0.06$)



(c) DR-POMDP ($c = 0.09$)

Figure 6: Statistics of the reward (mean, standard deviation, 5-percentile, 95-percentile) obtained by implementing DR-POMDP policies in in-sample tests under different ambiguity sets used by the nature.

when the DM takes Level 0 action, the transition probability of switching from an epidemic state to a non-epidemic state follows $p_0(N|E) = 0.99 + e \cdot x$, where $e \in \{0.0, 0.1, 0.2, 0.3\}$, and x follows a standard Normal distribution. (We ensure that $0 \leq p_0(N|E) \leq 1$ and re-sample if not.)

Figure 7 presents the statistics of the reward, including mean, standard deviation, 5-percentile and 95-percentile values, by implementing the DR-POMDP policies in out-of-sample scenarios under varying $\boldsymbol{p}$-values obtained at the end of each decision period. Similar to the previous section, we compare the reward statistics with the case when $e = 0.0$, i.e., the case when the DM can fully access the true $\boldsymbol{p}$-value at the end of each period. For different ambiguity sets ($c = 0.03, 0.06, 0.09$), the DR-POMDP solutions are not sensitive to the perturbation of $\boldsymbol{p}$-values obtained at the end of each period as we increase the noise. Moreover, all the statistics are within less than 2.5% differences from the results of $e = 0.0$, indicating that our assumption about the necessity of using side information to obtain the true $\boldsymbol{p}$-value at the end of each period is not strong.

6.2 Large-scale Dynamic Epidemic Control Problem

We demonstrate the algorithmic convergence and compare the computational-time difference for larger-sized instances when applying the HSVI algorithm. We increase the problem size and instance diversity by extending the previous two-state model. Specifically, we consider people who are susceptible to infection and people who have recovered, so that we can model the variation and dynamics in the infection rate. We utilize the SIR compartmental model in epidemiology (see Hethcote, 2000; Harko et al., 2014), where S, I, R represent the susceptible, infected and recovered population ratios, respectively. These quantities can be modeled using differential equations:

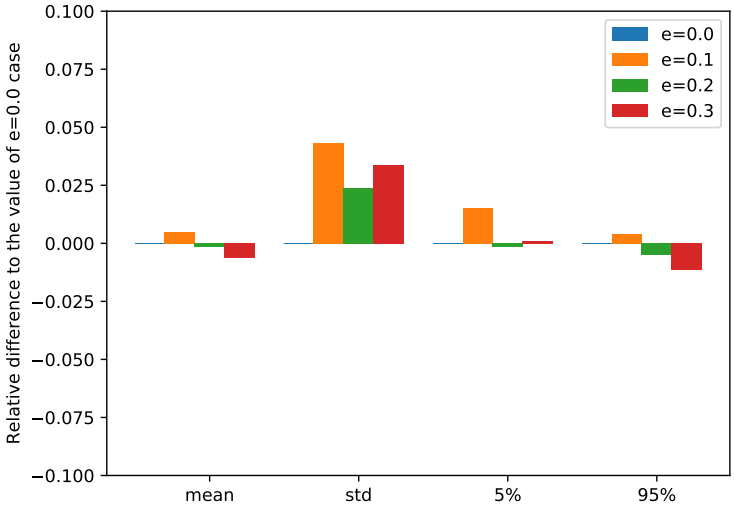
$$\begin{aligned}
\frac{dS(t)}{dt} &= -a_1 I(t) S(t), \\
\frac{dI(t)}{dt} &= a_1 I(t) S(t) - a_0 I(t), \\
\frac{dR(t)}{dt} &= a_0 I(t),
\end{aligned}$$

where a_0 is the rate of recovery, and a_1 is the average number of contacts per person per time. In this problem setting, we assume that these quantities can be controlled by the DM. We discretize the time horizon and consider discretized states $\tilde{S}$, $\tilde{I}$, $\tilde{R}$. Furthermore, we take a first-order approximation and define the transition probabilities such that they satisfy

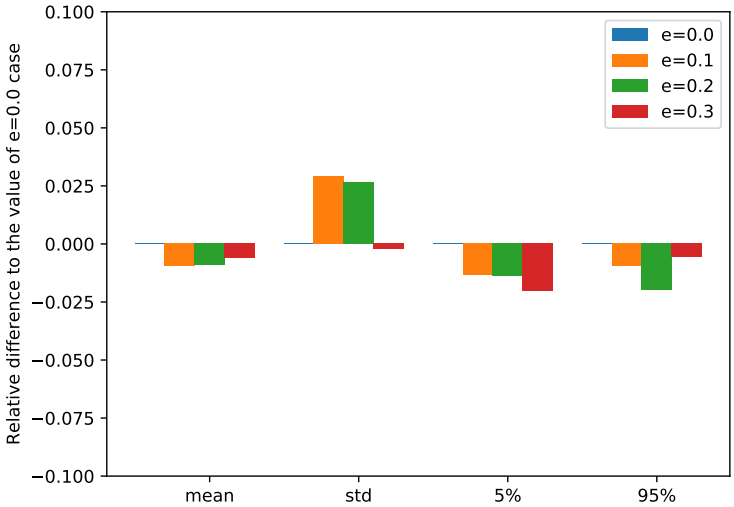
$$\begin{aligned}
\mathbb{E}\left[\tilde{S}^{t+1} | \tilde{S}^t\right] &= \tilde{S}^t - a_1 \tilde{I}^t \tilde{S}^t dt, \\
\mathbb{E}\left[\tilde{I}^{t+1} | \tilde{I}^t\right] &= \tilde{I}^t + a_1 \tilde{I}^t \tilde{S}^t dt - a_0 \tilde{I}^t dt, \\
\mathbb{E}\left[\tilde{R}^{t+1} | \tilde{R}^t\right] &= \tilde{R}^t + a_0 \tilde{I}^t dt.
\end{aligned}$$

We further assume that the states can only transition to its neighboring states, and the quantity of $\tilde{S}$ cannot increase. (Similarly, the quantity of $\tilde{R}$ cannot decrease.) We assume $dt = 1$ in the subsequent discussion.

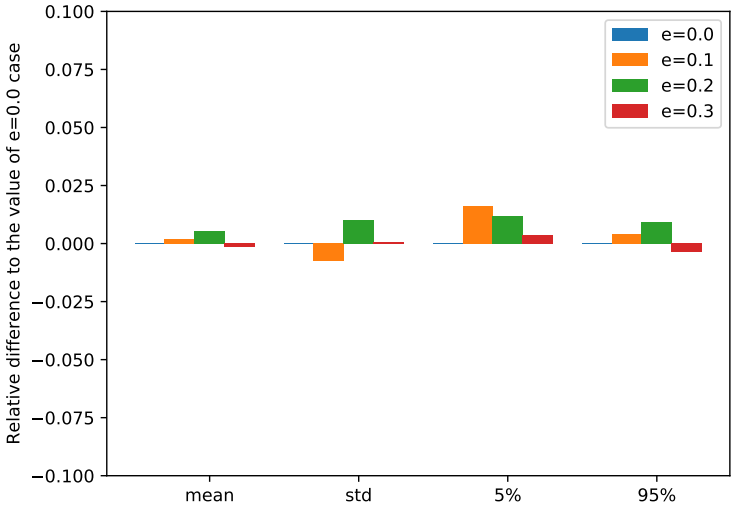
The DM is able to make an imperfect observation of the state $\tilde{I}^t$. The outcome of the observation is typically less than or equal to the true state $\hat{I}$, and the accuracy depends on the quality of the test. We assume that the observation outcome follows a Normal distribution with mean $a_2 \times \hat{I}$ (with a_2 being a parameter that the DM can control) and standard deviation $0.25 \times \hat{I}$, and is further discretized by allocating the probability mass to the closest discrete observation outcome.



(a) DR-POMDP ($c = 0.03$)



(b) DR-POMDP ($c = 0.06$)



(c) DR-POMDP ($c = 0.09$)

Figure 7: Statistics of the reward (mean, standard deviation, 5-percentile, 95-percentile) obtained by performing DR-POMDP policies in out-of-sample tests with noisy $\boldsymbol{p}$-values.

Moreover, the DM can implement certain epidemic control policies to vary $a_1 \in [0.1, 1.0]$ and $a_2 \in [0, 1]$, and we fix $a_0 = 0.25$. Choosing a low value of a_1 results in high cost due to its economic impact for a strict measure, and choosing a high value of a_2 results in high cost due to operating an expensive test process. We set the goal to minimize the number of infected people and preventing it from exceeding the treatment capacity, which is set as 0.2% of the overall population. Each percentage of population being infected will result in 10 units of cost, while 15 units of cost is incurred when the total infection is more than treatment capacity. Varying one unit of the a_1- and a_2-values costs 10 and 3 units, respectively. Additionally, when the total infection is more than 0.5% of the population, a reward = 20 will be given for performing the most strict measure in a_1. Therefore,

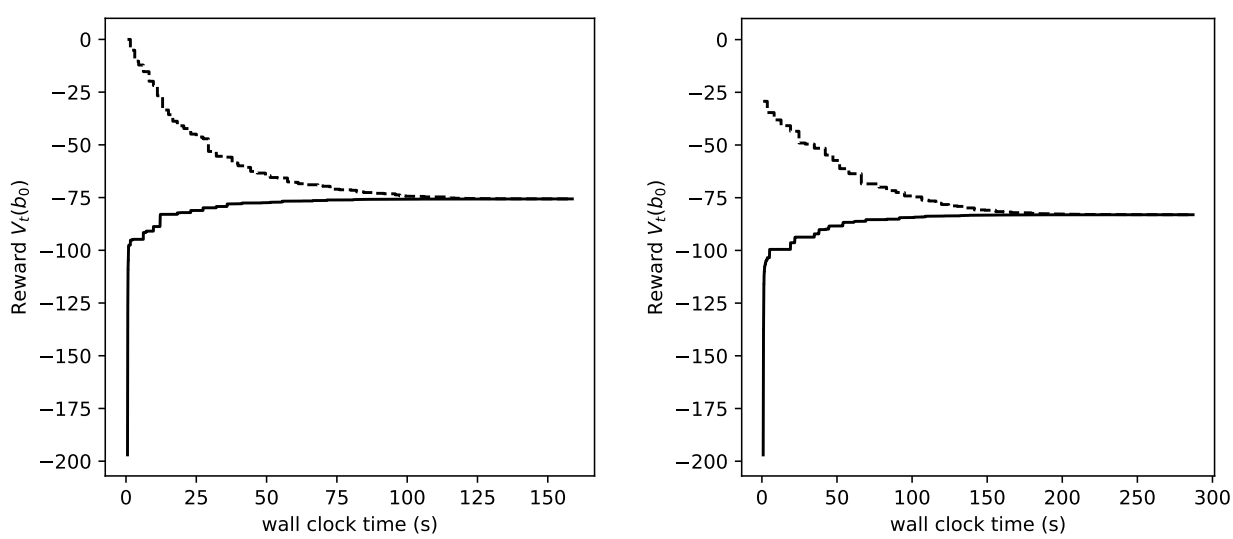
$$r_{as} = \begin{cases} -1000 \times \tilde{I} - 10 \times (1.0 - a_1) - 3 \times a_2, & \text{if } \hat{I} < 0.002 \\ -2500 \times \tilde{I} - 10 \times (1.0 - a_1) - 3 \times a_2, & \text{if } \hat{I} \geq 0.002, \end{cases}$$
$$+ 20 \text{ if } \hat{I} \geq 0.005 \text{ and } a_1 \text{ is the lowest value.}$$

where $a \in \{a_1, a_2\}$ and $s \in \{\tilde{S}, \tilde{I}, \tilde{R}\}$.

6.2.1 Computational Time for Varying Numbers of States

Let $\tilde{I} = 0.001$ and 0.005, representing the 'Non-epidemic' state and 'Epidemic' state, respectively. We consider the following discretization schemes for the states $\tilde{S}$: $\{0.90, 0.95\}$, $\{0.50, 0.70, 0.90, 0.95\}$, and $\{0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95\}$.

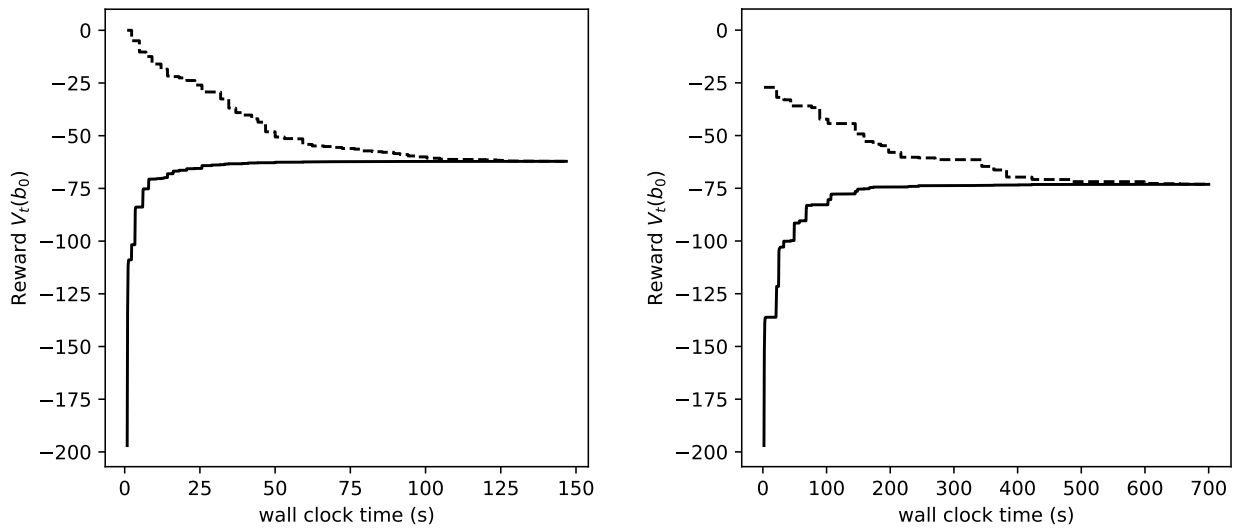
In the numerical experiment, we only consider ambiguities in the action $a_1 = 1.0$, corresponding to implementing the least strict control policy for reducing the infection rate. We set the radius of the ambiguity set as $c = 0.02$. Thus, the different problem sizes are $(s4, a4, z3, u8)$, $(s8, a4, z3, u16)$, and $(s16, a4, z3, u32)$. We set the initial belief to be totally in the non-epidemic state, and allow a tolerance $\epsilon = 1.0$. The computational time limit is 3600 seconds.



(a) POMDP ($c = 0.00$)

(b) DR-POMDP ($c = 0.02$)

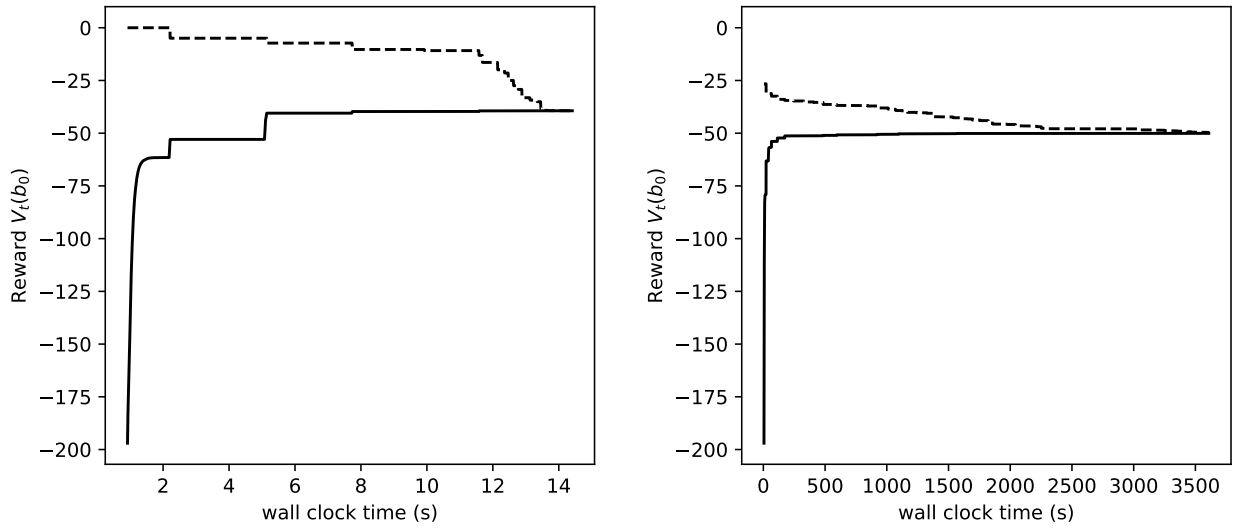
Figure 8: Dynamic epidemic control problem instance $(s4, a4, z3, u8)$. Solid line: lower bound, dashed line: upper bound



(a) POMDP ($c = 0.00$)

(b) DR-POMDP ($c = 0.02$)

Figure 9: Dynamic epidemic control problem instance $(s8, a4, z3, u16)$. Solid line: lower bound, dashed line: upper bound



(a) POMDP ($c = 0.00$)

(b) DR-POMDP ($c = 0.02$)

Figure 10: Dynamic epidemic control problem instance $(s16, a4, z3, u32)$. Solid line: lower bound, dashed line: upper bound

In Figures 8, 9, 10, we depict how the upper bound and lower bound of POMDP ($c = 0.00$) and DR-POMDP ($c = 0.02$) policies converge as functions of time for the above three problem sizes, respectively. We observe that the computational time for POMDP does not correlate with the number of states. When the number of states are 4 and 8, the corresponding instances take about 150 seconds to converge, as compared to the instances having 16 states take about 14 seconds to converge. On the other hand, the computational time for DR-POMDP increases as the number of states and ambiguity sets increase. We also point out that the value function for DR-POMDP evaluated at b_0 is lower than that of POMDP, which is expected since DR-POMDP is more conservative.

6.2.2 Computation Time for Varying Uncertainty Sizes

We change the number of ambiguity sets and compare their solutions and computation time. The states are $\tilde{S} \in \{0.50, 0.70, 0.90, 0.95\}$ and $\tilde{I} \in \{0.001, 0.005\}$, and actions are $(a_1, a_2) \in \{(0.1, 0.1), (0.1, 1.0), (1.0, 0.1), (1.0, 1.0)\}$. We increase the number of actions that are associated with ambiguity sets from 1 to 4. Since there are 8 states in total, the number of ambiguity sets are 8, 16, 32, and 64, respectively. The results are shown in Figure 11. The solution time are 614, 625, 1012, 1497 seconds, respectively and increase as the number of ambiguity sets increases. The optimal objective values are -62.64, -64.58, -71.71, -72.99, respectively, and decrease monotonically.

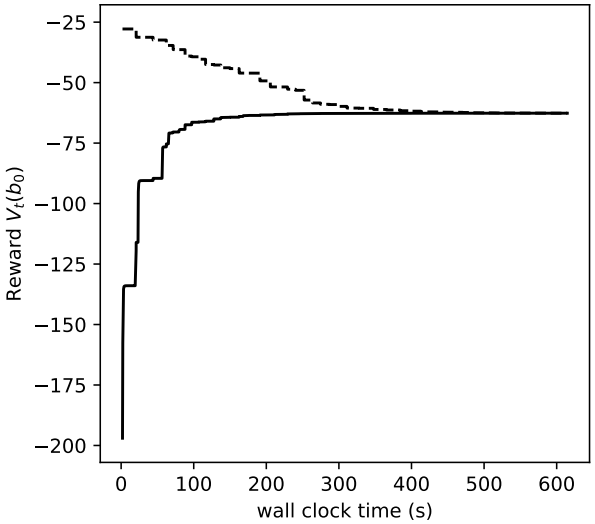
7 Conclusion

In this paper, we developed new models and algorithms for POMDP when the transition probability and the observation probability are uncertain, and the probability distribution is not perfectly known. We presented a scalable approximation algorithm and numerically compared DR-POMDP optimal policies with the ones of the standard POMDP and robust POMDP, in both in-sample and out-of-sample tests. Although due to the more complicated model and problem settings, DR-POMDP is much harder to solve, it produces more conservative and robust results than POMDP. It is also not sensitive to the misspecified ambiguity set and true transition-observation probability values obtained at the end of each decision period.

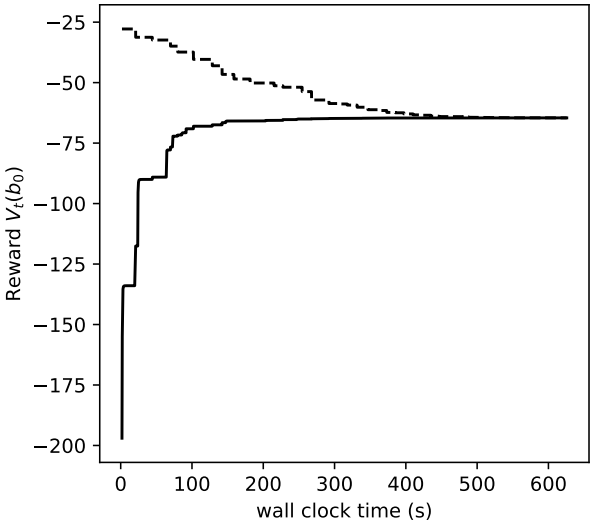
In the future research, we aim to solve DR-POMDP when the outcomes of the transition-observation probabilities are not observable to the DM at the end of each time. In such a case, the value function is dependent on a set of belief states, where the characterization of the value function becomes much more challenging. We are also interested in designing randomized policy or time-dependent policy for DR-POMDP when we relax the condition that the nature is able to perfectly observe the DM's action, or when the nature is not completely adversarial. We will compare the performance of different types of policies on diverse instances.

Acknowledgments

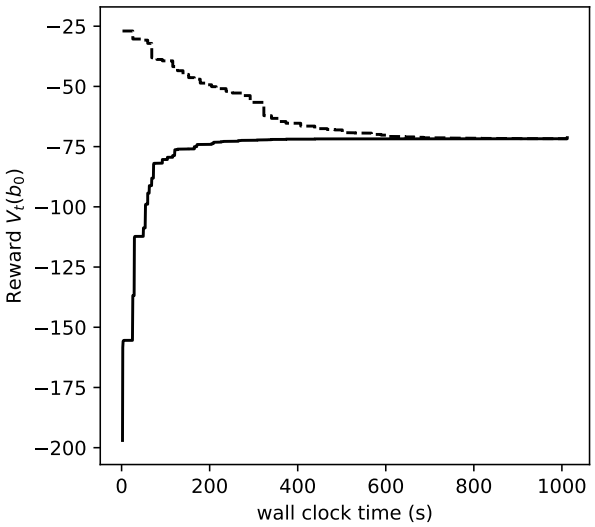
The authors thank the referees and the Associate Editor for their constructive comments and helpful suggestions. The authors gratefully acknowledge the support from the U.S. Department of Engineering (DoE) grant # DE-SC0018018 and National Science Foundation (NSF) grant # CMMI-1727618.



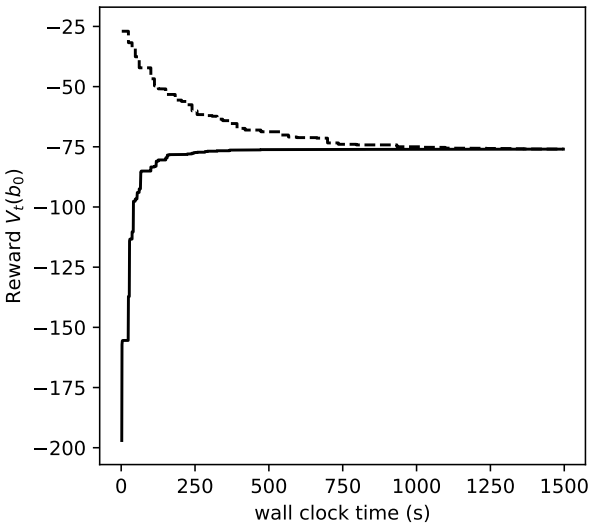
(a) DR-POMDP $(s8, a4, z3, u8)$



(b) DR-POMDP $(s8, a4, z3, u16)$



(c) POMDP $(s8, a4, z3, u32)$



(d) DR-POMDP $(s8, a4, z3, u64)$

Figure 11: Dynamic epidemic control problem instances with varying number of ambiguity sets. Solid line: lower bound, dashed line: upper bound

References

Abbad, M. and Filar, J. A. (1992). Perturbation and stability theory for Markov control problems. *IEEE Transactions on Automatic Control*, 37(9):1415–1420.

Abbad, M., Filar, J. A., and Bielecki, T. R. (1990). Algorithms for singularly perturbed limiting average Markov control problems. In *Decision and Control, 1990., Proceedings of the 29th IEEE Conference on*, pages 1402–1407. IEEE.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357.

Cassandra, A. R. (1998). A survey of POMDP applications. In *Working notes of AAAI 1998 Fall Symposium on planning with partially observable Markov decision processes*, pages 17–24.

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

Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612.

Du, D.-Z. and Pardalos, P. M. (2013). *Minimax and Applications*, volume 4. Springer Science & Business Media.

Du, X., King, A. A., Woods, R. J., and Pascual, M. (2017). Evolution-informed forecasting of seasonal influenza a (h3n2). *Science translational medicine*, 9(413):eaan5325.

Esfahani, P. M. and Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166.

Gao, R. and Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

Harko, T., Lobo, F. S., and Mak, M. (2014). Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the sir model with equal death and birth rates. *Applied Mathematics and Computation*, 236:184–194.

Harrell, F. E. and Davis, C. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3):635–640.

Hauskrecht, M. and Fraser, H. (2000). Planning treatment of ischemic heart disease with partially observable Markov decision processes. *Artificial Intelligence in Medicine*, 18(3):221–244.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM review*, 42(4):599–653.

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

Jiang, R. and Guan, Y. (2016). Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1-2):291–327.

Kumar, P. R. and Varaiya, P. (2015). *Stochastic Systems: Estimation, Identification, and Adaptive Control*, volume 75. SIAM.

Le Strat, Y. and Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden markov models. *Statistics in medicine*, 18(24):3463–3478.

Mannor, S., Mebel, O., and Xu, H. (2016). Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509.

Nakao, Hideaki and Jiang, Ruiwei and Shen, Siqian (2020). Online Supplement for "Distributionally Robust Partially Observable Markov Decision Process with Moment-based Ambiguity". `http://www-personal.umich.edu/~siqian/dataset.html`.

Nilim, A. and El Ghaoui, L. (2005). Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798.

Osogami, T. (2015). Robust partially observable Markov decision process. In *International Conference on Machine Learning (ICML)*, pages 106–115.

Pineau, J., Gordon, G., and Thrun, S. (2003). Point-based value iteration: An anytime algorithm for POMDPs. In *The Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI)*, volume 3, pages 1025–1032.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

Rasouli, M. and Saghafian, S. (2018). Robust partially observable Markov decision processes. *Working paper*.

Rath, T. M., Carreras, M., and Sebastiani, P. (2003). Automated detection of influenza epidemics with hidden markov models. In *International Symposium on Intelligent Data Analysis*, pages 521–532. Springer.

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

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

Smith, T. and Simmons, R. (2004). Heuristic search value iteration for POMDPs. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 520–527. UAI Press.

Treharne, J. T. and Sox, C. R. (2002). Adaptive inventory control for nonstationary demand and partial information. *Management Science*, 48(5):607–624.

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

Wiesemann, W., Kuhn, D., and Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376.

Xu, H. and Mannor, S. (2012). Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300.

Yang, I. (2017). A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169.

Yu, P. and Xu, H. (2016). Distributionally robust counterpart in Markov decision processes. *IEEE Transactions on Automatic Control*, 61(9):2538–2543.

Zhang, J. and Denton, B. T. (2018). Partially observable markov decision processes for prostate cancer screening, surveillance, and treatment: A budgeted sampling approximation method. *Decision Analytics and Optimization in Disease Prevention and Treatment*, pages 201–222.

Zymler, S., Kuhn, D., and Rustem, B. (2013). Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 137(1-2):167–198.

A Relaxation of a-rectangularity

In this section, we investigate a variant of DR-POMDP where we relax the rectangularity condition of the ambiguity set in the actions. So far, we have only considered the setting where the ambiguity set is rectangular in terms of the states in $\mathcal{S}$ and the actions in $\mathcal{A}$. This is known as (s, a)-rectangular set in the literature of Wiesemann et al. (2013), who defined the term in the context of robust MDP. Ref. Wiesemann et al. (2013) also considered s-rectangular set in robust POMDP, which is only rectangular in terms of the states $\mathcal{S}$. This setting has randomized policy as the optimal policy. We take a similar approach and formulate the Bellman equation:

$$V^t(\boldsymbol{b}) = \max_{\boldsymbol{\phi}\in\Delta(\mathcal{A})} \min_{\mu\in\mathcal{D}} \mathbb{E}_{P\sim\mu}\left[\sum_{a\in\mathcal{A}} \phi_a \sum_{s\in\mathcal{S}} b_s \left(r_{as} + \beta \sum_{z\in\mathcal{Z}} J_z \boldsymbol{p}_{as} V^{t+1}\left(\boldsymbol{f}\left(\boldsymbol{b}, a, \boldsymbol{p}_a, z\right)\right)\right)\right], \tag{24}$$

where ϕ_a is the probability for selecting action a. We define the ambiguity set to be

$$\tilde{\mathcal{D}}_s = \left\{ \tilde{\mu}_s \begin{pmatrix} \boldsymbol{p}_s \\ \boldsymbol{r}_s \\ \tilde{\boldsymbol{u}}_s \end{pmatrix} \middle| \begin{array}{l} \mathbb{E}_{(\boldsymbol{p}_s, \boldsymbol{r}_s, \tilde{\boldsymbol{u}}_s)\sim\tilde{\mu}_s}\left[F_s \boldsymbol{p}_s + G_s \boldsymbol{r}_s + H_s \tilde{\boldsymbol{u}}_s\right] = \boldsymbol{c}_s, \\ \tilde{\mu}_s\left(\mathcal{X}_s\right) = 1 \end{array} \right\}, \tag{25}$$

where $\tilde{\boldsymbol{u}}_s \in \mathbb{R}^Q$ is a vector of auxiliary variables, and

$$\mathcal{X}_s = \left\{ \begin{pmatrix} \boldsymbol{p}_s \\ \boldsymbol{r}_s \\ \tilde{\boldsymbol{u}}_s \end{pmatrix} \in \begin{matrix} \mathbb{R}^{|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{Z}|} \\ \mathbb{R}^{|\mathcal{A}|} \\ \mathbb{R}^L \end{matrix} \;\middle|\; B_s\boldsymbol{p}_s + C_s r_s + E_s\tilde{\boldsymbol{u}}_s \preceq_{K_s} \boldsymbol{d}_s \right\}. \tag{26}$$

Here, $F_s \in \mathbb{R}^{k\times(|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{Z}|)}$, $G_s \in \mathbb{R}^{k\times|\mathcal{A}|}$, $H_s \in \mathbb{R}^{k\times L}$, $\boldsymbol{c}_s \in \mathbb{R}^k$, $B_s \in \mathbb{R}^{\ell\times(|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{Z}|)}$, $C_s \in \mathbb{R}^{\ell\times|\mathcal{A}|}$, $E_s \in \mathbb{R}^{\ell\times L}$, and $\boldsymbol{d}_s \in \mathbb{R}^\ell$.

The value function is also convex in the form (10), since for $t < T$,

$$\begin{aligned} V^t(\boldsymbol{b}) = \max_{\boldsymbol{\phi}\in\Delta(\mathcal{A})} \max_{\substack{\boldsymbol{\alpha}_{az}\in\mathrm{Conv}(\Lambda^{t+1}) \\ \forall a\in\mathcal{A},\ z\in\mathcal{Z}}} \sum_{s\in\mathcal{S}} b_s \min_{(\hat{\boldsymbol{p}}_s,\hat{\boldsymbol{r}}_s,\hat{\tilde{\boldsymbol{u}}}_s)} \quad & \boldsymbol{\phi}^\top \left(\beta \sum_{z\in\mathcal{Z}} \left[\left(\boldsymbol{\alpha}_{az}^\top J_{az} \right)^\top,\ a\in\mathcal{A} \right]^\top \hat{\boldsymbol{p}}_s + \hat{\boldsymbol{r}}_s \right) \\ \text{s.t.} \quad & F_s\hat{\boldsymbol{p}}_s + G_s\hat{\boldsymbol{r}}_s + H_s\hat{\tilde{\boldsymbol{u}}}_s = \boldsymbol{c}_s, \qquad \forall s\in\mathcal{S} \\ & B_s\hat{\boldsymbol{p}}_s + C_s\hat{\boldsymbol{r}}_s + E_s\hat{\tilde{\boldsymbol{u}}}_s \preceq_{K_s} \boldsymbol{d}_s, \qquad \forall s\in\mathcal{S} \end{aligned}$$

where $J_{az} \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{Z}|)}$ is a matrix of zeros and ones that maps $\boldsymbol{p}_s$ to $\boldsymbol{p}_{asz}$. For an exact algorithm, we solve the inner minimization problem for all $\phi \in \Delta(\mathcal{A})$, $\boldsymbol{\alpha}_{az} \in \mathrm{Conv}(\Lambda^{t+1})$, $\forall z \in \mathcal{Z}$, $a \in \mathcal{A}$. The optimal objective is used for constructing the set Λ^t, at each time step t.

B General Ambiguity Set

In this section, we provide a general form of the ambiguity set where the mean values are on an affine manifold, and the supports are conic representable. For all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, we define a non-empty ambiguity set

$$\tilde{\mathcal{D}}_{as} = \left\{ \tilde{\mu}_{as} \begin{pmatrix} \boldsymbol{p}_{as} \\ r_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \;\middle|\; \begin{matrix} \mathbb{E}_{(\boldsymbol{p}_{as},r_{as},\tilde{\boldsymbol{u}}_{as})\sim\tilde{\mu}_{as}} \left[F_{as}\boldsymbol{p}_{as} + G_{as}r_{as} + H_{as}\tilde{\boldsymbol{u}}_{as} \right] = \boldsymbol{c}_{as}, \\ \tilde{\mu}_{as}\left(\mathcal{X}_{as}\right) = 1 \end{matrix} \right\}, \tag{27}$$

where $\tilde{\boldsymbol{u}}_{as} \in \mathbb{R}^L$ is a vector of auxiliary variables, and a support with a non-empty relative interior

$$\mathcal{X}_{as} = \left\{ \begin{pmatrix} \boldsymbol{p}_{as} \\ r_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \in \begin{matrix} \mathbb{R}^{|\mathcal{S}|\times|\mathcal{Z}|} \\ \mathbb{R} \\ \mathbb{R}^L \end{matrix} \;\middle|\; B_{as}\boldsymbol{p}_{as} + C_{as}r_{as} + E_{as}\tilde{\boldsymbol{u}}_{as} \preceq_{K_{as}} \boldsymbol{d}_{as} \right\}. \tag{28}$$

Here, $F_{as} \in \mathbb{R}^{k\times(|\mathcal{S}|\times|\mathcal{Z}|)}$, $G_{as} \in \mathbb{R}^{k\times 1}$, $H_{as} \in \mathbb{R}^{k\times L}$, $\boldsymbol{c}_{as} \in \mathbb{R}^k$, $B_{as} \in \mathbb{R}^{\ell\times(|\mathcal{S}|\times|\mathcal{Z}|)}$, $C_{as} \in \mathbb{R}^{\ell\times 1}$, $E_{as} \in \mathbb{R}^{\ell\times L}$, and $\boldsymbol{d}_{as} \in \mathbb{R}^\ell$. The symbol $\preceq_{K_{as}}$ represents a generalized inequality with respect to a proper cone K_{as}. We denote the marginal distribution by $\mu_{as} = \prod_{(\boldsymbol{p}_{as},r_{as})} \tilde{\mu}_{as}$, and also extend the definition to the ambiguity set so that $\mathcal{D}_{as} = \prod_{(\boldsymbol{p}_{as},r_{as})} \tilde{\mathcal{D}}_{as} = \bigcup_{\tilde{\mu}_{as}\in\tilde{D}_{as}} \prod_{(\boldsymbol{p}_{as},r_{as})} \tilde{\mu}_{as}$. The auxiliary variables $\tilde{\boldsymbol{u}}_{as}$ are used for "lifting" techniques, enabling the representation of nonlinear constraints to linear ones.

C Proofs of Theorems 1 and 2

First, we provide a detailed proof for Theorem 1 below. *Proof:* We show the result by induction.

When $t = T$, $V^T(\boldsymbol{b}) = 0$ satisfies (10). For $t < T$, the inner problem $Q^t(\boldsymbol{b}, a)$ described in (7) becomes

$$\min_{\tilde{\mu}_a \in \mathcal{P}(\tilde{\mathcal{X}}_a)} \quad \mathbb{E}_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \sim \tilde{\mu}_a} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1} \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) \right] \tag{29a}$$

$$\text{s.t.} \quad \mathbb{E}_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \sim \tilde{\mu}_a} \left[\tilde{\boldsymbol{u}}_{as} \right] = \boldsymbol{c}_{as}, \qquad \forall s \in \mathcal{S} \tag{29b}$$

$$\mathbb{E}_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \sim \tilde{\mu}_a} \left[I \left((\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as}) \in \tilde{\mathcal{X}}_{as} \right) \right] = 1, \qquad \forall s \in \mathcal{S} \tag{29c}$$

for all $a \in \mathcal{A}$. Here $I(\cdot)$ is an indicator function, such that if event $\cdot$ is true, it returns value 1 and 0 otherwise. Associating the dual variables $\boldsymbol{\rho}_{as}$ and ω_{as} with constraints (29b) and (29c), respectively, we formulate the dual of (29) as

$$\max_{\boldsymbol{\rho}_a, \boldsymbol{\omega}_a} \quad \sum_{s \in \mathcal{S}} \boldsymbol{c}_{as}^\top \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} \omega_{as} \tag{30a}$$

$$\text{s.t.} \quad \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} \omega_{as} \tag{30b}$$

$$\leq \sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1} \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) \qquad \forall (\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \in \tilde{\mathcal{X}}_a$$

$$\boldsymbol{\rho}_{as} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Z}|}, \ \omega_{as} \in \mathbb{R} \qquad \forall s \in \mathcal{S}. \tag{30c}$$

Constraints (30b) are further equivalent to the following inequality with a minimization problem on the right-hand side (RHS).

$$\sum_{s \in \mathcal{S}} \omega_{as} \leq \tag{31a}$$

$$\min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad \sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1} \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) - \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as}$$

$$\text{s.t.} \quad \tilde{\boldsymbol{u}}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as} \qquad \forall s \in \mathcal{S} \tag{31b}$$

$$\tilde{\boldsymbol{u}}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as} \qquad \forall s \in \mathcal{S} \tag{31c}$$

$$\mathbf{1}^\top \boldsymbol{p}_{as} = 1 \qquad \forall s \in \mathcal{S} \tag{31d}$$

$$\boldsymbol{p}_{as} \geq 0 \qquad \forall s \in \mathcal{S}. \tag{31e}$$

Substituting (10) for V^{t+1} and (1) for $\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z)$, we obtain

$$\text{RHS of (31)} = \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad \sum_{s \in \mathcal{S}} b_s r_{as} + \beta \sum_{z \in \mathcal{Z}} \max_{\boldsymbol{\alpha}_{az} \in \Lambda^{t+1}} \left[\boldsymbol{\alpha}_{az}^\top \sum_{s \in \mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s \right] - \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} \tag{32}$$

$$\text{s.t.} \quad \text{(31b)–(31e)}.$$

Since the objective of the maximization problem is linear in terms of $\boldsymbol{\alpha}_{az}, \forall z \in \mathcal{Z}$, the optimal objective value does not change by taking the convex hull of Λ^{t+1}, denoted as $\text{Conv}\left(\Lambda^{t+1}\right)$. Bringing the maximization to the front, we have

$$(32) = \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad \max_{\substack{\boldsymbol{\alpha}_{az} \in \text{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \left[\sum_{s \in \mathcal{S}} b_s r_{as} + \beta \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_{az}^\top \sum_{s \in \mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s - \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} \right] \tag{33}$$

$$\text{s.t.} \quad \text{(31b)–(31e)}$$

The expression in the bracket is convex (linear) in $(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)$ for fixed $\boldsymbol{\alpha}_{az}$, $z \in \mathcal{Z}$, and concave (affine) in $\boldsymbol{\alpha}_{az}$, $z \in \mathcal{Z}$ given fixed values of $(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)$. Moreover, (31b)–(31e) and $\text{Conv}\left(\Lambda^{t+1}\right)$ are

convex sets. The minimax theorem (see, e.g., Osogami (2015), Du and Pardalos (2013)) ensures that the problem is equivalent to

$$(33) = \max_{\substack{\boldsymbol{\alpha}_{az} \in \mathrm{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad \sum_{s \in \mathcal{S}} b_s r_{as} + \beta \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_{az}^\top \sum_{s \in \mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s - \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} \tag{34}$$
$$\text{s.t.} \quad (31b)\text{–}(31e)$$

We take the dual of the inner minimization by associating dual variables $\boldsymbol{\kappa}_{as}^1$, $\boldsymbol{\kappa}_{as}^2$, σ_{as} with constraints (31b)–(31d), respectively. We thus have the following equivalence:

$$(34) = \max_{\substack{\boldsymbol{\alpha}_{az} \in \mathrm{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \max_{\boldsymbol{\kappa}_a^1, \boldsymbol{\kappa}_a^2, \boldsymbol{\sigma}_a} \quad \sum_{s \in \mathcal{S}} b_s r_{as} + \sum_{s \in \mathcal{S}} \left(-\bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^1 + \bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^2 + \sigma_{as} \right) \tag{35a}$$
$$\text{s.t.} \quad \beta b_s \sum_{z \in \mathcal{Z}} \boldsymbol{J}_z{}^\top \boldsymbol{\alpha}_{az} + \boldsymbol{\kappa}_{as}^1 - \boldsymbol{\kappa}_{as}^2 - \mathbf{1}\sigma_{as} \geq 0, \qquad \forall s \in \mathcal{S} \tag{35b}$$
$$\boldsymbol{\kappa}_{as}^1 + \boldsymbol{\kappa}_{as}^2 + \boldsymbol{\rho}_{as} = 0, \qquad \forall s \in \mathcal{S} \tag{35c}$$
$$\boldsymbol{\kappa}_{as}^1, \boldsymbol{\kappa}_{as}^2 \in \mathbb{R}_+^{|\mathcal{S}| \times |\mathcal{Z}|}, \sigma_{as} \in \mathbb{R}, \qquad \forall s \in \mathcal{S}, \tag{35d}$$

Due to (31), we substitute $\sum_{s \in \mathcal{S}} \omega_{as}$ in the objective function (30a) with (35). As a result, the value function (5) is equivalent to

$$V^t(\boldsymbol{b}) = \max_{a \in \mathcal{A}} \max_{\substack{\boldsymbol{\alpha}_{az} \in \mathrm{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \tag{36a}$$
$$\max_{\boldsymbol{\rho}_a, \boldsymbol{\kappa}_a^1, \boldsymbol{\kappa}_a^2, \boldsymbol{\sigma}_a} \quad \sum_{s \in \mathcal{S}} \boldsymbol{c}_{as}^\top \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} b_s r_{as} + \sum_{s \in \mathcal{S}} \left(-\bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^1 + \bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^2 + \sigma_{as} \right)$$
$$\text{s.t.} \quad (35b)\text{–}(35d)$$
$$\boldsymbol{\rho}_{as} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Z}|} \ \forall s \in \mathcal{S}, \tag{36b}$$

and after taking the dual of the most inner maximization problem, we have

$$V^t(\boldsymbol{b}) = \max_{a \in \mathcal{A}} \max_{\substack{\boldsymbol{\alpha}_{az} \in \mathrm{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \sum_{s \in \mathcal{S}} b_s \times \Xi(a, \boldsymbol{\alpha}_{az} \ \forall z \in \mathcal{Z}, s), \tag{37}$$

where

$$\Xi(a, \boldsymbol{\alpha}_{az} \ \forall z \in \mathcal{Z}, s) = \min_{(\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as})} \quad \beta \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_{az}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} + r_{as} \tag{38a}$$
$$\text{s.t.} \quad \boldsymbol{c}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as} \tag{38b}$$
$$\boldsymbol{c}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as} \tag{38c}$$
$$\mathbf{1}^\top \boldsymbol{p}_{as} = 1 \tag{38d}$$
$$\boldsymbol{p}_{as} \geq 0. \tag{38e}$$

Defining set Λ^t as

$$\left\{ (\Xi(a, \boldsymbol{\alpha}_{az} \ \forall z \in \mathcal{Z}, s), \ s \in \mathcal{S})^\top \middle| \begin{array}{c} \forall a \in \mathcal{A}, \\ \forall \boldsymbol{\alpha}_{az} \in \mathrm{Conv}\left(\Lambda^{t+1}\right), \ \forall z \in \mathcal{Z} \end{array} \right\},$$

it follows that the above value function in (37) is of the form (10). Furthermore, by induction, this is true for all t. This completes the proof.

The proof of Theorem 2 is given as follows. *Proof:* Consider two arbitrary value functions V_1 and V_2. Given belief state $\boldsymbol{b}$, let

$$a_i^\star = \arg\max_{a \in \mathcal{A}} \min_{\mu_a \in \tilde{\mathcal{D}}_a} \mathbb{E}_{(\boldsymbol{p}_a, \boldsymbol{r}_a) \sim \mu_a} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V_i \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) \right],$$

for $i = 1, 2$, and for all actions $a \in \mathcal{A}$, denote

$$\mu^\star_{a,i} = \arg\min_{\mu_a \in \tilde{\mathcal{D}}_a} \mathbb{E}_{(\boldsymbol{p}_a, \boldsymbol{r}_a)\sim\mu_a} \left[\sum_{s\in\mathcal{S}} b_s \left(r_{as} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V_i \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) \right]$$

for $i = 1, 2$. First, suppose that $\mathcal{L}V_1(\boldsymbol{b}) \geq \mathcal{L}V_2(\boldsymbol{b})$. Then,

$$\begin{aligned}
0 &\leq \mathcal{L}V_1(\boldsymbol{b}) - \mathcal{L}V_2(\boldsymbol{b}) \\
&= \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,1}} \left[\sum_{s\in\mathcal{S}} b_s \left(r_{a_1^\star s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right] \\
&\quad - \mathbb{E}_{(\boldsymbol{p}_{a_2^\star}, \boldsymbol{r}_{a_2^\star})\sim\mu^\star_{a_2^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \left(r_{a_2^\star s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_2^\star s} V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_2^\star, \boldsymbol{p}_{a_2^\star}, z) \right) \right) \right] \\
&\leq \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \left(r_{a_1^\star s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right] \\
&\quad - \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \left(r_{a_1^\star s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right] \\
&= \beta \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \times \left(V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, z, \boldsymbol{p}_{a_1^\star}) \right) - V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right]. \qquad (39)
\end{aligned}$$

The inequality follows that we replace the nature's optimal decision $\mu^\star_{a_1^\star,1}$ for V_1 by $\mu^\star_{a_1^\star,2}$, and replace the DM's optimal solution $a_2^\star$ for V_2 by $a_1^\star$. Then, by changing the difference between V_1 and V_2 to the absolute value of the difference, we have

$$\begin{aligned}
(39) &\leq \beta \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \times \left| V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) - V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, z, \boldsymbol{p}_{a_1^\star}) \right) \right| \right] \\
&\leq \beta \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \sup_{\boldsymbol{b}'\in\Delta(\mathcal{S})} \left| V_1(\boldsymbol{b}') - V_2(\boldsymbol{b}') \right| \right] \\
&= \beta \sup_{\boldsymbol{b}'\in\Delta(\mathcal{S})} \left| V_1(\boldsymbol{b}') - V_2(\boldsymbol{b}') \right|.
\end{aligned}$$

The second inequality follows that we take the supremum for all belief states $\boldsymbol{b}' \in \Delta(\mathcal{S})$, and the last equality is because $\mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star})\sim\mu^\star_{a_1^\star,2}} \left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \right] = 1$.

The same result holds for the case where $\mathcal{L}V_1(\boldsymbol{b}) < \mathcal{L}V_2(\boldsymbol{b})$. Thus, for any belief state value $\boldsymbol{b}$, it follows that

$$\left| \mathcal{L}V_1(\boldsymbol{b}) - \mathcal{L}V_2(\boldsymbol{b}) \right| \leq \beta \sup_{\boldsymbol{b}'\in\Delta(\mathcal{S})} \left| V_1(\boldsymbol{b}') - V_2(\boldsymbol{b}') \right|,$$

and therefore,

$$\sup_{\boldsymbol{b}\in\Delta(\mathcal{S})} \left| \mathcal{L}V_1(\boldsymbol{b}) - \mathcal{L}V_2(\boldsymbol{b}) \right| \leq \beta \sup_{\boldsymbol{b}'\in\Delta(\mathcal{S})} \left| V_1(\boldsymbol{b}') - V_2(\boldsymbol{b}') \right|,$$

yielding that $\mathcal{L}$ is a contraction under $0 < \beta < 1$. This completes the proof.