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

Hideaki Nakao∗  Ruiwei Jiang†  Siqian Shen‡

Abstract

We consider a distributionally robust (DR) formulation of partially observable Markov decision process (POMDP), where the transition probabilities and observation probabilities are random and unknown, only revealed at the end of every time step. We construct the ambiguity set of the joint distribution of the two types of probabilities using moment information bounded via conic constraints and show that the value function of DR-POMDP is convex with respect to the belief state. We propose a heuristic search value iteration method to solve DR-POMDP, which finds lower and upper bounds of the optimal value function. Computational analysis is conducted to compare DR-POMDP with the standard POMDP using random instances of dynamic machine repair and a ROCKSAMPLE benchmark.

Keywords: POMDP, distributionally robust optimization, moment-based ambiguity set, heuristic value search iteration algorithm

1 Introduction

Partially observable Markov decision processes (POMDPs) model sequential decision making problems where a decision maker (DM) is only able to obtain partial information about the present state of a targeted system. Similar to the Markov decision processes (MDPs), the transition probabilities in between the states of the system are only dependent 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 according to (i) the action of the DM and (ii) the state into

∗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.
which the system has transitioned. As opposed to MDP where the DM is able to directly observe
the current state of the system, in POMDP the DM can only view the observation of the state.
This generalization allows modeling cases where measurement information is not entirely reliable.
Applications of POMDPs span over multiple research areas including controlling robots, clinical
decision making, and dynamic inventory control, which are described in e.g., Cassandra (1998),
Hauskrecht and Fraser (2000), and Treharne and Sox (2002).

A general objective of sequential decision making problems is to devise a policy of taking
dynamic actions to maximize (minimize) the expected value of the cumulative reward (cost). In
MDP, the DM gains a reward (or pays a cost) for each action made on a state of the system. In
POMDP, since the DM has no access to the true state, he/she is uncertain about the reward (cost)
received. Instead, the DM retains his/her belief of the state of the system based on the past history
of the actions and observation outcomes, and gains an expected value of the reward (or pays an
expected value of the 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 (see, e.g., Kumar and Varaiya (2015)).

POMDP assumes that the exact values of the transition and observation probabilities are known
to the DM. In practice, the transition-observation probabilities may not have fixed values, and can
change over time. In this paper, we consider a situation where the DM obtains the exact values
of the time-varying transition-observation probabilities after the action at each time step has been
taken. We propose a concept of ambiguity set to describe possible values of transition-observation
probabilities, and employ distributionally robust optimization to seek an optimal policy against the
worst-case realization of the transition-observation probability in the ambiguity set.

We use an example of dynamic machine repair problem to justify the above problem setting.
A machine has two states, functioning (F) and broken (B), and two possible observation outcomes
corresponding to F and B. At each step, the DM can perform one of the four actions: (i) simple-
repair (S) which has a high fixing rate (i.e., a rate of transitioning from B to F) but also has a high
rate of breaking a functioning machine (i.e., transitioning from F to B), (ii) overhaul (O) which
is capable of detecting the true state information with high accuracy but has a low fixing rate,
(iii) testing (T) which has a moderate detection rate but does not change the state, and (iv) doing
nothing (N).

Suppose that the probability of a machine breakdown (i.e., the transition probability from F to
B for each action) is dependent on the temperature during each time step. The relation between the
transition probability and the temperature is well known to the DM, but future temperature values are hard to predict. Let us further assume that the temperature is i.i.d., following a distribution unknown to the DM. Instead, the DM has an information about what family of distributions the true distribution of temperature is contained in. To model this, we take a distributionally robust (DR) optimization approach (see, e.g., Delage and Ye, 2010; Esfahani and Kuhn, 2018) and formulate DR-POMDP. In DR-POMDP, the DM assumes that the unknown parameter (i.e., the transition-observation probability matrix) follows a certain unknown distribution, which is included in a family of distributions, namely, an ambiguity set. The goal is to maximize (minimize) the total expected cumulative reward (cost) for the worst-case distribution in the ambiguity set.

Figure 1 depicts the sequence of events that occur during a single time step. Here, we consider another agent (the “nature”), who realizes the distribution of transition-observation probability at each time step. The DM expects that the nature acts adversely against his/her action \( a \) chosen at the beginning of each time step. Afterwards, the nature chooses a distribution \( \mu \) in the ambiguity set to realize the worst-case outcome of the reward. Next, the transition-observation probability \( p \) is realized from the true distribution \( \mu \). The state makes a transition according to \( p \), and the observation outcome \( z \) is realized. Finally, the DM obtains the values of \( z \) and \( p \) and we move on to the next time step to repeat the foregoing process.

![Figure 1: Sequence of events in a DR-POMDP](image.png)

In dynamic machine repair, the machine is either in state \( s_t = F \) or \( B \) at the beginning of time \( t \). The DM does not know the exact transition-observation probabilities when choosing an action \( a \) from \( S, O, T, \) and \( N \). Afterwards, the nature realizes a distribution \( \mu' \) of the temperature, from an ambiguity set known to the DM. Since the DM knows the relations between the temperature and the transition-observation probabilities \( p \), this is equivalent to nature choosing a distribution
\( \mu \) of \( p \). Then, from distribution \( \mu' \), the value of the current temperature is realized, as well the value of \( p \). Following \( p \), the value of the observation outcome \( z = F \) or \( B \) and the value of the next state \( s_{t+1} = F \) or \( B \) are realized. The DM then finds out the actual temperature surrounding the machine and therefore finds out the value of \( p \), as well as the observation outcome \( z \). The machine moves to state \( s_{t+1} \), and the process repeats again.

Another salient example to justify the assumed DR-POMDP is dynamic inventory control (or production planning) (see Treharne and Sox (2002)). Suppose that there are finite number of states corresponding to trends of a market at each time period. The trend makes a transition according to a probability mass function that is unknown to the DM. Each trend is associated with a certain distribution of product demand, which the DM tries to satisfy. The DM maintains a certain level of inventory corresponding to a discrete state. Therefore, the entire system can be described as a Cartesian product of trend state and inventory state. Suppose that the demand distribution and thus the transition probability is correlated with climate, such as temperature and precipitation. At the beginning of each period, the DM makes a production plan and then nature “decides” the distribution of temperature and precipitation. The DM observes the realizations of temperature and precipitation, to identify the value of \( p \), and also observes the value of demand. The state then makes a transition according to \( p \) and DM’s production decision, and the DM receives a reward according to how much demand he/she is able to satisfy, or pays a stocking cost. Assuming a family of distributions of unknown climate, the DM aims to maximize the worst-case revenue given the nature being an adversary. The above problem is especially important for planning orders or production in agriculture.

The main research in this paper is to develop a DR formulation of POMDP and analyze its properties, as well as to investigate efficient computational methods, when assuming the accessibility of transition-observation probability at the end of each time. Section 2 provides a comprehensive review of the related literature in MDP, POMDP, and distributionally robust optimization. In Section 3, we formulate the DR Bellman equation and show that the value function is convex, when the ambiguity set is characterized by moments as in Yu and Xu (2016), and introduce several examples of moment-based ambiguity set. In Section 4, we present an approximation algorithm for DR-POMDP for infinite-horizon case by using a DR variant of the heuristic value search iteration (HVSI) algorithm. Numerical studies are presented in Section 5 to compare DR-POMDP with POMDP, and to demonstrate properties of DR-POMDP solutions based on randomly generated observation outcomes. We conclude the paper and describe future research in Section 6.
2 Literature Review

2.1 POMDP

Although there exist strong modeling connections between MDP and POMDP, the technique applied to solve the standard MDP, where the states are discrete, is not directly applicable to solving POMDP since belief states are continuous. Smallwood and Sondik (1973) show that the value function of POMDP is piecewise linear and convex (PWLC) with respect to the belief state, and present 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 at each time step. Pineau et al. (2003) propose a point-based value iteration (PBVI) algorithm, which only keeps the characterizing vectors for a subset of belief states, and thus maintains the lower bound of the true value function. 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 as the subset of belief states are sampled densely in the reachable set of belief states. Smith and Simmons (2004) develop a heuristic search value iteration (HSVI) algorithm to generate an upper bound for the value function by finding the reachable belief through Monte-Carlo simulation. Smith and Simmons (2004) show that HSVI is guaranteed to terminate after the gap between the upper and lower bounds converges to below a certain threshold.

2.2 DR Optimization

DR optimization is an emerging field in Operations Research where solutions are chosen to optimize the worst-case objective outcome via a risk measure among possible distributions contained in an ambiguity set. Compared with robust optimization, which accounts for the worst-case outcome among all possible realizations of uncertain parameters in an uncertainty set, optimal solutions to DR optimization models are less conservative and can be adjusted through the amount of data/information we have. Different ambiguity sets yielding tractable convex models have been proposed and studied in the past decade. Delage and Ye (2010) develop 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. Standardization of ambiguity sets via conic representable sets is proposed by Wiesemann et al. (2014). Ben-Tal et al. (2013) formulate a DR counterpart for models with distributions bounded with a known $\phi$-divergence to a given distribution. Esfahani and Kuhn (2018) and Gao and Kleywegt (2016) consider an ambiguity set characterized by a set
of distributions whose Wasserstein distance to a nominal distribution is less than a pre-determined value. Zymler et al. (2013) consider tractable reformulations of DR chance-constrained programs using moment-based ambiguity set, while Jiang and Guan (2016) employ $\phi$-divergence-based ambiguity sets to derive reformulations of DR chance constraints. Recently, Chen et al. (2018) and Xie (2018) demonstrate tractable reformulations of DR chance-constrained models under Wasserstein-distance-based ambiguity set.

2.3 Robust and Distributionally Robust MDP

Nilim and El Ghaoui (2005) consider robust MDP, where the transition probabilities are uncertain. This is motivated from the fact that the estimation errors of the transition matrices may have a significant impact to the quality of solution (see, e.g., Abbad and Filar (1992), and Abbad et al. (1990)). As discussed in Delage and Mannor (2010), the use of robust optimization on MDP may lead to overly conservative solutions, as it does not incorporate some prior knowledge on the distribution of uncertain parameters. Based on this idea, Xu and Mannor (2012) present a model of DR-MDP, 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. (Note that the robust MDP is a special case of the aforementioned DR-MDP as the robust MDP can be described as having a single set with 100% confidence level.) Yu and Xu (2016) generalize DR-MDP to include multi-modal distributions and information of mean and variance. In a different approach, Yang (2017) propose a DR-MDP model, which considers an ambiguity set of distributions on transition probability characterized by a Wasserstein ball centered around a nominal distribution. The use of Wasserstein ball ambiguity set results in a Kantorovich-duality-based convex reformulation for DR-MDP.

2.4 Robust POMDP

Saghafian (2018) presents the framework of ambiguous POMDP (called APOMDP), which generalizes the robust POMDP. APOMDP optimizes over the $\alpha$-maxmin expected utility, which leads to a policy that achieves the intermediate performance of the worst case and the best case in the uncertainty set of parameters. Saghafian (2018) presents the conditions for which the value function of the APOMDP is PWLC, as it is not convex in the general case.

Meanwhile, Rasouli and Saghafian (2018) consider a general setting of robust POMDP, where the DM may not be able to obtain the exact information of the nature’s choice of decision. 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 time must be taken into account to realize a policy that is robust in terms of the entire cumulative expected reward. Rasouli and Saghafian (2018) derives an exact algorithm for the case where the uncertainty set is discrete. Here we note that robust POMDP for continuous support of uncertainty sets is computationally challenging even in a very simple setting (Nakao et al., 2019). Nakao et al. (2019) provide an algorithm to compute an upper bound of the value function for robust POMDP under the assumptions made above.

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 decision of the nature is assumed to be observable to the DM, which is similar to the setting of our 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 piece-wise linear, as there are possibly infinitely many supporting hyperplanes. Osogami (2015) then derives an efficient algorithm using the PBVI method to approximate the exact solution, and discusses the method to conduct a robust belief update.

3 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 time step. In Section 3.1, we describe the notation used throughout this paper. Then in Section 3.2, we formulate DR-POMDP as an optimization problem and construct a Bellman equation to derive the optimal policy. In Section 3.3, we show that the value function satisfying the Bellman equation is PWLC. Finally, in Section 3.4, we consider the infinite horizon case, and demonstrate that the value function converges under the Bellman update operation.

3.1 Notation

Let $S$ be the set of states, $A$ be the set of actions, and $Z$ be 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 the choice of action $a$. For $(s, a) \in S \times A$, let $r_{as}$ be the reward for taking action $a$ at state $s$.

To model the case where we neither know the exact probabilities for the states to make transi-
tions to another, nor the exact probability distribution for the measurement outcome, we consider a set of distributions on the transition-observation probability matrix. For all \( s \in \mathcal{S} \), \( a \in \mathcal{A} \), we define a vector of probabilities \( p_{as} = (p_{as}(s', z), (s', z) \in \mathcal{S} \times \mathcal{Z})^\top \) and assume that the Cartesian product \( (p_{as}, r_{as}) \) 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 further denote \( p_a = (p_{as}(s', z), (s, s', z) \in \mathcal{S}^2 \times \mathcal{Z})^\top \) and \( r_a = (r_{as}, s \in \mathcal{S})^\top \) for all \( a \in \mathcal{A} \). We assume that \( (p_{as}, r_{as}) \) follows a distribution \( \mu_{as} \), which is unknown but is included in an ambiguity set \( \mathcal{D}_{as} \subseteq \mathcal{P}(\mathcal{X}_{as}) \), where \( \mathcal{P}(\cdot) \) is a set of all probability distributions with support \( \cdot \).

Note that the support of the distribution is a set of values of the reward, and a set of joint probabilities of transition and observation. Note that for a convex support of the transition-observation probability, we cannot consider an individual convex support on transition probability and observation probability, because transition-observation probability is a product of the two probabilities. 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 \( \mathcal{D} = \bigotimes_{a \in \mathcal{A}} \bigotimes_{s \in \mathcal{S}} \mathcal{D}_{as}. \) This assumption is analogous to the \((s, a)\)-rectangularity in Wiesemann et al. (2013). These conditions increase the conservativeness of the model in general, as realizing the worst case scenario for all \((s, a)\) is typically rare. In Appendix A, we discuss a relaxation of the \( a \)-rectangularity assumption for DR-POMDP.

### 3.2 Problem Formulation and DR Bellman Equation

We formulate a dynamic game involving two players: The DM selects \( a \in \mathcal{A} \) to maximize the expected reward, and the nature selects \( \mu_{as} = \bigotimes_{s \in \mathcal{S}} \mu_{as} \) from the ambiguity set \( D_a = \bigotimes_{s \in \mathcal{S}} D_{as} \) to minimize the expected reward given the DM’s action \( a \). Let \( a^t, \ p_{a^t}^t, \ z^t \) be the action, transition-observation probability outcome, and observation at time \( t \), and \( h^t = (a^1, p_{a^1}^1, z^1, \ldots, a^{t-1}, p_{a^{t-1}}^{t-1}, z^{t-1}) \) be a history at time \( t \), with \( \mathcal{H}^t \) being the set of all histories up to time \( t \). The objective for the DM is to find an optimal policy of selecting an action \( a \in \mathcal{A} \) based on the history for \( t = 1 \) to \( T \), i.e., finding the best \( \pi = (\pi^1, \ldots, \pi^{T-1}) \), with \( \pi^t : \mathcal{H}^t \to \mathcal{A} \). Denote the set of all such policies as \( \Pi \). Define an extended history \( \tilde{h}^t = (a^1, p_{a^1}^1, z^1, \ldots, a^{t-1}, p_{a^{t-1}}^{t-1}, z^{t-1}, a^t) \in \mathcal{H}^t \), on which the nature bases its decision for choosing \( \mu_{a^t} \). The objective for the nature is therefore to find the best policy (from the nature’s perspective) \( \gamma = (\gamma^1, \ldots, \gamma^{T-1}) \), with \( \gamma^t : \mathcal{H}^t \to \mathcal{D}_{a^t} \). Similarly, we denote the set of all the nature’s policies as \( \Gamma \).

The belief state is a sufficient statistic for DR-POMDP in the case where the DM is able to obtain the value of transition-observation probability at the end of each time (see Rasouli and Saghaian, 2018). Let the belief state at time \( t \) be \( \{b^t_s, s \in \mathcal{S}\} = b^t \in \Delta(\mathcal{S}) \). Given action \( a \),
transition-observation probability $p_a$, and observation outcome $z$, the sufficient statistic for the history $h^{t+1} = (h^t, a, p_a, z)$, or the belief state at time $t + 1$ is given by

$$b^{t+1} = f(b, a, p_a, z) = \frac{\sum_{s \in S} J_s p_{as} b_s}{\sum_{s \in S} 1^\top J_s p_{as} b_s},$$

where we define $1$ as a vector of ones with length $|S|$, and $J_s \in \mathbb{R}^{|S| \times (|S| \times |Z|)}$ as a matrix of zeros and ones, which projects the vector $p_{as}$ to a vector $p_{asz} = (p_{as}(s', z), s' \in S)^\top$, whose components are associated with outcome $z$. That is, $p_{asz} = J_s p_{as}$. With slight abuse of notation, let $\pi$ be a policy that maps belief states to the actions, i.e., $\pi_t : \Delta(S) \rightarrow A$ for all $t \in \{1, \ldots, T-1\}$. Similarly, let $\gamma_t : \Delta(S) \times A \rightarrow D_{a_t}$ 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.

Given the nature’s choice of distribution, the expected value of the instantaneous reward given belief state $b$ and action $a$ is denoted as $\mathbb{E}_{(p_a, r_a) \sim \mu_a} [b^\top r_a]$, 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 which maximizes the worst-case cumulative discounted expected reward of all of the nature’s policies choosing the distribution $\mu$. That is, DR-POMDP aims to solve

$$\max_{\pi \in \Pi} \min_{\gamma \in \Gamma} \mathbb{E} \left[ \sum_{t=1}^{T-1} \beta^t b^\top r_{a_t}^t \right] \tag{2a}$$

s.t. $a^t = \pi^t(b^t)$, $\forall t \in \{1, \ldots, T-1\}$

$$\mu_{a_t}^t = \gamma^t(b^t, a^t), \quad \forall t \in \{1, \ldots, T-1\} \tag{2b}$$

$$(p_{as}^t, r_{a_t}^t) \sim \mu_{a_t}^t, \quad \forall t \in \{1, \ldots, T-1\} \tag{2c}$$

$$(s^{t+1}, z^t) \sim p_{asz}^t, \quad \forall t \in \{1, \ldots, T-1\} \tag{2d}$$

$$b^{t+1} = f(b^t, a^t, p_{asz}^t, z^t), \quad \forall t \in \{1, \ldots, T-1\} \tag{2e}$$

where the terminal reward is zero without loss of generality. The initial belief state is given as $b$.

Alternatively, we denote the problem (2) as

$$\max_{\pi \in \Pi} \min_{\gamma \in \Gamma} \mathbb{E}^\pi \mathbb{E}^\gamma \left[ \sum_{t=1}^{T-1} \beta^t b^\top r_{a_t}^t \left| b^1 = b \right. \right]. \tag{3}$$

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

**Proposition 1** Let $\pi^{t:T-1} = (\pi^t, \pi^{t+1}, \ldots, \pi^{T-1})$, and $\gamma^{t:T-1} = (\gamma^t, \gamma^{t+1}, \ldots, \gamma^{T-1})$ be a sequence of policies from $t$ to $T-1$, and $\Pi^{t:T-1}, \Gamma^{t:T-1}$ be sets of all policies of $\pi^{t:T-1}$ and $\gamma^{t:T-1}$, respectively.
Define
\[
V^t(b) = \max_{\pi^t:T-1 \in \Pi^t:T-1, \gamma^t:T-1 \in \Gamma^t:T-1} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} V^{t+1} \left( f (b, a, p_{a}, z) \right) \right\} + \sum_{s \in S} b_s 1^T J_s p_{\pi^{t}(b)s} \right].
\]
(4)

Then,
\[
V^t(b) = \max_{a \in A} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} V^{t+1} \left( f (b, a, p_{a}, z) \right) \right\} \right].
\]
(5)

Proof: We begin by isolating the term associated to time \( t \) inside the expectation of (4) as follows.
\[
V^t(b) = \max_{\pi^t:T-1 \in \Pi^t:T-1, \gamma^t:T-1 \in \Gamma^t:T-1} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} V^{t+1} \left( f (b, a, p_{a}, z) \right) \right\} \right].
\]

Given \( a^t = \pi^t(b), \ p^t_a = p_{\pi^t(b)}, \ z^t = z \), the probability of observing \( z \) is
\[
\sum_{s \in S} b_s 1^T J_s p_{\pi^{t}(b)s}.
\]
Thus the expectation conditioned on the values of \( a^t, \ p^t_a, \ z^t \) is
\[
V^t(b) = \max_{\pi^{t-1} \in \Pi^{t-1}, \gamma^{t-1} \in \Gamma^{t-1}} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} \right\} + \sum_{s \in S} b_s 1^T J_s p_{\pi^{t}(b)s} \right]
\]
\[
= \max_{\pi^{t-1} \in \Pi^{t-1}, \gamma^{t-1} \in \Gamma^{t-1}} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} \right\} + \sum_{s \in S} b_s 1^T J_s p_{\pi^{t}(b)s} \right]
\]
where the second equality is due to rearranging the terms and the fact that \( b \) is an information state. Because policies beyond time \( t \) do not affect \( (p^t_a, r^t_a) \), we have
\[
V^t(b) = \max_{a \in A} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} \right\} + \sum_{s \in S} b_s 1^T J_s p_{\pi^{t}(b)s} \right]
\]
\[
= \max_{a \in A} \min_{\mu_{a} \in \mathcal{D}_a} E_{(p_{a}, r_{a}) \sim \mu_{a}} \left[ \sum_{n \in S} b_n \left\{ r_{a_n} + \beta \sum_{z \in Z} 1^T J_z p_{a_n} \right\} + \sum_{s \in S} b_s 1^T J_s p_{\pi^{t+1}(b)s} \right]
\]
(5).

The final equality follows the definition of \( V^{t+1} \). This completes the proof.

By Proposition 1, the policies optimal to (3) can be determined by recursively solving (5) from time \( T \) to \( t = 1 \).
Now denote
\[
U^t(b, a, µ_a) = \mathbb{E}_{(p_a, r_a) \sim µ_a}\left[ \sum_{s \in S} b_s \left\{ r_{as} + \beta \sum_{z \in Z} 1^T J_z p_as V^{t+1} (f(b, a, p_a, z)) \right\} \right],
\]
and
\[
Q^t(b, a) = \min_{µ_a \in D_a} U^t(b, a, µ_a).
\]
The solution to the Bellman equation provides the optimal action given belief state \(b\). That is, an optimal action for the DM at time \(t\) is
\[
\arg \max_{a \in A} Q^t(b, a),
\] whereas the optimal choice of distribution for the nature, provided the value of the belief \(b\) and the DM’s action \(a\), is
\[
\arg \min_{µ_a \in D_a} U^t(b, a, µ_a).
\]

### 3.3 Solving DR Bellman Equation (5)

In the following sections, we consider a special case of an ambiguity set considered in Wiesemann et al. (2014) and Yu and Xu (2016), where we do not assume “nested set” structure characterizing the confidence level of unknown parameters. Instead, we only consider the set of distributions 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
\[
\bar{D}_{as} = \left\{ \begin{pmatrix} p_{as} \\ r_{as} \\ u_{as} \end{pmatrix} \mid \mathbb{E}_{(p_{as}, r_{as}, u_{as}) \sim \bar{µ}_as} [F_{as} p_{as} + G_{as} r_{as} + H_{as} u_{as}] = c_{as}, \bar{µ}_as (X_{as}) = 1 \right\},
\]
where \(u_{as} \in \mathbb{R}^L\) is a vector of auxiliary variables, and a support with a non-empty relative interior
\[
X_{as} = \left\{ \begin{pmatrix} p_{as} \\ r_{as} \\ u_{as} \end{pmatrix} \in \mathbb{R}^{|S| \times |Z|} \times \mathbb{R} \times \mathbb{R}^L \mid B_{as} p_{as} + C_{as} r_{as} + E_{as} u_{as} \preceq_{K_{as}} d_{as} \right\}.
\]

Here, \(F_{as} \in \mathbb{R}^{k \times (|S| \times |Z|)}, G_{as} \in \mathbb{R}^{k \times 1}, H_{as} \in \mathbb{R}^{k \times L}, C_{as} \in \mathbb{R}^k, B_{as} \in \mathbb{R}^{\ell \times (|S| \times |Z|)}, C_{as} \in \mathbb{R}^{\ell \times 1}, E_{as} \in \mathbb{R}^{\ell \times L},\) and \(d_{as} \in \mathbb{R}^\ell\). The symbol \(\preceq_{K_{as}}\) represents a generalized inequality with respect to a proper cone \(K_{as}\). We denote the marginal distribution by \(µ_{as} = \prod_{(p_{as}, r_{as}) \sim \bar{µ}_{as}}\bar{µ}_{as}\), and also extend the definition to the ambiguity set so that \(D_{as} = \prod_{(p_{as}, r_{as}) \sim \bar{µ}_{as}}\bar{D}_{as} = \bigcup_{\bar{µ}_{as} \in \bar{D}_{as}} \prod_{(p_{as}, r_{as})} \bar{µ}_{as} \). The
auxiliary variables $\tilde{u}_as$ are used for “lifting” techniques, enabling the representation of nonlinear constraints to linear ones. We introduce some examples of ambiguity sets which can be modelled using (8) and (9):

1. **Mean absolute deviation**: Suppose that the expected value of the deviation of the transition-observation probability from $\bar{p}_as$ is at most $c_as$. Then, $\mu_as$ satisfies $E_{p_as \sim \tilde{\mu}_as} \left[ | p_as - \bar{p}_as | \right] \leq c_as$. This can be reformulated as

$$
E_{p_as \sim \tilde{\mu}_as} [\tilde{u}_as] = c_as,
$$

$$
\tilde{\mu}_as \left( \begin{array}{c}
\tilde{u}_as \geq p_as - \bar{p}_as, \\
1^\top p_as = 1,
\end{array} \right) = 1.
$$

2. **Mean**: Suppose that the true value of the mean of $p_as$ is constrained as $F_as \ E_{p_as \sim \tilde{\mu}_as} [p_as] \preceq K_as \ c_as$, for some proper cone $K_as$. This can be reformulated as

$$
E_{p_as \sim \tilde{\mu}_as} [\tilde{u}_as] = c_as,
$$

$$
\tilde{\mu}_as \left( \begin{array}{c}
F_as p_as \preceq K_as \tilde{u}_as, \\
1^\top p_as = 1, \\
p_as \geq 0
\end{array} \right) = 1.
$$

For ambiguity sets and supports defined in terms of (8) and (9), respectively, we show that the value function is convex with respect to the belief state $b$.

**Theorem 1** For all $a \in A$ and $s \in S$, let the ambiguity set and support be (8) and (9). For all $t \in \{1, \ldots, T\}$, the value function can be expressed in the form

$$
V^t(b) = \max_{\alpha \in \Lambda^t} \alpha^\top b,
$$

where the set $\Lambda^t$ has possibly infinite elements.

**Proof:** When $t = T$, $V^T(b) = 0$ satisfies (10). For $t < T$, the value of $Q^t(b,a)$ is

$$
\min_{\tilde{\mu}_a \in \mathcal{P}(X_a)} \int_{X_a} \sum_{s \in S} b_s \left( r_as + \beta \sum_{z \in Z} 1^\top J_z p_as V^{t+1} (f(b,a,p_a,z)) \right) d\tilde{\mu}_a (p_a, r_a, \tilde{u}_a) \tag{11a}
$$

s.t.

$$
\int_{X_a} (F_as p_as + G_as r_as + H_as \tilde{u}_as) d\tilde{\mu}_a (p_a, r_a, \tilde{u}_a) = c_as, \quad \forall s \in S \tag{11b}
$$

$$
\int_{X_a} I ((p_as, r_as, \tilde{u}_as) \in X_as) d\tilde{\mu}_a (p_a, r_a, \tilde{u}_a) = 1, \quad \forall s \in S \tag{11c}
$$
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. Defining the dual variables $\rho_{as}$, $\omega_{as}$ corresponding to (11b) and (11c) respectively, the dual of (11) is

$$
\max_{\rho_{as}, \omega_{as}} \sum_{s \in S} c_{as}^\top \rho_{as} + \sum_{s \in S} \omega_{as}
$$

subject to

$$
\sum_{s \in S} (F_{as}p_{as} + G_{as}r_{as} + H_{as}\tilde{u}_{as})^\top \rho_{as} + \sum_{s \in S} \omega_{as} \leq \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^\top J_z p_{as} V^{t+1} (f(b, a, p_a, z)) \right), \quad \forall (p_a, r_a, \tilde{u}_a) \in \mathcal{X}_a
$$

$$
\rho_{as} \in \mathbb{R}^k, \quad \omega_{as} \in \mathbb{R} \quad \forall s \in S,
$$

where the strong duality holds under the assumptions we made on the ambiguity set and support.

The constraints (12b) correspond to an inequality with a minimization problem

$$
\sum_{s \in S} \omega_{as} \leq \min_{(p_a, r_a, u_a)} \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^\top J_z p_{as} V^{t+1} (f(b, a, p_a, z)) \right) - \sum_{s \in S} (F_{as}p_{as} + G_{as}r_{as} + H_{as}\tilde{u}_{as})^\top \rho_{as}
$$

subject to

$$
B_{as}p_{as} + C_{as}r_{as} + E_{as}\tilde{u}_{as} \leq K_{as}d_{as}, \quad \forall s \in S.
$$

Substituting (10) for $V^{t+1}$ and (1) for $f(b, a, p_a, z)$, we get

$$
(13) = \min_{(p_a, r_a, u_a)} \sum_{s \in S} b_s r_{as} + \beta \sum_{z \in Z} \max_{\alpha_{az} \in \Lambda^{t+1}} \left[ \alpha_{az}^\top \sum_{s \in S} J_z p_{as} b_s \right] - \sum_{s \in S} (F_{as}p_{as} + G_{as}r_{as} + H_{as}\tilde{u}_{as})^\top \rho_{as}
$$

subject to (13b)

Since the objective 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 get

$$
(14) = \min_{(p_a, r_a, u_a)} \left[ \sum_{s \in S} b_s r_{as} + \beta \sum_{z \in Z} \max_{\alpha_{az} \in \Lambda^{t+1}} \sum_{s \in S} J_z p_{as} b_s \right] - \sum_{s \in S} (F_{as}p_{as} + G_{as}r_{as} + H_{as}\tilde{u}_{as})^\top \rho_{as}
$$

subject to (13b)
The expression in the bracket is convex (linear) in \((p_a, r_a, \tilde{u}_a)\) for fixed \(\alpha_{az}, z \in Z\), and concave (affine) in \(\alpha_{az}, z \in Z\) for fixed \((p_a, r_a, \tilde{u}_a)\). Moreover, (13b) and \(\text{Conv}(\Lambda^{t+1})\) are convex sets. The minimax theorem (see Osogami (2015), Du and Pardalos (2013)) ensures that the problem is equivalent to

\[
(15) = \max_{\alpha \in \text{Conv}(\Lambda^{t+1}) (p_a, r_a, \tilde{u}_a)} \min_{\forall z \in Z} \sum_{s \in S} b_s r_{as} + \beta \sum_{z \in Z} \alpha_{az}^T \sum_{s \in S} J_z p_{as} b_s
\]

\[-\sum_{s \in S} (F_{as} p_{as} + G_{as} r_{as} + H_{as} \tilde{u}_{as})^T \rho_{as}\]

s.t. (13b)

Since the relative interior of (9) is non-empty, the strong duality holds when taking the dual of the minimization. Thus, we have

\[
(16) = \max_{\alpha \in \text{Conv}(\Lambda^{t+1})} \max_{\forall z \in Z} \kappa_a - \sum_{s \in S} d_{as}^T \kappa_{as}\]

s.t. \(\beta b_s \sum_{z \in Z} J_z^T \alpha_{az} - F_{as}^T \rho_{as} + B_{as}^T \kappa_{as} = 0, \forall s \in S\) \hfill (17b)

\(b_s - G_{as}^T \rho_{as} + C_{as}^T \kappa_{as} = 0, \forall s \in S\) \hfill (17c)

\(-H_{as}^T \rho_{as} + E_{as}^T \kappa_{as} = 0, \forall s \in S\) \hfill (17d)

\(\kappa_{as} \in K_{as}^*, \forall s \in S\) \hfill (17e)

where \(K_{as}^*\) is a dual cone of \(K_{as}\). Now, since (17) is equivalent to the maximum value of \(\sum_{s \in S} \omega_{as}\), we eliminate the variable \(\omega_{as}\) and substitute the second component of the objective of (12a) with (17). Thus, the value function (5) is equivalent to

\[
V^t(b) = \max_{a \in A} \max_{\alpha \in \text{Conv}(\Lambda^{t+1})} \max_{\forall z \in Z} \rho_a, \kappa_a \sum_{s \in S} c_{as}^T \rho_{as} - \sum_{s \in S} d_{as}^T \kappa_{as}\]

s.t. \(F_{as}^T \rho_{as} - D_{as}^T \kappa_{as} = \beta b_s \sum_{z \in Z} J_z^T \alpha_{az}, \forall s \in S\) \hfill (18b)

\(G_{as}^T \rho_{as} - C_{as}^T \kappa_{as} = b_s, \forall s \in S\) \hfill (18c)

\(H_{as}^T \rho_{as} - E_{as}^T \kappa_{as} = 0, \forall s \in S\) \hfill (18d)

\(\kappa_{as} \in K_{as}^*, \forall s \in S\) \hfill (18e)

\(\rho_{as} \in \mathbb{R}^k, \forall s \in S\) \hfill (18f)
and after taking the dual of the third maximization, we obtain

\[
V'(b) = \max_{a \in A} \max_{\alpha_{az} \in \text{Conv}(\Lambda^{t+1})} \left( \sum_{s \in S} b_s \min_{(p_s, r_s, \tilde{u}_s)} \beta \sum_{z \in Z} \alpha_{az}^\top J_z p_{as} + r_{as} \right) \tag{19a}
\]

\[
\text{s.t. } F_{as} p_{as} + G_{as} r_{as} + H_{as} \tilde{u}_{as} = c_{as}, \quad \forall s \in S \tag{19b}
\]

\[
B_{as} p_{as} + C_{as} r_{as} + E_{as} \tilde{u}_{as} \preceq K_{as} d_{as}, \quad \forall s \in S. \tag{19c}
\]

Defining \( \Lambda^t \) to be

\[
\begin{cases}
\min \beta \sum_{z \in Z} \alpha_{az}^\top J_z p_{as} + r_{as} \\
\text{s.t. } F_{as} p_{as} + G_{as} r_{as} + H_{as} \tilde{u}_{as} = c_{as}, \quad s \in S \\
B_{as} p_{as} + C_{as} r_{as} + E_{as} \tilde{u}_{as} \preceq K_{as} d_{as} \quad \forall a \in A, \quad \forall \alpha_{az} \in \text{Conv}(\Lambda^{t+1}), \quad \forall z \in Z
\end{cases}
\]

it follows that the value function is of the form (10). By induction, this is true for all \( t \). This completes the proof.

Having provided the the values of \( a \) and \( \alpha_{az} \), the inner minimization of (19) can be solved efficiently using barrier methods when \( K_{as} \) are positive semidefinite cones, or second order cones. The issue, however, is that there are possibly infinitely many elements in \( \text{Conv}(\Lambda^{t+1}) \), 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 time \( T \) to 1.

### 3.4 Case of Infinite Horizon

By the Banach fixed point theorem (see, e.g., Puterman, 2014), we show that by repeatedly updating the value function with (5), it converges to a unique function, corresponding to the optimal value function \( V^* \) of the infinite horizon problem.

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

\[
\mathcal{L}V(b) = \max_{a \in A} \min_{\mu_a \in D_a} \mathbb{E}(p_a, r_a) \sim \mu_a \left[ \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^\top J_z p_{as} V_i(f(b, a, p_a, z)) \right) \right] \tag{20}
\]

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

**Proof:** Consider an arbitrary function \( V_1 \) and \( V_2 \), and for fixed \( b \), let

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

and

\[
\mu^*_{a,i} = \arg \min_{\mu_a \in D_a} \mathbb{E}(p_a, r_a) \sim \mu_a \left[ \sum_{s \in S} b_s \left( r_{as} + \beta \sum_{z \in Z} 1^\top J_z p_{as} V_i(f(b, a, p_a, z)) \right) \right], \quad \forall a \in A
\]
for \( i = 1, 2 \). Suppose that \( \mathcal{L}V_1(b) \geq \mathcal{L}V_2(b) \). Then,

\[
0 \leq \mathcal{L}V_1(b) - \mathcal{L}V_2(b) = \mathbb{E}(p_{a_1^*, r_{a_1^*}}) \sum_{s \in S} b_s \left( r_{a_1^*} + \beta \sum_{z \in Z} 1^T J_z p_{a_1^*} V_1(f(b, a_1^*, p_{a_1^*}, z)) \right) - \mathbb{E}(p_{a_2^*, r_{a_2^*}}) \sum_{s \in S} b_s \left( r_{a_2^*} + \beta \sum_{z \in Z} 1^T J_z p_{a_2^*} V_2(f(b, a_2^*, p_{a_2^*}, z)) \right)
\]

The inequality is by changing to a sub-optimal nature’s decision \( \mu_{a_1^*, i}^* \to \mu_{a_2^*, i}^* \) for \( V_1 \) and sub-optimal DM’s action \( a_2^* \to a_1^* \) for \( V_2 \). Then, by changing the difference of \( V_1 \) and \( V_2 \) to the absolute value of the difference,

\[
(21) \leq \beta \mathbb{E}(p_{a_1^*, r_{a_1^*}}) \mu_{a_2^*, 2}^* \left( \sum_{s \in S} b_s \sum_{z \in Z} 1^T J_z p_{a_1^*} \left| V_1(f(b, a_1^*, p_{a_1^*}, z)) - V_2(f(b, a_1^*, z, p_{a_1^*})) \right| \right)
\]

The second inequality is by taking the supremum for all belief states \( b \in \Delta(S) \), and the last equality is because \( \mathbb{E}(p_{a_1^*, r_{a_1^*}}) \mu_{a_2^*, 2}^* \left[ \sum_{s \in S} b_s \sum_{z \in Z} 1^T J_z p_{a_1^*} \right] = 1. \)

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

\[
|\mathcal{L}V_1(b) - \mathcal{L}V_2(b)| \leq \beta \sup_{b'} \left| V_1(b') - V_2(b') \right|,
\]

and therefore,

\[
\sup_b |\mathcal{L}V_1(b) - \mathcal{L}V_2(b)| \leq \beta \sup_{b'} \left| V_1(b') - V_2(b') \right|,
\]

which yields that \( \mathcal{L} \) is a contraction under \( 0 < \beta < 1 \). This completes the proof. \( \square \)

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 \( \beta \) by iteratively performing the Bellman operator \( \mathcal{L} \).
4 Solution Method

As POMDP is a special case of DR-POMDP, the exact solution method is intractable. In this section, we present a variant of the HSVI algorithm proposed in Smith and Simmons (2004) 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 $\Upsilon_V$ whose elements are $(b, v)$, which is a combination of a belief $b$ and an upper bound $v$ of the true value function at belief $b$. Therefore, the resulting PWLC function $\overline{V}$ bounds the value function from above. The upper bound $v$ corresponding to a belief $b$ is obtained through Monte Carlo 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. We present the main part of the algorithmic steps in Algorithm 1. Step 4 is an initial step of a recursive algorithm to explore future steps until either the gap between upper and lower bounds falls below a tolerance $\epsilon$, or time limit is reached. In Section 4.1, we explain how the upper and lower bounds of the value function are initialized, and in Section 4.2, we present an exploration strategy (DR-BoundExplore) to close the gap to a pre-determined tolerance level. Finally, in Section 4.3, we discuss how the value functions are updated, given a belief state $b$.

**Algorithm 1 HSVI**

1: **Input:** initial belief state $b^0$, tolerance $\epsilon$
2: **Initialize:** $\overline{V}$, $\underline{V}$ (see details in Section 4.1)
3: **while** $\overline{V}(b^0) - \underline{V}(b^0) > \epsilon$, or time limit is reached **do**
4: $DR$-BoundExplore$(b^0, 0)$ (see details in Algorithm 2)
5: **end while**
6: **Output:** $\overline{V}$, $\underline{V}$

4.1 Initialization

Recall the ambiguity set and support defined in (8) and (9). In the initialization step, the lower bound for the true value function is computed by taking the best action for obtaining the worst expected rewards at every time step. That is, for each action $a$, we solve

$$R_a = \sum_{t=0}^{\infty} \beta^t \min_{s \in S} \min_{\mu_{as} \in D_{as}} \mathbb{E}(p_{as}, r_{as}) \sim \mu_{as} [r_{as}] = \frac{1}{1 - \beta} \min_{s \in S} \min_{\mu_{as} \in D_{as}} \mathbb{E}(p_{as}, r_{as}) \sim \mu_{as} [r_{as}],$$
where for each $s \in S$, the inner minimization problem is solved by

$$\min_{p_{as}, r_{as}, \tilde{u}_{as}} r_{as}$$

$$\text{s.t. } F_{as} p_{as} + G_{as} r_{as} + H_{as} \tilde{u}_{as} = c_{as}$$

$$B_{as} p_{as} + C_{as} r_{as} + E_{as} \tilde{u}_{as} \preceq K_{as} d_{as},$$

and the minimum value for all $s \in S$ is taken by enumeration. We then define $\alpha_{s}^{\text{init}} = \max_{a \in A} R_{a}, \forall s \in S$, and let $\Lambda_{V} = \{\alpha^{\text{init}} \}^{T}$.

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 $V_{MDP} \in R^{|S|}$ be a value function for the distributionally-optimistic MDP. The value function $V_{MDP}$ satisfies

$$V_{MDP}(s) = \max_{a \in A} \max_{\mu_{as} \in D_{as}} E_{(p_{as}, r_{as}) \sim \mu_{as}} \left[ r_{as} + \beta V_{MDP}^{T} \sum_{z \in Z} J_{z} p_{as} \right], \forall s \in S.$$

To solve this, we take a linear programming approach and formulate

$$\min_{\rho, \kappa, V_{MDP}} \quad 1^{T} V_{MDP}$$

$$\text{s.t. } V_{MDP}(s) + c_{as}^{T} \rho_{as} - d_{as}^{T} \kappa_{as} \geq 0, \quad \forall s \in S, a \in A$$

$$\beta \sum_{z \in Z} J_{z}^{T} V_{MDP} + F_{as}^{T} \rho_{as} - B_{as}^{T} \kappa_{as} = 0, \quad \forall s \in S, a \in A$$

$$G_{as}^{T} \rho_{as} - C_{as}^{T} \kappa_{as} = -1, \quad \forall s \in S, a \in A$$

$$H_{as}^{T} \rho_{as} - E_{as}^{T} \kappa_{as} = 0, \quad \forall s \in S, a \in A$$

$$\kappa_{as} \in K_{as}^{*}, \rho_{as} \in R^{k}, \quad \forall s \in S, a \in A$$

$$V_{MDP} \in R^{|S|}. \quad (22g)$$

After the optimal solution is discovered, we initialize $\Upsilon_{V} = \{(e_{s}, V_{MDP}(s)) \}, \forall s \in S$, where $e_{s}$ is a column vector with 1 in the $s^{th}$ element and zero elsewhere. The initialization step consists of solving a polynomial number of convex optimization problems.

To obtain $V(b)$, we solve

$$\max \left\{ \alpha^{T} b \mid \forall \alpha \in \Lambda_{V} \right\}$$

by enumerating all the values of $\alpha^{T} b$. To obtain $V(b)$, we consider a convex combination of points $(b^{i}, v^{i}) \in \Upsilon_{V}$, 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

\[
v = \min \left\{ \sum_{i \in |\Upsilon_V|} w^i v^i \left| \sum_{i \in |\Upsilon_V|} w^i b^i = b, \sum_{i \in |\Upsilon_V|} w^i = 1, w^i \geq 0, \forall i \in |\Upsilon_V| \right. \right\}, \tag{23}
\]

where \([N]\) denotes a set \(\{1, \ldots, N\}\), for some integer \(N\). The idea of the upper bounding and lower bounding of the value function is illustrated in Figure 2. Here, the solid line represents the true value function in a two-state system. 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.

![Figure 2: An example of upper and lower bounds of a value function](image)

The computation complexity of the lower bound value is linear in the number of elements in \(\Lambda_V\), and the computation complexity of the upper bound value is polynomial in the number of elements in \(\Upsilon_V\) as we solve a linear programming model. Both \(|\Lambda_V|\) and \(|\Upsilon_V|\) increase monotonically, but most elements in the two sets are dominated by others. In the later numerical studies, we follow a pruning heuristic, where the dominated elements are pruned every time when the total number of elements is increased by 10%.
4.2 Forward Exploration Heuristics

The purpose of the algorithmic steps we describe in this section is to provide a greedy Monte Carlo simulation strategy to close the gap between the upper and lower bounds of the value function. The scenario of the simulation is branched by DM’s actions $a$ and nature’s distribution choices $\mu_a$, and their outcomes $z$ and $p_a$. Here, we describe a greedy strategy to select the branches.

Let us define $U_V, Q_V, U_L, Q_L$ as follows:

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

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

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

$$Q_L(b, a) = \min_{\mu_a \in D_a} U_L(b, a, \mu_a).$$

To take a sample, we take an action which maximizes $Q_V$, corresponding to the greatest upper bound. If $a^* = \arg \max_{a \in A} Q_V(b, a)$ is a suboptimal action, the upper bound is eventually replaced by another action. By (20), the relation between the gap for $Q_V$ and $Q_L$ at state $b$ and action $a^*$ is

$$Q_V(b, a^*) - Q_L(b, a^*) = \min_{\mu_a^* \in D_a} E_{(p_a^*, r_a^*)} \left[ \sum_{s \in S} b_s \left( r_{a^*s} + \beta \sum_{z \in Z} 1^T J_z p_{a^*z} V(f(b, a^*, p_{a^*}, z)) \right) \right] - \min_{\mu_a^* \in D_a} E_{(p_a^*, r_a^*)} \left[ \sum_{s \in S} b_s \left( r_{a^*s} + \beta \sum_{z \in Z} 1^T J_z p_{a^*z} V(f(b, a^*, p_{a^*}, z)) \right) \right].$$

Let us denote $\mu_{a^*} = \arg \min_{\mu_a \in D_a} U_L(b, a^*, \mu_a)$. Then, since $\mu_{a^*}$ is suboptimal to the first term of the above, we have

$$Q_V(b, a^*) - Q_L(b, a^*) \leq \beta E_{(p_a^*, r_a^*)} \left[ \sum_{s \in S} \sum_{z \in Z} 1^T J_z p_{a^*z} V(f(b, a^*, p_{a^*}, z)) - V(f(b, a^*, p_{a^*}, z)) \right].$$

(24)

To achieve a gap of $\epsilon$ at the initial state $b_0$, the condition for the gap at depth $t$ from the initial state is only $\epsilon \beta^{-t}$. This is used as a terminal condition for the Monte Carlo simulation. Here, we define

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

Then,

$$\text{excess}(b, t) \leq E_{(p_a^*, r_a^*)} \left[ \sum_{s \in S} \sum_{z \in Z} 1^T J_z p_{a^*z} \text{excess}(f(b, a^*, p_{a^*}, z), t + 1) \right].$$

(25)
To have an early termination, we choose $z^*, p_{a^*}^*$, so that it has the maximum contribution of the expected value in the right hand side of (25):

$$z^*, p_{a^*}^* = \arg \max_{z \in \mathcal{Z}, p_{a^*} \in \mathcal{X}_{a^*}} \mu^*_{a^*}(p_{a^*}) \times \sum_{s \in \mathcal{S}} b_s 1^T J_z p^*_{a^*} s \text{excess}(f(b, a^*, p_{a^*}, z), t + 1).$$

(26)

Note that since the worst-case distribution under ambiguity set (8) is a point mass distribution, obtaining $p_{a^*}^*$ is trivial. We present the algorithmic steps in Algorithm 2. Here, Algorithm 2 is called recursively to make decisions on which branch to choose in the next depth $t + 1$. 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($b, t$)

1: **Input:** belief state $b$, depth $t$
2: if $V(b) - V(b) > \epsilon \beta^{-t}$ then
3: $a^* \leftarrow \arg \max_{a \in \mathcal{A}} Q_{\mathcal{V}}(b, a)$
4: $\mu^*_{a^*} \leftarrow \arg \min_{\mu_{a^*} \in D_a} U_{\mathcal{V}}(b, a^*, \mu_{a^*})$
5: $z^*, p_{a^*}^* \leftarrow \arg \max_{z \in \mathcal{Z}, p_{a^*} \in \mathcal{X}_{a^*}} \mu^*_{a^*}(p_{a^*}) \times \sum_{s \in \mathcal{S}} b_s 1^T J_z p^*_{a^*} s \text{excess}(f(b, a^*, p_{a^*}, z), t + 1)$
6: DR-BoundExplore($f(b, a^*, p_{a^*}^*, z^*$), $t + 1$)
7: $\Lambda_{\mathcal{V}} \leftarrow \Lambda_{\mathcal{V}} \cup DR\text{-backup}(b, \Lambda_{\mathcal{V}})$ (see the details of DR-backup in Algorithm 3)
8: $\Upsilon_{\mathcal{V}} \leftarrow \Upsilon_{\mathcal{V}} \cup DR\text{-update}(b, \Upsilon_{\mathcal{V}})$ (see the details of DR-update in Algorithm 4)
9: end if

4.3 Local Updates

In this section, we describe the 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 inner two maximization problems of (18) provided $a$ and $b$, where $\Lambda_{\mathcal{V}}$ is used in place of $\Lambda^{t+1}$. The convex hull of $\Lambda_{\mathcal{V}}$ is therefore,

$$\text{Conv} (\Lambda_{\mathcal{V}}) = \left\{ \sum_{i \in [||\Lambda_{\mathcal{V}}||]} w^i \alpha^i \right\} \bigg| \sum_{i \in [||\Lambda_{\mathcal{V}}||]} w^i = 1, \ \alpha^i \in \Lambda_{\mathcal{V}}, \ w^i \geq 0, \ i \in [||\Lambda_{\mathcal{V}}||] \right\}. \quad (27)$$

Thus, we combine the inner two maximization problems of (18) as

$$\max_{\rho_a, \kappa_a} \sum_{s \in \mathcal{S}} c_{as}^T \rho_{as} - \sum_{s \in \mathcal{S}} d_{as}^T \kappa_{as} \quad (28a)$$
\[ \text{s.t. } F_{as}^\top \rho_{as} - B_{as}^\top \kappa_{as} - \beta b_s \sum_{i \in [|\Lambda_V|]} w_{az}^i J_z^\top \alpha_{az}^i = 0, \quad \forall s \in S \]  

(28b)

\[ \sum_{i \in [|\Lambda_V|]} w_{az}^i = 1, \quad \forall z \in \mathcal{Z} \]  

(28c)

\[ w_{az}^i \in \mathbb{R}^+, \quad \forall i \in [|\Lambda_V|], z \in \mathcal{Z} \]  

(28d)

(18c)–(18f).

Denote the optimal solution of (28) using a superscript with ⋆, and let the optimal dual variables of (28b) and (18c) be \( \hat{p}_{az}^\star \) and \( \hat{r}_{az}^\star \). For each \( a \in \mathcal{A} \), we generate a lower bounding hyperplane

\[ \alpha' = \left( \hat{r}_{az}^\star + \beta \sum_{z \in \mathcal{Z}} \alpha_{az}^\star J_z \hat{p}_{az}^\star, \ s \in S \right)^\top, \]  

(29)

where \( \alpha_{az}^\star = \sum_{i \in [|\mathcal{V}|]} w_{az}^i \alpha_{az}^i \). The detailed algorithmic steps are presented in Algorithm 3.

**Algorithm 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 (28) for action \( a \)

4: \( \mathcal{L}(a) \leftarrow \alpha' \) (value of (29))

5: **end for**

6: **Output**: \( \text{argmax}_{\alpha \in \mathcal{L}} \alpha^\top b \)

In Algorithm 4, we discuss how the upper bound is updated in DR-update. Combining (18) and the dual representation of (23), we solve

\[ \max_{\rho, \kappa} \sum_{s \in S} c_{as}^\top \rho_{as} - \sum_{s \in S} d_{as}^\top \kappa_{as} \]  

(30a)

\[ \text{s.t. } F_{as}^\top \rho_{as} - B_{as}^\top \kappa_{as} - \beta b_s \sum_{i \in [|\Lambda_V|]} J_z^\top \varphi_{az} - \beta b_s \sum_{i \in [|\Lambda_V|]} J_z^\top 1 = 0, \quad \forall s \in S \]  

(30b)

\[ b_s^\top \varphi_{az} \leq v_i, \quad \forall z \in \mathcal{Z}, i \in [|\mathcal{Y}_V|] \]  

(30c)

\[ \varphi_{az} \in \mathbb{R}^{|\mathcal{S}|}, \ \psi_{az} \in \mathbb{R}, \quad \forall z \in \mathcal{Z}, i \in [|\mathcal{Y}_V|] \]  

(30d)

(18c)–(18f)

for all \( a \in \mathcal{A} \). Here, the variables \( \varphi_{az} \) and \( \psi_{az} \) correspond to dual variables of constraints

\[ \sum_{i \in [|\mathcal{Y}_V|]} w_{az}^i b_i = b, \]  

(31)
\[ \sum_{i \in \left| \mathcal{Y}_V \right|} w^i = 1. \] (32)

The maximum objective value of all \( a \in \mathcal{A} \) is added to \( \mathcal{Y}_V \).

Algorithm 4 DR-update\((b, \mathcal{Y}_V)\)

1: **Input:** belief \( b \), upper bounding points \( \mathcal{Y}_V \)
2: **for** \( \forall a \in \mathcal{A} \) **do**
3: \[ Q(a) \leftarrow \text{(optimal objective value of (30) for action } a \text{)} \]
4: **end for**
5: **Output:** \((b, \max(Q))\)

5 Numerical Studies

To demonstrate the above work, we first test DR-POMDP policies for dynamic machine repair, and compare them with the ones of POMDP without uncertainty. We then vary the size of the ambiguity set and analyze the result sensitivity. We also generate and test instances in which the given transition-observation probability is inaccurate. Finally, we compare solution time of different approaches for instances with diverse sizes.

5.1 Policy Comparison

To observe the effect of the distributionally robust policy, we focus on a simple setting of the machine repair problem discussed in Section 1, and consider the ambiguity only in the transition-observation probability of the overhaul (O), and when the machine is in the broken (B) state. For the ambiguity set, we consider a mean absolute deviation of the transition-observation probabilities,

\[ \mathbb{E}_{p_{as} \sim \mu_{as}} \left[ |p_{as} - \bar{p}_{as}| \right] \leq c_{as} \text{ for } a = O \text{ and } s = B, \]

where \( p_{as} \in \Delta(S \times Z) \) and \( c_{as} \in \mathbb{R}^{|S \times Z|} \). We fix the bound of the deviation \( c_{as} \) to be \( c \cdot 1 \) for some \( c \in \mathbb{R} \). We consider only the case of mean absolute deviation in the computational study and set the discount factor \( \beta = 0.95 \), gap tolerance \( \epsilon = 1.0 \). The computation is terminated when the gap between the upper bound and the lower bound is less than \( \epsilon \), at the initial states \( b^0_B = 0.8 \), \( b^0_F = 0.2 \). We also compute the POMDP policy for which the transition-observation probabilities are \( \bar{p}_{as} \) for all \( a \) and \( s \). Figure 3 compares the DR-POMDP and POMDP policies. The horizontal axis corresponds to the belief of B denoted as \( b_B \), and the vertical axis is the value function. The solid and dashed lines are related to lower and
upper bounds of the value function, respectively. The region of the belief in red (horizontal shade) corresponds to action S, green (cross shade) for O, blue (dotted shade) for T, and white (diagonal shade) for N. The policy given by DR-POMDP has the expected behaviour being less dependent on the action O, as the ambiguity of the transition probability given by action O increases.

5.2 Sensitivity Analysis for the Worst-Case Distribution

We simulate the policies developed in Section 5.1 on instances with B chosen with probability 80% as the initial state. The nature considers a different ambiguity set from the one tested in Section 4.1 and selects their worst-case distribution after the DM chooses his/her action at each time step. The number of simulated instances is 3000 each, and the mean value of the realized reward is presented in Table 1, with the standard error of mean (SEM) shown to the next. As expected, DR-POMDP assuming larger ambiguity outperforms POMDP ($c = 0.00$) in the case where the nature’s ambiguity set is large. However, when the transition-observation probability does not change, POMDP ($c = 0.00$) performs better than DR-POMDP. Except for the case where the nature’s policy is derived from the setting $c = 0.09$, the sample mean of the expected cumulative reward is the greatest when the DM’s policy matches with the nature’s policy.

Table 1: Expected reward for different policies under different nature policies

<table>
<thead>
<tr>
<th>DM’s policy</th>
<th>Nature’s policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>mean</td>
</tr>
<tr>
<td>0.00</td>
<td>−140.04</td>
</tr>
<tr>
<td>0.03</td>
<td>−145.48</td>
</tr>
<tr>
<td>0.06</td>
<td>−167.77</td>
</tr>
<tr>
<td>0.09</td>
<td>−163.89</td>
</tr>
</tbody>
</table>

5.3 Sensitivity Analysis for the Case of Misinformed Probabilities

Adding to the setting of the previous section, we assume that the DM receives incorrect values of the transition-observation probabilities, but the nature receives a correct one. After the nature has decided the worst-case distribution and the transition-observation probability is realized, we make a perturbation following a uniform distribution centered around the outcome, bounded with a fixed
Figure 3: Value functions for different ambiguity levels. Solid line: value function lower bound, dashed line: value function upper bound. Corresponding lower bound actions: S – (red, horizontal), O – (green, cross), T – (blue, dot), N – (white, diagonal)
$L_2$ norm. For the case where the radius of the perturbation is 0.02, we obtain the results in Table 2. When the radius is 0.05, we obtain the results in Table 3. We observe that the performance of both policies are not affected by perturbations of size $0.02 \sim 0.05$.

Table 2: Expected reward for different policies with outcome perturbed by at most 0.02

<table>
<thead>
<tr>
<th>DM’s policy</th>
<th>Nature’s policy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>0.03</td>
<td>0.06</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>SEM</td>
<td>mean</td>
<td>SEM</td>
<td>mean</td>
</tr>
<tr>
<td>0.00</td>
<td>$-154.17$</td>
<td>2.21</td>
<td>$-175.93$</td>
<td>2.33</td>
</tr>
<tr>
<td>0.03</td>
<td>$-160.17$</td>
<td>2.39</td>
<td>$-174.04$</td>
<td>2.22</td>
</tr>
<tr>
<td>0.06</td>
<td>$-165.22$</td>
<td>2.14</td>
<td>$-166.46$</td>
<td>2.17</td>
</tr>
<tr>
<td>0.09</td>
<td>$-165.54$</td>
<td>2.13</td>
<td>$-165.91$</td>
<td>2.17</td>
</tr>
</tbody>
</table>

Table 3: Expected reward for different policies with outcome perturbed by at most 0.05

<table>
<thead>
<tr>
<th>DM’s policy</th>
<th>Nature’s policy</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$c$</td>
<td>0.03</td>
<td>0.06</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>SEM</td>
<td>mean</td>
<td>SEM</td>
<td>mean</td>
</tr>
<tr>
<td>0.00</td>
<td>$-160.97$</td>
<td>2.29</td>
<td>$-171.97$</td>
<td>2.29</td>
</tr>
<tr>
<td>0.03</td>
<td>$-156.71$</td>
<td>2.32</td>
<td>$-174.26$</td>
<td>2.31</td>
</tr>
<tr>
<td>0.06</td>
<td>$-164.51$</td>
<td>2.14</td>
<td>$-168.00$</td>
<td>2.23</td>
</tr>
<tr>
<td>0.09</td>
<td>$-170.14$</td>
<td>2.27</td>
<td>$-170.60$</td>
<td>2.27</td>
</tr>
</tbody>
</table>

5.4 Computation Time

In this section, we analyze the change in computation time for different sizes of problem instances. The machine repair problem is a two-state, four-action, two-observation, one-ambiguous state-action pair problem. We denote this as $(s_2,a_4,z_2,u_1)$ for notational convenience. We also test the ROCKSAMPLE problem in Smith and Simmons (2004), for the case 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 \{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 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 problems we experiment for the ROCKSAMPLE 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, and all the computation is performed on 2.9 GHz Intel Core i5 processor. We show how the upper bound and lower bound of POMDP policies and DR-POMDP policies converge as functions of time in Figures 4, 5, 6 for the three problem sizes, respectively. While the solving time between POMDP and DR-POMDP was similar for the machine repair problem, we observe that DR-POMDP requires more time than POMDP for larger problem size. For the ROCKSAMPLE problem $(s9, a6, z2, u9)$, it took only 4.4 seconds for POMDP to find a policy with gap less than 1.0, while DR-POMDP policy required 287.6 seconds. At the computational time limit for the ROCKSAMPLE problem $(s37, a7, z2, u74)$, the DR-POMDP policy has the gap of 12.3 between the upper and lower bounds, while the POMDP policy has the gap of 1.46.

6 Conclusion

In this paper, we developed new models and algorithms for POMDP when the transition probability and the observation probability is uncertain, and the probability distribution is not perfectly known. We presented a scalable approximation algorithm and numerically shown that it outperforms the standard POMDP without uncertainty in occasions where the distribution takes the worst value. However, due to the more complicated model and problem settings, DR-POMDP is much harder to solve. Future research includes solving 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
Figure 4: Machine repair problem ($s_2, a_4, z_2, u_1$). Solid line: lower bound, dashed line: upper bound

Figure 5: ROCKSAMPLE problem ($s_9, a_6, z_2, u_9$). Solid line: lower bound, dashed line: upper bound
Figure 6: ROCKSAMPLE problem ($s37,a7,z2,u74$). Solid line: lower bound, dashed line: upper bound value function becomes much more challenging.

Acknowledgments

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.

References


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. 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:

$$V^t(b) = \max_{\phi \in \Delta(\mathcal{A})} \min_{\mu \in \mathcal{D}} \mathbb{E}_{\mu \sim P} \left[ \sum_{a \in \mathcal{A}} \phi_a \sum_{s \in \mathcal{S}} b_s \left( r_{as} + \beta \sum_{z \in \mathcal{Z}} J_z p_{az} V^{t+1}(f(b, a, p_a, z)) \right) \right], \quad (33)$$
where $\phi_a$ is the probability for selecting action $a$. We define the ambiguity set to be

$$
\hat{D}_s = \left\{ \left( \hat{p}_s, \hat{r}_s, \hat{u}_s \right) \mid \tilde{\mu}_s \left( \hat{X}_s \right) = 1, \quad \tilde{\mu}_s \left( E (p_s, r_s, \hat{u}_s) \right) \sim \tilde{\mu}_s \left( F_s p_s + G_s r_s + H_s \hat{u}_s \right) = c_s \right\},
$$

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

$$
\hat{X}_s = \left\{ \left( p_s, r_s, \hat{u}_s \right) \in \mathbb{R}^{|A| \times |S| \times |Z|}, \quad B_s p_s + C_s r_s + E_s \hat{u}_s \preceq_{K_s} d_s \right\}.
$$

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

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

$$
V^t(b) = \max_{\phi \in \Delta(A)} \max_{\forall a \in A, \forall z \in Z} \sum_{s \in S} b_s \min_{(\hat{p}_s, \hat{r}_s, \hat{u}_s)} \phi^\top \left( \beta \sum_{z \in Z} \left( \alpha_{az}^\top J_{az} \right)^\top, \ a \in A \right)^\top \hat{p}_s + \hat{r}_s
$$

s.t. $F_s \hat{p}_s + G_s \hat{r}_s + H_s \hat{u}_s = c_s, \quad \forall s \in S$

$$
B_s \hat{p}_s + C_s \hat{r}_s + E_s \hat{u}_s \preceq_{K_s} d_s, \quad \forall s \in S
$$

where $J_{az} \in \mathbb{R}^{(|S| \times (|A| \times |S| \times |Z|))}$ is a matrix of zeros and ones that maps $p_s$ to $p_{aosz}$. For an exact algorithm, we solve the inner minimization problem for all $\phi \in \Delta(A)$, $\alpha_{az} \in \text{Conv}(\Lambda_{t+1})$, $\forall z \in Z$, $a \in A$. The optimal objective is used for constructing the set $\Lambda_t$, at each time step $t$. 