DISTRIBUTIONALLY ROBUST PARTIALLY OBSERVABLE
MARKOV DECISION PROCESS WITH MOMENT-BASED
AMBIGUITY∗

HIDEAKI NAKAO†, RUIWEI JIANG‡, AND 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 conservative and 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.

Key words. Partially Observable Markov Decision Process (POMDP), distributionally robust optimization, moment-based ambiguity set, heuristic value search iteration (HVSI), epidemic control

AMS subject classifications. 90C39, 90C40, 90C47, 93E20

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 [4, 11, 26].

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 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 gains an expected value of the reward (or pays the expected cost) based on the belief. The DM’s belief

∗Funding: The U.S. Department of Engineering (DoE) grant # DE-SC0018018.
†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.
is represented by a probability mass associated with each state of the system, which needs to be a sufficient statistic of the history of past actions and observations [14].

In POMDPs, a critical assumption is that the exact transition and observation probabilities are known to the DM for each action-and-state combination. In practice, there may exist estimation errors about either the transition or observation probability values, to handle which, [21] 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 the results of robust POMDP (see [18]) with 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. 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. This setting makes the belief a sufficient statistic of the history of past actions and observations, and it can be justified by a wide range of applications with time-varying, ambiguous transition-observation probabilities. We will provide a few examples to illustrate and justify the assumption in Section 3.

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. [6] 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 [28]. Ref. [33] considers tractable reformulations of DR chance-constrained programs using moment-based ambiguity set. Other types of ambiguity sets used in DRO models bound the $\phi$-divergence [3, 13] or Wasserstein distance [9, 10] 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 property 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 [31]. 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, the technique applied to solve MDP models where the states are discrete, is not directly applicable to solving POMDP since belief states are continuous. Ref. [24] 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. [19] 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. [25] develops a heuristic search value iteration (HSVI) algorithm to derive an upper bound of the value function via finding the reachable set through Monte Carlo simulation. Ref. [25] 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., [1, 2]). In [27], 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. [17] 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. [12] 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. [5] argues that robust MDP models may produce over-conservative solutions, as they do not incorporate the distributional information of uncertain parameters. Then [29] 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. [31] generalizes the distributionally robust MDP to include multimodal distributions and the information of mean and variance. Ref. [30] 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. [23] presents a modeling framework of ambiguous POMDP (called APOMDP), which generalizes the robust POMDP in [21]. APOMDP optimizes over the $\alpha$-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. [23] describes conditions under which the value function of APOMDP is PWLC. Meanwhile, [21] 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 (see an example provided in [16]). Moreover, [18] 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 $\mu$ of the transition-observation probabilities from a pre-assumed ambiguity set. The DM expects that the nature acts adversely against her action $a$ taken at the beginning of the period, and therefore the distribution $\mu$ is expected to lead to the worst-case expected reward. Next, the joint transition-observation probability $p$ is realized from the distribution $\mu$. The state makes a transition according to $p$, and the observation outcome $z$ is shown. Finally, the DM obtains the values of $z$ and $p$ at the end of the period.

We denote $S$ as the set of states, $A$ as the set of actions, and $Z$ as the set of observation outcomes. For all $(s, s', z, a) \in S^2 \times Z \times A$, 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 $(s, a) \in S \times A$, let $r_{as}$ be the reward for taking action $a$ at state $s$. For all $s \in S$, $a \in A$, we define a vector of probabilities $p_{as} = (p_{as}(s', z), (s', z) \in S \times Z)^\top$ and assume that the Cartesian product $(p_{as}, r_{as})$ is a member of a set $\mathcal{X}_{as} \subseteq \Delta(S \times Z) \times \mathbb{R}$, where $\Delta(\cdot)$ is a probability simplex of set $\cdot$. We denote $p_{a} = (p_{as}(s', z), (s, s', z) \in S^2 \times Z)^\top$ and $r_{a} = (r_{as}, s \in S)^\top$ for all $a \in A$. We assume that $(p_{as}, r_{as})$ follows a distribution $\mu_{as}$, which is unknown but is included in an ambiguity set $D_{as} \subseteq \mathcal{P}(\mathcal{X}_{as})$, 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 $A$ and the set of states $S$, i.e., the overall ambiguity set is $D = \bigotimes_{a \in A} D_{as}$. This assumption is analogous to
the \((s, a)\)-rectangularity in [27]. The above conditions increase the conservativeness of the model in general. In the online supplement SM1, we discuss a relaxation of the \(a\)-rectangularity assumption for DR-POMDP.

Below we describe several examples in which the settings of DR-POMDP can be justified and thus our approach can be applied to optimize corresponding policies. The key is to justify whether the DM can obtain the true value of \(p\) 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 tests even we add noise to the true \(p\)-value and thus the assumption is relatively weak.

First, we consider epidemic surveillance and control. During flu season, Centers for Disease Control and Prevention reports the number of weekly visits of patients who show influenza-like illness (ILI) symptoms. 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, [22] studies a two-state MDP model (i.e., epidemic vs. non-epidemic) and states that the ILI rate follows a Gaussian and an exponential distribution for the epidemic and non-epidemic state, respectively; [15] 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., 8]. 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 [32], 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 [32], 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.

Lastly, one can also consider planning production or maintaining inventory in highly seasonal industries such as agriculture [26], 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 \( p \).

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 \( \mathcal{D}_a = \bigotimes_{s \in \mathcal{S}} \mathcal{D}_{as} \) to minimize the expected reward given the DM’s action \( a \). Let \( a^t, p_{at}, z^t \) 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 \( h^t = (a^1, p_{a1}^1, z^1, \ldots , a^{t-1}, p_{a(t-1)}^{t-1}, z^{t-1}) \) 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 \( \pi^t : \mathcal{H}^t \rightarrow \mathcal{A} \). We denote the set of all such policies as \( \Pi \), and define an extended history \( \tilde{h}^t = (a^1, p_{a1}^1, z^1, \ldots , a^{t-1}, p_{a(t-1)}^{t-1}, z^{t-1}, a^t) \in \tilde{\mathcal{H}}^t \), on which the nature bases its decision for choosing \( \mu_{at} \). The nature’s objective is to find the best policy (from the nature’s perspective) \( \gamma = (\gamma^1, \ldots , \gamma^{T-1}) \), with \( \gamma^t : \tilde{\mathcal{H}}^t \rightarrow \mathcal{D}_{at} \) to minimize the expected reward. Similarly, we denote the set of all the nature’s policies as \( \Gamma \).

The belief state is a sufficient statistic for DR-POMDP when the DM can obtain the value of transition-observation probability at the end of each decision period (see, e.g., [21]). Let the belief state in period \( t \) be \( (b^t, s) = b^t \in \Delta(\mathcal{S}) \). Given action \( a \), transition-observation probability \( p_a \), and observation outcome \( z \), the sufficient statistic for the history \( \tilde{h}^{t+1} = (\tilde{h}^t, a, p_a, z) \), or the belief state in period \( t+1 \) is given by

\[
(4.1) \quad b^{t+1} = f(b, a, p_a, z) = \frac{\sum_{s \in \mathcal{S}} J_z p_{as} b_s}{\sum_{s \in \mathcal{S}} 1^T J_z p_{as} b_s},
\]

where \( 1 \) represents a vector of ones having the length \( |\mathcal{S}| \), \( J_z \in \mathbb{R}^{|\mathcal{S}| \times (|\mathcal{S}| \times |\mathcal{Z}|) } \) is a matrix of zeros and ones that projects the vector \( p_{as} \) to a vector \( p_{asz} = (p_{as}(s', z), s' \in \mathcal{S})^T \) whose entries correspond to the outcome \( z \). That is, \( p_{asz} = J_z p_{as} \), \( \forall a, s, z \). Note that the belief state cannot be updated using (4.1) and will not be a sufficient statistic of the history of past actions and observations if we do not have the true values of \( p_{as} \).

With slight abuse of notation, let \( \pi \) be a policy that maps belief states to the actions, i.e., \( \pi^t : \Delta(\mathcal{S}) \rightarrow \mathcal{A} \) for all \( t \in \{1, \ldots , T-1\} \). Similarly, let \( \gamma^t : \Delta(\mathcal{S}) \times \mathcal{A} \rightarrow \mathcal{D}_{at} \) 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.

Given the nature’s choice of distribution \( \mu_a \), the expected value of the instantaneous reward given belief state \( b \) and action \( a \) is denoted as \( \mathbb{E}_{(p_a, r_a) \sim \mu_a} \left[ b^t r_a \right] \), 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
DISTRIBUTIONALLY ROBUST POMDP WITH MOMENT AMBIGUITY

policies (i.e., distributions of transition-observation probabilities) by the nature. That is, DR-POMDP aims to solve

\[
\max_{\pi \in \Pi} \min_{\gamma \in \mathcal{F}} \mathbb{E} \left[ \sum_{t=1}^{T-1} \beta^t b_t^\top r_{a_t}^t \right]
\]

(4.2a)

s.t. \( a_t = \pi_t(b_t) \), \( \forall t \in \{1, \ldots, T - 1\} \)

(4.2b)

\( \mu_{a_t}^t = \gamma_t(b_t, a_t) \), \( \forall t \in \{1, \ldots, T - 1\} \)

(4.2c)

\( (p_{a_t}^t, r_{a_t}^t) \sim \mu_{a_t}^t \), \( \forall t \in \{1, \ldots, T - 1\} \)

(4.2d)

\( (s_t^{t+1}, z^t) \sim p_{a_t}^t s_{t+1}, \gamma_t, p_{a_t}^t, z^t \), \( \forall t \in \{1, \ldots, T - 1\} \)

(4.2e)

\( b_t^{t+1} = f(b_t, a_t, p_{a_t}^t, z_t) \), \( \forall t \in \{1, \ldots, T - 1\} \)

where the terminal reward is zero without loss of generality. The initial belief state is given as \( b \). Alternatively, we denote the problem (4.2) as

\[
\max_{\pi \in \Pi} \min_{\gamma \in \mathcal{F}} \mathbb{E} \left[ \sum_{t=1}^{T-1} \beta^t b_t^\top r_{a_t}^t \bigg| b^t = b \right]
\]

(4.3)

Here we omit all the constraints in (4.2) for presentation simplicity.

To solve (4.2), we propose to use dynamic programming, and derive the Bellman equation below.

PROPOSITION 4.1. 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

\[
V^t(b) = \max_{\pi^{t:T-1} \in \Pi^{t:T-1}, \gamma^{t:T-1} \in \Gamma^{t:T-1}} \min_{\mu_{a_t}^t \in \mathcal{D}_a} \mathbb{E} \left[ \sum_{n=t}^{T-1} \beta^{n-t} b_{n}^\top r_{a_n}^n \bigg| b^t = b \right]
\]

(4.4)

Then,

\[
V^t(b) = \max_{\pi^{t:T-1} \in \Pi^{t:T-1}, \gamma^{t:T-1} \in \Gamma^{t:T-1}} \min_{\mu_{a_t}^t \in \mathcal{D}_a} \mathbb{E} \left[ \sum_{n=t}^{T-1} \beta^{n-t} b_{n}^\top r_{a_n}^n \bigg| b^t = b \right]
\]

(4.5)

\[
V^t(b) = \max_{\pi^{t:T-1} \in \Pi^{t:T-1}, \gamma^{t:T-1} \in \Gamma^{t:T-1}} \min_{\mu_{a_t}^t \in \mathcal{D}_a} \mathbb{E} \left[ \sum_{n=t}^{T-1} \beta^{n-t} b_{n}^\top r_{a_n}^n \bigg| b^t = b \right]
\]

(4.6)

Proof. We first isolate the term associated with period \( t \) inside the expectation of (4.4) as follows.

\[
V^t(b) = \max_{\pi^{t:T-1} \in \Pi^{t:T-1}, \gamma^{t:T-1} \in \Gamma^{t:T-1}} \min_{\mu_{a_t}^t \in \mathcal{D}_a} \mathbb{E} \left[ \sum_{n=t}^{T-1} \beta^{n-t} b_{n}^\top r_{a_n}^n \bigg| b^t = b \right]
\]

(4.7)

Given \( a_t = \pi_t(b_t) \), \( p_{a_t}^t = p_{\pi_t(b_t)} \), \( z^t = z \), the probability of observing \( z \) is

\[
\sum_{\pi \in S} b_a \mathbb{1}^\top J_s p_{\pi_t(b_t)}
\]

(4.8)

Thus, we can calculate the expectation conditioned on the values of \( a_t, p_{a_t}^t, z^t \) in the value function as:

\[
V^t(b) = \max_{\pi^{t:T-1} \in \Pi^{t:T-1}, \gamma^{t:T-1} \in \Gamma^{t:T-1}} \min_{\mu_{a_t}^t \in \mathcal{D}_a} \mathbb{E} \left[ \sum_{n=t}^{T-1} \beta^{n-t} b_{n}^\top r_{a_n}^n \bigg| b^t = b \right]
\]

(4.9)

This manuscript is for review purposes only.
Following Proposition 4.1, the policies optimal to (4.3) can be determined by recur-
sing to the Bellman equation provided the optimal action given belief state \( b \) is an information state. Because policies beyond period \( t \) do not affect \((p_{a,t}^t, r_{a,t}^t)\), we have
\[
V^t(b) = \max_{a \in A} \min_{\mu_a \in D_{\mu_a}} \left\{ \sum_{s \in S} \left[ r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V^{t+1}(f(b, a, p_{a}, z)) \right] \right\},
\]
where the second equality is due to rearranging the terms and the fact that \( b \) is an information state. The solution to the Bellman equation provides the optimal action given belief state \( b \) as:
\[
\arg \max_{a \in A} Q^t(b, a),
\]
whereas the optimal distribution chosen by the nature, under belief state \( b \) and the DM’s action \( a \), is
\[
\arg \min_{\mu_a \in D_{\mu_a}} U^t(b, a, \mu_a).
\]

### 4.2. Properties of Distributionally Robust Bellman Equation (4.5)

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 SM2 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 \( \bar{p}_{as} \) is at most \( c_{as} \). Then for all \( a \in A \) and \( s \in S \), the unknown distribution \( \mu_{as} \) satisfies \( \mathbb{E}_{p_{as} \sim \mu_{as}} [ |p_{as} - \bar{p}_{as} | ] \leq c_{as} \), which is reformulated as:
\[
\mathbb{E}_{p_{as} \sim \tilde{\mu}_{as}} [ \tilde{u}_{as} ] = c_{as},
\]
\[
\tilde{u}_{as} \begin{cases} 
  \geq p_{as} - \bar{p}_{as}, & 1^T p_{as} = 1 \\
  \geq p_{as} - \bar{p}_{as}, & p_{as} \geq 0 
\end{cases} = 1.
\]
Here, $\bar{u}_{as} \in \mathbb{R}^{\mid S \mid \times |Z|}$ denotes a vector of auxiliary variables, and $\bar{\mu}_{as}$ is a joint distribution of $(p_{as}, \bar{u}_{as})$. This notation is introduced to differentiate from $\mu_{as}$, which represents the true distribution of $p_{as}$. The ambiguity set for distribution $\bar{\mu}_{as}$ is therefore

$$\mathcal{D}_{as} = \left\{ \bar{\mu}_{as} \left( p_{as}, \bar{u}_{as} \right) \mid E(p_{as}, \bar{u}_{as}) \sim \bar{\mu}_{as}, \bar{u}_{as} = c_{as} \right\},$$

while the support $\tilde{X}_{as}$ for $(p_{as}, \bar{u}_{as})$ is given by

$$\tilde{X}_{as} = \left\{ \left( p_{as}, \bar{u}_{as} \right) \in \mathbb{R}^{\mid S \mid \times |Z|} \mid \bar{u}_{as} \geq p_{as} - \bar{p}_{as}, \bar{u}_{as} \geq p_{as} - p_{as}, 1^T p_{as} = 1 \right\}.$$  

For ambiguity sets and supports respectively defined in terms of (4.8) and (4.9), we show that the value function is convex with respect to the belief state $b$ for each decision period.

**Theorem 4.2.** For all $a \in A$ and $s \in S$, let the ambiguity set and support be (4.8) and (4.9), respectively. For all $t \in \{1, \ldots, T\}$, there exists a set $\Lambda^t$ of slopes such that the value function can be expressed as follows.

$$V^t(b) = \max_{\alpha \in \Lambda^t} \alpha^T b. \tag{4.10}$$

**Proof.** We show the result by induction. When $t = T$, $V^T(b) = 0$ satisfies (4.10).

For $t < T$, the inner problem $Q^t(b, a)$ described in (4.7) becomes

$$\begin{align*}
\min_{\bar{\mu}_{as} \in \mathcal{D}(\tilde{X}_{as})} & \quad E(p_{as}, \bar{u}_{as}) \sim \bar{\mu}_{as} \left[ \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V^{t+1} (f(b, a, p_{as}, z)) \right) \right] \\
\text{s.t.} & \quad E(p_{as}, \bar{u}_{as}) \sim \bar{\mu}_{as}, \quad \forall s \in S \\
& \quad E(p_{as}, \bar{u}_{as}) \sim \bar{\mu}_{as}, \quad I \left( (p_{as}, \bar{u}_{as}) \in \tilde{X}_{as} \right) = 1, \quad \forall s \in S \\
& \quad \forall (p_{as}, \bar{u}_{as}) \in \tilde{X}_{as} \\
& \quad \forall (p_{as}, \bar{u}_{as}) \in \tilde{X}_{as} \\
& \quad \forall (p_{as}, \bar{u}_{as}) \in \tilde{X}_{as} \\
\end{align*} \tag{4.11a}$$

for all $a \in A$. 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 $\rho_{as}$ and $\omega_{as}$ with constraints (4.11b) and (4.11c), respectively, we formulate the dual of (4.11) as

$$\begin{align*}
\max_{\rho_{as}, \omega_{as}} & \quad \sum_{s \in S} c_{as} \rho_{as} + \sum_{s \in S} \omega_{as} \\
\text{s.t.} & \quad \sum_{s \in S} \bar{u}_{as}^T \rho_{as} + \sum_{s \in S} \omega_{as} \\
& \quad \leq \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V^{t+1} (f(b, a, p_{as}, z)) \right), \quad \forall (p_{as}, \bar{u}_{as}) \in \tilde{X}_{as} \\
\end{align*} \tag{4.12a}$$

Constraints (4.12b) are further equivalent to the following inequality with a minimization problem on the right-hand side (RHS).

$$\begin{align*}
\sum_{s \in S} \omega_{as} \leq & \quad \max_{\rho_{as}} \sum_{s \in S} c_{as} \rho_{as} + \sum_{s \in S} \omega_{as} \\
\text{s.t.} & \quad \sum_{s \in S} \bar{u}_{as}^T \rho_{as} + \sum_{s \in S} \omega_{as} \\
& \quad \leq \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V^{t+1} (f(b, a, p_{as}, z)) \right), \quad \forall (p_{as}, \bar{u}_{as}) \in \tilde{X}_{as} \\
\end{align*} \tag{4.13a}$$

This manuscript is for review purposes only.
Substituting (4.10) for \( V^{t+1} \) and (4.1) for \( f(b, a, p_a, z) \), we obtain

\[
(4.14) \quad \text{RHS of (4.13)} = \min_{(p_a, \mu_a)} \max_{\alpha_a \in \text{Conv}(\Lambda^{t+1})} \left[ \sum_{s \in S} b_s r_{as} + \beta \sum_{s \in S} \alpha_a \sum_{z \in Z} J_{z} p_{as} - \sum_{s \in S} \tilde{u}_{as}^\top p_{as} \right]
\]

\[\text{s.t. } (4.13b) - (4.13e).\]

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

\[
(4.15) \quad (4.14) = \min_{(p_a, \mu_a)} \max_{\alpha_{az} \in \text{Conv}(\Lambda^{t+1})} \left[ \sum_{s \in S} b_s r_{as} + \beta \sum_{z \in Z} \max_{\alpha_{az} \in \Lambda^{t+1}} \alpha_{az} \sum_{s \in S} J_{z} p_{as} - \sum_{s \in S} \tilde{u}_{as}^\top p_{as} \right]
\]

\[\text{s.t. } (4.13b) - (4.13e).\]

The expression in the bracket is convex (linear) in \( (p_a, \tilde{u}_{a}) \) for fixed \( \alpha_{az}, z \in Z \), and concave (affine) in \( \alpha_{az}, z \in Z \) given fixed values of \( (p_a, \tilde{u}_{a}) \). Moreover, (4.13b)–(4.13e) and \( \text{Conv}(\Lambda^{t+1}) \) are convex sets. The minimax theorem (see, e.g., [18], [7]) ensures that the problem is equivalent to

\[
(4.16) \quad (4.15) = \max_{\alpha_{az} \in \text{Conv}(\Lambda^{t+1})} \min_{(p_a, \mu_a)} \sum_{s \in S} b_s r_{as} + \beta \sum_{z \in Z} \max_{\alpha_{az} \in \Lambda^{t+1}} \alpha_{az} \sum_{s \in S} J_{z} p_{as} - \sum_{s \in S} \tilde{u}_{as}^\top p_{as}
\]

\[\text{s.t. } (4.13b) - (4.13e).\]

We take the dual of the inner minimization by associating dual variables \( \kappa^1_{az}, \kappa^2_{az} \) with constraints (4.13b)–(4.13d), respectively. We thus have the following equivalence:

\[
(4.17a) \quad (4.16) = \max_{\alpha_{az} \in \text{Conv}(\Lambda^{t+1})} \max_{\kappa^1_{az}, \kappa^2_{az}, \sigma_{as}} \sum_{s \in S} b_s r_{as} + \sum_{z \in Z} \left( -\beta_{az} \kappa^1_{az} + \beta_{az}^\top \sigma_{as} + \kappa^2_{az} + \sigma_{as} \right)
\]

\[\text{s.t. } \beta_{az} \sum_{z \in Z} J_{z}^\top \alpha_{az} + \kappa^1_{az} - \kappa^2_{az} - 1 \sigma_{as} \geq 0, \quad \forall s \in S\]

\[
(4.17b) \quad \kappa^1_{az} + \kappa^2_{az} + \rho_{as} = 0, \quad \forall s \in S
\]

\[
(4.17c) \quad \kappa^1_{az} \geq 0, \kappa^2_{az} \geq 0, \quad \sigma_{as} \in \mathbb{R}, \quad \forall s \in S
\]

Due to (4.13), we substitute \( \sum_{s \in S} \omega_{az} \) in the objective function (4.12a) with (4.17). As a result, the value function (4.5) is equivalent to

\[
(4.18a) \quad V^t(b) = \max_{\alpha_{az} \in \text{Conv}(\Lambda^{t+1})} \max_{\kappa^1_{az}, \kappa^2_{az}, \sigma_{as}} \sum_{s \in S} c_{s} \rho_{as} + \sum_{s \in S} b_s r_{as} + \sum_{z \in Z} \left( -\beta_{az}^\top \kappa^1_{az} + \beta_{az}^\top \sigma_{as} + \kappa^2_{az} + \sigma_{as} \right)
\]

\[\text{s.t. } (4.17b) - (4.17d)\]

\[
(4.18b) \quad \rho_{as} \in \mathbb{R}^{[2|S|]} \quad \forall s \in S
\]
and after taking the dual of the most inner maximization problem, we have

\begin{equation}
V^t(b) = \max_{a \in A} \max_{\alpha_{az} \in \text{Conv}(\Lambda^{t+1})} \sum_{s \in S} b_s \times \Xi(a, \alpha_{az}, \forall z \in Z, s),
\end{equation}

where

\begin{equation}
\Xi(a, \alpha_{az}, \forall z \in Z, s) = \min_{(p_{as}, \tilde{a}_s)} \beta \sum_{z \in Z} \alpha_{az}^* J_z p_{as} + r_{as}
\end{equation}

\begin{align}
&\text{s.t. } c_{as} \geq p_{as} - \tilde{p}_{as} \\
&c_{as} \geq \bar{p}_{as} - \bar{p}_{as} \\
&1^T p_{as} = 1 \\
&p_{as} \geq 0.
\end{align}

Defining set \( \Lambda^t \) as

\begin{equation}
\left\{ (\Xi(a, \alpha_{az}, \forall z \in Z, s), s \in S) \right\}^{\top} \forall \alpha_{az} \in \text{Conv}(\Lambda^{t+1}), \forall z \in Z
\end{equation}

it follows that the above value function in (4.19) is of the form (4.10). Furthermore, by induction, this is true for all \( t \). This completes the proof. \( \square \)

Having provided the values of \( a \) and \( \alpha_{az} \), the inner minimization in (4.19) can be solved efficiently using linear programming. The issue, however, is that there are possibly infinitely many elements in Conv(\( \Lambda^{t+1} \)), and even if there are finitely many, the number of supporting hyperplanes \( \alpha \) inside \( \Lambda^t \) increases exponentially as the value functions are calculated from period \( t = T \) to \( t = 1 \). We describe in Section 5 a heuristic value search iteration (HVSI) algorithm for efficiently computing optimal policies in DR-POMDP.

### 4.3. Case of Infinite Horizon

In this section, 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., [20]), and show that by repeatedly updating the value function in (4.5), it converges to a unique function corresponding to the optimal value \( V^* \) of the infinite-horizon DR-POMDP problem.

**Theorem 4.3.** The operator \( \mathcal{L} \) defined as

\begin{equation}
\mathcal{L} V(b) = \max_{a \in A} \min_{\mu_a \in \mathcal{D}_a} \mathbb{E}_{(p_a, r_a) \sim \mu_a} \left[ \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V(f(b, a, p_a, z)) \right) \right]
\end{equation}

is a contraction for \( 0 < \beta < 1 \).

**Proof.** Consider two arbitrary value functions \( V_1 \) and \( V_2 \). Given belief state \( b \), let

\begin{equation}
a^*_i = \arg \max_{a \in A} \min_{\mu_a \in \mathcal{D}_a} \mathbb{E}_{(p_a, r_a) \sim \mu_a} \left[ \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V_i(f(b, a, p_a, z)) \right) \right],
\end{equation}

for \( i = 1, 2 \), and for all actions \( a \in A \), denote

\begin{equation}
\mu^*_{a,i} = \arg \min_{\mu_a \in \mathcal{D}_a} \mathbb{E}_{(p_a, r_a) \sim \mu_a} \left[ \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^T J_z p_{as} V_i(f(b, a, p_a, z)) \right) \right]
\end{equation}
for $i = 1, 2$. First, suppose that $\mathcal{L}V_1(b) \geq \mathcal{L}V_2(b)$. Then,

$$0 \leq \mathcal{L}V_1(b) - \mathcal{L}V_2(b)$$

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

$$0 \leq \mathcal{L}V_1(b) - \mathcal{L}V_2(b)$$

by the Bellman operator.

The inequality follows that we replace the nature’s optimal decision $\mu^*_{s_1, 1}$ for $V_1$ by $\mu^*_{s_1, 1}z$, and replace the DM’s optimal solution $a^*_2$ for $V_2$ by $a^*_1$. Then, by changing the difference between $V_1$ and $V_2$ to the absolute value of the difference, we have

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

and therefore,

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

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

Theorem 4.3 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 $\beta$ 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.

The inequality follows that we replace the nature’s optimal decision $\mu^*_{s_1, 1}$ for $V_1$ by $\mu^*_{s_1, 1}z$, and replace the DM’s optimal solution $a^*_2$ for $V_2$ by $a^*_1$. Then, by changing the difference between $V_1$ and $V_2$ to the absolute value of the difference, we have

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

and therefore,

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

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

Theorem 4.3 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 $\beta$ 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 [25] (originally for solving POMDP) for efficiently computing upper and lower bounds for DR-POMDP. We maintain a set of finite number of hyperplanes $\Lambda_{V'}$, where the resulting PWLC function $V$ bounds the true value function from below. We also maintain a set of points $T^{-1}$ whose elements are $(b, v)$, which is a combination of a belief $b$ and an upper bound $v$ of the true value function at the belief $b$. Therefore,
the resulting PWLC function $\mathbf{V}$ bounds the value function from above. The upper bound $\mathbf{V}$ corresponding to a belief $\mathbf{b}$ is obtained through Monte Carlo sampling. The sampling follows a greedy strategy to close the gap between the upper bound $\mathbf{V}$ and the lower bound $\mathbf{V}$ for the belief points that are reachable from the initial belief.

Algorithm 5.1 Heuristic Search Value Iteration (HSVII)

1: Input: initial belief state $\mathbf{b}^0$, tolerance $\epsilon$
2: Initialize: $\mathbf{V}$, $\mathbf{V}$ (see details in Section 5.1)
3: while $\mathbf{V}(\mathbf{b}^0) - \mathbf{V}(\mathbf{b}^0) > \epsilon$ or time limit is reached do
4:    DR-BoundExplore($\mathbf{b}^0$, 0) (see details in Algorithm 5.2)
5: end while
6: Output: $\mathbf{V}$, $\mathbf{V}$

Algorithm 5.1 presents the main algorithmic steps in HSVII, where the details of Step 4 are later provided in Algorithm 5.2. 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 $\mathbf{b}$.

5.1. Initialization. Recall the ambiguity set and support defined in (4.8) and (4.9), respectively. In the initialization step, we compute the lower bound for the true value function by taking the best action for obtaining the worst-case expected reward in each decision period. That is, for each action $a$, we solve

$$R_a = \sum_{t=0}^{\infty} a^t \min_{s \in S} \min_{\mu \in \mathcal{A}_a} \mathbb{E}_{p_{as}, r_{as}}[r_{as}] = \frac{1}{1 - \beta} \min_{s \in S} \min_{\mu \in \mathcal{A}_a} \mathbb{E}_{p_{as}, r_{as}}[r_{as}].$$

In the case of mean absolute deviation based ambiguity set (4.8), the second minimization is trivial as $r_{as}$ is fixed. The minimum value for all $s \in S$ is computed by enumeration. We then define an initial lower bounding hyperplane $\alpha'_s = \max_{a \in A} R_a(s), \forall s \in S$ and set $\Lambda = \{\alpha'_s, a \in S\}^T$.

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 $\mathbf{V}^{MDP} \in \mathbb{R}^{|S|}$ be a value function for the distributionally-optimistic MDP. It satisfies

$$V^{MDP}_s = \max_{a \in A} \max_{\mu \in \mathcal{A}_a} \mathbb{E}_{p_{as}, r_{as}}[r_{as} + \beta \mathbf{V}^{MDP}\top \sum_{z \in Z} J_z p_{as}], \forall s \in S.$$ 

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

$$\min_{\mathbf{V}^{MDP}} \mathbf{V}^{MDP}$$

$$\text{s.t. } V^{MDP}_s \geq \max_{\mu \in \mathcal{A}_a} \mathbb{E}_{p_{as}, r_{as}}[r_{as} + \beta \mathbf{V}^{MDP}\top \sum_{z \in Z} J_z p_{as}], \forall a \in A, s \in S.$$ 

In the case of ambiguity set (4.8), model (5.1) becomes

$$\min_{\rho, \mathbf{V}^{MDP}} \mathbf{V}^{MDP}$$

$$\text{s.t. } V^{MDP}_s - c_{as} \rho_{as} - \mathbf{p}_{as} \mathbf{k}_{as}^1 + \mathbf{p}_{as} \mathbf{k}_{as}^2 - \sigma_{as} \geq r_{as}, \forall s \in S, a \in A$$

This manuscript is for review purposes only.
After the optimal solution is discovered, we initialize $\Upsilon_{\uparrow} = \{ (e_s, V_{s\uparrow}^{MDP}) \mid \forall s \in S \}$, where $e_s$ 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 $V(b)$, we solve

$$\max \{ \alpha^T b \mid \forall \alpha \in \Lambda_{\Upsilon} \}$$

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

$$(5.3) \quad v = \min \left\{ \sum_{i \in [\Upsilon_{\downarrow}]} w^i v^i \mid \sum_{i \in [\Upsilon_{\downarrow}]} w^i b^i = b, \sum_{i \in [\Upsilon_{\downarrow}]} w^i = 1, w^i \geq 0, \forall i \in [\Upsilon_{\downarrow}] \right\},$$

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

We refer the readers to the online supplement SM3 for an illustration of the lower and upper bounds for a specific DR-POMDP value function computed for a two-state system.

### 5.2. Forward Exploration Heuristics

The forward heuristics follow from the HSVI algorithm from [25], 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 Monte Carlo simulation 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 $\mu_a$, and their outcomes $z$ and $p_a$. Here we describe a greedy strategy to select the branches. We consider the following two functions:

$$U_V(b, a, \mu_a) = E_{(p_a, r_a) \sim \mu_a} \left[ \sum_{s \in S} b_s \left\{ r_{as} + \beta \sum_{z \in Z} 1_z^T J_z p_a V(f(b, a, p_a, z)) \right\} \right],$$

$$Q_V(b, a) = \min_{\mu_a \in D_a} U_V(b, a, \mu_a).$$

Then we can obtain $(U_{\uparrow}, Q_{\uparrow})$ and $(U_{\downarrow}, Q_{\downarrow})$ by letting $V = \overline{V}$ and $V = V$, respectively.

To sample, we take an action that maximizes $Q_{\uparrow}$ corresponding to the best upper bound. If $a^* = \arg\max_{a \in A} Q_{\uparrow}(b, a)$ is suboptimal, the upper bound will eventually be replaced by a better one associated with another action. Following (4.21), the gap between $Q_{\uparrow}$ and $Q_{\downarrow}$ at belief state $b$ and action $a^*$ is

$$Q_{\uparrow}(b, a^*) - Q_{\downarrow}(b, a^*)$$
Recall that action \(a\) starting from the initial one is only \(\epsilon \beta\). Thus, we combine the two inner maximization problems in (4.18) as

\[
Q_T(b, a^*) - Q_V(b, a^*) \leq \epsilon \beta^{-t}.
\]

To achieve the gap \(\epsilon\) at the initial state \(b_0\), the condition for the gap at depth level \(t\) starting from the initial one is only \(\epsilon \beta^{-t}\). This is used as a terminal condition for the Monte Carlo simulation. Specifically, we define

\[
\text{excess}(b, t) = V(b) - \bar{V}(b) - \epsilon \beta^{-t}.
\]

Recall that action \(a^*\) maximizes \(Q_T(b, a)\) for all \(a \in A\), but it does not necessarily maximize \(Q_V(b, a)\). Therefore,

\[
V(b) - \bar{V}(b) \leq Q_T(b, a^*) - Q_V(b, a^*).
\]

Further applying inequality (5.4), we have

\[
\text{excess}(b, t) \leq \mathbb{E}(p_{a^*} | a^*, r_{a^*} | b_{a^*}) \sum_{\pi \in \Sigma} b_{a^*} \sum_{\pi \in \Sigma} 1^T J_b p_{a^*} (V(f(b, a^*, p_{a^*}, z)) - V(f(b, a^*, p_{a^*}, z))) + 1^T J_b p_{a^*} \text{excess}(f(b, a^*, p_{a^*}, z), t + 1).\]

To have an early termination, we choose \((z^*, p_{a^*}^*)\) so that they can lead to the maximum expected value on the right-hand side of (5.5), i.e.,

\[
(z^*, p_{a^*}^*) \in \arg \max_{z^*, p_{a^*}} \sum_{a^*} b_{a^*} 1^T J_b p_{a^*} \text{excess}(f(b, a^*, p_{a^*}, z), t + 1).
\]

Note that because the worst-case distribution under ambiguity set (4.8) is a point mass distribution, obtaining \(p_{a^*}^*\) is trivial. Algorithm 5.2 describes the detailed algorithmic steps. In the HSVI approach, Algorithm 5.2 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.

### 5.3. Local Updates

In this section, we describe the details of \(DR\)-backup and \(DR\)-update steps in Algorithm 5.2. We first illustrate how the lower bound is updated in \(DR\)-backup. For each \(a \in A\), we solve the two inner maximization problems in (4.18) provided \(a\) and \(b\), where we set \(\Lambda^{t+1} = \Lambda_L\). The convex hull of \(\Lambda_L\) is therefore

\[
\text{Conv}(\Lambda_L) = \left\{ \sum_{i \in [||\Lambda_L||]} w_i \alpha_i \right\} \sum_{i \in [||\Lambda_L||]} w_i = 1, \alpha_i \in \Lambda_L, w_i \geq 0, i \in [||\Lambda_L||]\right\}.
\]

Thus, we combine the two inner maximization problems in (4.18) as

\[
\max_{\rho_a, \kappa_a^1, \kappa_a^2, \sigma_a} \sum_{s \in S} c_{as} \rho_{as} + \sum_{s \in S} b_{as} r_{as} + \sum_{s \in S} \left(-p_{as}^T \kappa_a^1 + \bar{p}_{as}^T \kappa_a^2 + \sigma_{as}\right).
\]
Algorithm 5.2 DR-BoundExplore(b, t)

1: Input: belief state b, depth level t
2: if \( V(b) - \hat{V}(b) > \epsilon \beta_{-t} \) then
3: \( a^* \leftarrow \text{argmax}_{a \in \mathcal{A}} Q_{\text{DR}}(b, a) \)
4: \( \mu_{a^*} \leftarrow \text{argmin}_{\mu_{a^*} \in \mathcal{D}_{\text{DR}}(b, a^*, \mu_{a^*})} \)
5: \( z^*, \hat{p}_{a^*}^* \leftarrow \max_{z \in Z, \mu_{a^*} \in \mathcal{D}_{\text{DR}}(b, a^*, \mu_{a^*})} \sum_{s \in S} b_s 1^T J_s \hat{p}_{a^*}^* \times \) excess(f(b, a^*, p_{a^*}, z), t + 1)
6: \( \text{DR-BoundExplore}(f(b, a^*, p_{a^*}^*, z^*), t + 1) \)
7: \( \Lambda_V \leftarrow \Lambda_V \cup \text{DR-backup}(b, \Lambda_V) \) (see the details in Algorithm 5.3)
8: \( \Upsilon_{\text{DR}} \leftarrow \Upsilon_{\text{DR}} \cup \text{DR-update}(b, \Upsilon_{\text{DR}}) \) (see the details in Algorithm 5.4)
9: end if

We denote the optimal solutions to (5.8) using a superscript * and let the optimal dual solutions associated with constraints (5.8b) be \( \hat{p}_{a^*}^* \). For each action \( a \in \mathcal{A} \), we can generate a lower bounding hyperplane

\[
\alpha' = \left( r_{a^*} + \beta \sum_{z \in Z} \alpha_{a^*}^z J_z \hat{p}_{a^*}^*, s \in S \right)^	op,
\]

where \( \alpha_{a^*}^z = \sum_{i \in [N]} w_i^a \alpha_{a^*}^i \). We present the detailed algorithmic steps in Algorithm 5.3.

Algorithm 5.3 DR-backup(b, \( \Lambda_V \))

1: Input: belief b, lower bounding hyperplanes \( \Lambda_V \)
2: for \( \forall a \in \mathcal{A} \) do
3: solve (5.8) for action \( a \)
4: \( \mathcal{L}(a) \leftarrow \alpha' \) (calculated using (5.9))
5: end for
6: Output: \( \text{argmax}_{\alpha \in \mathcal{L}} \alpha^T b \)

Next, we discuss how to update the upper bound and describe the algorithmic steps of DR-update in Algorithm 5.4. Combining (4.18) and the dual representation of (5.3), for each \( a \in \mathcal{A} \), we solve

\[
\max_{\mu_{a^*} \in \mathcal{D}_{\text{DR}}(b, a^*, \mu_{a^*})} \sum_{s \in S} c_s^a \rho_{a^*} + \sum_{s \in S} b_s r_{a^*} + \sum_{s \in S} \left( -\hat{p}_{a^*}^T \kappa_{a^*}^1 + \hat{p}_{a^*}^T \kappa_{a^*}^2 + \sigma_{a^*} \right)
\]

s.t.

\[
\begin{align}
\beta b_s &\sum_{z \in Z} J_z \varphi_{a^*} + \beta b_s \sum_{z \in Z} \psi_{a^*} J_z 1 + \kappa_{a^*}^1 - \kappa_{a^*}^2 - 1 \sigma_{a^*} \geq 0, \quad \forall s \in S \\
\varphi_{a^*} &\in [\mathbb{R}^{|S|}], \psi_{a^*} \in \mathbb{R}, \quad \forall s \in Z, i \in ||\Upsilon_{\text{DR}}||
\end{align}
\]

where (4.17c), (4.17d), (4.18b).
Here $\phi_{az}$ and $\psi_{az}$ are the dual variables associated with the following two sets of constraints, respectively.

\begin{align}
\sum_{i \in [\Upsilon_{TV}]} w_i b^i &= b, \\
\sum_{i \in [\Upsilon_{TV}]} w_i &= 1.
\end{align}

The maximum objective value among all $a \in \mathcal{A}$ is added to $\Upsilon_{TV}$.

**Algorithm 5.4 DR-update($b, \Upsilon_{TV}$)**

1: **Input:** belief $b$, upper bounding points $\Upsilon_{TV}$
2: for $\forall a \in \mathcal{A}$ do
3: $Q(a) \leftarrow$ (optimal objective value of (5.10) for action $a$)
4: end for
5: **Output:** ($b, \max_{a \in \mathcal{A}} \{ Q(a) \}$)

### 6. Numerical Studies.

We test DR-POMDP policies for an influenza epidemic control problem (Section 6.1), and compare the results with the ones given by POMDP and by robust POMDP (Section 6.2). We also vary the parameter to test the robustness and sensitivity of DR-POMDP policies (i) under various types of ambiguity sets used in the in-sample tests (Sections 6.3, 6.4) and (ii) given certain noise added to the transition-observation probability values obtained at the end of each decision period in out-of-sample tests (Sections 6.5, 6.6). In Section 6.7, we compare solution time of different approaches for solving instances with diverse sizes.

#### 6.1. Experimental Design.

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 $a \in \{\text{Level 0, Level 1, Level 2, Inspection}\}$. 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 such as obtaining more accurate ILI rate.

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

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

When $a = 0$ (i.e., the DM does nothing), the above transition probabilities follow studies on influenza epidemics (see, e.g., [15]). The setting of the matrix (6.1) 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 $a = \text{‘Inspection’} (\Gamma)$ 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 [22] and assume that the ILI rate follows a Gaussian distribution with mean value $\mu_E = 2 - 0.5a$ and variance $\text{Var}_E = 30 - \mu_E^2$ for $s = \text{‘Epidemic’} (\Upsilon)$, and with mean $\mu_N = 0.2 - 0.05a$ and variance $\text{Var}_N = 10 - \mu_N^2$ for $s = \text{‘Non-Epidemic’} (\Lambda)$.
variance $\text{Var}_N = 2 - \mu_N^2$ for $s = \text{‘Non-epidemic’ (‘N’)}$. We discretize the observation outcome into five levels as $\{(\infty, 0), (0, 1/3), (10/3, 20/3), (20/3, 10), (10, \infty)\}$. For $a = \text{‘I’,}$ the probabilities of observing the five outcomes are $\{0.01, 0.1/3, 0.1/3, 0.1/3, 0.89\}$ when $s = \text{‘E’,}$ and the ILI rate follows the same distribution as the one of $a = 0$ if $s = \text{‘N’,}$ 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.

| 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, the discount factor $\beta = 0.95$ and we set the gap tolerance $\epsilon = 1.0$ in the HSVI algorithm. The computation is terminated when the gap between the upper and lower bounds is less than $\epsilon$, at the initial states $b_0^E = 0.5$, $b_0^N = 0.5$. 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 $\Lambda_V$, and the complexity of computing the upper bound is polynomial in the size of set $\Upsilon_V$, as we need to solve linear programs. Both $|\Lambda_V|$ and $|\Upsilon_V|$ 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%.

6.2. 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., $s = \text{‘E’,}$) by keeping all the values the same as the base setting in (6.1) 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, and also scale the radius of the ambiguity set up by 1.5 to adjust the conservatism of DR-POMDP policy.

| Nature’s policy | POMDP | DR-POMDP | DR-1.5radius | Robust |
|-----------------|-------|----------|--------------|--------|
| POMDP           | −542.72 | −560.92 | −574.34 | −576.86 |
| DR-POMDP        | −611.16 | −588.00 | −587.88 | −592.04 |
| DR-1.5radius    | −604.57 | −598.76 | −592.82 | −592.78 |
| Robust          | −603.66 | −599.83 | −594.10 | −595.71 |
Table 3: Five-percentile values of the cross-tested rewards

| Nature’s policy | POMDP       | DR-POMDP   | DR-1.5radius | Robust       |
|-----------------|-------------|------------|--------------|--------------|
| POMDP           | –656.73     | –665.95    | –689.70      | –694.83      |
| DR-POMDP        | –696.94     | –676.56    | –691.61      | –692.42      |
| DR-1.5radius    | –715.04     | –697.06    | –691.88      | –698.54      |
| Robust          | –715.26     | –704.87    | –710.32      | –690.29      |

We implement the DM’s optimal policies given by different approaches in out-of-sample environments where the nature follows the settings of POMDP, DR-POMDP, DR-POMDP with 1.5 times the mean absolute deviation radius, and robust POMDP to realize the transition probabilities in each period. The number of simulated instances is 1000 each. We report the mean and the 5-percentile values of the reward in each case in Tables 2 and 3, respectively. 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. As we increase the size of the ambiguity set (in DR-1.5radius), the mean values become even more robust with misspecified transition probabilities. Lastly, the robust POMDP policy yields worse mean value and tail performance when the true environment is POMDP or DR-POMDP.

6.3. 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

\[ \mathbb{E}_{p_{as} \sim \mu_{as}} [||p_{as} - \bar{p}_{as}||] \leq c_{as} \] for \( a = 0 \) and \( s = 'E' \), where \( \bar{p}_{as} \in \Delta(S \times Z) \) is the mean value of given probability samples and \( c_{as} \in \mathbb{R}^{|S\times Z|} \). We let \( c_{as} = c \cdot 1 \) 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 \( \bar{p}_{as} \) as the transition-observation probabilities for all \( a \) and \( s \), which corresponds to a special case of DR-POMDP with \( c = 0.00 \). Figure 2 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.
6.4. 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 3a, 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 3b, 3c, 3d.

We observe that the reward becomes smaller as we increase the number of action-
Fig. 3: 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).

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.

6.5. 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 6.3 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 4 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.

6.6. Solution Sensitivity under Noise Added to the Realized Transition-Observation Probabilities. In this section, we demonstrate 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 $p$-value obtained at the end of each period. Specifically, 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 5 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 $p$-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 $p$-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 $p$-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 $p$-value at the end of each period is not strong.

6.7. Results of Computational Time. Finally, we present the computational-time difference for different sizes of problem instances when applying the HSVI algorithm. In the base-case setting, the influenza epidemic control problem is a two-state, four-action, five-observation, one-ambiguous state-action pair problem, and we denote it as $(s2, a4, z5, a1)$ for notation convenience. We also test the ROCKSAMPLE problem in [25] for cases of $2 \times 2$ grid with one rock, and $3 \times 3$ grid with two rocks. The task of the ROCKSAMPLE problem is to move a rover in a given grid to sample the rocks whose locations are known. Each rock has binary features $\{\text{Good, Bad}\}$ representing its scientific value, and therefore the entire state space is the cross product of the location of the rover and features of the rocks. Sampling the rocks are expensive, and thus the rover is equipped with a noisy long-range sensor which identifies whether the rock has scientifically Good feature. The rover may choose an action among (i) moving towards one of the four directions, (ii) checking the feature of one of the rocks using a sensor, and (iii) sampling the rock. A reward value 10 is obtained when the
Fig. 4: 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.
Fig. 5: Statistics of the reward (mean, standard deviation, 5-percentile, 95-percentile) obtained by performing DR-POMDP policies in out-of-sample tests with noisy $p$-values.
rover successfully samples a *Good* rock, which is followed by a state transition of the sampled rock from *Good* to *Bad*. On the other hand, a $-10$ penalty is given when the rover samples a *Bad* rock. The accuracy of the sensor is dependent on the Euclidean distance of the rock and the rover, and becomes higher as the rover gets closer to the rock. Finally, the rover is given a reward $10$ when the rover reaches a terminating state located outside of the boundary on the right of the grid.

In the numerical experiment, we only consider ambiguities in the checking actions, for which we set $c = 0.02$. Thus, the ROCKSAMPLE problem sizes are $(s9, a6, z2, u9)$, and $(s37, a7, z2, u74)$. We set the initial belief to having no information about the state of the rock, and allow a tolerance $\epsilon = 1.0$. The computational time limit is $3600$ seconds.

![Graph](image1)

(a) POMDP ($c = 0.00$)  
(b) DR-POMDP ($c = 0.02$)

Fig. 6: Influenza epidemic control problem $(s2, a4, z5, u1)$. Solid line: lower bound, dashed line: upper bound

In Figures 6, 7, 8, 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. While the computational time between POMDP and DR-POMDP is similar for the influenza epidemic control problem, we observe that DR-POMDP requires more time than POMDP for larger-sized ROCKSAMPLE instances. For the ROCKSAMPLE problem $(s9, a6, z2, u9)$, it took only $0.05$ seconds for POMDP to find a policy with gap less than $1.0$, while DR-POMDP policy required $100.5$ seconds. When terminating the computation for the ROCKSAMPLE problem $(s37, a7, z2, u74)$, the DR-POMDP policy has the gap of $10.6$ between the upper and lower bounds, while the POMDP policy finds an optimal policy in $59.8$ seconds.

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 set-

This manuscript is for review purposes only.
tings, 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

This manuscript is for review purposes only.
interested in designing randomized policy or time-dependent policy for DR-POMDP
instead of a deterministic strategy which our method is based on in this paper. 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 ac-
knowledge the support from the U.S. Department of Engineering (DoE) grant #
DE-SC0018018 and National Science Foundation (NSF) grant # CMMI-1727618.

References.
[1] M. Abbad and J. A. Filar, Perturbation and stability theory for Markov con-
trol problems, IEEE Transactions on Automatic Control, 37 (1992), pp. 1415–
1420.
[2] M. Abbad, J. A. Filar, and T. R. Bielecki, Algorithms for singularly per-
turbed limiting average Markov control problems, in Decision and Control, 1990.,
Proceedings of the 29th IEEE Conference on, IEEE, 1990, pp. 1402–1407.
[3] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Meelenberg, and
G. Rennen, Robust solutions of optimization problems affected by uncertain
probabilities, Management Science, 59 (2013), pp. 341–357.
[4] A. R. Cassandra, A survey of POMDP applications, in Working notes of AAAI
1998 Fall Symposium on planning with partially observable Markov decision pro-
cesses, 1998, pp. 17–24.
[5] E. Delage and S. Mannor, Percentile optimization for Markov decision pro-
cesses with parameter uncertainty, Operations Research, 58 (2010), pp. 203–213.
[6] E. Delage and Y. Ye, Distributionally robust optimization under moment
uncertainty with application to data-driven problems, Operations Research, 58
(2010), pp. 595–612.
[7] D.-Z. Du and P. M. Pardalos, Minimax and Applications, vol. 4, Springer
Science & Business Media, 2013.
[8] X. Du, A. A. King, R. J. Woods, and M. Pascual, Evolution-informed fore-
casting of seasonal influenza a (h3n2), Science translational medicine, 9 (2017),
p. eaan5325.
[9] P. M. Esfahani and D. Kuhn, Data-driven distributionally robust optimization
using the Wasserstein metric: Performance guarantees and tractable reformula-
tions, Mathematical Programming, 171 (2018), pp. 115–166.
[10] R. Gao and A. J. Kleywegt, Distributionally robust stochastic optimization
with Wasserstein distance, arXiv preprint arXiv:1604.02199, (2016).
[11] M. Hauskrecht and H. Fraser, Planning treatment of ischemic heart dis-
ease with partially observable Markov decision processes, Artificial Intelligence in
Medicine, 18 (2000), pp. 221–244.
[12] G. N. Iyengar, Robust dynamic programming, Mathematics of Operations Re-
search, 30 (2005), pp. 257–280.
[13] R. Jiang and Y. Guan, Data-driven chance constrained stochastic program,
Mathematical Programming, 158 (2016), pp. 291–327.
[14] P. R. Kumar and P. Varaiya, Stochastic Systems: Estimation, Identification,
and Adaptive Control, vol. 75, SIAM, 2015.
[15] Y. Le Strat and F. Carrat, Monitoring epidemiologic surveillance data using
hidden markov models, Statistics in medicine, 18 (1999), pp. 3463–3478.
[16] H. Nakao, R. Jiang, and S. Shen, An approximate upper bounding approach
for robust partially observable Markov decision process, IEEE Conference on De-
cision and Control, submitted, (2019).
[17] A. Nilim and L. El Ghaoui, Robust control of Markov decision processes with uncertain transition matrices, Operations Research, 53 (2005), pp. 780–798.
[18] T. Osogami, Robust partially observable Markov decision process, in International Conference on Machine Learning (ICML), 2015, pp. 106–115.
[19] J. Pineau, G. Gordon, and S. Thrun, Point-based value iteration: An anytime algorithm for POMDPs, in The Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI), vol. 3, 2003, pp. 1025–1032.
[20] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.
[21] M. Rasouli and S. Saghafian, Robust partially observable Markov decision processes, Working paper, (2018).
[22] T. M. Rath, M. Carreras, and P. Sebastiani, Automated detection of influenza epidemics with hidden markov models, in International Symposium on Intelligent Data Analysis, Springer, 2003, pp. 521–532.
[23] S. Saghafian, Ambiguous partially observable Markov decision processes: Structural results and applications, Journal of Economic Theory, 178 (2018).
[24] R. D. Smallwood and E. J. Sondik, The optimal control of partially observable Markov processes over a finite horizon, Operations Research, 21 (1973), pp. 1071–1088.
[25] T. Smith and R. Simmons, Heuristic search value iteration for POMDPs, in Proceedings of the 20th conference on Uncertainty in artificial intelligence, UAI Press, 2004, pp. 520–527.
[26] J. T. Treharne and C. R. Sox, Adaptive inventory control for nonstationary demand and partial information, Management Science, 48 (2002), pp. 607–624.
[27] W. Wiesemann, D. Kuhn, and B. Rustem, Robust Markov decision processes, Mathematics of Operations Research, 38 (2013), pp. 153–183.
[28] W. Wiesemann, D. Kuhn, and M. Sim, Distributionally robust convex optimization, Operations Research, 62 (2014), pp. 1358–1376.
[29] H. Xu and S. Mannor, Distributionally robust Markov decision processes, Mathematics of Operations Research, 37 (2012), pp. 288–300.
[30] I. Yang, A convex optimization approach to distributionally robust Markov decision processes with Wasserstein distance, IEEE Control Systems Letters, 1 (2017), pp. 164–169.
[31] P. Yu and H. Xu, Distributionally robust counterpart in Markov decision processes, IEEE Transactions on Automatic Control, 61 (2016), pp. 2538–2543.
[32] J. Zhang and B. T. Denton, 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, (2018), pp. 201–222.
[33] S. Zymler, D. Kuhn, and B. Rustem, Distributionally robust joint chance constraints with second-order moment information, Mathematical Programming, 137 (2013), pp. 167–198.
SUPPLEMENTARY MATERIALS: DISTRIBUTIONALLY ROBUST PARTIALLY OBSERVABLE MARKOV DECISION PROCESS WITH MOMENT-BASED AMBIGUITY

HIDEAKI NAKAO†, RUIWEI JIANG‡, AND SIQIAN SHEN§

SM1. Relaxation of \(\alpha\)-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 [1], who defined the term in the context of robust MDP. Ref. [1] 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:

\[
V'(b) = \max_{\phi \in \Delta(\mathcal{A})} \min_{s \in \mathcal{S}} \mathbb{E}_{p \sim \mu} \left[ \sum_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} b_{s'} \left( r_{s,a} + \beta \sum_{z \in \mathcal{Z}} J_z p_{sa} V'^{t+1} (f(b,a,p_a,z)) \right) \right],
\]

where \(\phi_a\) is the probability for selecting action \(a\). We define the ambiguity set to be

\[
\hat{\mathcal{D}}_s = \left\{ \tilde{\mu}_s \left( \begin{array}{c} p_s \\ r_s \\ \hat{u}_s \end{array} \right) \mid \mathbb{E}_{p \sim \tilde{p}_s} \left[ E_s p_s + G_s r_s + H_s \hat{u}_s \right] = c_s, \right\}
\]

where \(\hat{u}_s \in \mathbb{R}^Q\) is a vector of auxiliary variables, and

\[
\mathcal{X}_s = \left\{ \left( \begin{array}{c} p_s \\ r_s \\ \hat{u}_s \end{array} \right) \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{Z}|} \mid B_s p_s + C_s r_s + E_s \hat{u}_s \preceq d_s \right\}.
\]

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

The value function is also convex in the form (4.10), since for \(t < T\),

\[
V'(b) = \max_{\phi \in \Delta(\mathcal{A})} \max_{\alpha \in \text{Conv}(\Lambda^{t+1})} \sum_{s \in \mathcal{S}} b_s \min_{(p_s,r_s,\hat{u}_s)} \phi^{\top} \left( \beta \sum_{z \in \mathcal{Z}} \left( \alpha_{az} J_{az} \right)^{\top} + 1 \right) \hat{p}_s + \hat{r}_s \]

subject to \(F_s \hat{p}_s + G_s \hat{r}_s + H_s \hat{u}_s = c_s, \) \(\forall s \in \mathcal{S}\)

\(B_s \hat{p}_s + C_s \hat{r}_s + E_s \hat{u}_s \preceq d_s, \) \(\forall s \in \mathcal{S}\)

where \(J_{az} \in \mathbb{R}^{(|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{Z}|)}\) is a matrix of zeros and ones that maps \(p_s\) to \(p_{azs}\).

For an exact algorithm, we solve the inner minimization problem for all \(\phi \in \Delta(\mathcal{A})\),

---

\*Funding: The U.S. Department of Engineering (DoE) grant # DE-SC0018018.
†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.
The optimal objective is used for constructing the set $\Lambda^t$, at each time step $t$.

**SM2. 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 $a \in A$ and $s \in S$, we define a non-empty ambiguity set

$$\tilde{D}_{as} = \left\{ (\tilde{\mu}_{as}, \tilde{u}_{as}) \mid \begin{bmatrix} F_{as} p_{as} + G_{as} r_{as} + H_{as} \tilde{u}_{as} \end{bmatrix} = c_{as}, \tilde{\mu}_{as}(\mathcal{X}_{as}) = 1 \right\},$$

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

$$\mathcal{X}_{as} = \left\{ \begin{pmatrix} p_{as} \\ r_{as} \\ u_{as} \end{pmatrix} \in \mathbb{R}^{(|S| \times |Z|)} \mid \begin{bmatrix} B_{as} p_{as} + C_{as} r_{as} + E_{as} \tilde{u}_{as} \end{bmatrix} \preceq_{K_{as}} d_{as} \right\}.$$

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

**SM3. Illustration of the Lower and Upper Bounds in HSVI.** Figure SM3.1 illustrates the upper and lower bounding ideas introduced in Section 5.1 when running a HSVI algorithm for solving DR-POMDP in a two-state system. The solid
line represents the true value function, the dots are the elements of \( \Upsilon_V \), and the dotted lines are the hyperplanes in \( \Lambda_V \). The dashed lines and the thicker dotted lines are upper and lower bounds, respectively. Note that some hyperplanes in \( \Lambda_V \) are completely dominated by some of the others, and some points are greater than the convex combination of other points. These are the dominated elements of \( \Lambda_V \) and \( \Upsilon_V \), and need to be periodically eliminated for better computational performance.

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