Distributionally Robust Optimization in Sequential Decision Making

by

Hideaki Nakao

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in The University of Michigan
2021

Doctoral Committee:

Associate Professor Siqian Shen, Chair
Assistant Professor Ruiwei Jiang
Professor Brian Denton
Dr. Kibaek Kim, Argonne National Laboratory
Associate Professor Jun Li
Associate Professor Cong Shi

Hideaki Nakao

nakaoh@umich.edu

ORCID iD: 0000-0001-6534-1330

This dissertation is dedicated to my mother Mari and my sister Risako for their support throughout my life
and in memory of my father Akio who inspired me to pursue my career in science and engineering.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my special gratitude to my advisor Professor Siqian Shen for giving me a great opportunity to study operations research. I would like to thank her for her patience in working with me on multiple projects and providing suitable advice. I am also very grateful to Professor Ruiwei Jiang, who spent many hours with me discussing the key components of my dissertation. Besides my advisors, I would like to thank the rest of my dissertation committee: Professor Brian Denton, Professor Cong Shi, Professor Jun Li, and Dr. Kibaek Kim, for their insightful comments and suggestions.

It was also my privilege to learn from professors in the Department of Industrial and Operations Engineering, who built me a solid foundation in Operations Research and enabled me to continue my studies.

I also want to express my thanks to all my colleagues who supported me through my Ph.D. journey: Dr. Yiling Zhang, Dr. Miao Yu, Xian Yu, Huiwen Jia, Kati Moug, Matthew-Remy Aguirre, Timothy Williams, Dr. Geunyeong Byeon, Dr. Minseok Ryu, Weiyu Li, Luze Xu, and Junhong Guo.

Special thanks to the members of the Japanese Ph.D. student group “Mamao” for the mutual support in pursuing the Ph.D. degree, and to my friend and life mentor Dwight Clark.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

ABSTRACT

Distributionally Robust Optimization in Sequential Decision Making

by

Hideaki Nakao

Chair: Siqian Shen

Distributionally robust optimization (DRO) is an effective modeling paradigm for making optimal decisions under uncertainty, where distributional information about the random parameters in a problem of interest is hardly available at the time when decisions are made. DRO encompasses conventional modeling approaches such as stochastic programming and robust optimization for decision making under uncertainty. The former requires perfect or near-perfect knowledge about the statistics of the random parameters for accurate decision making, while the latter only assumes that the supports of the random parameters are known, which often leads to overly conservative solutions. DRO overcomes these concerns by optimizing the expected value or a risk measure of the worst-case distribution in a set of distributions where the true distribution is contained with high probability. In this dissertation, we apply the DRO techniques to various types of sequential decision-making models and explore the capability of the new models for producing reliable and also economic decisions under different settings of data-decision interactions.

In Chapter II, we consider a distributionally robust variant of a partially observable Markov decision process (POMDP), where the transition-observation probabilities are uncertain. We assume that these parameters differ over time and are revealed at the end of each time step. We construct an algorithm to find an optimal policy by iteratively updating the upper and lower bounds of the value function. We demonstrate the use of distributionally robust POMDP in an application of epidemic control when the probability of true infection status is unknown as well as prevention and/or intervention decisions have to be made sequentially and robustly with updated information. In Chapter III, we derive a Wasserstein distance to bound between the true and an empirical distribution when the states and actions of a dynamic sequential decision-making process are finite. We further apply the approach to a regret-based reinforcement learning problem that uses the principle of optimism under uncertainty, and compare the empirical performance of the optimal solution to our model with the conventional approach by testing instances of an ambulance dispatch problem. Finally, in Chapter IV, we focus on a multistage mixed-integer stochastic programming model, and employ a dual decomposition algorithm for solving a distributionally robust variant of the model. We analyze the numerical performance through instances of a transmission expansion problem in power systems under the uncertainty of loads and renewable generation capabilities.

Overall, the contributions of this dissertation are threefold. First, we develop mathematical models of various distributionally robust sequential decision making problems, some of which involve discrete decision variables and are generally NP-hard. Second, we derive efficient solution algorithms to solve the proposed models via relaxation and decomposition techniques. Third, we evaluate the performance of solution approaches and their results via extensive numerical experiments based on epidemic control, healthcare, and energy applications. The models and solution algorithms developed in this work can be used by practitioners to solve a variety of sequential decision making problems in different business contexts, and thus can generate significant societal and economic impacts.

CHAPTER I

Introduction

Sequential decision making arises in many engineering problems including transportation, energy, healthcare operations management, finance, and medicine. The most challenging aspect of sequential decision making is the presence of data and system uncertainties whose values are revealed to the decision maker (DM) gradually and iteratively over time. Due to the increasing number of combinations of the possible outcomes in the future, obtaining an optimal decision that overlooks all the potential scenarios is often a very difficult task. In this thesis, we focus mainly on two approaches. The first is the Markov Decision Processes (MDP) approach, where the model is restricted to have finite states that evolve according to a Markov process, i.e., the probability for transitioning to the state in the next time period is only dependent on the current state-and-action pair. Because of this assumption/restriction, there exist some efficient ways for calculating the future expected reward. The second approach we focus on is the multistage stochastic programming approach, which involves continuous or discrete states and actions. The uncertain parameters are often driven by exogenous random variables, and modeling endogenous uncertainty is quite challenging. In this chapter, we will introduce the general mathematical formulations of sequential decision-making processes for different approaches.

The traditional literature in optimization under uncertainty commonly assumes that the full knowledge of underlying true distributions of uncertain parameters. However, in practice, there are situations where only a small amount of data is available to fully characterize the statistics of the uncertain parameters when decisions need to be made. For example, only a few samples of parameters about a system of interest may be given to the DM initially. The data can then be used to construct a set of distributions, namely the ambiguity set, which contains the true distribution with high probability. In distributionally robust optimization (DRO), the decisions are made against the worst-case expected value or a certain risk measure of the objective function over all possible distributions characterized by the ambiguity set. We will describe the concept and general models of DRO in Section 1.3.

1.1 Markov Decision Processes

In this section, we briefly introduce formulations of MDP and partially observable Markov decision processes (POMDP).

1.1.1 Formulating Markov Decision Processes

A finite horizon MDP is a 5−tuple $(\mathcal{S}, \mathcal{A}, P_t^a, R_t^a, T)$, where

- $\mathcal{S}$ is a set of states, finite;
- $\mathcal{A}$ is a set of actions;
- $P_t^a(s, s')$ is the probability of transitioning from state $s \in \mathcal{S}$ at time t to $s' \in \mathcal{S}$ at time $t+1$ by taking action $a \in \mathcal{A}$ at time t;
- $R_t^a(s)$ is an immediate reward for taking an action $a \in \mathcal{A}$ at time t;
- T is the total number of periods in the overall time horizon.

The objective for the decision maker (DM) is to find a policy that maximizes the cumulative expected reward. This can be obtained by solving the Bellman equation

(*Puterman*, 2014):

$$V_t(s) = \max_{a \in \mathcal{A}} \left[R_t^a(s) + \sum_{s' \in \mathcal{S}} P_t^a(s, s') V_{t+1}(s') \right], \ \forall s \in \mathcal{S}, t \in [T-1], \tag{1.1}$$

where $[\cdot] := \{1, 2, \ldots, \cdot\}$, and

$$V_T(s) = R_T(s), \ \forall s \in \mathcal{S}. \tag{1.2}$$

1.1.2 Partially Observable Markov Decision Processes

A finite horizon POMDP is a 7−tuple $(\mathcal{S}, \mathcal{A}, P_t^a, R_t^a, \Omega, O_t^a, T)$, where $\mathcal{S}$, $\mathcal{A}$, P_t^a, R_t^a, T are the same as MDP, and

- Ω is a set that contains all possible observations;
- $O_t^a(s', o)$ is a conditional probability of observation $o \in \Omega$ given a state $s' \in \mathcal{S}$ at time $t+1$ and action $a \in \mathcal{A}$ at time t.

The sufficient statistic of POMDP is a belief $\boldsymbol{b}_t \in \Delta(\mathcal{S})$, where $\Delta(\cdot)$ is a probability simplex of $\cdot$. That is, it is sufficient to maintain the DM's subjective probability of the state which the system is in, rather than keeping all the sequence of the actions and observations to come up with an optimal policy. We can consider the belief as an information state (*Kumar and Varaiya*, 2015), since it can be iteratively updated with incoming data of action and observation:

$$\boldsymbol{b}_{t+1}^{\boldsymbol{b}_t, a, o}(s') := \frac{\sum_{s \in \mathcal{S}} P_t^a(s, s') O_t^a(s', o) \boldsymbol{b}_t(s)}{\sum_{s'' \in \mathcal{S}} \sum_{s \in \mathcal{S}} P_t^a(s, s'') O_t^a(s', o) \boldsymbol{b}_t(s)}, \ \forall s' \in \mathcal{S}. \tag{1.3}$$

Here, $\boldsymbol{b}_{t+1}^{\boldsymbol{b}_t, a, o}$ is the posterior probability after taking an action a and observing an outcome o, when the prior probability is $\boldsymbol{b}_t$.

The optimal policy is obtained by solving

$$V_t(\boldsymbol{b}_t) = \max_{a \in \mathcal{A}} \left[\sum_{s \in \mathcal{S}} R_t^a(s) \boldsymbol{b}_t(s) + \sum_{o \in \Omega} \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} P_t^a(s, s') O_t^a(s', o) \boldsymbol{b}_t(s) V_{t+1} \left(\boldsymbol{b}_{t+1}^{\boldsymbol{b}_t, a, o} \right) \right], \tag{1.4}$$

for all $\boldsymbol{b}_t \in \Delta(\mathcal{S})$ and $t \in [T-1]$, and

$$V_T(\boldsymbol{b}_T) = \sum_{s \in \mathcal{S}} R_T(s) \boldsymbol{b}_T(s).$$

The value function is piecewise-linear and convex (PWLC) with respect to the belief state (*Smallwood and Sondik*, 1973), but the exact solution of the optimal policy is known to be highly intractable to obtain (*Papadimitriou and Tsitsiklis*, 1987). Recent developments that use approximation algorithms for solving POMDPs are summarized in *Shani et al.* (2013).

1.2 Optimization under Uncertainty with Full Distributional Information

1.2.1 Stochastic Programming

We first introduce the basic concepts of stochastic programming, as an approach that is ubiquitously used for optimization under uncertainty when distributional information is fully known to the DM. We let $\boldsymbol{x}$ be the decision variable, and assume that the feasible region $\mathcal{X} \subseteq \mathbb{R}^n$ is non-empty, for ease of exposition. We are interested in solving a problem of the form

$$\min_{\boldsymbol{x} \in \mathcal{X}} h(\boldsymbol{x}, \boldsymbol{\xi}), \tag{1.5}$$

where $\boldsymbol{\xi}$ is an uncertain parameter. This is an ill-posed problem, since we cannot minimize the random outcome $h(\boldsymbol{x}, \boldsymbol{\xi})$ when the value of $\boldsymbol{\xi}$ is uncertain. Instead, we consider a certain function of the random variables $\boldsymbol{\xi}$:

$$\min_{\boldsymbol{x} \in \mathcal{X}} \varrho \left(h(\boldsymbol{x}, \boldsymbol{\xi}) \right), \tag{1.6}$$

Typically, the expected value is chosen for ϱ:

$$\min_{\boldsymbol{x} \in \mathcal{X}} \mathbb{E} \left[h(\boldsymbol{x}, \boldsymbol{\xi}) \right], \tag{1.7}$$

When the DM is risk-averse, we use a different function that puts more weight on the realizations of random objective having greater values. We may be interested in the β-quantile of the objective value, namely the value-at-risk (VaR). However, conditional value-at-risk (CVaR) (*Rockafellar and Uryasev*, 2002) is often used instead of VaR due to its convex property. CVaR is the mean value of the realizations of the objective value that are greater than the value of the VaR, and is an upper bound approximation of VaR. It also satisfies other desirable properties of risk measures described in *Artzner et al.* (1999). It is formulated as

$$\min_{\boldsymbol{x} \in \mathcal{X}, \alpha \in \mathbb{R}} \alpha + \frac{1}{1-\beta} \mathbb{E} \left[[h(\boldsymbol{x}, \boldsymbol{\xi}) - \alpha]^{+} \right], \tag{1.8}$$

where $[a]^{+} = \max\{0, a\}$. Here, α is the VaR, and the second term in the objective function represents the conditional average of the margins above α.

For continuously distributed random parameter $\boldsymbol{\xi}$, the computation of the expected value is often a difficult task since it requires integrating the objective function and the full knowledge of the probability distribution of the uncertain $\boldsymbol{\xi}$. To circumvent this, a sample average approximation (SAA) algorithm is employed and we briefly describe its steps as follows. Suppose that $\boldsymbol{\xi}$ follows a distribution F and has a support Ξ. We take N identically and independently distributed (i.i.d.) samples from the distribution F and set $\hat{\Xi} = \left\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^N\right\}$ as the sample set. Then, (1.7) is approximated as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{s=1}^{N} h(\boldsymbol{x}, \boldsymbol{\xi}^s). \tag{1.9}$$

By the Law of Large Numbers, the optimal objective value converges pointwise w.p. 1 to the true expected value, and so does the optimal solution under some regularity conditions (*Shapiro et al.*, 2009).

1.2.2 Two-stage Stochastic Programming

In two-stage stochastic programs, some decisions are made before the values of uncertain parameters are revealed (i.e., here-and-now decision variables), and afterward (i.e., wait-and-see variables). Benders decomposition (*Birge and Louveaux*, 2011) is a well-known technique that exploits the sparsity of the constraints in the large-scale linear optimization problem. A generic formulation of a two-stage stochastic program is given by

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{s=1}^{N} Q(\boldsymbol{x}, \boldsymbol{\xi}^s) && \text{(1.10a)} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}, && \text{(1.10b)} \\ & \boldsymbol{x} \in \mathbb{R}^{n_1}_+, && \text{(1.10c)} \end{aligned}$$

where for each scenario $s \in [N]$, we let $\boldsymbol{\xi}^s = (\boldsymbol{q}^s, \boldsymbol{W}^s, \boldsymbol{T}^s, \boldsymbol{h}^s)$ and define

$$\begin{aligned} Q(\boldsymbol{x}, \boldsymbol{\xi}^s) = \min_{\boldsymbol{y}} \quad & \left(\boldsymbol{q}^s\right)^\top \boldsymbol{y} && \text{(1.11a)} \\ \text{s.t.} \quad & \boldsymbol{W}^s \boldsymbol{y} = \boldsymbol{h}^s - \boldsymbol{T}^s \boldsymbol{x}, && \text{(1.11b)} \\ & \boldsymbol{y} \in \mathbb{R}^{n_2}_+. && \text{(1.11c)} \end{aligned}$$

We apply strong duality to the second-stage problem and obtain

$$\begin{aligned} Q(\boldsymbol{x},\boldsymbol{\xi}^s) = \max_{\boldsymbol{\pi}} \quad & \boldsymbol{\pi}^\top \left(\boldsymbol{h}^s - \boldsymbol{T}^s\boldsymbol{x}\right) && \text{(1.12a)} \\ \text{s.t.} \quad & \boldsymbol{\pi}^\top \boldsymbol{W}^s \leq \boldsymbol{q}^s. && \text{(1.12b)} \end{aligned}$$

Suppose that for a feasible $\boldsymbol{x}$ obtained from solving (1.10), the problem (1.11) is feasible. Then, $Q(\boldsymbol{x},\boldsymbol{\xi}^s) \geq \boldsymbol{\pi}^\top \left(\boldsymbol{h}^s - \boldsymbol{T}^s\boldsymbol{x}\right)$ holds for all $\boldsymbol{\pi}$-values that satisfy (1.12b). Because the feasible region (1.12b) is a polyhedron, it is sufficient to consider the extreme points of (1.12b). If for a given $\boldsymbol{x}$ from (1.10), (1.11) is infeasible, then there exists an extreme ray $\boldsymbol{r}$ which the objective value of (1.12) increases indefinitely. To suppress this, we add a constraint $\boldsymbol{r}^\top \left(\boldsymbol{h}^s - \boldsymbol{T}^s\boldsymbol{x}\right) \leq 0$ for all extreme rays. In practice, only a subset of these constraints are necessary to obtain the optimal solution $\boldsymbol{x}$ to (1.10). In the Benders decomposition algorithm, the first-stage problem is approximated from below as a relaxed master problem, and these constraints are added as needed from iteratively solving the second-stage problem (1.12). The master problem is

$$\begin{aligned} \min_{\boldsymbol{x},\boldsymbol{\theta}} \quad & \boldsymbol{c}^\top \boldsymbol{x} + \frac{1}{N}\sum_{s=1}^{N} \theta^s && && \text{(1.13a)} \\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}, && && \text{(1.13b)} \\ & \boldsymbol{x} \in \mathbb{R}_+^{n_1}, && && \text{(1.13c)} \\ & \theta^s \geq \left(\hat{\boldsymbol{\pi}}^s\right)^\top \left(\boldsymbol{h}^s - \boldsymbol{T}^s\boldsymbol{x}\right), && \forall \hat{\boldsymbol{\pi}}^s \in \mathcal{V}^s,\ s = 1,\ldots,N, && \text{(1.13d)} \\ & \left(\hat{\boldsymbol{r}}^s\right)^\top \left(\boldsymbol{h}^s - \boldsymbol{T}^s\boldsymbol{x}\right) \leq 0, && \forall \hat{\boldsymbol{r}}^s \in \mathcal{R}^s,\ s = 1,\ldots,N. && \text{(1.13e)} \end{aligned}$$

Here, $\mathcal{V}^s$ is a subset of extreme points of (1.12b), where new points are added when a solution $(\boldsymbol{x}^\star, \boldsymbol{\theta}^\star)$ is discovered and scenario s of the second stage problem is feasible with $\theta^{s\star} < Q(\boldsymbol{x}^\star, \boldsymbol{\xi}^s)$. This is called the *optimality cut*. Similarly, $\mathcal{R}^s$ is a subset of extreme rays of (1.12b), where new rays are added when a solution $(\boldsymbol{x}^\star, \boldsymbol{\theta}^\star)$ is infeasible for scenario s of the second stage problem. This is called the *feasibility cut.*

A special case where for any first-stage feasible solution $\boldsymbol{x}$, the second-stage problem is feasible for all scenarios is called the *relatively complete recourse* problem and feasibility cuts are not needed in this case.

1.2.3 Multistage Stochastic Programming

A multistage stochastic program is a generalization of the aforementioned two-stage stochastic program with K stages ($K \geq 2$). It is formulated as

$$\begin{aligned} \min_{\boldsymbol{x}_1} \quad & \boldsymbol{c}_1^\top \boldsymbol{x}_1 + \mathbb{E}\left[Q_2(\boldsymbol{x}_1, \boldsymbol{\xi}_2)\right] && (1.14) \\ \text{s.t.} \quad & \boldsymbol{x}_1 \in \mathcal{X}_1, && (1.15) \end{aligned}$$

where

$$\begin{aligned} Q_k(\boldsymbol{x}_{k-1}, \boldsymbol{\xi}_k) := \min_{\boldsymbol{x}_k} \quad & \boldsymbol{c}_k^\top \boldsymbol{x}_k + \mathbb{E}\left[Q_{k+1}(\boldsymbol{x}_k, \boldsymbol{\xi}_{k+1})\right] && (1.16a) \\ \text{s.t.} \quad & \boldsymbol{W}_k \boldsymbol{x}_k = \boldsymbol{h}_k - \boldsymbol{T}_k \boldsymbol{x}_{k-1}, && (1.16b) \\ & \boldsymbol{x}_k \in \mathbb{R}^{n_k}, && (1.16c) \end{aligned}$$

for all $k = 2, \ldots, K-1$, and

$$\begin{aligned} Q_K(\boldsymbol{x}_{K-1}, \boldsymbol{\xi}_K) := \min_{\boldsymbol{x}_K} \quad & \boldsymbol{c}_K^\top \boldsymbol{x}_K && (1.17a) \\ \text{s.t.} \quad & \boldsymbol{W}_K \boldsymbol{x}_K = \boldsymbol{h}_K - \boldsymbol{T}_K \boldsymbol{x}_{K-1}, && (1.17b) \\ & \boldsymbol{x}_K \in \mathbb{R}^{n_K}. && (1.17c) \end{aligned}$$

The algorithm works very similarly to the two-stage stochastic problems, although there is a notion of scenario trees in the multistage case to cause the "curse of dimensionality" issue in computation. When the uncertain parameters are stage-wise independent, the computation can be greatly reduced using stochastic dual dynamic programming (SDDP) (*Pereira and Pinto*, 1991).

1.3 Distributionally Robust Optimization

For ease of exposition, we focus on the case where the objective function is piecewise linear and convex (PWLC), i.e.,

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}\left[\max_{\ell=1,\dots,L} \boldsymbol{\xi}_\ell^\top \boldsymbol{x}\right]. \tag{1.18}$$

Here, we use the same notation of the decision variable $\boldsymbol{x} \in \mathcal{X}$, and $\boldsymbol{\xi}_\ell$ is one of the random cost vectors associated with decision $\boldsymbol{x}$.

Different from the stochastic programming approaches reviewed in Section 1.2, here we do not assume sufficiently many data, i.e., only a small number of samples is available to the DM, and therefore it is difficult to justify the use of the sample average approximation. Using a robust optimization method by constructing a bounded support ignores statistical information that can be obtained from data if not the full knowledge about the true distribution, and often leads to an overly conservative solution. DRO generalizes these approaches by constructing a set of distributions, namely the ambiguity set, using data and thus become data-driven. Then, we optimize for the worst-case expected value of the objective among all the distributions in the ambiguity set:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{F\in\mathcal{D}} \mathbb{E}_F\left[\max_{\ell=1,\dots,L} \boldsymbol{\xi}_\ell^\top \boldsymbol{x}\right]. \tag{1.19}$$

Here, F is a distribution and $\mathcal{D}$ is an ambiguity set. Below, we introduce some of the ambiguity sets that can yield a tractable reformulation of (1.19).

1.3.1 Moment-based Ambiguity Set

Let $\boldsymbol{\xi} = (\boldsymbol{\xi}_1 \ldots, \boldsymbol{\xi}_L)$, and suppose that the support Ξ is bounded and convex. We let $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ be the sample mean and covariance matrix of N data samples. Let γ_1 and γ_2 be some functions of N and a small probability δ given in *Delage and Ye* (2010). Then, with probability at least $1 - \delta$, the true distribution contains in an ambiguity set

$$\mathcal{D} = \left\{ F \in \mathcal{M} \;\middle|\; \begin{array}{l} \mathbb{P}(\boldsymbol{\xi} \in \Xi) = 1 \\ \left(\mathbb{E}[\boldsymbol{\xi}] - \hat{\boldsymbol{\mu}}\right)^\top \hat{\boldsymbol{\Sigma}}^{-1} \left(\mathbb{E}[\boldsymbol{\xi}] - \hat{\boldsymbol{\mu}}\right) \leq \gamma_1 \\ \mathbb{E}\left[\left(\boldsymbol{\xi} - \hat{\boldsymbol{\mu}}\right)\left(\boldsymbol{\xi} - \hat{\boldsymbol{\mu}}\right)^\top\right] \preceq \gamma_2 \hat{\boldsymbol{\Sigma}} \end{array} \right\}, \tag{1.20}$$

where $\mathcal{M}$ is a set of probability measures. The first condition indicates that the uncertain parameter must lie inside the support Ξ, and the second condition restricts the true mean to be inside an ellipsoid characterized by the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$. The third condition requires the difference of the right-hand-side (RHS) matrix and the left-hand-side (LHS) matrix to be positive semidefinite. This gives a condition on the proximity of the events to the sampled average $\hat{\boldsymbol{\mu}}$ in terms of the sampled covariance matrix $\hat{\boldsymbol{\Sigma}}$.

The DRO problem (1.19) can be expressed as

$$\begin{aligned}
\min_{\boldsymbol{x}\in\mathcal{X}}\max_{F} \quad & \int\limits_{\Xi} \left(\max_{\ell=1,\dots,K} \boldsymbol{\xi}_\ell^\top \boldsymbol{x}\right) F(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} && \text{(1.21a)}\\
\text{s.t.} \quad & \int\limits_{\Xi} F(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = 1, && \text{(1.21b)}\\
& \int\limits_{\Xi} \begin{bmatrix} \hat{\boldsymbol{\Sigma}} & (\boldsymbol{\xi}-\hat{\boldsymbol{\mu}}) \\ (\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})^\top & \gamma_1 \end{bmatrix} F(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \succeq 0, && \text{(1.21c)}\\
& \int\limits_{\Xi} (\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})(\boldsymbol{\xi}-\hat{\boldsymbol{\mu}})^\top F(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \preceq \gamma_2\hat{\boldsymbol{\Sigma}}, && \text{(1.21d)}\\
& F \in \mathcal{M}. && \text{(1.21e)}
\end{aligned}$$

The constraint (1.21c) is by Schur complement. This problem involves an infinite number of variables $F(\boldsymbol{\xi})$, but under some regularity conditions, one can use strong duality to obtain a reformulation of (1.21):

$$\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{Q},\boldsymbol{q},r,t} \quad & r+t && \text{(1.22a)}\\
\text{s.t.} \quad & r \geq \boldsymbol{\xi}_\ell^\top\boldsymbol{x} - \boldsymbol{\xi}^\top\boldsymbol{Q}\boldsymbol{\xi} + \boldsymbol{\xi}^\top\boldsymbol{q},\ \forall\boldsymbol{\xi}\in\Xi,\ \ell=1,\dots,L, && \text{(1.22b)}\\
& t \geq \left(\gamma_2\hat{\boldsymbol{\Sigma}} + \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^\top\right) \bullet \boldsymbol{Q} + \hat{\boldsymbol{\mu}}^\top\boldsymbol{q} + \sqrt{\gamma_1}\left|\left|\hat{\boldsymbol{\Sigma}}^{1/2}(\boldsymbol{q}+2\boldsymbol{Q}\hat{\boldsymbol{\mu}})\right|\right|_2, && \text{(1.22c)}\\
& \boldsymbol{Q} \succeq 0, && \text{(1.22d)}\\
& \boldsymbol{x} \in \mathcal{X}, && \text{(1.22e)}
\end{aligned}$$

where $\bullet$ is a Frobenius product operation. There are infintiely many constraints in (1.22b) as it needs to be considered for all $\boldsymbol{\xi} \in \Xi$, which renders the problem intractable. To solve (1.22), we employ a cutting-plane-based decomposition algorithm by relaxing this set of constraints and add cuts as needed. Given solutions

$(\boldsymbol{x}^\star, \boldsymbol{Q}^\star, \boldsymbol{q}^\star, r^\star)$, we solve the following problem for all $\ell = 1, \ldots, L$:

$$\max_{\boldsymbol{\xi}} \quad s_\ell \tag{1.23a}$$

$$\text{s.t.} \quad s_\ell \leq \boldsymbol{\xi}_\ell^\top \boldsymbol{x}^\star - \boldsymbol{\xi}^\top \boldsymbol{Q}^\star \boldsymbol{\xi} + \boldsymbol{\xi}^\top \boldsymbol{q}^\star, \tag{1.23b}$$

$$\boldsymbol{\xi} \in \Xi. \tag{1.23c}$$

If the solution $(s_\ell^\star, \boldsymbol{\xi}^\star)$ satisfies $s_\ell^\star > r^\star$, then it indicates that there exists $\boldsymbol{\xi}^\star$ such that it violates (1.22b), and therefore we add a cut

$$r \geq \boldsymbol{\xi}_\ell^{\star\top} \boldsymbol{x} - \boldsymbol{\xi}^{\star\top} \boldsymbol{Q} \boldsymbol{\xi}^\star + \boldsymbol{\xi}^{\star\top} \boldsymbol{q} \tag{1.24}$$

to (1.22) until a certain precision is satisfied.

1.3.2 Wasserstein-based Ambiguity Set

The 1-Wasserstein distance between two distributions P and Q is formally defined as follows:

$$W(P, Q) = \inf_{\Pi} \int_{\Xi^2} ||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2|| \Pi(\mathrm{d}\boldsymbol{\xi}_1, \mathrm{d}\boldsymbol{\xi}_2) \tag{1.25a}$$

$$\text{s.t} \quad \Pi \text{ is a joint distribution with marginals } P \text{ and } Q. \tag{1.25b}$$

1-Wasserstein distance is also known as the earth mover's distance (*Villani*, 2008; *Gao and Kleywegt*, 2016), where the cost to move a unit probability mass from $\boldsymbol{\xi}_1$ to $\boldsymbol{\xi}_2$ is given by some norm $||\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2||$. The formulation (1.25) has a dual form (*Esfahani and Kuhn*, 2018)

$$W(P, Q) = \sup_{f \in \mathcal{L}} \left\{ \int_{\Xi} f(\boldsymbol{\xi}) P(\mathrm{d}\boldsymbol{\xi}) - \int_{\Xi} f(\boldsymbol{\xi}) Q(\mathrm{d}\boldsymbol{\xi}) \right\}, \tag{1.26}$$

where $\mathcal{L}$ is the Lipschitz set

$$\mathcal{L} = \{f \ : \ ||f(\boldsymbol{\xi}_1) - f(\boldsymbol{\xi}_2)|| \leq |\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2|\}\,. \tag{1.27}$$

The N samples are used to construct an empirical distribution $\hat{F}$, and the true distribution F is expected to have a small Wasserstein distance from $\hat{F}$. In fact, when the true distribution is light-tailed, i.e., there exists a such that

$$A := \mathbb{E}^{\mathbb{P}}\left[\exp\left(\|\xi\|^a\right)\right] = \int\limits_{\Xi} \exp\left(\|\xi\|^a\right)\mathbb{P}(\mathrm{d}\xi) < \infty. \tag{1.28}$$

Then, with probability at least $1-\delta$, the Wasserstein distance between F and $\hat{F}$ is less than

$$\varepsilon_N(\delta) := \left\{ \begin{array}{ll} \left(\frac{\log\left(c_1\delta^{-1}\right)}{c_2 N}\right)^{1/\max\{m,2\}} & \text{if } N \geq \frac{\log\left(c_1\delta^{-1}\right)}{c_2} \\ \left(\frac{\log\left(c_1\delta^{-1}\right)}{c_2 N}\right)^{1/a} & \text{if } N < \frac{\log\left(c_1\delta^{-1}\right)}{c_2} \end{array} \right\}, \tag{1.29}$$

where m is the dimension of $\boldsymbol{\xi}$, and c_1, c_2 are some constants that only depend on a, A, and m (*Fournier and Guillin*, 2015). We can therefore define an ambiguity set having a finite Wasserstein distance from the empirical distribution, namely the Wasserstein ball:

$$\mathcal{D} = \left\{F \in \mathcal{M} \ : \ W(F, \hat{F}_M) \leq \epsilon_N(\delta)\right\}. \tag{1.30}$$

For simplicity, suppose that the support Ξ is a polytope

$$\Xi = \left\{\boldsymbol{\xi} \in \mathbb{R}^m \ : \ \boldsymbol{\xi}^\top \boldsymbol{E} \leq \boldsymbol{d}^\top\right\}. \tag{1.31}$$

The DRO model (1.19) can be reformulated as

$$
\begin{aligned}
\min_{\boldsymbol{x},\lambda,\boldsymbol{s},\boldsymbol{\nu}} \quad & \lambda\epsilon_N(\delta) + \frac{1}{N}\sum_{i=1}^{N} s_i && \text{(1.32a)}\\
\text{s.t.} \quad & \boldsymbol{\xi}_\ell^{i\top}\boldsymbol{x} + (\boldsymbol{d}^\top - \boldsymbol{\xi}^{i\top}\boldsymbol{E})\boldsymbol{\nu}_{ik} \le s_i,\ \forall i = 1,\dots,N,\ \ell = 1,\dots,L, && \text{(1.32b)}\\
& ||\boldsymbol{E}\boldsymbol{\nu}_{i\ell} - \boldsymbol{I}_\ell\boldsymbol{x}||_{\_}* \le \lambda,\ \forall i = 1,\dots,N,\ \ell = 1,\dots,L, && \text{(1.32c)}\\
& \boldsymbol{\nu}_{i\ell} \ge 0,\ \forall i = 1,\dots,N,\ \ell = 1,\dots,L, && \text{(1.32d)}\\
& \boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n, && \text{(1.32e)}
\end{aligned}
$$

where $\boldsymbol{I}_\ell$ is a zero matrix of size $\mathbb{R}^{nL\times n}$, except for the ℓth block which is an identity matrix.

The rest of the dissertation is outlined as follows.

Chapter II is joint work with Dr. Siqian Shen and Dr. Ruiwei Jiang. This chapter considers a distributionally robust variant of POMDP which was introduced in Section 1.1.2. It also considers an ambiguity set related to the one introduced in Section 1.3.1.

Chapter III is joint work with Dr. Siqian Shen. I would also like to acknowledge Dr. Ruiwei Jiang and Dr. Cong Shi for constructive discussions on the theory of Wasserstein distance and concentration inequality. This chapter considers a slightly different variant of MDP introduced in Section 1.1.1, and uses a discrete version of the Wasserstein-based ambiguity set introduced in Section 1.3.2.

Chapter IV is joint work with Dr. Kibaek Kim and Dr. Siqian Shen. I would also like to express my gratitude to Dr. Miao Yu for the discussions on the application in the transmission expansion problem. This chapter is related to Section 1.2 and considers the Wasserstein-based ambiguity set introduced in Section 1.3.2.

Finally, the conclusion is given in Chapter V.

CHAPTER II

Distributionally Robust Partially Observable Markov Decision Process

2.1 Introductory Remarks

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

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

In POMDPs, a critical assumption is that the exact transition and observation probabilities are known to the DM for each action-state combination. In practice, there may exist estimation errors about either the transition or observation probability values, to handle which, *Rasouli and Saghafian* (2018) builds an uncertainty set of probabilities and develops an exact algorithm for the problem of maximizing the expected reward in the worst-case realization of the unknown probabilities in POMDPs. We will numerically compare actions of robust POMDP (see *Osogami* (2015)) with decision policies of DR-POMDP and POMDP in Section 2.6.

In this chapter, using bounded moments, we construct an ambiguity set of the unknown joint distribution of the transition-observation probabilities, in which the true joint distribution lies with high probability. We consider a distributionally robust optimization framework of POMDPs (called DR-POMDP) to seek an optimal policy against the worst-case distribution in the ambiguity set, when realizations of the transition and observation probabilities in each decision period are generated from this distribution. Moreover, we allow transition-observation probabilities to vary in different decision periods, and assume that at the end of each period, the DM can gather side information to infer the true values of the transition-observation probabilities realized in that period, even these values were unknown to the DM when decisions were made. Admittedly, it is rather restrictive to have this assumption where the transition-observation probabilities can be observed retrospectively. However, there exist a wide range of applications where the underlying dynamics are understood and can be simulated to produce unknown parameters (i.e., transition-observation probabilities) once values of some exogenous parameters are gained after the decisions are made. For example, *Mannor et al.* (2016) justify the electric power system as one case where the system performance can be reliably simulated when environmental factors, such as wind and solar radiation levels, are known. In Section 2.3, we provide a few examples to further illustrate and justify this assumption and in Section 2.6, we conduct numerical tests on dynamic epidemic control problem instances, which satisfy the assumption.

In distributionally robust optimization (DRO), we seek solutions to optimize the worst-case objective given by possible distributions contained in an ambiguity set. Compared with robust optimization that accounts for the worst-case objective outcome given by all possible realizations of uncertain parameters in an uncertainty set, optimal solutions to DRO models are less conservative and can be adjusted through the amount of data/information we have. *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, to derive tractable reformulations of several distributionally robust convex programs. Standardization of ambiguity sets via conic representable sets is proposed by *Wiesemann et al.* (2014). *Zymler et al.* (2013) consider tractable reformulations of DR chance-constrained programs using moment-based ambiguity set. Other types of ambiguity sets used in DRO models bound the ϕ-divergence (*Ben-Tal et al.*, 2013; *Jiang and Guan*, 2016) or Wasserstein distance (*Esfahani and Kuhn*, 2018; *Gao and Kleywegt*, 2016) in between possible distributions to a nominal distribution. In this chapter, 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 chapter is organized as follows. In Section 2.2, we review the most relevant POMDP, robust MDP/POMDP, and DRO literature. In Section 2.3, we formally present DR-POMDP and provide a few examples to show possible applications. In Section 2.4, we formulate the Bellman equation and show that the value function is piecewise linear convex under general moment-based ambiguity sets described in *Yu and Xu* (2016). In Section 2.5, we develop an approximation algorithm for DR-POMDP based on a distributionally robust variant of the heuristic value search iteration algorithm. In Section 2.6, 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 2.7 concludes the chapter and presents future research directions.

2.2 Literature Review

Although strong modeling connections exist in between MDP and POMDP, techniques applied to solve MDP models where the states are discrete, are not directly applicable to solving POMDP since belief states are continuous. *Smallwood and Sondik* (1973) show 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. *Pineau et al.* (2003) propose 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. *Smith and Simmons* (2004) develop a heuristic search value iteration (HSVI) algorithm to derive an upper bound of the value function via finding the reachable set through simulation. *Smith and Simmons* (2004) show that HSVI is guaranteed to terminate after the gap between the upper and lower bounds converges within a certain threshold.

The research on robust MDP is motivated by possible estimation errors of transition matrices and how they may have a significant impact to the solution quality (see, e.g., *Abbad and Filar* (1992); *Abbad et al.* (1990)). In *Wiesemann et al.* (2013), the authors show probabilistic guarantees for solutions to robust MDPs by building an uncertainty set using fully observable history. By construction, their robust policy achieves or exceeds its worst-case performance with a certain confidence. *Nilim and El Ghaoui* (2005) consider 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. *Iyengar* (2005) analyzes a robust formulation for discrete-time dynamic programming where the transition probabilities are uncertain and ambiguously known, and shows that it is equivalent to stochastic zero-sum games with perfect information. *Delage and Mannor* (2010) argue that robust MDP models may produce over-conservative solutions, as they do not incorporate the distributional information of uncertain parameters. Then *Xu and Mannor* (2012) present 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. *Yu and Xu* (2016) generalize the distributionally robust MDP to include multi-modal distributions and the information of mean and variance. *Yang* (2017) proposes a distributionally robust MDP model by building an ambiguity set of distributions on transition probability using a Wasserstein ball centered around a nominal distribution. The use of Wasserstein ball ambiguity set results in a Kantorovich-duality-based convex reformulation for distributionally robust MDP.

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

2.3 Problem Description

Figure 2.1 depicts the sequence of events that occur during one decision period. In a distributionally robust setting, we consider another agent (the "nature"), who chooses a distribution μ of the transition-observation probabilities from a pre-assumed ambiguity set. The DM expects that the nature may access to the same information as the DM and acts adversarially against the DM's action a taken at the beginning of each period. Therefore, the distribution μ is expected to lead to the worst-case expected reward. Next, the joint transition-observation probability $\boldsymbol{p}$ is realized from the distribution μ. The state makes a transition according to $\boldsymbol{p}$, and the observation outcome z is shown. Finally, the DM obtains the values of z and $\boldsymbol{p}$ at the end of the period.

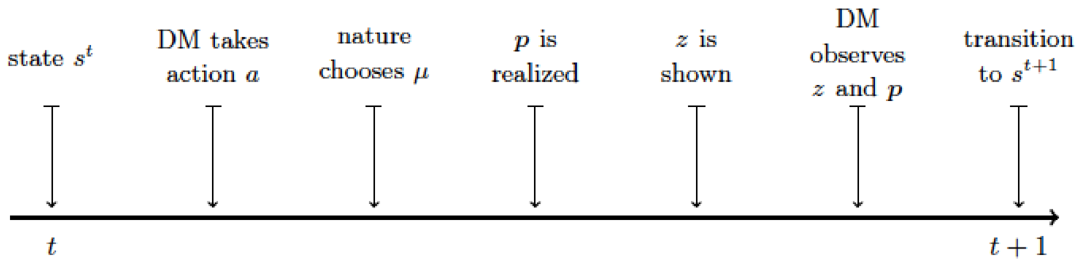


Figure 2.1: Sequence of events during one decision period in a DR-POMDP

We denote $\mathcal{S}$ as the set of states, $\mathcal{A}$ as the set of actions, and $\mathcal{Z}$ as the set of observation outcomes. For all $(s, s', z, a) \in \mathcal{S}^2 \times \mathcal{Z} \times \mathcal{A}$, we define $p_{as}(s', z) = \Pr(s', z|s, a)$ as the probability of transitioning between (s, s') and observing z, given action a. For $(s, a) \in \mathcal{S}\times\mathcal{A}$, let r_{as} be the reward for taking action a at state s. For all $s \in \mathcal{S},\ a \in \mathcal{A}$, we define a vector of probabilities $\boldsymbol{p}_{as} = (p_{as}(s', z),\ (s', z) \in \mathcal{S} \times \mathcal{Z})^\top$ and assume that the Cartesian product $(\boldsymbol{p}_{as}, r_{as})$ is a member of a set $\mathcal{X}_{as} \subseteq \Delta(\mathcal{S} \times \mathcal{Z}) \times \mathbb{R}$, where $\Delta(\cdot)$ is a probability simplex of set $\cdot$. We denote $\boldsymbol{p}_a = (p_{as}(s', z),\ (s, s', z) \in \mathcal{S}^2 \times \mathcal{Z})^\top$ and $\boldsymbol{r}_a = (r_{as},\ s \in \mathcal{S})^\top$ for all $a \in \mathcal{A}$. We assume that $(\boldsymbol{p}_{as}, r_{as})$ follows a distribution

μ_{as}, which is unknown but is included in an ambiguity set $\mathcal{D}_{as} \subseteq \mathcal{P}(\mathcal{X}_{as})$, where $\mathcal{P}(\cdot)$ represents a set of all probability distributions with support $\cdot$. Furthermore, the set of distributions is rectangular with respect to the set of actions $\mathcal{A}$ and the set of states $\mathcal{S}$, i.e., the overall ambiguity set is $\mathcal{D} = \bigotimes_{\substack{a\in\mathcal{A}\\ s\in\mathcal{S}}} D_{as}$. This assumption is analogous to the (s, a)-rectangularity in *Wiesemann et al.* (2013). The above conditions increase the conservativeness of the model in general. In Section A.1, we discuss a relaxation of the a-rectangularity assumption for DR-POMDP.

Below we describe several examples in which the above settings of DR-POMDP can be justified, and therefore our approach can be applied to optimize corresponding policies. The key is to justify whether the DM can obtain the true value of $\boldsymbol{p}$ using side information at the end of each decision period. In Section 2.6, we also numerically show that our approach can produce quite stable reward in out-of-sample simulation tests even we add noise to the true $\boldsymbol{p}$-value obtained at the end of each period and thus the assumption is relatively weak.

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

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

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

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

2.4 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 2.4.1, we formulate DR-POMDP as an optimization problem and construct the Bellman equation to derive the optimal policy. In Section 2.4.2, we show that the value function satisfying the Bellman equation is PWLC. Finally, in Section 2.4.3, we consider the infinite-horizon case, and demonstrate that the value function converges under the Bellman update operation.

2.4.1 Distributionally Robust Bellman Equation

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

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

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

$$\boldsymbol{b}^{t+1} = \boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) = \frac{\sum_{s\in\mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s}{\sum_{s\in\mathcal{S}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} b_s}, \tag{2.1}$$

where $\mathbf{1}$ represents a vector of ones having the length $|\mathcal{S}|$; $\boldsymbol{J}_z \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\times|\mathcal{Z}|)}$ is a matrix of zeros and ones that projects the vector $\boldsymbol{p}_{as}$ to a vector $\boldsymbol{p}_{asz} = (p_{as}(s', z),\ s' \in \mathcal{S})^\top$, whose entries correspond to the outcome z. That is, $\boldsymbol{p}_{asz} = \boldsymbol{J}_z \boldsymbol{p}_{as},\ \forall a,\ s,\ z$. Note that the belief state cannot be updated using (2.1) and will not be a sufficient statistic of the history of past actions and observations if we do not have the true values of

$\boldsymbol{p}_{as}$.

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

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

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

$$\begin{aligned}
\max_{\pi \in \Pi} \min_{\gamma \in \Gamma} \quad & \mathbb{E}\left[\sum_{t=1}^{T-1} \beta^t {\boldsymbol{b}^t}^\top \boldsymbol{r}_{a^t}^t \right] && && \text{(2.2a)} \\
\text{s.t.} \quad & a^t = \pi^t(\boldsymbol{b}^t), && \forall t \in \{1, \ldots, T-1\} && \text{(2.2b)} \\
& \mu_{a^t}^t = \gamma^t(\boldsymbol{b}^t, a^t), && \forall t \in \{1, \ldots, T-1\} && \text{(2.2c)} \\
& (\boldsymbol{p}_{a^t}^t, \boldsymbol{r}_{a^t}^t) \sim \mu_{a^t}^t, && \forall t \in \{1, \ldots, T-1\} && \text{(2.2d)} \\
& (s^{t+1}, z^t) \sim \boldsymbol{p}_{a^t s^t}^t, && \forall t \in \{1, \ldots, T-1\} && \text{(2.2e)} \\
& \boldsymbol{b}^{t+1} = \boldsymbol{f}(\boldsymbol{b}^t, a^t, \boldsymbol{p}_{a^t}^t, z^t), && \forall t \in \{1, \ldots, T-1\} && \text{(2.2f)}
\end{aligned}$$

where the terminal reward is zero without loss of generality. The initial belief state is given as $\boldsymbol{b}$. Alternatively, we denote the problem (2.2) as

$$\max_{\pi\in\Pi}\min_{\gamma\in\Gamma}\mathbb{E}\left[\sum_{t=1}^{T-1}\beta^t \boldsymbol{b}^{t\top}\boldsymbol{r}_{a^t}^t \,\middle|\, \boldsymbol{b}^1=\boldsymbol{b}\right]. \tag{2.3}$$

Here we omit all the constraints in (2.2) for presentation simplicity.

To solve (2.2), we propose to use dynamic programming, and derive the Bellman equation below.

Proposition II.2. *Denote $\pi^{t:T-1} = (\pi^t, \pi^{t+1}, \ldots, \pi^{T-1})$ and $\gamma^{t:T-1} = (\gamma^t, \gamma^{t+1}, \ldots, \gamma^{T-1})$ as sequences of policies from t to $T-1$. Let $\Pi^{t:T-1}$ and $\Gamma^{t:T-1}$ be the sets of all policies $\pi^{t:T-1}$ and $\gamma^{t:T-1}$, respectively. Consider the value function in period t as*

$$V^t(\boldsymbol{b}) = \max_{\pi^{t:T-1}\in\Pi^{t:T-1}}\min_{\gamma^{t:T-1}\in\Gamma^{t:T-1}}\mathbb{E}\left[\sum_{n=t}^{T-1}\beta^{n-t}\boldsymbol{b}^{n\top}\boldsymbol{r}_{a^n}^n \,\middle|\, \boldsymbol{b}^t=\boldsymbol{b}\right]. \tag{2.4}$$

Then,

$$V^t(\boldsymbol{b}) = \max_{a\in\mathcal{A}}\min_{\mu_a\in\mathcal{D}_a}\mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a}\left[\sum_{s\in\mathcal{S}} b_s\left\{r_{as}+\beta\sum_{z\in\mathcal{Z}}\mathbf{1}^\top\boldsymbol{J}_z\boldsymbol{p}_{as}V^{t+1}\left(\boldsymbol{f}\left(\boldsymbol{b},a,\boldsymbol{p}_a,z\right)\right)\right\}\right]. \tag{2.5}$$

Proof. We first isolate the term associated with period t inside the expectation of (2.4) as follows.

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

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

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

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

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

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

$$\begin{aligned}
V^t(\boldsymbol{b}) = &\max_{a\in\mathcal{A}} \min_{\mu_a\in\mathcal{D}_a} \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a} \Bigg[\sum_{s\in\mathcal{S}} b_s \Bigg\{ r_{as} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} \\
&\times \max_{\pi^{t+1:T-1}\in\Pi^{t+1:T-1}} \min_{\gamma^{t+1:T-1}\in\Gamma^{t+1:T-1}} \mathbb{E}\Bigg[\sum_{n=t+1}^{T-1} \beta^{n-(t+1)} \boldsymbol{b}^{n\top} \boldsymbol{r}^n_{a^n} \Bigg| \boldsymbol{b}^{t+1} = \boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z)\Bigg]\Bigg\}\Bigg] \\
= &\ (2.5).
\end{aligned}$$

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

Following Proposition II.2, the policies optimal to (2.3) can be determined by recursively solving (2.5) from period T to $t = 1$.

Now define two functions:

$$U^t(\boldsymbol{b}, a, \mu_a) = \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a} \left[\sum_{s\in\mathcal{S}} b_s \left\{ r_{as} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1}\left(\boldsymbol{f}\left(\boldsymbol{b}, a, \boldsymbol{p}_a, z\right)\right)\right\}\right], \tag{2.6}$$

$$Q^t(\boldsymbol{b}, a) = \min_{\mu_a\in\mathcal{D}_a} U^t(\boldsymbol{b}, a, \mu_a). \tag{2.7}$$

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

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

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

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

2.4.2 Properties of Distributionally Robust Bellman Equation (2.5)

We consider an ambiguity set based on mean absolute deviation of transition-observation probabilities as described below. We refer the readers to Section A.2 for a more general ambiguity set that can also involve ambiguity in the reward, and the mean values are on an affine manifold with conic representable support. The same property here holds for DR-POMDP with the general ambiguity set and we omit the details for presentation simplicity.

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

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

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

$$\tilde{\mathcal{D}}_{as} = \left\{ \tilde{\mu}_{as} \begin{pmatrix} \boldsymbol{p}_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \middle| \begin{array}{l} \mathbb{E}_{(\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as}) \sim \tilde{\mu}_{as}} \left[\tilde{\boldsymbol{u}}_{as} \right] = \boldsymbol{c}_{as} \\ \tilde{\mu}_{as} \left(\tilde{\mathcal{X}}_{as} \right) = 1 \end{array} \right\}, \tag{2.8}$$

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

$$\tilde{\mathcal{X}}_{as} = \left\{ \begin{pmatrix} \boldsymbol{p}_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \in \begin{array}{c} \mathbb{R}_{+}^{|\mathcal{S}| \times |\mathcal{Z}|} \\ \mathbb{R}^{L} \end{array} \middle| \begin{array}{c} \tilde{\boldsymbol{u}}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as} \\ \tilde{\boldsymbol{u}}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as} \\ \mathbf{1}^{\top} \boldsymbol{p}_{as} = 1 \end{array} \right\}. \tag{2.9}$$

For ambiguity sets and supports respectively defined in terms of (2.8) and (2.9), we show that the value function is convex with respect to the belief state $\boldsymbol{b}$ for each decision period.

Theorem II.3. *For all $a \in \mathcal{A}$ and $s \in \mathcal{S}$, let the ambiguity set and support be* (2.8) *and* (2.9)*, respectively. For all $t \in \{1, \ldots, T\}$, there exists a set Λ^t of slopes such that the value function can be expressed as follows.*

$$V^t(\boldsymbol{b}) = \max_{\boldsymbol{\alpha} \in \Lambda^t} \boldsymbol{\alpha}^{\top} \boldsymbol{b}. \tag{2.10}$$

A detailed proof of Theorem II.3 is shown in Section A.3. Following this result, having provided the values of a and $\boldsymbol{\alpha}_{az}$, the inner minimization in (A.14) can be solved efficiently using linear programming. The issue, however, is that there are possibly infinitely many elements in $\operatorname{Conv}(\Lambda^{t+1})$, and even if there are finitely many, the number of supporting hyperplanes $\boldsymbol{\alpha}$ inside Λ^t increases exponentially as the value functions are calculated from period $t = T$ to $t = 1$. We describe in Section 2.5 a heuristic search value iteration (HSVI) algorithm for efficiently computing optimal policies in DR-POMDP.

2.4.3 Case of Infinite Horizon

We show that the PWLC property of the value function can be extended to the case with infinite horizon. We prove the result by following the Banach fixed point theorem (see, e.g., *Puterman* (2014)), and show that by repeatedly updating the value function in (2.5), it converges to a unique function corresponding to the optimal value V^* of the infinite-horizon DR-POMDP problem.

Theorem II.4. *The operator $\mathcal{L}$ defined as*

$$\mathcal{L}V(\boldsymbol{b}) = \max_{a\in\mathcal{A}} \min_{\mu_a\in\tilde{\mathcal{D}}_a} \mathbb{E}_{(\boldsymbol{p}_a,\boldsymbol{r}_a)\sim\mu_a}\left[\sum_{s\in\mathcal{S}} b_s\left(r_{as} + \beta\sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V\left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z)\right)\right)\right] \tag{2.11}$$

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

We refer the readers to a detailed proof provided in Section A.3. Theorem II.4 suggests that by employing the exact algorithm discussed in the finite horizon case, starting from any initial value function, the value function V converges to an optimal function V^* with rate β by iteratively performing the Bellman operator $\mathcal{L}$. Therefore, we can use the same solution approach to be discussed in Section 2.5 for handling both finite-horizon and infinite-horizon cases of DR-POMDP.

2.5 Solution Method

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

Algorithm 1 Heuristic Search Value Iteration (HSVI)

```
Input: initial belief state $\boldsymbol{b}^0$, tolerance $\epsilon$
Initialize: $\overline{V}$, $\underline{V}$ (see details in Section 2.5.1)
while $\overline{V}(\boldsymbol{b}^0) - \underline{V}(\boldsymbol{b}^0) > \epsilon$ or time limit is reached do
    DR-BoundExplore($\boldsymbol{b}^0, 0$) (see details in Algorithm 2)
end while
Output: $\overline{V}$, $\underline{V}$
```

Algorithm 1 presents the main algorithmic steps in HSVI, where the details of Step 4 are later provided in Algorithm 2. During Step 4, one sample path of DM, the nature's action and the observation outcomes are greedily selected, and then the bounds are updated using Bellman equations. Figure 2.2 demonstrates how the lower bound of the value function can be described as the maximum of the lower bounding hyperplanes, and the upper bound can be described as a convex hull of the upper bounding points. Figure 2.3 illustrates an example of how newly discovered bounding hyperplanes and points can be used to locally update the bounds.

In Section 2.5.1, we explain how the upper and lower bounds of the value function are initialized (i.e., the details for Step 2), and in Section 2.5.2, we present an exploration strategy to close the gap to a pre-determined tolerance level. Finally, in Section 2.5.3, we discuss how the value functions are updated given a belief state $\boldsymbol{b}$.

2.5.1 Initialization

Recall the ambiguity set and support defined in (2.8) and (2.9), 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

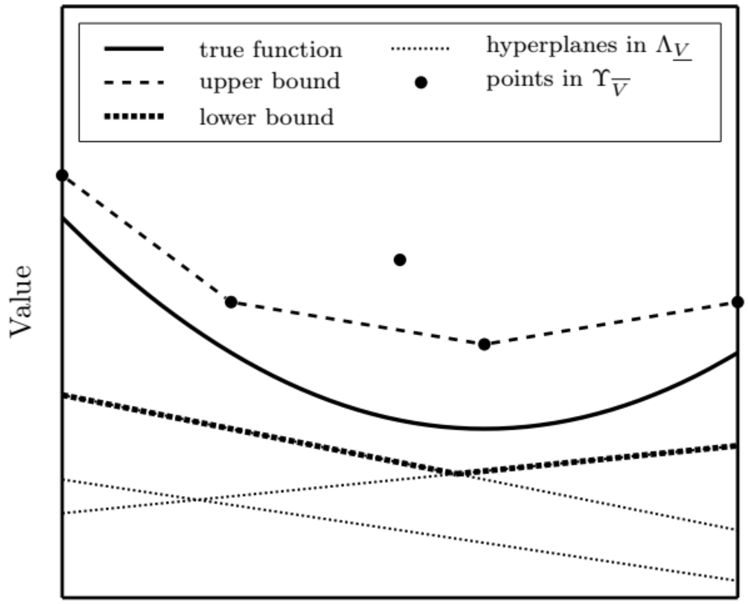


Figure 2.2: An example of upper- and lower-bounds of a value function

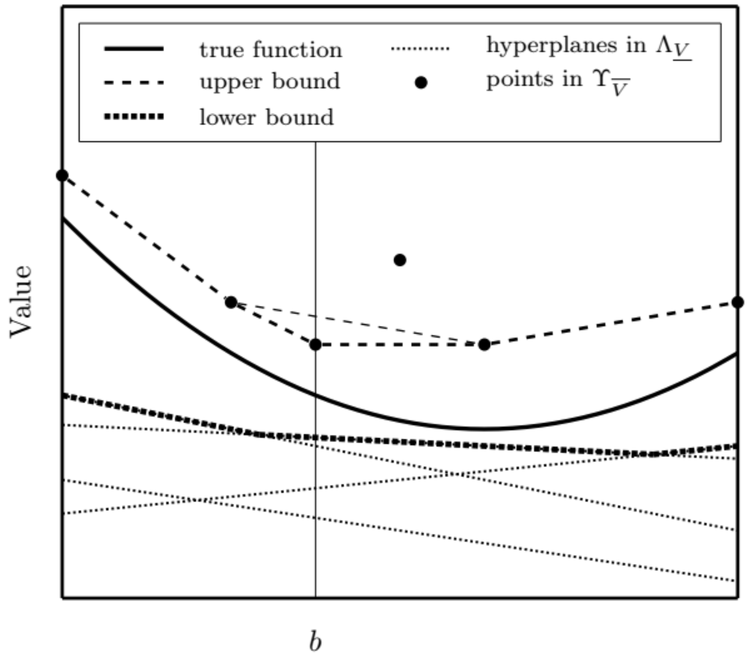


Figure 2.3: An example of updated upper- and lower-bounds

period. That is, for each action a, we solve

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

In the case of mean absolute deviation based ambiguity set (2.8), the second minimization is trivial as r_{as} is fixed. The minimum value for all $s \in \mathcal{S}$ is computed by enumeration. We then define an initial lower bounding hyperplane $\alpha'_s = \max_{a \in \mathcal{A}} \underline{R}_a$, $\forall s \in \mathcal{S}$ and set $\Lambda_{\underline{V}} = \{\boldsymbol{\alpha}'\}$, where $\boldsymbol{\alpha}' = (\alpha'_s, s \in \mathcal{S})^\top$.

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

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

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

$$\min_{\boldsymbol{V}^{MDP}} \quad \mathbf{1}^\top \boldsymbol{V}^{MDP} \tag{2.12a}$$

$$\text{s.t.} \quad V_s^{MDP} \geq \max_{\mu_{as} \in \mathcal{D}_{as}} \mathbb{E}_{(\boldsymbol{p}_{as}, r_{as}) \sim \mu_{as}} \left[r_{as} + \beta \boldsymbol{V}^{MDP\top} \sum_{z \in \mathcal{Z}} \boldsymbol{J}_z \boldsymbol{p}_{as} \right], \ \forall a \in \mathcal{A}, s \in \mathcal{S}. \tag{2.12b}$$

In the case of ambiguity set (2.8), model (2.12) becomes

$$\min_{\boldsymbol{\rho},\boldsymbol{\kappa},\boldsymbol{V}^{MDP}} \quad \mathbf{1}^\top \boldsymbol{V}^{MDP} \tag{2.13a}$$

$$\text{s.t.} \quad V_s^{MDP} - \boldsymbol{c}_{as}^\top \boldsymbol{\rho}_{as} - \bar{\boldsymbol{p}}_{as}^\top \boldsymbol{\kappa}_{as}^1 + \bar{\boldsymbol{p}}_{as}^\top \boldsymbol{\kappa}_{as}^2 - \sigma_{as} \geq r_{as}, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{2.13b}$$

$$\beta \sum_{z \in \mathcal{Z}} \boldsymbol{J}_z{}^\top \boldsymbol{V}^{MDP} - \boldsymbol{\kappa}_{as}^1 + \boldsymbol{\kappa}_{as}^2 - \mathbf{1}\sigma_{as} \leq 0, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{2.13c}$$

$$\boldsymbol{\kappa}_{as}^1 + \boldsymbol{\kappa}_{as}^2 - \boldsymbol{\rho}_{as} = 0, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{2.13d}$$

$$\boldsymbol{\kappa}_{as}^1, \kappa_{as}^2 \in \mathbb{R}_+^{|\mathcal{S}| \times |\mathcal{A}|}, \sigma_{as} \in \mathbb{R}\ \boldsymbol{\rho}_{as} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}, \quad \forall s \in \mathcal{S},\ a \in \mathcal{A} \tag{2.13e}$$

$$\boldsymbol{V}^{MDP} \in \mathbb{R}^{|\mathcal{S}|}. \tag{2.13f}$$

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

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

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

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

$$v = \min \left\{ \sum_{i \in [|\Upsilon_{\overline{V}}|]} w^i v^i \;\middle|\; \sum_{i \in [|\Upsilon_{\overline{V}}|]} w^i \boldsymbol{b}^i = \boldsymbol{b},\ \sum_{i \in [|\Upsilon_{\overline{V}}|]} w^i = 1,\ w^i \geq 0,\ \forall i \in [|\Upsilon_{\overline{V}}|] \right\}, \tag{2.14}$$

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

2.5.2 Forward Exploration Heuristics

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

We consider the following function:

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

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

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

$$\begin{aligned}
& U_{\overline{V}}(\boldsymbol{b}, a^*, \mu^*_{a^*}) - U_{\underline{V}}(\boldsymbol{b}, a^*, \mu^*_{a^*}) \\
= \ & \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \left(r_{a^*s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \overline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right)\right)\right] \\
& \quad - \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \left(r_{a^*s} + \beta \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \underline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right)\right)\right] \\
= \ & \beta \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*})\sim\mu^*_{a^*}}\left[\sum_{s\in\mathcal{S}} b_s \sum_{z\in\mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^*s} \left(\overline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right) - \underline{V}\left(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z)\right)\right)\right].
\end{aligned} \tag{2.15}$$

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

$$\begin{aligned}
\overline{V}(\boldsymbol{b}) - \underline{V}(\boldsymbol{b}) &= \max_{a \in \mathcal{A}} \min_{\mu_a \in \tilde{\mathcal{D}}_a} U_{\overline{V}}(\boldsymbol{b}, a, \mu_a) - \max_{a \in \mathcal{A}} \min_{\mu_a \in \tilde{\mathcal{D}}_a} U_{\underline{V}}(\boldsymbol{b}, a, \mu_a) \\
&\leq \max_{a \in \mathcal{A}} U_{\overline{V}}(\boldsymbol{b}, a, \mu_a^*) - \max_{a \in \mathcal{A}} U_{\underline{V}}(\boldsymbol{b}, a, \mu_a^*) \\
&\leq U_{\overline{V}}(\boldsymbol{b}, a^*, \mu_{a^*}^*) - U_{\underline{V}}(\boldsymbol{b}, a^*, \mu_{a^*}^*). \qquad (2.16)
\end{aligned}$$

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

To achieve the gap ϵ at the initial state $\boldsymbol{b}_0$, the condition for the gap at depth level t starting from the initial one is only $\epsilon\beta^{-t}$, which can readily be seen from (2.15) and (2.16). We define the difference of the gap and the required condition as the excess uncertainty, which is

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

Using (2.16) and applying the identity (2.15), we have

$$\text{excess}(\boldsymbol{b}, t) \leq \beta \mathbb{E}_{(\boldsymbol{p}_{a^*}, \boldsymbol{r}_{a^*}) \sim \mu_{a^*}^*} \left[\sum_{s \in \mathcal{S}} b_s \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^* s} \text{excess}(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z), t+1) \right]. \qquad (2.17)$$

Next, we greedily choose $(z^*, p_{a^*}^*)$ so that the quantity associated to the pair in RHS of (2.17) has the maximum expected value, i.e.,

$$(z^*, \boldsymbol{p}_{a^*}^*) \in \underset{z \in \mathcal{Z},\ \boldsymbol{p}_{a^*} \in \mathcal{X}_{a^*}}{\arg\max}\ \mu_{a^*}^*(\boldsymbol{p}_{a^*}) \times \sum_{s \in \mathcal{S}} b_s \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a^* s}^* \text{excess}(\boldsymbol{f}(\boldsymbol{b}, a^*, \boldsymbol{p}_{a^*}, z), t+1). \qquad (2.18)$$

Note that because the worst-case distribution under ambiguity set (2.8) is a point mass distribution, obtaining $\boldsymbol{p}_{a^*}^*$ is trivial. Algorithm 2 describes the detailed algorithmic steps. In the HSVI approach, Algorithm 2 is called recursively to make decisions on which branch to choose in the next depth level $t+1$. After the simulation is terminated, the updates on the lower and upper bounds are made for the belief states that are discovered through the simulation.

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

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

2.5.3 Local Updates

In this section, we describe the details of *DR-backup* and *DR-update* steps in Algorithm 2. We first illustrate how the lower bound is updated in *DR-backup*. For each $a \in \mathcal{A}$, we solve the two inner maximization problems in (A.13) provided a and $\boldsymbol{b}$, where we set $\Lambda^{t+1} = \Lambda_{\underline{V}}$. The convex hull of $\Lambda_{\underline{V}}$ is therefore,

$$\text{Conv}\left(\Lambda_{\underline{V}}\right) = \left\{ \sum_{i \in [|\Lambda_{\underline{V}}|]} w^i \boldsymbol{\alpha}^i \,\middle|\, \sum_{i \in [|\Lambda_{\underline{V}}|]} w^i = 1,\ \boldsymbol{\alpha}^i \in \Lambda_{\underline{V}},\ w^i \geq 0,\ i \in [|\Lambda_{\underline{V}}|] \right\}. \tag{2.19}$$

Thus, we combine the two inner maximization problems in (A.13) as

$$\begin{aligned}
\max_{\boldsymbol{\rho}_a, \boldsymbol{\kappa}^1_a, \boldsymbol{\kappa}^2_a, \boldsymbol{\sigma}_a} \quad & \sum_{s \in \mathcal{S}} \boldsymbol{c}^\top_{as} \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} b_s r_{as} + \sum_{s \in \mathcal{S}} \left(-\bar{p}^\top_{as} \boldsymbol{\kappa}^1_{as} + \bar{p}^\top_{as} \boldsymbol{\kappa}^2_{as} + \sigma_{as} \right) && && \text{(2.20a)} \\
\text{s.t.} \quad & \beta b_s \sum_{z \in \mathcal{Z}} \sum_{i \in [|\Lambda_{\underline{V}}|]} w^i_{az} \boldsymbol{J}_z{}^\top \boldsymbol{\alpha}^i_{az} + \boldsymbol{\kappa}^1_{as} - \boldsymbol{\kappa}^2_{as} - \mathbf{1}\sigma_{as} \geq 0, && \forall s \in \mathcal{S} && \text{(2.20b)} \\
& \sum_{i \in [|\Lambda_{\underline{V}}|]} w^i_{az} = 1, && \forall z \in \mathcal{Z} && \text{(2.20c)} \\
& w^i_{az} \in \mathbb{R}_+, && \forall i \in [|\Lambda_{\underline{V}}|],\ z \in \mathcal{Z} && \text{(2.20d)} \\
& \text{(A.12c)}, \text{(A.12d)}, \text{(A.13b)}.
\end{aligned}$$

We denote the optimal solutions to (2.20) using a superscript $\star$, and let the optimal dual solutions associated with constraints (2.20b) be $\hat{\boldsymbol{p}}^{\star}_{as}$. For each action $a \in \mathcal{A}$, we can generate a lower bounding hyperplane

$$\boldsymbol{\alpha}' = \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}^{\star\top}_{az} \boldsymbol{J}_z \hat{\boldsymbol{p}}^{\star}_{as},\ s \in \mathcal{S} \right)^{\top}, \tag{2.21}$$

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

```
Algorithm 3 DR-backup($\boldsymbol{b}, \Lambda_{\underline{V}}$)
Input: belief $\boldsymbol{b}$, lower bounding hyperplanes $\Lambda_{\underline{V}}$
for $\forall a \in \mathcal{A}$ do
    solve (2.20) for action $a$
    $\mathcal{L}(a) \leftarrow \boldsymbol{\alpha}'$ (calculated using (2.21))
end for
Output: $\mathrm{argmax}_{\boldsymbol{\alpha} \in \mathcal{L}} \boldsymbol{\alpha}^{\top}\boldsymbol{b}$
```

Next, we discuss how to update the upper bound and describe the algorithmic steps of *DR-update* in Algorithm 4. Combining (A.13) and the dual representation of (2.14), for each $a \in \mathcal{A}$, we solve

$$\max_{\boldsymbol{\rho}_a, \boldsymbol{\kappa}^1_a, \boldsymbol{\kappa}^2_a, \boldsymbol{\sigma}_a} \quad \sum_{s \in \mathcal{S}} \boldsymbol{c}^{\top}_{as} \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} b_s r_{as} + \sum_{s \in \mathcal{S}} \left(-\bar{p}^{\top}_{as} \boldsymbol{\kappa}^1_{as} + \bar{p}^{\top}_{as} \boldsymbol{\kappa}^2_{as} + \sigma_{as} \right) \tag{2.22a}$$

$$\text{s.t.} \quad \beta b_s \sum_{z \in \mathcal{Z}} \boldsymbol{J}_z{}^{\top} \varphi_{az} + \beta b_s \sum_{z \in \mathcal{Z}} \psi_{az} \boldsymbol{J}_z{}^{\top} \mathbf{1} + \boldsymbol{\kappa}^1_{as} - \boldsymbol{\kappa}^2_{as} - \mathbf{1}\sigma_{as} \geq 0, \quad \forall s \in \mathcal{S} \tag{2.22b}$$

$$\boldsymbol{b}^{i\top} \varphi_{az} + \psi_{az} \leq v_i, \qquad \forall z \in \mathcal{Z}, i \in [|\Upsilon_{\overline{V}}|] \tag{2.22c}$$

$$\varphi_{az} \in \mathbb{R}^{|\mathcal{S}|},\ \psi_{az} \in \mathbb{R}, \qquad \forall z \in \mathcal{Z}, i \in [|\Upsilon_{\overline{V}}|] \tag{2.22d}$$

$$\text{(A.12c)}, \text{(A.12d)}, \text{(A.13b)}.$$

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

Remark II.5. The complexity of the related algorithm presented in *Smith and Sim-*

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

```
Input: belief $\boldsymbol{b}$, upper bounding points $\Upsilon_{\overline{V}}$
for $\forall a \in \mathcal{A}$ do
    $\mathcal{Q}(a) \leftarrow$(optimal objective value of (2.22) for action $a$)
end for
Output: $(\boldsymbol{b}, \max_{a \in \mathcal{A}}\{\mathcal{Q}(a)\})$
```

mons (2004) is based on the finiteness of the scenario tree up to a tolerance level ϵ. In the DR-HSVI algorithm, the scenario tree is not finite as the nature is able to choose from a continuous ambiguity set of distributions, and therefore the scenario tree has an infinite number of elements. Later we numerically demonstrate the convergence of the DR-HSVI algorithm in Section 2.6 for different combinations of parameter choices.

2.6 Numerical Studies

We test DR-POMDP policies for dynamic epidemic control (Sections 2.6.1 and 2.6.2), and compare the results of a two-state epidemic control problem with the ones given by POMDP and robust POMDP (Section 2.6.1.1). We vary parameter choices to test the robustness and sensitivity of DR-POMDP policies (i) under various types of ambiguity sets used in the in-sample tests (Sections 2.6.1.2, 2.6.1.3) and (ii) given certain noise added to the transition-observation probability value obtained at the end of each decision period in out-of-sample tests (Sections 2.6.1.4, 2.6.1.5). In Sections 2.6.2.1 and 2.6.2.2, we increase the sizes of the two-state influenza epidemic control instances in Section 2.6.1, demonstrate the algorithmic convergence, and present computational time results of using POMDP and DR-POMDP for solving larger-scale epidemic control instances.

2.6.1 Two-state Influenza Epidemic Control Problem

We study the problem of influenza epidemic control mentioned in Section 2.3. 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 to obtain more accurate ILI rate.

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

$$\begin{pmatrix} 0.99 - 0.1a & 0.01 + 0.1a \\ 0.3 - 0.1a & 0.7 + 0.1a \end{pmatrix}. \tag{2.23}$$

When $a = 0$ (i.e., the DM does nothing), the above transition probabilities follow studies on influenza epidemics (see, e.g., *Le Strat and Carrat* (1999)). The setting of the matrix (2.23) indicates that higher-level actions (i.e., more restrictive control strategies) will lead to greater chances that an epidemic state turns into non-epidemic and that a non-epidemic state remains itself. The transition probability for $a =$ 'Inspection' ('I') is the same as the one for $a = 0$. The observation outcome is the ILI rate, calculated as the number of ILI patients per 1000 population. For actions $a \in \{0, 1, 2\}$, we follow *Rath et al.* (2003) and assume that the ILI rate follows a Gaussian distribution with mean value $\mu_E = 2 - 0.5a$ and variance $\text{Var}_E = 30 - \mu_E^2$ for $s =$ 'Epidemic' ('E'), and with mean $\mu_N = 0.2 - 0.05a$ and variance $\text{Var}_N = 2 - \mu_N^2$ for $s =$ 'Non-epidemic' ('N'). We discretize the observation outcome into five levels as $\{(-\infty, 0], (0, 1/3], (10/3, 20/3], (20/3, 10], (10, \infty)\}$. For $a =$ 'I', the probabilities of observing the five outcomes are $\{0.01, 0.1/3, 0.1/3, 0.1/3, 0.89\}$ when $s =$ 'E', and the ILI rate follows the same distribution as the one of $a = 0$ if $s =$ 'N', to model the situation where more careful inspection action can result in more ILI patients showing up. The rewards for each action-state combination are presented in Table 2.1, reflecting the negative number of total infections minus the effort paid for different actions in different states.

Table 2.1: Reward setting for each state-action pair

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

When implementing the HSVI algorithm in Section 2.5 for solving DR-POMDP, we set the discount factor $\beta = 0.95$ and the gap tolerance $\epsilon = 1.0$. The computation is terminated when the gap between the upper and lower bounds is less than ϵ, at the initial states $b^0_E = 0.5,\ b^0_N = 0.5$. 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 $\Lambda_{\underline{V}}$, and the complexity of computing the upper bound is polynomial in the size of set $\Upsilon_{\overline{V}}$ as we need to solve linear programs. Both $|\Lambda_{\underline{V}}|$ and $|\Upsilon_{\overline{V}}|$ increase monotonically, but most elements in the two sets are dominated by others. We follow a heuristic to prune all the dominated elements whenever the number of elements increases by 10%.

2.6.1.1 Policy Comparison

We compare DR-POMDP policies with the ones by POMDP and robust POMDP via cross testing. We randomly generate ten samples of the transition probability for Level 2 action (i.e., $a = 2$) and epidemic state (i.e., $s =$ 'E'), by keeping all the values the same as the base setting in (2.23) but letting the probability $p_2(N|E) =$

$0.99-0.1\times 2+0.1\times x$, where x follows a standard Normal distribution. (We make sure that $0 \leq p_2(N|E) \leq 1$ and re-sample if not.) For all three approaches, the mean value of the ten samples is used as the nominal transition probability. For robust POMDP, the maximum L1 norm from the mean defines an uncertainty set centered around the nominal probability. For DR-POMDP, we use the mean absolute deviation to define the ambiguity set.

Table 2.2: Estimated median values of the cross-tested rewards

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

Table 2.3: Estimated five-percentile values of the cross-tested rewards

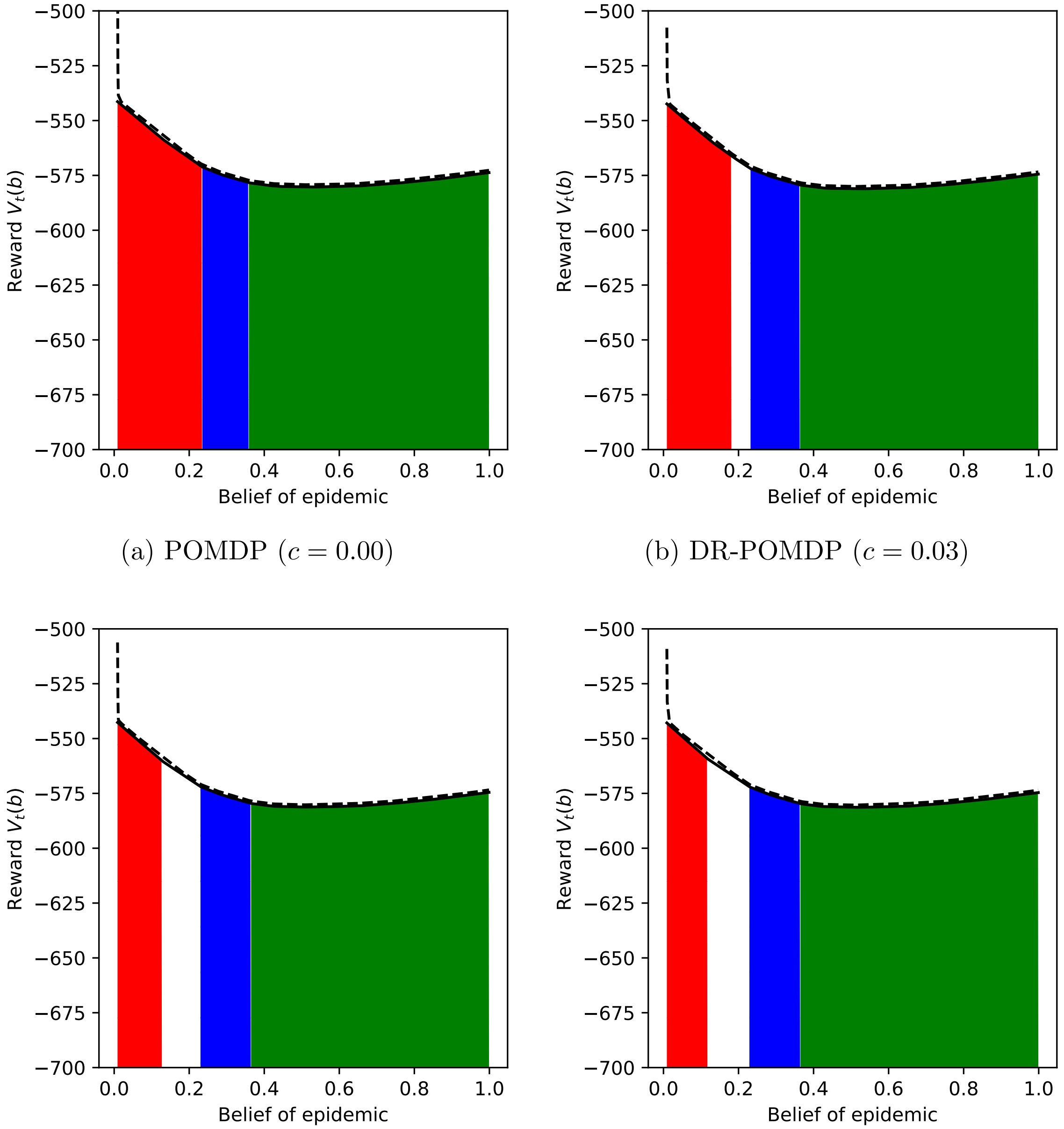
	Nature's policy		
DM's policy	POMDP(std)	DR-POMDP(std)	Robust(std)
POMDP	$-\mathbf{656.99}$ (2.39)	-696.34 (1.34)	-711.14 (1.43)
DR-POMDP	-669.26 (2.35)	$-\mathbf{677.87}$ (1.95)	-705.61 (1.60)
Robust	-689.26 (1.78)	-691.77 (2.07)	$-\mathbf{698.93}$ (2.19)

We implement the DM's optimal polices given by different approaches in out-of-sample environments where the nature follows the settings of POMDP, DR-POMDP, and robust POMDP to realize the transition probabilities in each period. The number of simulated instances is 5000 each. We report the estimated value of the median and the 5-percentile values of the reward in each case in Tables 2.2 and 2.3, respectively using Harrell-Davis quantile estimator (*Harrell and Davis*, 1982). We also include the standard deviation of the estimator. Note that the 5-percentile of the reward is equivalent to the 95-percentile of the cost, indicating the tail (worse) performance of different policies. Therefore, Tables 2.2 and 2.3 indicate that POMDP has the smallest reward when the nature agrees with the DM to pick the nominal transition probabilities at each decision period, but it can lead to much worse reward (both in terms of the mean value and tail performance) if the transition probabilities are realized as the worst-case (in robust POMDP) or from the worst-case distribution (in DR-POMDP). On the other hand, the performance of DR-POMDP solutions is quite stable and robust under all out-of-sample circumstances but the tail performance is worse than the mean results. Lastly, the robust POMDP policy yields worse mean value and tail performance when the true environment is POMDP or DR-POMDP.

2.6.1.2 Results of Varying Ambiguity Set Sizes

We first only consider an ambiguity in the transition-observation probabilities of Level 0 action and epidemic state. We build the ambiguity set based on the mean absolute deviation such that $\mathbb{E}_{\boldsymbol{p}_{as}\sim\mu_{as}}\left[|\boldsymbol{p}_{as}-\bar{\boldsymbol{p}}_{as}|\right]\leq \boldsymbol{c}_{as}$ for $a=0$ and $s=$ 'E', where $\bar{\boldsymbol{p}}_{as}\in\Delta(\mathcal{S}\times\mathcal{Z})$ is the mean value of given probability samples and $\boldsymbol{c}_{as}\in\mathbb{R}^{|\mathcal{S}\times\mathcal{Z}|}$. We let $\boldsymbol{c}_{as}$ be $c\cdot\mathbf{1}$ for some $c\in\mathbb{R}$ and vary the values of c in our tests to vary the size of the ambiguity set.

We vary $c=0.03, 0.06, 0.09$ for DR-POMDP and also compute the POMDP policy using $\bar{\boldsymbol{p}}_{as}$ as the transition-observation probabilities for all a and s, which corresponds to a special case of DR-POMDP with $c=0.00$. Figure 2.4 depicts the upper bound (dashed line) and the lower bound (solid line) of the value functions of POMDP and DR-POMDP, as well as optimal actions corresponding to different beliefs of the epidemic. The region of the belief in red (horizontal shade) corresponds to Level 0 action, blue (dotted shade) to Level 1 action, green (cross shade) to Level 2 action, and white (diagonal shade) to Inspection action. Because the ambiguity is in the transition-observation probabilities related to $a=0$, in all the subfigures, as compared to POMDP, the DR-POMDP policy relies less on Level 0 action and replaces it with the 'Inspection' action when the belief of epidemic is relatively higher. When the belief increases further, both DR-POMDP and POMDP agree on implementing Level 1 or Level 2 action. As the ambiguity set size increases (i.e., c increases), the DR-POMDP



(a) POMDP ($c = 0.00$)

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

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

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

Figure 2.4: Value functions for different ambiguity-set sizes. Solid line: lower bound, dashed line: upper bound. Corresponding actions: Level 0 – (red, horizontal), Level 1 – (blue, dot), Level 2 – (green, cross), Inspection – (white, diagonal)

policy becomes more conservative and shifts to the ‘Inspection’ action earlier, even in relatively low belief of epidemic.

2.6.1.3 Results of Multiple Ambiguities

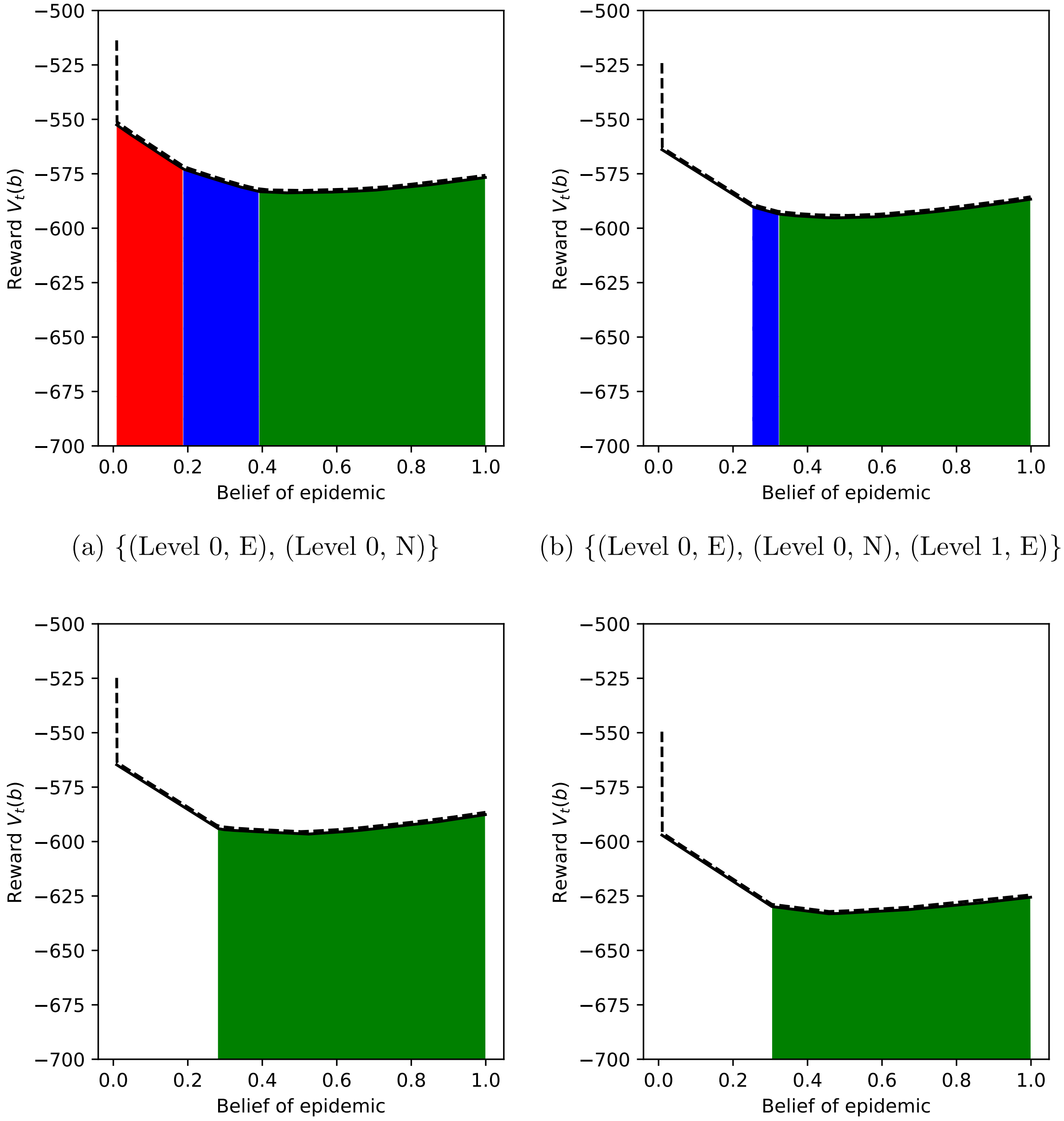
Next, we increase the number of action-state pairs that have distributional ambiguity in the transition-observation probabilities. We use $c = 0.05$ for all ambiguity sets and vary the number of action-state pairs among $\{2, 3, 4, 5\}$. In Figure 2.5a, action-state pairs (Level 0, E) and (Level 0, N) have ambiguous probability distributions and then we add pairs (Level 1, E), (Level 1, N), and (Level 2, E) one by one in the subsequent Figures 2.5b, 2.5c, 2.5d.

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

2.6.1.4 Solution Robustness under Different Ambiguity Sets

We simulate the DR-POMDP policies on instances with an initial state ‘E’ chosen with probability 50%. We use different sizes of ambiguity sets for the nature to choose the worst-case distributions in the in-sample computation. Specifically, we consider $c = 0.03, 0.06, 0.09$ to compute DR-POMDP policies using the ambiguity setting in Section 2.6.1.2 and then vary $c' = 0.00, 0.03, 0.06, 0.09$ to change the nature’s ambiguity set size for testing each DR-POMDP policy.

Figure 2.6 presents the statistics of the reward, including mean, standard deviation, 5-percentile and 95-percentile values, by implementing the DR-POMDP policies



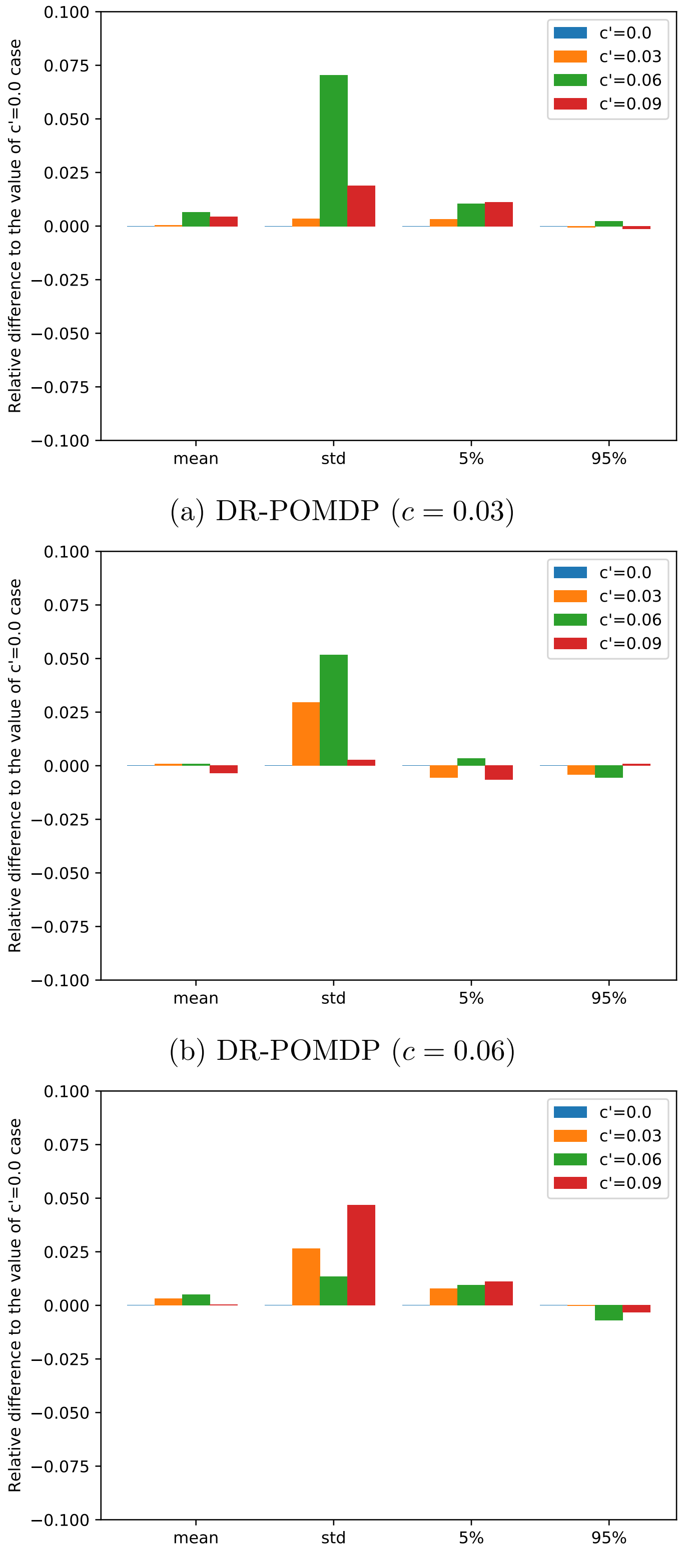
(a) {(Level 0, E), (Level 0, N)}

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

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

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

Figure 2.5: Value functions for increasing number of action-state pairs with distributional ambiguity. Solid line: lower bound, dashed line: upper bound. Corresponding actions: Level 0 – (red, horizontal), Level 1 – (blue, dot), Level 2 – (green, cross), Inspection – (white, diagonal)



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

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

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

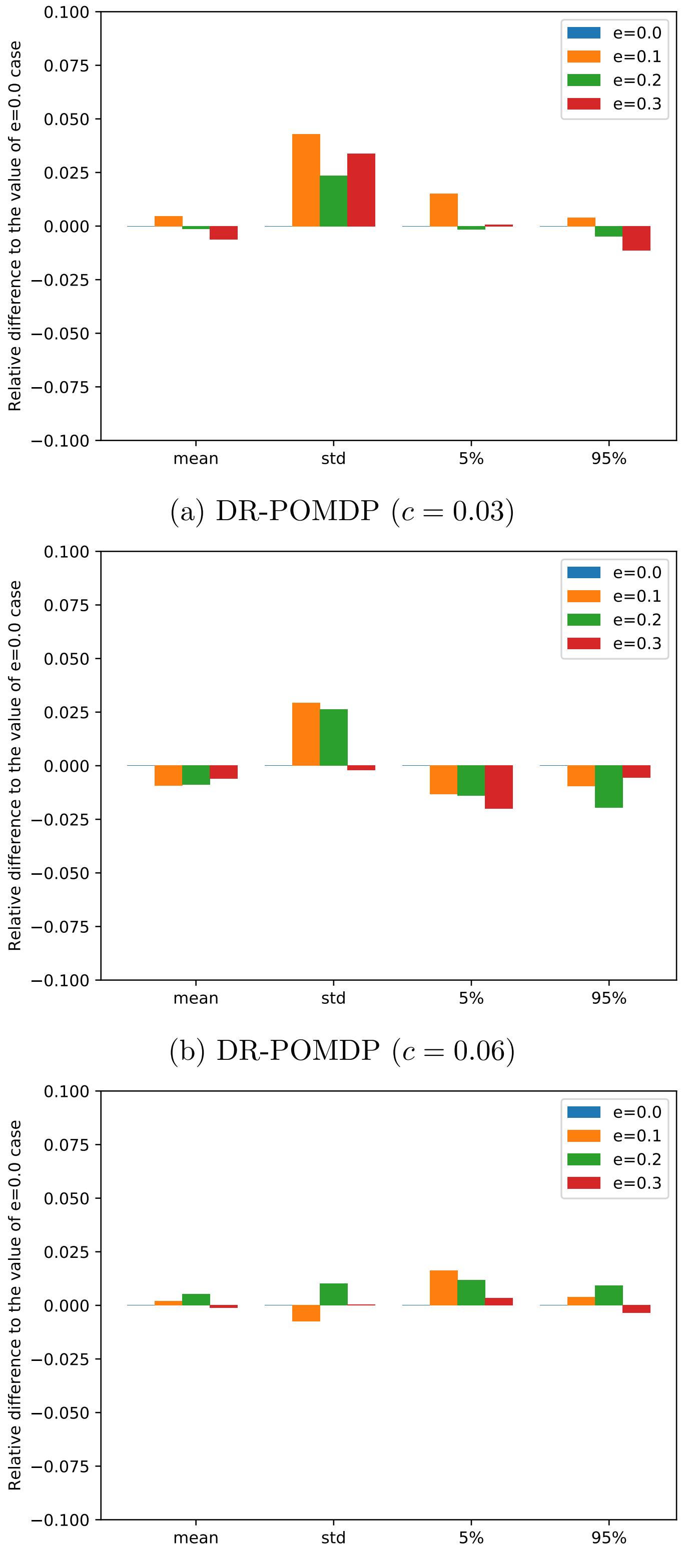
Figure 2.6: Statistics of the reward (mean, standard deviation, 5-percentile, 95-percentile) obtained by implementing DR-POMDP policies in in-sample tests under different ambiguity sets used by the nature.

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.

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

We argue that our assumption about the true transition-observation probabilities being accessible at the end of each decision period is relatively weak, by testing the DR-POMDP policies in out-of-sample scenarios while adding noise to the $\boldsymbol{p}$-value obtained at the end of each period. Specifically, 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 2.7 presents the statistics of the reward, including mean, standard deviation, 5-percentile and 95-percentile values, by implementing the DR-POMDP policies in out-of-sample scenarios under varying $\boldsymbol{p}$-values obtained at the end of each decision period. Similar to the previous section, we compare the reward statistics with the case when $e = 0.0$, i.e., the case when the DM can fully access the true $\boldsymbol{p}$-value at the end of each period. For different ambiguity sets ($c = 0.03, 0.06, 0.09$), the DR-POMDP solutions are not sensitive to the perturbation of $\boldsymbol{p}$-values obtained at the end of each period as we increase the noise. Moreover, all the statistics are within less than 2.5% differences from the results of $e = 0.0$, indicating that our assumption about the necessity of using side information to obtain the true $\boldsymbol{p}$-value at the end of each period is not strong.



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

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

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

Figure 2.7: Statistics of the reward (mean, standard deviation, 5-percentile, 95-percentile) obtained by performing DR-POMDP policies in out-of-sample tests with noisy $\boldsymbol{p}$-values.

2.6.2 Large-scale Dynamic Epidemic Control Problem

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

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

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

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

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

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

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

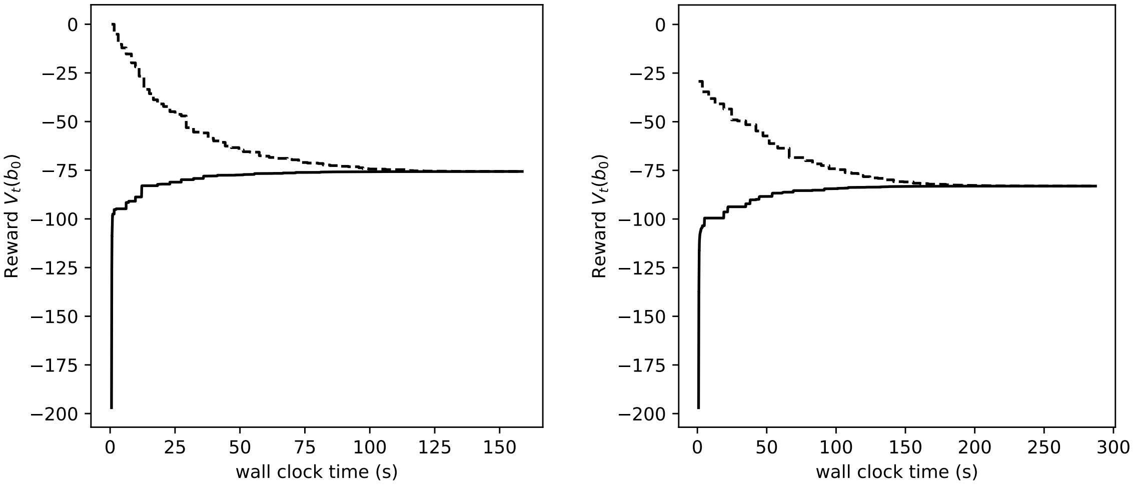
$$r_{as} = \begin{cases} -1000 \times \tilde{I} - 10 \times (1.0 - a_1) - 3 \times a_2, & \text{if } \hat{I} < 0.002 \\ -2500 \times \tilde{I} - 10 \times (1.0 - a_1) - 3 \times a_2, & \text{if } \hat{I} \geq 0.002, \end{cases}$$
$$+\, 20 \text{ if } \hat{I} \geq 0.005 \text{ and } a_1 \text{ is the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

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

2.6.2.1 Computational Time for Varying Numbers of States

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

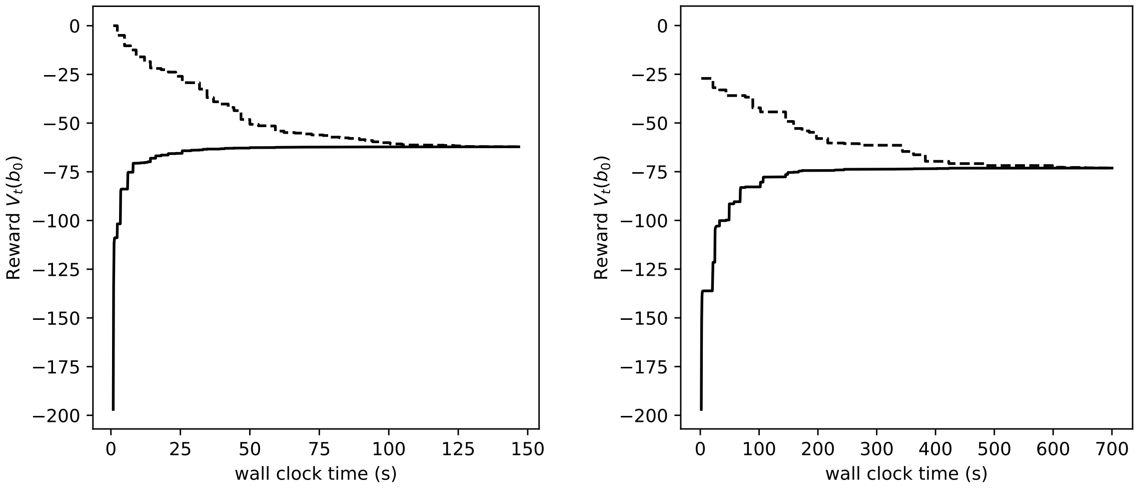
In the numerical experiment, we only consider ambiguities in the action $a_1 = 1.0$, corresponding to implementing the least strict control policy for reducing the infection rate. We set the radius of the ambiguity set as $c = 0.02$. Thus, the different problem sizes are $(s4, a4, z3, u8)$, $(s8, a4,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



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

Figure 2.8: Dynamic epidemic control problem instance $(s4, a4, z3, u8)$. Solid line: lower bound, dashed line: upper bound

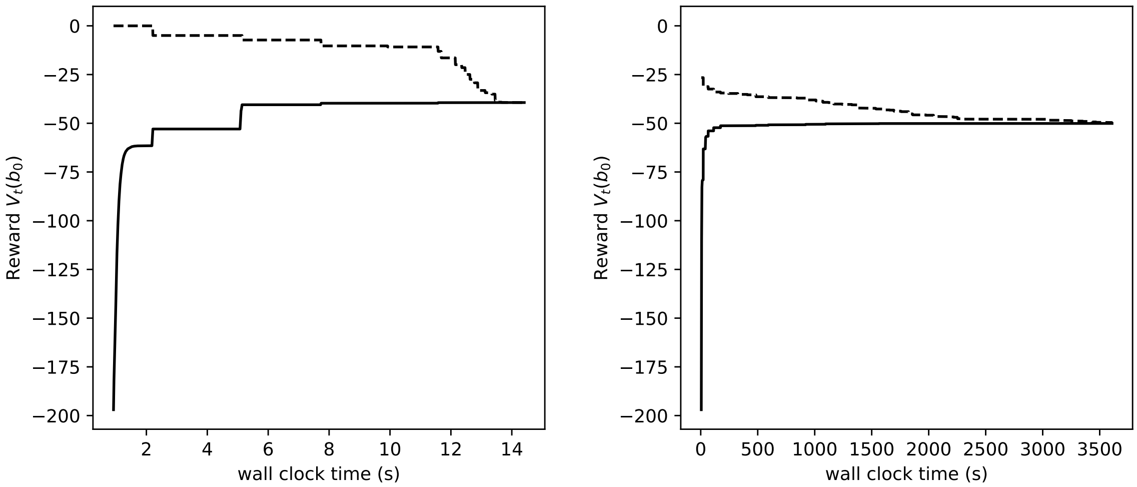
In Figures 2.8, 2.9, 2.10, we depict how the upper bound and lower bound of POMDP ($c = 0.00$) and DR-POMDP ($c = 0.02$) policies converge as functions of time for the above three problem sizes, respectively. We observe that the computational time for POMDP does not correlate with the number of states. When the number of states are 4 and 8, the corresponding instances take about 150 seconds to converge, as compared to the instances having 16 states take about 14 seconds to converge.



(a) POMDP ($c = 0.00$)

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

Figure 2.9: Dynamic epidemic control problem instance $(s8, a4, z3, u16)$. Solid line: lower bound, dashed line: upper bound



(a) POMDP ($c = 0.00$)

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

Figure 2.10: Dynamic epidemic control problem instance $(s16, a4, z3, u32)$. Solid line: lower bound, dashed line: upper bound

On the other hand, the computational time for DR-POMDP increases as the number of states and ambiguity sets increase. We also point out that the value function for DR-POMDP evaluated at b_0 is lower than that of POMDP, which is expected since DR-POMDP is more conservative.

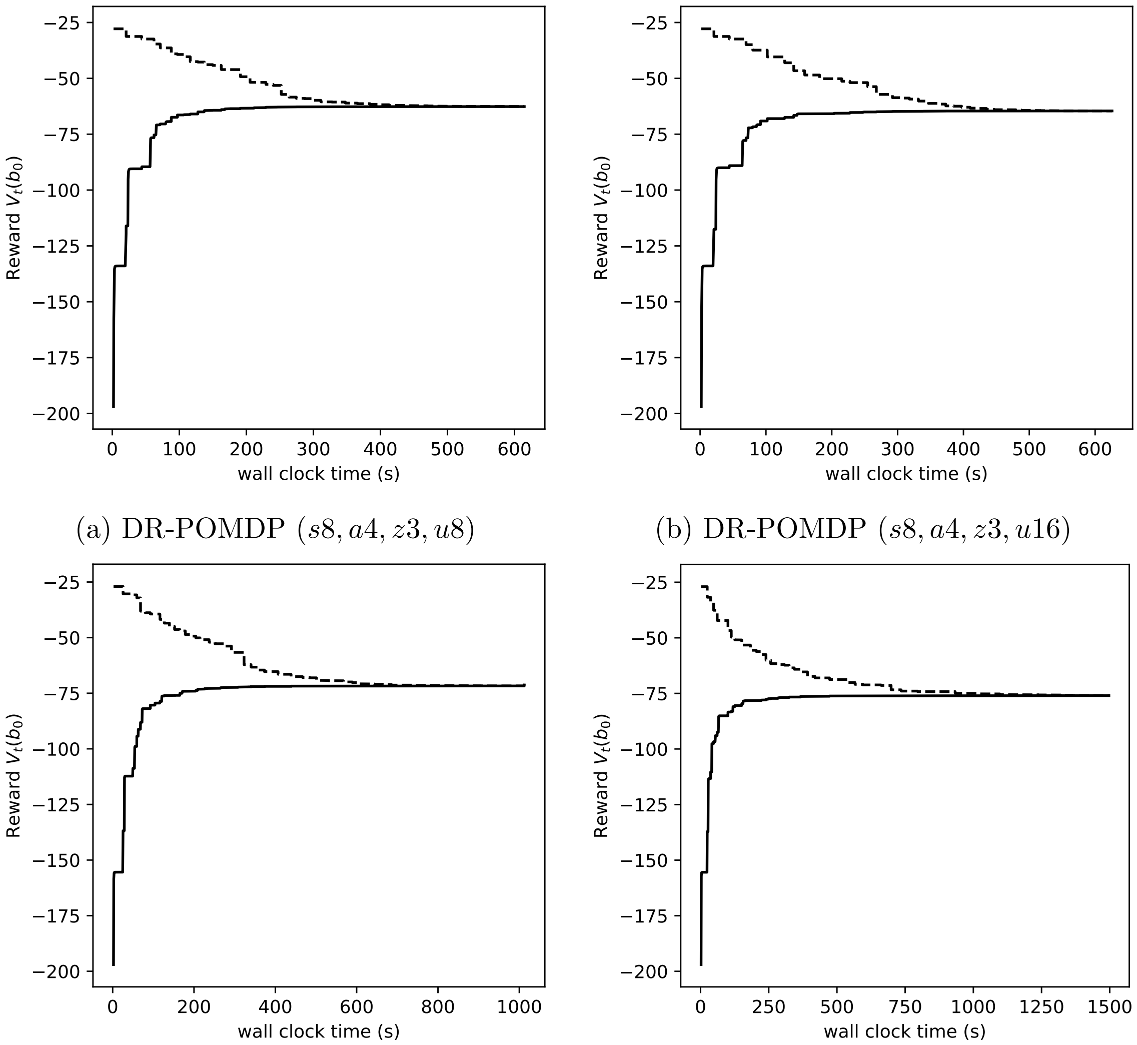
2.6.2.2 Computation Time for Varying Uncertainty Sizes

We change the number of ambiguity sets and compare their solutions and computation time. The states are $\tilde{S} \in \{0.50, 0.70, 0.90, 0.95\}$ and $\tilde{I} \in \{0.001, 0.005\}$, and actions are $(a_1, a_2) \in \{(0.1, 0.1), (0.1, 1.0), (1.0, 0.1), (1.0, 1.0)\}$. We increase the number of actions that are associated with ambiguity sets from 1 to 4. Since there are 8 states in total, the number of ambiguity sets are 8, 16, 32, and 64, respectively. The results are shown in Figure 2.11. The solution time are 614, 625, 1012, 1497 seconds, respectively and increase as the number of ambiguity sets increases. The optimal objective values are -62.64, -64.58, -71.71, -72.99, respectively, and decrease monotonically.

2.7 Concluding Remarks

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

In the future research, we aim to solve DR-POMDP when the outcomes of the



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

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

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

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

Figure 2.11: Dynamic epidemic control problem instances with varying number of ambiguity sets. Solid line: lower bound, dashed line: upper bound

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

CHAPTER III

Finite Sample Wasserstein Distance Bounds with an Application in Reinforcement Learning

3.1 Introductory Remarks

In this chapter, we discuss a theoretical development of the Wasserstein-based ambiguity set. As introduced in Section 1.3.2, a Wasserstein distance is defined as the minimum cost to transform one distribution to another, where the cost is determined by a distance measure between two events on a sample space. A Wasserstein ball is a set of distributions that are centered around a nominal distribution and having a Wasserstein distance bounded by a fixed quantity. When the nominal distribution is the sample distribution, we are interested in a bound which can guarantee that the true distribution is included in the Wasserstein ball with high probability. When the support of the random variable is continuous and the distance measure between the events are given by a norm, the bound which guarantees with probability at least $1-\delta$ is given by (1.29). However, this formulation involves constants c_1 and c_2 that cannot be easily estimated (*Ji and Lejeune*, 2018). In this chapter, we focus specifically on discrete distributions and derive Wasserstein distance bounds that can be computed with ease compared to the continuous case.

We use a Wasserstein-based formulation of regret minimization algorithm for reinforcement learning as an application of our result. In reinforcement learning, we assume that the transition probabilities for MDP are unknown, but the information will be collected throughout the process. The DM maintains an empirical distribution of the transition probabilities, based on the counts of transitions that occurred for each action at each state. We consider an ambiguity set surrounding the empirical distribution, and choose actions based on the distribution that will perform the best out of all the distributions in the ambiguity set. This is in contrast to the DRO where the worst-case distribution is considered. This is due to the balancing of exploration and exploitation of the uncertain transition probabilities, and thus taking the best-case distribution enables lowering the upper-confidence bound on the reward. In determining the policy, it is crucial to have a theoretical guarantee of the probability that the true distribution lies within the assumed ambiguity set. Our result on the theoretical bound of the Wasserstein distance is useful in this particular situation.

There is also an advantage in using the Wasserstein ball ambiguity set for reinforcement learning. While most approaches (e.g., *Jaksch et al.* (2010)) use ambiguity sets based on total variational distance, the Wasserstein-based ambiguity set is more general and can utilize the domain knowledge of the states. For example, if the states represent locations on a space such as grids, then it is more likely to transit to a state that is geometrically closer. Wasserstein distance is able to model certain penalties for moving the distribution to a geometrically distant location, so it is more suitable than total variational distance, where the distances between different pairs of states are uniformly distributed.

Our contribution in this chapter is twofold.

1. We derive concrete Wasserstein distance bounds between the true and empirical distributions with a probabilistic guarantee.

2. We apply the Wasserstein distance bounds to a reinforcement learning problem.

The rest of this chapter is organized as follows. In Section 3.2, we introduce the Wasserstein-ball ambiguity set and derive a finite confidence interval for Wasserstein distance. In Section 3.3, we introduce average-reward MDP and regret-based reinforcement learning. In Section 3.3.4, we provide a description of the algorithm and the performance guarantee. In Section 3.4, we demonstrate the computational performance of the Wasserstein-based ambiguity set using a simple numerical example of an ambulance dispatching problem. Finally, we provide concluding remarks in Section 3.5.

3.2 Wasserstein-based Ambiguity Set

In this section we will first introduce the preliminaries on Wasserstein ball ambiguity set. Then, we derive two types of Wasserstein distance bounds in Sections 3.2.2 and 3.2.3.

3.2.1 Preliminaries

We first discuss the formulation of the Wasserstein-ball ambiguity set. Consider discrete events $x \in \mathcal{X}$, where $\mathcal{X}$ is finite, and suppose that it takes cost $d(x, x')$ to move a unit of probability mass from event x to event x'. Then, the Wasserstein distance of order 1 from distribution p to distribution q is

$$\begin{aligned}
W(p,q) = \min \quad & \sum_{x\in\mathcal{X}} \sum_{x'\in\mathcal{X}} d(x,x')\kappa_{x,x'} & & \qquad (3.1\text{a}) \\
\text{s.t.} \quad & \sum_{x'\in\mathcal{X}} \kappa_{x,x'} = p(x), & \forall x \in \mathcal{X}, & \qquad (3.1\text{b}) \\
& \sum_{x\in\mathcal{X}} \kappa_{x,x'} = q(x'), & \forall x' \in \mathcal{X}, & \qquad (3.1\text{c}) \\
& \kappa_{x,x'} \geq 0, & \forall x, x' \in \mathcal{X}, & \qquad (3.1\text{d})
\end{aligned}$$

where κ can be interpreted as a joint distribution with marginals p and q as described in constraints (3.1b) and (3.1c).

Let us introduce the definition of empirical distribution for a multinomial distribution on a set of events $\mathcal{X}$.

Definition III.1. Let $\hat{x}_1, \ldots, \hat{x}_N$ be i.i.d. samples of random variables from a finite set $\mathcal{X}$ with true distribution p. We define the empirical distribution as

$$\hat{p}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{x}_i}, \tag{3.2}$$

where δ_x is a distribution which takes value 1 at $x \in \mathcal{X}$ and 0 otherwise.

The Wasserstein ball ambiguity set is defined as the set of all distributions having a Wasserstein distance that is less than or equal to θ from an empirical distribution $\hat{p}_N$. That is,

$$\mathcal{D}(\hat{p}_N, \theta) = \{p \in \Delta(\mathcal{X}) \mid W(p, \hat{p}_N) \leq \theta\}. \tag{3.3}$$

In the following sections, we obtain bounds for the probability where the Wasserstein distance between the true and the empirical distributions are less than or equal to θ. That is, we are interested in estimating the value

$$\mathbb{P}\left[p \in \mathcal{D}(\hat{p}_N, \theta)\right]. \tag{3.4}$$

This analysis can also be used in determining the value of θ with a fixed confidence level.

3.2.2 L_1 Distance Bound

We have the following relation between the Wasserstein distance and the weighted L_1 norm of the distribution.

Lemma III.2 (*Villani* (2008), Theorem 6.15)**.** *For some arbitrary $x_0 \in \mathcal{X}$, the Wasserstein distance between distributions p and q are bounded as follows:*

$$W(p,q) \leq \sum_{x \in \mathcal{S}} d(x_0, x)|p(x) - q(x)|, \tag{3.5}$$

where $d(x_0, x)$ is a cost to transport a unit probability mass from state x_0 to x.

Let us define $d^* = \min_{x \in \mathcal{X}} \max_{x' \in \mathcal{X}} d(x, x')$, which is a quantity corresponding to the smallest worst-case cost for transporting a unit probability mass from state x to x'. Lemma III.2 immediately leads to the relation with the L_1 distance of the probability measure

$$||p - q||_1 := \sum_{x \in \mathcal{X}} |p(x) - q(x)|. \tag{3.6}$$

Corollary III.3. *The Wasserstein distance between distributions p and q are bounded by*

$$W(p,q) \leq d^* ||p - q||_1. \tag{3.7}$$

We provide two probability bounds for the L_1 distance between the true and the empirical distributions. The first one is described in the following theorem.

Theorem III.4 (*Weissman et al.* (2003), Theorem 2.1)**.** *Let X be the cardinality of $\mathcal{X}$. The probability that the L_1 distance between the true distribution p and the empirical distribution $\hat{p}_N$ deviates more than θ is bounded by*

$$\mathbb{P}\left[||p - \hat{p}_N||_1 \geq \theta\right] \leq \left(2^X - 2\right) e^{-N\theta^2/2}. \tag{3.8}$$

This leads to the following remark:

Remark III.5. Using Corollary III.3 and Theorem III.4, the Wasserstein distance is bounded by

$$\theta = d^* \sqrt{\frac{2}{N} \log\left(\frac{2^X - 2}{\delta}\right)} \approx d^* \sqrt{\frac{2X}{N} \log\left(\frac{2}{\delta^{\frac{1}{X}}}\right)}, \tag{3.9}$$

with probability at least $1 - \delta$.

The second bound is one of our contributions in this chapter.

Theorem III.6. *The probability that the L_1 distance between the true distribution p and the empirical distribution $\hat{p}_N$ deviates more than θ is bounded by*

$$\mathbb{P}\left[||p - \hat{p}_N||_1 \geq \theta\right] \leq \left(\frac{eN\theta^2/2}{X-1}\right)^{X-1} e^{-N\theta^2/2} \tag{3.10}$$

The proof is immediate from the two lemmas we introduce below.

Lemma III.7 (Csiszar-Kullback-Pinsker inequality)**.** *The L_1 distance of distribution is bounded by the following:*

$$||p - q||_1 \leq \sqrt{2D(p||q)}, \tag{3.11}$$

where

$$D(p||q) := \begin{cases} \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}, & \textit{if } p \ll q \\ +\infty & \textit{otherwise} \end{cases}, \tag{3.12}$$

is a Kullback-Leibler divergence with $p \ll q$ indicating that p is absolutely continuous with respect to q.

Lemma III.8 (*Agrawal* (2020), Theorem I.2)**.** *The probability that the relative entropy between the true distribution p and the empirical distribution $\hat{p}_N$ deviates more* *than ϵ is bounded by*

$$\mathbb{P}\left[D(\hat{p}_N||p) \geq \epsilon\right] \leq \left(\frac{eN\epsilon}{X-1}\right)^{X-1} e^{-N\epsilon}, \tag{3.13}$$

if $\epsilon > \frac{X-1}{N}$.

We are rather interested in the value of θ that achieves the concentration inequality. The following corollary provides the criteria.

Corollary III.9. *Let $w^{-1}(y)$ be the inverse transformation of the fucntion*

$$w(z) = ze^{-z}, z \geq 1. \tag{3.14}$$

Define

$$\theta = \sqrt{\frac{2(X-1)}{N} w^{-1}\left(\frac{1}{e}\delta^{\frac{1}{X-1}}\right)}. \tag{3.15}$$

Then with probability at least $1-\delta$, the empirical distribution satisfies $||p-\hat{p}_N||_1 \leq \theta$.

Proof. From Lemma III.8, the following inequality holds with probability at least $1-\delta$.

$$D(\hat{p}_N||p) \leq \frac{(X-1)}{N} w^{-1}\left(\frac{1}{e}\delta^{\frac{1}{X-1}}\right). \tag{3.16}$$

Then, we substitute to (3.11), which provides a bound for $||p-\hat{p}_N||_1$. □

We conclude this section by providing the bound for the Wasserstein distance in the following remark.

Remark III.10. Define

$$\theta = d^* \sqrt{\frac{2(X-1)}{N} w^{-1}\left(\frac{1}{e}\delta^{\frac{1}{X-1}}\right)}. \tag{3.17}$$

Then the empirical distribution satisfies $W(p, \hat{p}_N) \leq \theta$ with probability at least $1 - \delta$.

Note, however, that this criterion ignores most of the information contained in d as only d^* is used, and bound can be very conservative.

3.2.3 Weighted L_1 Distance Bound

We formulate an alternative bound that includes information of d. First, we introduce the following lemma on the bounds of weighted L_1 distance.

Lemma III.11 (*Bolley and Villani* (2005), Theorem 2.1, Weighted Csiszar-Kullback-Pinsker inequality)**.** *For all $\alpha > 0$,*

$$\sum_{x \in \mathcal{X}} d(x_0, x)|p(x) - q(x)| \leq \sqrt{\frac{2}{\alpha}} \left(1 + \log \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0,x)} q(x) \right)^{1/2} \sqrt{D(p||q)}. \tag{3.18}$$

The following proposition gives a criterion for the bound θ.

Proposition III.12. *Let θ be a bound which satisfies*

$$\theta \geq \sqrt{\frac{1 + \log \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0,x)} p(x)}{\alpha}} \sqrt{\frac{2\,(X-1)}{N} w^{-1} \left(\frac{1}{e} \delta^{\frac{1}{X-1}} \right)}, \tag{3.19}$$

for some $x \in \mathcal{X}$ and $\alpha > 0$. Then with probability at least $1 - \delta$, the empirical distribution satisfies $W(p, \hat{p}_N) \leq \theta$.

Proof. For simplicity, we denote

$$K^q_{x_0,\alpha} = \log \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0,x)} q(x). \tag{3.20}$$

Combining Lemma III.2 and Lemma III.11,

$$W(p, \hat{p}_N) \leq \sqrt{\frac{2}{\alpha} \left(1 + K^p_{x_0,\alpha}\right) D(\hat{p}_N||p)}, \tag{3.21}$$

for some $x_0 \in \mathcal{X}$ and $\alpha > 0$.

Thus, by substituting (3.16), if

$$\theta \geq \sqrt{\frac{1 + K^p_{x_0,\alpha}}{\alpha}} \sqrt{\frac{2\,(X-1)}{N} w^{-1}\left(\frac{1}{e}\delta^{\frac{1}{X-1}}\right)}, \tag{3.22}$$

$W(p, \hat{p}_N) \leq \theta$ is satisfied with probability at least $1 - \delta$. □

Note, however, this bound involves a quantity $K^p_{x_0,\alpha}$ which we are not able to obtain. We would like to substitute it with an empirical $K^{\hat{p}_N}_{x_0,\alpha}$, but it is subject to some deviation from the true quantity. We, therefore, take a distributionally robust optimization approach to get the worst-case value of $K^p_{x_0,\alpha}$ to be conservative and then replace the RHS of (3.22) by a valid lower bound.

Let us introduce a set of distributions p characterized by the inverse-Kullback-Leibler divergence.

Definition III.13. A set of distributions p centered around the empirical distribution $\hat{p}_N$ with radius ϵ is defined as

$$\mathcal{D}_b(\hat{p}_N, \epsilon) := \{p \in \Delta(\mathcal{X}) \mid D(\hat{p}_N || p) \leq \epsilon\}\,. \tag{3.23}$$

For simplifying the notation, we define

$$\epsilon = \frac{X-1}{N} w^{-1}\left(\frac{1}{e}\delta^{\frac{1}{X-1}}\right). \tag{3.24}$$

Now, we present the main result of this chapter in the following theorem, which provides a value for θ with a probabilistic guarantee.

Theorem III.14. *For some $x \in \mathcal{X}$ and $\alpha > 0$, let*

$$\theta \geq \sqrt{\frac{1 + \sup_{p \in \mathcal{D}_b(\hat{p}_N, \epsilon)} \log \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0, x)} p(s)}{\alpha}} \sqrt{\frac{2\,(X-1)}{N} w^{-1} \left(\frac{1}{e} \delta^{\frac{1}{X-1}} \right)}. \tag{3.25}$$

Then with probability at least $1 - \delta$, the empirical distribution satisfies $W(p, \hat{p}_N) \leq \theta$.

Proof. If $D(\hat{p}_N || p) \leq \epsilon$ holds, then $K^p_{x_0,\alpha} \leq \sup_{p \in \mathcal{D}_b(\hat{p}_N, \epsilon)}$. Therefore, the inequality

$$\sqrt{\frac{2}{\alpha}\,(1 + K^p_{x_0,\alpha})\, D(\hat{p}_N || p)} \leq \sqrt{\frac{2}{\alpha} \left(1 + \sup_{p \in \mathcal{D}_b(\hat{p}_N, \epsilon)} K^p_{x_0,\alpha} \right) D(\hat{p}_N || p)} \tag{3.26}$$

holds. We obtain (3.25) by substituting the above inequality to (3.19). □

The first square root term of the RHS of (3.25) is dependent on x_0 and α, which we have the freedom to choose. We solve the following to lower the value of θ.

$$\inf_{x_0 \in \mathcal{X}, \alpha > 0} \sqrt{\frac{1 + \sup_{p \in \mathcal{D}_b(\hat{p}_N, \epsilon)} \log \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0, x)} p(x)}{\alpha}}. \tag{3.27}$$

We introduce the following theorem which describes the complexity of the optimization problem.

Theorem III.15. *The problem* (3.27) *can be solved in polynomial time.*

Proof. Notice that $K^p_{x_0,\alpha} = \log \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0, x)} p(x)$ is a convex function of α, since it can be interpreted as the cumulant generating function of random variable with realizations $d^2(x_0, x)$ with probability $p(x)$. Then, since taking the supremum for p over a convex set does not change the convexity of a convex function, so $\sup_{p \in \mathcal{D}_b(\hat{p}_N, \epsilon)} K^p_{x_0,\alpha}$ is convex. Now, let us suppose $f(z)$ is convex and consider $\sqrt{f(z)/z}$. This is not a transformation that preserves convexity, but the sublevel set

$$\left\{z \in \mathrm{dom} f \mid \sqrt{f(z)/z} \leq t\right\} \tag{3.28}$$

is convex for all $t \in \mathbb{R}$, indicating that $\sqrt{f(z)/z}$ is quasiconvex. Thus, the objective function of (3.27) is quasiconvex in α (see, e.g., *Boyd et al.* (2004)). Thus, a one-dimensional search algorithm such as golden section search is able to determine the optimal α in polynomial time.

The evaluation of the inner supremum can also be done in polynomial time. Notice that the optimal solution does not change after moving the supremum into the logarithm. That is, we solve

$$\begin{aligned}
\sup_{p} \quad & \sum_{x \in \mathcal{X}} e^{\alpha d^2(x_0,x)} p(x) && \text{(3.29a)}\\
\text{s.t.} \quad & \sum_{x \in \mathcal{X}} \hat{p}_N(s) \log \frac{\hat{p}_N(s)}{p(s)} \leq \epsilon && \text{(3.29b)}\\
& \sum_{x \in \mathcal{X}} p(s) = 1, \; p \in \mathbb{R}_+^{\mathcal{X}}, && \text{(3.29c)}
\end{aligned}$$

for a given α, which is a problem with a linear objective and a convex feasible region, which has a polynomial time complexity.

Finally, we perform this optimization for all $x_0 \in \mathcal{X}$, and choose the lowest value.

□

Rather than iterating over all $x_0 \in \mathcal{X}$ to obtain the optimal solution, we can formulate a heuristic algorithm by considering the following remark.

Remark III.16. For a given $x_0 \in \mathcal{X}$, the limit of the objective of (3.27) when $\alpha \to \infty$ is bounded by $\max_{x \in \mathcal{X}} d(x_0, x)$. Thus, by selecting $x_0 = \arg\min_{x \in \mathcal{X}} \max_{x' \in \mathcal{X}} d(x, x')$, we recover the bound in (3.17), since

$$\lim_{\alpha\to\infty} \sqrt{\frac{1 + \sup\limits_{p\in\mathcal{D}_b(\hat{p}_N,\epsilon)} \log \sum\limits_{x\in\mathcal{X}} e^{\alpha d^{*2}} p(s)}{\alpha}} = \lim_{\alpha\to\infty} \sqrt{\frac{1+\alpha d^{*2}}{\alpha}} = d^{*}. \tag{3.30}$$

3.3 Applications in Regret-based Reinforcement Learning

In Sections 3.3.1–3.3.3, we introduce some preliminaries for regret-based reinforcement learning. Then, in Section 3.3.4, we present the regret bound for the case where the Wasserstein-ball ambiguity set is used.

3.3.1 Average Reward Markov Decision Processes

Recall that $\mathcal{S}$ is a set of states, $\mathcal{A}$ is a set of actions, and $r(s,a)$ is a reward for taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$. Let S_t and A_t be a random state and action values at time t. We are interested in finding a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ such that it maximizes the average reward given by

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^{\pi} \left[r(S_t, A_t) \mid S_1 = s \right], \tag{3.31}$$

which is the time average reward in the long run initializing from state s. The optimal gain is the optimal objective value with respect the policy and the initial state, defined as

$$\rho^* := \max_{s\in\mathcal{S}} \sup_{\pi} \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}^{\pi} \left[r(S_t, A_t) \mid S_1 = s \right]. \tag{3.32}$$

Throughout this chapter, we assume that the MDP is strongly connected, i.e., for any pairs of states s and s', there exists a policy such that the probability for reaching s' from s eventually is nonzero. This guarantees the existence of the optimal solution in (3.32). The connectivity of the MDP can also be described by the diameter. Let us denote the MDP as $M = (\mathcal{S}, \mathcal{A}, p, r)$. The diameter of the MDP is defined as

$$D(M) = \max_{s \neq s'} \min_{\pi \in \Pi_{DM}} \mathbb{E}^{\pi} \left[\min\{t \geq 1 : S_t = s'\} \mid S_1 = s\right] - 1, \tag{3.33}$$

where Π_{DM} is a set of all policies that are deterministic and memory-less. The definition comes from the minimum expected amount of time it takes to reach between the worst combination of states s and s'. For strongly connected MDP, $D(M)$ is finite.

When the transition probabilities $p(s'|s,a)$ are known, the optimal policy can be found by solving the Bellman optimality equations

$$\rho + v(s) = \max_{a \in \mathcal{A}} \left(r(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a)v(s') \right), \forall s \in \mathcal{S}. \tag{3.34}$$

Here, the optimal solution of ρ is the optimal objective value of (3.32), and $v(s)$ is the differential value function indicating the relative advantage of the starting state s. The Bellman equation (3.34) can be reformulated as a linear program below.

$$\min \quad \rho \tag{3.35a}$$

$$\text{s.t.} \quad \rho + v(s) \geq r(s,a) + \sum_{s' \in \mathcal{S}} p(s'|s,a)v(s'), \qquad \forall s \in \mathcal{S},\ a \in \mathcal{A}. \tag{3.35b}$$

The optimal policy is gained by taking the optimal solutions of the dual variables associated with constraints (3.35b), and we denote the dual variables as $\pi(s,a)$. For each $s \in \mathcal{S}$, any $a \in \mathcal{A}$ having $\pi(s,a) > 0$ is optimal. If such action $a \in \mathcal{A}$ does not exist, then any action $a \in \mathcal{A}$ is optimal.

3.3.2 Optimism in the Face of Uncertainty

For notational convenience, we denote $p(\cdot|s,a)$ as p_{sa}. We consider a case where the transition probabilities p_{sa} are not fully known, and characterize a policy of optimism in the face of uncertainty (*Tewari and Bartlett*, 2008; *Jaksch et al.*, 2010). Here, we let $\mathcal{P}_{sa}(\hat{p}_{sa}, \theta_{sa})$ be an ambiguity set constructed from sample average $\hat{p}_{sa}$, and a distance (or divergence) measure θ_{sa}. The optimistic policy assumes that it takes a best-case distribution of p_{sa} out of all possible distributions in $\mathcal{P}_{sa}$.

For each state $s \in \mathcal{S}$, the corresponding distributionally optimistic Bellman equation is formulated as

$$v(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \max_{p_{sa} \in \mathcal{P}_{sa}(\hat{p}_{sa}, \theta_{sa})} \sum_{s' \in \mathcal{S}} p_{sa}(s') v(s') \right\}. \tag{3.36}$$

In *Jaksch et al.* (2010), the ambiguity set $\mathcal{P}_{sa}$ is based on a total variational distance, and in *Filippi et al.* (2010), it is based on a Kullback-Leibler divergence. In the chapter, we formulate the case when the ambiguity set is Wasserstein-based.

3.3.3 Regret-based Reinforcement Learning

A cumulative regret is a difference between the cumulative reward that is obtained and the cumulative reward that would have been obtained if the DM knew all the parameters of the MDP. At time T, this can be expressed as

$$T\rho^* - \sum_{t=1}^{T} r(S_t, A_t), \tag{3.37}$$

and the goal is to find a policy that minimizes the bound of regret.

Tewari and Bartlett (2008) consider a generalization of index policies using an optimistic linear programming algorithm, and achieve a regret bound that is asymptotically logarithmic in T steps. However, the bound is also known to be exponential in the number of states. *Jaksch et al.* (2010) solve this by proposing an algorithm UCRL2, showing that the upper bound of the regret is $\tilde{O}(D|\mathcal{S}|\sqrt{|\mathcal{A}|T})$, where D is the diameter of MDP. *Jaksch et al.* (2010) also prove that the lower bound of the regret is $\Omega(\sqrt{D|\mathcal{S}||\mathcal{A}|T})$. *Azar et al.* (2017) propose an algorithm with upper bound $\Omega(\sqrt{H|\mathcal{S}||\mathcal{A}|T})$ for a finite horizon MDP, where the MDP is repeated over again whenever the horizon H is reached.

In contrast to the theoretical work done in the literature above, we are interested in a problem assuming that some knowledge of the system is known as a form of transportation cost d.

3.3.4 Algorithm for Reinforcement Learning with Wasserstein Ball Ambiguity Set

We extend the UCRL2 algorithm in *Jaksch et al.* (2010); *Lattimore and Szepesvári* (2020). The empirical distribution at step t is given by

$$\hat{p}^t_{sa}(s') = \frac{\sum_{u=1}^{t} \mathbb{I}\left(S_u = s, A_u = a, S_{u+1} = s'\right)}{\max\left\{1, N^t_{sa}\right\}}, \tag{3.38}$$

where $N^t_{sa} = \sum_{u=1}^{n} \mathbb{I}\left(S_u = s, A_u = a\right)$ is the count of the realization of the state-action pair (s, a).

We define the sets of transition probabilities for each state-action pair (s, a) as

$$\mathcal{C}^t_{sa} = \left\{p_{sa} \in \Delta(\mathcal{S}) \mid W(\hat{p}_{sa}, p_{sa}) \leq \min\{\theta^{1t}_{sa}, \theta^{2t}_{sa}\}\right\}, \tag{3.39}$$

where

$$\theta^{1t}_{sa} = d^* \sqrt{\frac{2}{\max\left\{1, N^t_{sa}\right\}} \log\left(\frac{2^{|\mathcal{S}|} 15 |\mathcal{S}||\mathcal{A}| t^7}{\delta}\right)}, \tag{3.40}$$

and

$$\theta^{2t}_{sa} = \sqrt{\frac{1 + \sup\limits_{p \in \mathcal{D}_b(\hat{p}^t_{sa}, \epsilon)} \log \sum\limits_{s \in \mathcal{S}} e^{\alpha^\star d^2(s^\star_0, s)} p(s)}{\alpha^\star}} \sqrt{\frac{2\left(|\mathcal{S}| - 1\right)}{\max\left\{1, N^t_{sa}\right\}} w^{-1}\left(\frac{1}{e}\left(\frac{\delta}{15|\mathcal{S}||\mathcal{A}|t^7}\right)^{\frac{1}{|\mathcal{S}|-1}}\right)}. \tag{3.41}$$

Lemma III.17. *With probability at least* $1 - \frac{\delta}{15t^6}$*, the true MDP satisfies* $p_{sa} \in \mathcal{C}^t_{sa}$

for all state-action pair (s, a) up to stage t.

We dissect the time steps into episodes where the next episode begins when a visit to a state-action pair (s, a) doubles. That is, we define the beginning of the first episode as $\tau_1 = 1$, and the beginning of the $(k+1)$th episode as

$$\tau_{k+1} = 1 + \min\left\{t : \; N^t_{S_t A_t} \geq 2N^{\tau_k - 1}_{S_t A_t}\right\}. \tag{3.42}$$

At each episode, we update the policy. Let $\mathcal{M}_k$ be the set of plausible MDP at episode k. If the true MDP M is in $\mathcal{M}_k$, the optimistic solution $\tilde{\rho}_k$ is greater than or equal to ρ^*. Furthermore, The cumulative regret is bounded by

$$\hat{R}_T = \sum_{t=1}^{T} (\rho^\star - r(S_t, A_t)) \leq \sum_{k=1}^{K} \sum_{t \in E_k} (\tilde{\rho}_k - r(S_t, A_t)). \tag{3.43}$$

For all k, we have

$$\tilde{\rho}_k = r(S_t, A_t) - v_k(S_t) + \sum_{s' \in \mathcal{S}} \tilde{p}^k_{S_t A_t}(s') v_k(s'), \;\; \forall t \in E_k, \tag{3.44}$$

where $\tilde{p}^k_{sa}$ is the optimistic distribution at episode k. Thus, the regret for any single episode is bounded by

$$\begin{aligned}
\hat{R}_T &\leq \sum_{k=1}^{K} \sum_{t \in E_k} (\tilde{\rho}_k - r(S_t, A_t)) \\
&= \sum_{k=1}^{K} \sum_{t \in E_k} \left(-v_k(S_t) + \sum_{s' \in \mathcal{S}} \tilde{p}^k_{S_t A_t}(s') v_k(s') \right) \\
&= \sum_{k=1}^{K} \sum_{t \in E_k} \left(-v_k(S_t) + \sum_{s' \in \mathcal{S}} p_{S_t A_t}(s') v_k(s') \right) \\
&\qquad + \sum_{k=1}^{K} \sum_{t \in E_k} \sum_{s' \in \mathcal{S}} \left(\tilde{p}^k_{S_t A_t}(s') - p_{S_t A_t}(s') \right) v_k(s').
\end{aligned} \tag{3.45}$$

The first term of the RHS of the last equality can be bounded by

$$\sum_{k=1}^{K}\left(D(M)+\sum_{t\in E_k}\left(\mathbb{E}\left[v_k(S_{t+1})|S_t\right]-v_k(S_{t+1})\right)\right), \tag{3.46}$$

where we have used the fact that $v(s')-v(s)\leq D(M)$ for all $s,s'\in\mathcal{S}$. Using Azuma-Hoeffding inequality, *Jaksch et al.* (2010) showed that (3.46) can be bounded by

$$KD(M)+D(M)\sqrt{T\frac{5}{2}\log\left(\frac{8T}{\delta}\right)} \tag{3.47}$$

with probability at least $1-\frac{\delta}{12T^{5/4}}$. Furthermore, the number of episodes K can be bounded by $|\mathcal{S}||\mathcal{A}|\log_2\left(\frac{8T}{|\mathcal{S}||\mathcal{A}|}\right)$, and therefore the first term of (3.45) can be bounded by

$$D(M)\sqrt{T\frac{5}{2}\log\left(\frac{8T}{\delta}\right)}+D(M)|\mathcal{S}||\mathcal{A}|\log_2\left(\frac{8T}{|\mathcal{S}||\mathcal{A}|}\right), \tag{3.48}$$

with probability at least $1-\frac{\delta}{12T^{5/4}}$.

The second term of the RHS of (3.45) can be bounded by

$$\sum_{k=1}^{K}\frac{D(M)}{2}\sum_{t\in E_k}\sum_{s'\in\mathcal{S}}\left|\tilde{p}^{k}_{S_tA_t}(s')-p_{S_tA_t}(s')\right|, \tag{3.49}$$

which is bounded by

$$\begin{aligned}
&\sum_{k=1}^{K}\frac{D(M)}{2}\sum_{t\in E_k}\sqrt{\frac{2}{\max\left\{1,N^{\tau_k-1}_{S_tA_t}\right\}}\log\left(\frac{2^{|\mathcal{S}|}15|\mathcal{S}||\mathcal{A}|t^7}{\delta}\right)}\\
&\leq\sum_{k=1}^{K}\frac{D(M)}{2}\sqrt{2\log\left(\frac{2^{|\mathcal{S}|}15|\mathcal{S}||\mathcal{A}|T^7}{\delta}\right)}\sum_{t\in E_k}\sqrt{\frac{1}{\max\left\{1,N^{\tau_k-1}_{S_tA_t}\right\}}}\\
&=\sum_{k=1}^{K}\frac{D(M)}{2}\sqrt{2\log\left(\frac{2^{|\mathcal{S}|}15|\mathcal{S}||\mathcal{A}|T^7}{\delta}\right)}\sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\frac{N_{sa}(k)}{\sqrt{\max\left\{1,N^{\tau_k-1}_{sa}\right\}}},
\end{aligned} \tag{3.50}$$

where $N_{sa}(k)$ is the total count of state-action pair (s, a) at episode k. The first inequality is by substituting t with T, and the second equality is due to the definition of $N_{sa}(k)$. We further use the inequality

$$2 \log \left(\frac{2^{|\mathcal{S}|} 15 |\mathcal{S}||\mathcal{A}| T^7}{\delta} \right) \leq 14 |\mathcal{S}| \log \left(\frac{2|\mathcal{A}|T}{\delta} \right), \tag{3.51}$$

and

$$\begin{aligned} \sum_{k=1}^{K} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \frac{N_{sa}(k)}{\sqrt{\max\left\{1, N_{sa}^{\tau_k - 1}\right\}}} &\leq (\sqrt{2} + 1) \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \sqrt{N_{sa}^{T}} \\ &\leq (\sqrt{2} + 1) \sqrt{|\mathcal{S}||\mathcal{A}|T}. \end{aligned} \tag{3.52}$$

Therefore, the second term of (3.45) can be bounded by

$$\frac{D(M)}{2} \sqrt{14 |\mathcal{S}| \log \left(\frac{2|\mathcal{A}|T}{\delta} \right)} (\sqrt{2} + 1) \sqrt{|\mathcal{S}||\mathcal{A}|T} \tag{3.53}$$

Combining (3.48) and (3.53), the regret is bounded by

$$\begin{aligned} \hat{R}^T \leq & D(M) \sqrt{T \frac{5}{2} \log \left(\frac{8T}{\delta} \right)} + D(M) |\mathcal{S}||\mathcal{A}| \log_2 \left(\frac{8T}{|\mathcal{S}||\mathcal{A}|} \right) \\ & + \frac{D(M)}{2} \sqrt{14 |\mathcal{S}| \log \left(\frac{2|\mathcal{A}|T}{\delta} \right)} (\sqrt{2} + 1) \sqrt{|\mathcal{S}||\mathcal{A}|T}, \end{aligned} \tag{3.54}$$

with probability at least $1 - \frac{\delta}{6T^{5/4}} \geq 1 - \delta$. The upper bound of the cumulative regret can be simplified to $34 D(M) |\mathcal{S}| \sqrt{|\mathcal{A}| T \log \left(\frac{T}{\delta} \right)}$.

3.4 Computational Study

3.4.1 Ambulance Dispatching Problem

We consider a reinforcement learning variant of the optimal ambulance dispatching problem introduced in *Jagtenberg et al.* (2017). In this problem, there exists a set $\mathcal{V}$ of demand locations, and the task is to dispatch an ambulance from a set $\mathcal{B}$ of ambulances. Incidents occur at demand locations according to a Poisson distribution with a certain rate. When an ambulance arrives at the incident location, it takes a random amount of time to provide a service and decide whether a patient needs to be taken to a hospital. If the patient does not require immediate care, the ambulance becomes idle. Otherwise, the ambulance drives to the nearest hospital from $\mathcal{H} \subset \mathcal{V}$, and takes a random amount of time to serve at the hospital before becoming idle. Here, we assume that τ_{ij}, which is the time it takes to drive between $i, j \in \mathcal{V}$ is deterministic and known. The objective of this problem is to minimize the average response time and serve as many incidents as possible.

The state s is represented as a tuple

$$\left(Loc_{acc}, idle_1, \ldots, idle_{|\mathcal{B}|}\right), \tag{3.55}$$

where $Loc_{acc} \in \mathcal{V} \cup \{0\}$ represents a location of an incident that occurred in the previous time steps. The location 0 is a dummy node indicating that no incidents have ocurred. The element $idle_i \in \{true, false\}$ represents whether an ambulance i is idle or not. We denote $Loc_{acc}(s)$ and $idle_i(s)$ to indicate the value of the corresponding element for a given state s.

The set of actions is given by $\mathcal{A} = \mathcal{B} \cup \{0\}$, which is either dispatching an ambulance $a \in \mathcal{B}$, or a dummy action 0 indicating that no action is taken. There are certain restrictions for an action to be taken. For example, an ambulance that is not idle cannot be dispatched, an idle ambulance must be dispatched if an incident occurred, and actions cannot be taken if there are no incidents.

The transition probabilities $p_{sa}(s')$ are formulated as a product of two probabilities $p^1(s')$ and $p^2_{sa}(s')$, which are probabilities that an incident occurs at $Loc_{acc}(s')$, and probabilities that certain ambulances become available. Here, we have made a Markovian assumption such that the rate at which the incident occurs is independent of the previous realizations, and incidents that are not served immediately are lost.

We let the overall incidence rate be λ_c, and define a fraction d_v for all $v \in \mathcal{V}$. Then,

$$p^1(s') = \begin{cases} \lambda_c d_{Loc_{acc}(s')} & \text{if } Loc_{acc}(s') \in \mathcal{V} \\ 1 - \lambda_c & \text{otherwise.} \end{cases} \tag{3.56}$$

For tractability, we assume that the ambulances become idle following a geometric distribution with a fixed parameter $\frac{1}{r_c}$. This reflects the random travel time and the random service time averaged across all ambulances and the incidents. We also assume that the busy ambulances don't immediately become idle. In this case, $p^2_{sa}(s')$ becomes

$$p^2_{sa}(s') = \prod_{i \in \mathcal{B}} p^c_a(idle_i(s), idle_i(s')), \tag{3.57}$$

where if ambulance i is idle and the decision is to dispatch ambulance i, then

$$p^c_a(idle_i(s), idle_i(s')) = \begin{cases} 1 & \text{if } idle_i(s') = false \\ 0 & \text{if } idle_i(s') = true \end{cases}. \tag{3.58}$$

If ambulance i is busy, then

$$p_a^c(idle_i(s), idle_i(s')) = \begin{cases} r_c & \text{if } idle_i(s') = true \\ 1 - r_c & \text{if } idle_i(s') = false \end{cases}. \tag{3.59}$$

A cost is generated when there is an incident but there are no idle ambulances, or the ambulances are dispatched but the travel time is long. We convert this to a reward maximization problem and normalize it so that reward 1 is gained when there are no incidents, and a reward $1 - \tau_{ij}/M$ is gained for dispatching ambulance at node i to incident and node j. M is a normalization term indicating the worst-case travel time.

The major between the original formulation in *Jagtenberg et al.* (2017) and the reinforcement learning formulation is that the rates at which the incidents occur are unknown. However, we can postulate that the rates are correlated geographically: i.e., the closer demand locations have similar incident rates, possibly due to the amount of traffic, age distributions, etc. Because of the construct of the states, we are also familiar with the neighboring relations of the states. For example, the probability that all the busy ambulances become idle at the same time is low. We can incorporate this background knowledge to distance measure between the states. Under this setting, we can justify the use of a Wasserstein-based ambiguity.

3.4.2 Experimental Design and Setup

We scatter 10 incident locations in a unit square plane, of which 2 of them are also hospital locations. We plot the configuration in Fig. 3.1, where the orange dots represent hospital locations. We assign one ambulance to each hospital, making it a 44 state and 3 action MDP. The probability that an incident occurs is 25%, and the rate for each location are distributed unevenly in Fig. 3.1. The rate is higher for the locations that are close to the lower-left corner of the square. The probability that the ambulance becomes idle is 20%.

The distance between two locations is given by the Manhattan distance, i.e., the $L1$ norm. We use the Manhattan distance as the distance of the states and multiply with a constant term 0.1 when the transition probability is nonzero.

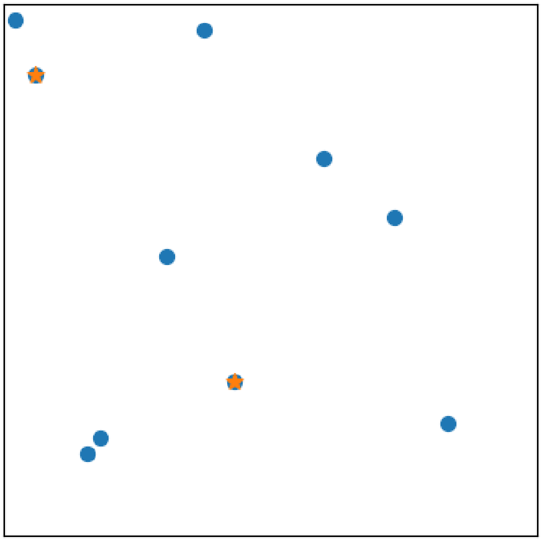


Figure 3.1: Incident locations (blue: incident locations, orange: hospitals and incident locations)

3.4.3 Computational Results

3.4.3.1 Regret

We tested three reinforcement learning algorithms over 1,000,000 steps. The cumulative regret is shown in Fig. 3.2 and the average regret is shown in 3.3. The blue line corresponds to the using (3.9) as the distance bound and the orange line corresponds to using (3.25) as the distance bound. The green line corresponds to the conventional UCRL2 algorithm (*Jaksch et al.*, 2010) using a total variational distance.

We note that the performance of Wasserstein bound (3.25) outperforms the bound

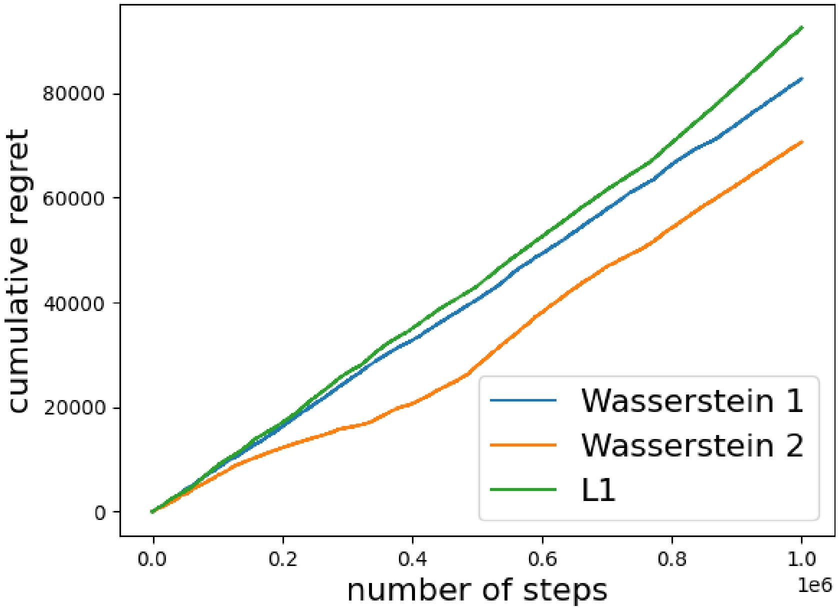


Figure 3.2: Cumulative regret

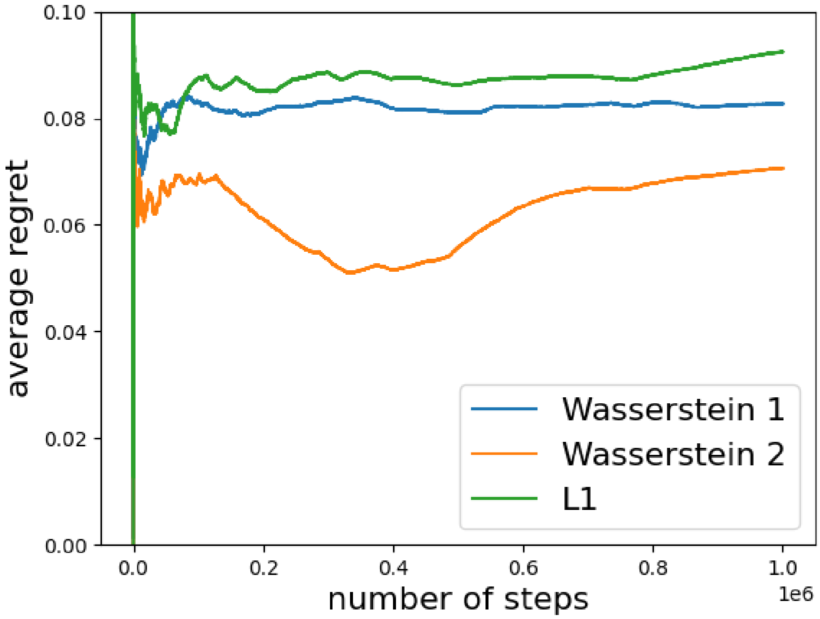


Figure 3.3: Average regret

(3.9) and the total variational distance, indicating the advantage of using all the information of the state distances and Wasserstein ball ambiguity set.

3.4.3.2 Bounds θ

We plot how the Wasserstein distance bounds change as the number of samples increases. The bounds are compared between (3.9), (3.25), and the case where the bound (3.25) is used, but not optimized over α. We find that the optimization over α is necessary to obtain a stronger Wasserstein distance bound.

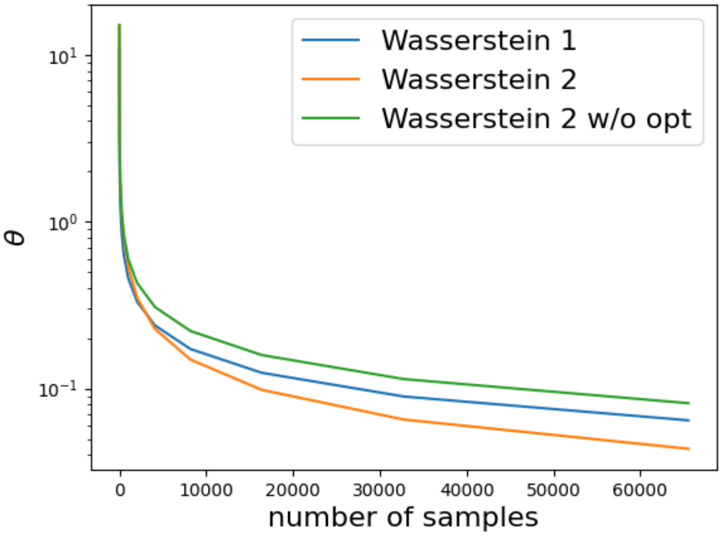


Figure 3.4: Bounds

3.5 Concluding Remarks

In this chapter, we derived concrete Wasserstein distance bounds for true and empirical distributions when the set of events are finite. We then applied the result in a reinforcement learning application, where the notion of optimism in the face of uncertainty matches the concept of ambiguity sets. In the future, we will improve the computational efficiency of the algorithm as problems with similar structures are being solved repeatedly as the information is gained over time.

CHAPTER IV

Multistage Distributionally Robust Mixed-integer Programming under the Wasserstein Ambiguity Set

4.1 Introductory Remarks

Multistage stochastic programming extends the two-stage stochastic programming formulation where there are more than two sequences of decisions to be made. The sequences of realized random variables are expressed using a scenario tree which increases exponentially in size as the number of stages increases. The basic approach for solving multistage stochastic programming is nested Benders decomposition, which extends the Benders decomposition algorithm used in two-stage stochastic programming (*Gassmann*, 1990; *Birge and Louveaux*, 2011). It begins with a relaxed formulation and alternates between the forward pass which chooses a sample path in the scenario tree and the backward pass which generates valid cuts to strengthen the relaxation. Stochastic dual dynamic programming (SDDP) (*Pereira and Pinto*, 1991) further assumes that the scenarios are stage-wise independent, allowing a more efficient algorithm where the generated cuts can be shared across different sample paths having the common future realization of uncertain variables. However, these two methods are only applicable for cases where the variables are continuous. Recently, *Zou et al.* (2019) propose stochastic dual dynamic integer programming (SDDiP) which solves problems that have binary state variables by utilizing a stronger set of cuts derived from a particular reformulation of the problem. Meanwhile, *Philpott et al.* (2018) and *Duque and Morton* (2020) consider a distributionally robust variant of SDDiP, where *Philpott et al.* (2018) assume ambiguity sets based on χ^2 distance centered around a nominal distribution, and *Duque and Morton* (2020) assume ambiguity sets based on Wasserstein distance. Furthermore, *Yu and Shen* (2020) extend distributionally robust SDDiP to cases where the random variables are endogenous, i.e., dependent on the previous decisions, using moment-based ambiguity sets.

In this chapter, we discuss a dual decomposition approach to multistage distributionally robust programming. *Carøe and Schultz* (1999) develop a dual decomposition formulation for two-stage stochastic programming by taking a Lagrangian relaxation of the non-anticipativity constraints. The main advantage of this approach is that it is able to handle mixed-integer variables and the subproblems can be solved in parallel. Recently, *Kim* (2020) apply dual decomposition method to two-stage distributionally robust mixed-integer programming. This chapter extends the dual decomposition techniques in *Kim* (2020) to the case of multistage stochastic programming, and further implement a branch-and-bound algorithm to obtain an optimal solution.

In Section 4.2, we introduce the notations and the problem formulation of the multistage distributionally robust program. In Section 4.3, we present the deterministic equivalent formulation of multistage distributionally robust program which is used to derive the Lagrange dual formulation in Section 4.4. We present the algorithmic formulation in Section 4.5. In Section 4.6, we discuss the application of dual decomposition algorithm to a transmission expansion problem. Finally, we note our concluding remarks in Section 4.7.

4.2 Preliminaries

4.2.1 Notations

We let $[N]$ be a set of integers $\{1, \dots, N\}$. Consider arbitrary sets Ξ_k indexed by $k \in [K]$, where each elements are denoted by ξ_k. For indices $1 \le i \le j \le K$, we define $\Xi_{i:j}$ as $\Xi_i \times \Xi_{i+1} \times \cdots \times \Xi_j$. We denote the elements of $\Xi_{i:j}$ by $\xi_{i:j}$, which is equivalent to $(\xi_i, \xi_{i+1}, \dots, \xi_j)$. When $i > j$, we define $\Xi_{i:j} := \emptyset$.

4.2.2 Wasserstein Ambiguity Set

Define a set of probability measures on support Ξ as

$$\mathcal{M}(\Xi) := \left\{ P \in \mathbb{M} : \int\limits_{\Xi} dP(\xi) = 1 \right\}, \tag{4.1}$$

where $\mathbb{M}$ is a set of all nonnegative measures $P : \Xi \to \mathbb{R}_+$. Let $\{\hat{\xi}^1, \dots, \hat{\xi}^N\}$ be the set of empirical observations with probability estimates $\hat{p}^1, \dots, \hat{p}^N$, where $\hat{p}^s > 0$ for all $s \in [N]$. The Wasserstein ball ambiguity set is

$$\mathcal{P} := \left\{ P \in \mathcal{M}(\Xi) : \begin{array}{l} \int_\Xi \sum_{s=1}^N u^s(\xi) ||\hat{\xi}^s - \xi|| d\xi \le \epsilon, \\ \int_\Xi u^s(\xi) d\xi = \hat{p}^s, \ \forall s \in [N], \\ \sum_{s=1}^N u^s(\xi) = P(\xi), \ \forall \xi \in \Xi, \\ u^s(\xi) \ge 0, \ \forall \xi \in \Xi, \ s \in [N] \end{array} \right\}, \tag{4.2}$$

where ϵ is the Wasserstein distance limit.

4.2.3 Problem Statement

In this section, we introduce several models of distributionally robust multistage mixed-integer program. For notational simplicity, we assume that the first stage is subject to a deterministic variable $\xi_1 \in \Xi_1$, where the cardinality of Ξ_1 is 1.

We first formulate the stage-wise independent case where the ambiguity set is independent of any previous realizations. We will focus on this formulation throughout this chapter. However, we note that the dual decomposition method can be extended to the stage-wise dependent ambiguity set without difficulty. Subsequently, we assume relatively complete recourse to simplify the argument.

4.2.3.1 Stage-wise independent case

The stage-wise independent case of distributionally robust multistage stochastic mixed-integer programming can be stated as

$$\min_{x_1 \in X_1} \left\{ c_1^\top(\xi_1)x_1 + \max_{P_2 \in \mathcal{P}_2} \mathbb{E}_{\xi_2 \sim P_2} \left[Q_2(x_1, \xi_2) \right] \right\}, \tag{4.3}$$

where

$$\begin{aligned} Q_k(x_{k-1}, \xi_k) := \min_{x_k \in X_k} c_k^\top(\xi_k)x_k + \max_{P_{k+1} \in \mathcal{P}_{k+1}} \mathbb{E}_{\xi_{k+1} \sim P_{k+1}} \left[Q_{k+1}(x_k, \xi_{k+1}) \right] && \text{(4.4a)} \\ \text{s.t. } W_k(\xi_k)x_k \geq h_k(\xi_k) - T_k(\xi_k)x_{k-1}, && \text{(4.4b)} \end{aligned}$$

for all $k = 2, \dots, K-1$, and

$$\begin{aligned} Q_K(x_{K-1}, \xi_K) := \min_{x_K \in X_K} c_K^\top(\xi_K)x_K && \text{(4.5a)} \\ \text{s.t. } W_K(\xi_K)x_K \geq h_K(\xi_K) - T_K(\xi_K)x_{K-1}. && \text{(4.5b)} \end{aligned}$$

Here, $X_k \subseteq \mathbb{R}^{n_k}$ can be mixed-integer sets. The ambiguity sets are

$$\mathcal{P}_{k+1} := \left\{ P_{k+1} \in \mathcal{M}(\Xi_{k+1}) : \begin{array}{l} \int_{\Xi_{k+1}} \sum_{s=1}^{N_{k+1}} u_k^s(\xi_{k+1}) ||\hat{\xi}_{k+1}^s - \xi_{k+1}|| d\xi_{k+1} \leq \epsilon_{k+1}, \\ \int_{\Xi_{k+1}} u_k^s(\xi_{k+1}) d\xi_{k+1} = \hat{p}_{k+1}^s, \ \forall s \in [N_{k+1}], \\ \sum_{s=1}^{N_{k+1}} u_k^s(\xi_{k+1}) = P_{k+1}(\xi_{k+1}), \ \forall \xi_{k+1} \in \Xi_{k+1}, \\ u_k^s(\xi_{k+1}) \geq 0, \ \forall \xi_{k+1} \in \Xi_{k+1}, \ s \in [N_{k+1}] \end{array} \right\}. \tag{4.6}$$

4.2.3.2 Stage-wise dependent ambiguity set

A more general case where the ambiguity sets, as well as the costs and constraints, are dependent on realizations of scenarios in the previous stages are modeled as

$$\min_{x_1 \in X_1} \left\{ c_1^\top(\xi_1) x_1 + \max_{P_2 \in \mathcal{P}_2(\xi_1)} \mathbb{E}_{\xi_2 \sim P_2} \left[Q_2(x_1, \xi_{1:2}) \right] \right\}, \tag{4.7}$$

where

$$Q_k(x_{k-1}, \xi_{1:k}) := \min_{x_k \in X_k(\xi_{1:k})} c_k^\top(\xi_{1:k}) x_k + \max_{P_{k+1} \in \mathcal{P}_{k+1}(\xi_{1:k})} \mathbb{E}_{\xi_{k+1} \sim P_{k+1}} \left[Q_{k+1}(x_k, \xi_{1:k+1}) \right] \tag{4.8a}$$

$$\text{s.t. } W_k(\xi_{1:k}) x_k \geq h_k(\xi_{1:k}) - T_k(\xi_{1:k}) x_{k-1}, \tag{4.8b}$$

for all $k = 2, \ldots, K-1$, and

$$Q_K(x_{K-1}, \xi_{1:K}) := \min_{x_K \in X_K(\xi_{1:K})} c_K^\top(\xi_{1:K}) x_K \tag{4.9a}$$

$$\text{s.t. } W_K(\xi_{1:K}) x_K \geq h_K(\xi_{1:K}) - T_K(\xi_{1:K}) x_{K-1}. \tag{4.9b}$$

Given samples $\left(\hat{\xi}_{k+1}^{\xi_{1:k},s}, \ s \in [N_{k+1}^{\xi_{1:k}}] \right)$ from $\Xi_{k+1}^{\xi_{1:k}}$, the stage-wise dependent ambiguity set is

$$\mathcal{P}_{k+1}(\xi_{1:k}) := \left\{ P_{k+1}^{\xi_{1:k}} \in \mathcal{M}(\Xi_{k+1}^{\xi_{1:k}}) : \begin{array}{l} \int_{\Xi_{k+1}^{\xi_{1:k}}} \sum_{s=1}^{N_{k+1}^{\xi_{1:k}}} u_k^s(\xi_{k+1}) || \hat{\xi}_{k+1}^{\xi_{1:k},s} - \xi_{k+1} || d\xi_{k+1} \leq \epsilon_{k+1}^{\xi_{1:k}}, \\ \int_{\Xi_{k+1}^{\xi_{1:k}}} u_k^s(\xi_{k+1}) d\xi_{k+1} = \hat{p}_{k+1}^{\xi_{1:k},s}, \ \forall s \in [N_{k+1}^{\xi_{1:k}}], \\ \sum_{s=1}^{N_{k+1}^{\xi_{1:k}}} u_k^s(\xi_{k+1}) = P_{k+1}^{\xi_{1:k}}(\xi_{k+1}), \ \forall \xi_{k+1} \in \Xi_{k+1}^{\xi_{1:k}}, \\ u_k^s(\xi_{k+1}) \geq 0, \ \forall \xi_{k+1} \in \Xi_{k+1}^{\xi_{1:k}}, \ s \in [N_{k+1}^{\xi_{1:k}}] \end{array} \right\}. \tag{4.10}$$

A particularly interesting case is when $\xi_2, \ldots, \xi_K$ is an i.i.d. random variable sampled from a common support Ξ, and consider a setting where we learn about the distribution over time. Suppose we have N initial samples at stage one, which we denote as $\hat{\boldsymbol{\xi}} := \left\{ \hat{\xi}^1, \ldots, \hat{\xi}^N \right\}$. Then, at stage $k \geq 2$, we have $N + k - 2$ samples of ξ. At stage $k+1$, we have samples $(\hat{\boldsymbol{\xi}}, \xi_{2:k})$, where $\xi_{2:k}$ are realizations that are observed during as the stages progress. The ambiguity sets are expressed as

$$\mathcal{P}_{k+1}(\hat{\boldsymbol{\xi}}, \xi_{2:k}) := \left\{ P_{k+1}^{\xi_{2:k}} \in \mathcal{M}(\Xi) : \begin{array}{l} \int_{\Xi} \sum_{s=1}^{N+k-1} u_k^s(\xi_{k+1}) || \hat{\xi}^s - \xi_{k+1} || d\xi_{k+1} \leq \epsilon_{N+k-1}, \\ \int_{\Xi} u_k^s(\xi_{k+1}) d\xi_{k+1} = \frac{1}{N+k-1}, \ \forall s = 1, \ldots, N+k-1, \\ \sum_{s=1}^{N+k-1} u_k^s(\xi_{k+1}) = P_{k+1}^{\xi_{2:k}}(\xi_{k+1}), \ \forall \xi_{k+1} \in \Xi, \\ u_k^s(\xi_{k+1}) \geq 0, \ \forall \xi_{k+1} \in \Xi, \ s = 1, \ldots, N+k-1 \\ \hat{\xi}^s = \xi_{s-N+1}, \ \forall s = N+1, \ldots, N+k-1 \end{array} \right\}, \tag{4.11}$$

where the radius of the Wasserstein ball ϵ_{N+k-1} are given in *Esfahani and Kuhn* (2018) which becomes smaller as the number of samples increases.

The decisions based on ambiguity sets (4.11) are conservative at the beginning of the time horizon, but get progressively accurate as the data are collected. However, due to the immense difficulty of solving multistage stochastic programs in general, we have not been able to find any practical application concerning this type of problem.

4.3 Deterministic Equivalent Formulation

We present the deterministic formulation of multistage DRMIP. The deterministic formulation is the first step in deriving the decomposition scheme using Lagrangian duality. To reduce the min-max structure of the problem in (4.3), we use the following duality property of the Wasserstein-based ambiguity set:

Lemma IV.1 (*Kim* (2020))**.** *For any random variable $f(\xi) \in \mathbb{R}$, the strong duality property holds for the following problem:*

$$\max_{P \in \mathcal{P}} \mathbb{E}_P[f(\xi)]. \tag{4.12}$$

Furthermore, its dual is given as the following semi-infinite program:

$$\min_{\alpha \geq 0, \beta^s} \epsilon\alpha + \sum_{s=1}^{N} \hat{p}^s \beta^s \tag{4.13a}$$

$$s.t. \left\| \hat{\xi}^s - \xi \right\| \alpha + \beta^s \geq f(\xi) \quad \forall \xi \in \Xi, \;\; s \in [N]. \tag{4.13b}$$

We now present the deterministic formulation of (4.3).

Proposition IV.2. *The multistage DRMIP* (4.3) *can be reformulated as*

$$
\begin{aligned}
\min \quad & c_1^\top(\xi_1)x_1(\xi_1) + \epsilon_2\alpha_1(\xi_1) + \sum_{s=1}^{N_2} \hat{p}_2^s \beta_1^s(\xi_1) && (4.14a)\\
s.t. \quad & \left\|\hat{\xi}_k^s - \xi_k\right\| \alpha_{k-1}(\xi_{1:k-1}) + \beta_{k-1}^s(\xi_{1:k-1}) \geq c_k^\top(\xi_k)x_k(\xi_{1:k}) + \epsilon_{k+1}\alpha_k(\xi_{1:k}) \\
& \quad + \sum_{s'=1}^{N_{k+1}} \hat{p}_{k+1}^{s'} \beta_k^{s'}(\xi_{1:k}),\ \forall \xi_{1:k} \in \Xi_{1:k},\ s \in [N_k],\ k = 2, \ldots, K-1, && (4.14b)\\
& \left\|\hat{\xi}_K^s - \xi_K\right\| \alpha_{K-1}(\xi_{1:K-1}) + \beta_{K-1}^s(\xi_{1:K-1}) \geq c_K^\top(\xi_K)x_K(\xi_{1:K}), \\
& \qquad \forall \xi_{1:K} \in \Xi_{1:K},\ s \in [N_K], && (4.14c)\\
& T_k(\xi_k)x_{k-1}(\xi_{1:k-1}) + W_k(\xi_k)x_k(\xi_{1:k}) \geq h_k(\xi_k), \quad \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \ldots, K, \\
& && (4.14d)\\
& x_k(\xi_{1:k}) \in X_k, \quad \forall \xi_{1:k} \in \Xi_{1:k},\ k = 1, \ldots, K && (4.14e)\\
& \alpha_k(\xi_{1:k}) \geq 0, \quad \forall \xi_{1:k} \in \Xi_{1:k},\ k = 1, \ldots, K-1, && (4.14f)\\
& \beta_k^s(\xi_{1:k}) \in \mathbb{R}, \quad \forall s \in [N_{k+1}],\ \xi_{1:k} \in \Xi_{1:k},\ k = 1, \ldots, K-1. && (4.14g)
\end{aligned}
$$

Proof. Problem (4.3) can be rewritten as the following semi-infinite program:

$$
\begin{aligned}
\min \quad & c_1^\top(\xi_1)x_1 + q_1 && (4.15a)\\
\text{s.t.} \quad & q_1 \geq \mathbb{E}_{P_2}\left[c_2^\top(\xi_2)x_2(\xi_2) + \max_{P_3 \in \mathcal{P}_3} \mathbb{E}_{P_3}\left[Q_3(x_2(\xi_2), \xi_3)\right]\right], \quad \forall P_2 \in \mathcal{P}_2 && (4.15b)\\
& T_2(\xi_2)x_1 + W_2(\xi_2)x_2(\xi_2) \geq h_2(\xi_2), \quad \forall \xi_2 \in \Xi_2, && (4.15c)\\
& x_1 \in X_1,\ q_1 \in \mathbb{R}, && (4.15d)\\
& x_2(\xi_2) \in X_2, \quad \forall \xi_2 \in \Xi_2. && (4.15e)
\end{aligned}
$$

Using Lemma IV.1, constraint (4.15b) can be rewritten as

$$q_1 \geq \epsilon_2 \alpha_1 + \sum_{s_2=1}^{N_2} \hat{p}_2^{s_2} \beta_1^{s_2}, \tag{4.16a}$$

$$\left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| \alpha_1 + \beta_1^{s_2} \geq c_2^\top(\xi_2) x_2(\xi_2) + \max_{P_3 \in \mathcal{P}_3} \mathbb{E}_{P_3}\left[Q_3(x_2, \xi_3) \right] \quad \forall \xi_2 \in \Xi_2,\ s_2 \in [N_2], \tag{4.16b}$$

where $\alpha_1 \in \mathbb{R}_+$ and $\beta_1^{s_2} \in \mathbb{R}$, for all $s_2 \in [N_2]$. This results in a reformulation

$$\min \quad c_1^\top(\xi_1) x_1 + \epsilon_2 \alpha_1 + \sum_{s=1}^{N_2} \hat{p}_2^s \beta_1^s \tag{4.17a}$$

$$\text{s.t.} \quad \left\| \hat{\xi}_2^s - \xi_2 \right\| \alpha_1 + \beta_1^s \geq c_2^\top(\xi_2) x_2(\xi_2) + \max_{P_3 \in \mathcal{P}_3} \mathbb{E}_{P_3}\left[Q_3(x_2(\xi_2), \xi_3) \right], \quad \forall \xi_2 \in \Xi_2,\ s \in [N_2], \tag{4.17b}$$

$$T_2(\xi_2) x_1 + W_2(\xi_2) x_2(\xi_2) \geq h_2(\xi_2), \quad \forall \xi_2 \in \Xi_2, \tag{4.17c}$$

$$x_1 \in X_1, \tag{4.17d}$$

$$x_2(\xi_2) \in X_2, \quad \forall \xi_2 \in \Xi_2, \tag{4.17e}$$

$$\alpha_1 \geq 0, \tag{4.17f}$$

$$\beta_1^s \in \mathbb{R}, \quad \forall s \in [N_2], \tag{4.17g}$$

which is equivalent to

$$\begin{aligned}
\min \quad & c_1^\top(\xi_1)x_1 + \epsilon_2\alpha_1 + \sum_{s=1}^{N_2} \hat{p}_2^s \beta_1^s && \text{(4.18a)}\\
\text{s.t.} \quad & \left\|\hat{\xi}_2^s - \xi_2\right\| \alpha_1 + \beta_1^s \geq c_2^\top(\xi_2)x_2(\xi_2) + q_2(\xi_2),\ \forall \xi_2 \in \Xi_2,\ s \in [N_2], && \text{(4.18b)}\\
& q_2(\xi_2) \geq \mathbb{E}_{P_3}\left[c_3^\top(\xi_3)x_3(\xi_{2:3}) + \max_{P_4 \in \mathcal{P}_4} \mathbb{E}_{P_4}\left[Q_4(x_3(\xi_{2:3}), \xi_4)\right]\right],\ \forall \xi_2 \in \Xi_2,\ P_3 \in \mathcal{P}_3, && \\
& && \text{(4.18c)}\\
& T_2(\xi_2)x_1 + W_2(\xi_2)x_2(\xi_2) \geq h_2(\xi_2), \quad \forall \xi_2 \in \Xi_2, && \text{(4.18d)}\\
& T_3(\xi_3)x_2(\xi_2) + W_3(\xi_3)x_3(\xi_{2:3}) \geq h_3(\xi_3), \quad \forall \xi_{2:3} \in \Xi_{2:3}, && \text{(4.18e)}\\
& x_1 \in X_1, && \text{(4.18f)}\\
& x_2(\xi_2) \in X_2,\ q_2(\xi_2) \in \mathbb{R}, \quad \forall \xi_2 \in \Xi_2, && \text{(4.18g)}\\
& x_3(\xi_{2:3}) \in X_3, \quad \forall \xi_{2:3} \in \Xi_{2:3}, && \text{(4.18h)}\\
& \alpha_1 \geq 0, && \text{(4.18i)}\\
& \beta_1^s \in \mathbb{R}, \quad \forall s \in [N_2], && \text{(4.18j)}
\end{aligned}$$

where we have substituted the maximization problem with respect to P_3 using Lemma

IV.1. Repeating this substitution process until the terminal stage K yields

$$\min \quad c_1^\top(\xi_1)x_1 + \epsilon_2\alpha_1 + \sum_{s=1}^{N_2} \hat{p}_2^s \beta_1^s \tag{4.19a}$$

$$\text{s.t.} \quad \left\|\hat{\xi}_2^s - \xi_2\right\| \alpha_1 + \beta_1^s \geq c_2^\top(\xi_2)x_2(\xi_2) + \epsilon_3\alpha_2(\xi_2) + \sum_{s'=1}^{N_3} \hat{p}_3^{s'} \beta_2^{s'}(\xi_2),$$

$$\forall \xi_2 \in \Xi_2,\ s \in [N_2], \tag{4.19b}$$

$$\left\|\hat{\xi}_k^s - \xi_k\right\| \alpha_{k-1}(\xi_{2:k-1}) + \beta_{k-1}^s(\xi_{2:k-1}) \geq c_k^\top(\xi_k)x_k(\xi_{2:k}) + \epsilon_{k+1}\alpha_k(\xi_{2:k})$$

$$+ \sum_{s'=1}^{N_{k+1}} \hat{p}_{k+1}^{s'} \beta_k^{s'}(\xi_{2:k}),\ \forall \xi_{2:k} \in \Xi_{2:k},\ s \in [N_k],\ k = 3, \ldots, K-1, \tag{4.19c}$$

$$\left\|\hat{\xi}_K^s - \xi_K\right\| \alpha_{K-1}(\xi_{2:K-1}) + \beta_{K-1}^s(\xi_{2:K-1}) \geq c_k^\top(\xi_K)x_K(\xi_{2:K}),$$

$$\forall \xi_{2:K} \in \Xi_{2:K},\ s \in [N_K], \tag{4.19d}$$

$$T_2(\xi_2)x_1 + W_2(\xi_2)x_2(\xi_2) \geq h_2(\xi_2), \quad \forall \xi_2 \in \Xi_2, \tag{4.19e}$$

$$T_k(\xi_k)x_{k-1}(\xi_{2:k-1}) + W_k(\xi_k)x_k(\xi_{2:k}) \geq h_k(\xi_k), \quad \forall \xi_{2:k} \in \Xi_{2:k},\ k = 3, \ldots, K, \tag{4.19f}$$

$$x_1 \in X_1, \tag{4.19g}$$

$$x_k(\xi_{2:k}) \in X_k, \quad \forall \xi_{2:k} \in \Xi_{2:k},\ k = 2, \ldots, K \tag{4.19h}$$

$$\alpha_1 \geq 0, \tag{4.19i}$$

$$\alpha_k(\xi_{2:k}) \geq 0, \quad \forall \xi_{2:k} \in \Xi_{2:k},\ k = 2, \ldots, K-1, \tag{4.19j}$$

$$\beta_1^s \in \mathbb{R}, \quad \forall s \in [N_2], \tag{4.19k}$$

$$\beta_k^s(\xi_{2:k}) \in \mathbb{R}, \quad \forall s \in [N_{k+1}],\ \xi_{2:k} \in \Xi_{2:k},\ k = 2, \ldots, K-1. \tag{4.19l}$$

By simplifying the notation using the convention that Ξ_1 is a singleton set, we have (4.14). This completes the proof.

□

We use (4.14) to derive the dual decomposition formulation in the following sections.

4.4 Lagrangian Dual of DRMSMIP

In this section, we first derive the Lagrangian dual of problem (4.3), which does not assume any form of ambiguity sets. We then formulate the Lagrangian dual of problem (4.14) where the ambiguity sets are Wasserstein-based and compare the two forms of Lagrangian duals.

4.4.1 Lagrangian dual for general ambiguity set

Proposition IV.3. *The Lagrangian dual of problem* (4.3) *is given by*

$$\max \quad \int_{\Xi_{1:K}} D'(\mu_{1:K}(\xi_{1:K}), \xi_{1:K}) d\xi_{1:K} \tag{4.20a}$$

$$s.t. \quad \int_{\Xi_{2:K}} \mu_1(\xi_{1:K}) d\xi_{2:K} = c_1(\xi_1), \tag{4.20b}$$

$$\int_{\Xi_{k+1:K}} \mu_k(\xi_{1:K}) d\xi_{k+1:K} = P_2^{\xi_1}(\xi_2) P_3^{\xi_{1:2}}(\xi_3) \cdots P_k^{\xi_{1:k-1}}(\xi_k) c_k(\xi_k),$$

$$\forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \ldots, K, \tag{4.20c}$$

$$P_{k+1}^{\xi_{1:k}} \in \mathcal{P}_{k+1},\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \ldots, K-1, \tag{4.20d}$$

where the subproblems are

$$D'(\mu_{1:K}(\xi_{1:K}), \xi_{1:K}) := \min \quad \sum_{k=1}^{K} \mu_k^\top(\xi_{1:K}) x_k \tag{4.21a}$$

$$s.t. \quad W_k(\xi_k) x_k \geq h_k(\xi_k) - T_k(\xi_k) x_{k-1},\ \forall k = 2, \ldots, K, \tag{4.21b}$$

$$x_k \in X_k,\ \forall k = 1, \ldots K. \tag{4.21c}$$

Proof. Using min-max inequality and repeatedly exchanging the minimization and maximization, the following problem bounds the original problem (4.3) from below.

$$\max_{\substack{P_k^{\xi_{1:k-1}} \in \mathcal{P}_{k+1},\\ \forall \xi_{1:k} \in \Xi_{1:k}\\ k=2,\ldots,K}} \min_{x_1(\xi_1) \in X_1} \quad c_1^\top(\xi_1) x_1(\xi_1) + \mathbb{E}_{P_2^{\xi_1}} \left[Q_2'(x_1(\xi_1), P_{1:K}, \xi_{1:2}) \right], \tag{4.22}$$

where

$$\begin{aligned} & Q_k'(x_{k-1}(\xi_{1:k-1}), P_{1:K}, \xi_{1:k}) \\ & \qquad := \min_{x_k(\xi_{1:k}) \in X_k} \quad c_k^\top(\xi_k) x_k(\xi_{1:k}) + \mathbb{E}_{P_{k+1}^{\xi_{1:k}}} \left[Q_{k+1}'(x_k(\xi_{1:k}), P_{1:K}, \xi_{1:k+1}) \right] && \text{(4.23a)} \\ & \qquad\qquad \text{s.t.} \quad W_k(\xi_k) x_k(\xi_{1:k}) \geq h_k(\xi_k) - T_k(\xi_k) x_{k-1}(\xi_{1:k-1}), && \text{(4.23b)} \end{aligned}$$

and

$$\begin{aligned} & Q_K'(x_{K-1}(\xi_{1:K-1}), P_{1:K}, \xi_{1:K}) \\ & \qquad := \min_{x_K(\xi_{1:K}) \in X_K} \quad c_K^\top(\xi_K) x_K(\xi_{1:K}) && \text{(4.24a)} \\ & \qquad\qquad \text{s.t.} \quad W_K(\xi_K) x_K(\xi_{1:K}) \geq h_K(\xi_K) - T_K(\xi_K) x_{K-1}(\xi_{1:K-1}). && \text{(4.24b)} \end{aligned}$$

Notice that all the maximization with respect to the unknown probability $P_k^{\xi_{1:k-1}}$ is combined at the beginning of (4.22). By aggregating the multistage formulation to a single optimization problem, we obtain an equivalent form:

$$\max_{\substack{P_k^{\xi_{1:k-1}} \in \mathcal{P}_{k+1}, \\ \forall \xi_{1:k} \in \Xi_{1:k} \\ k=2,\ldots,K}} \min \quad c_1^\top(\xi_1)x_1(\xi_1) + \sum_{k=2}^{K} \mathbb{E}_{P_2^{\xi_1}} \left[\mathbb{E}_{P_3^{\xi_{1:2}}} \left[\cdots \mathbb{E}_{P_k^{\xi_{1:k}}} \left[c_k^\top(\xi_k)x_k(\xi_{1:k}) \right] \cdots \right] \right] \tag{4.25a}$$

$$\text{s.t.} \quad W_k(\xi_k)x_k(\xi_{1:k}) \geq h_k(\xi_k) - T_k(\xi_k)x_{k-1}(\xi_{1:k-1}),$$

$$\forall \xi_{1:k} \in \Xi_{1:k}, \ k = 2, \ldots, K, \tag{4.25b}$$

$$x_k(\xi_{1:k}) \in X_k, \ \forall \xi_{1:k} \in \Xi_{1:k}, \ k = 1, \ldots, K. \tag{4.25c}$$

After simplifying the objective function and introducing the non-anticipativity constraints to (4.25), we obtain

$$\max_{\substack{P_k^{\xi_{1:k-1}} \in \mathcal{P}_{k+1}, \\ \forall \xi_{1:k} \in \Xi_{1:k} \\ k=2,\ldots,K}} \min \quad \mathbb{E}_{P_2^{\xi_1}} \left[\mathbb{E}_{P_3^{\xi_{1:2}}} \left[\cdots \mathbb{E}_{P_K^{\xi_{1:K}}} \left[\sum_{k=1}^{K} c_k^\top(\xi_k)x_k(\xi_{1:k}) \right] \cdots \right] \right] \tag{4.26a}$$

$$\text{s.t.} \quad x_k(\xi_{1:k}) = \breve{x}_k(\xi'_{1:K}), \ \forall (\xi_{1:k}, \xi'_{1:K}) \text{ such that } \xi_{1:k} = \xi'_{1:k}, \ k = 1, \ldots, K \tag{4.26b}$$

$$W_k(\xi_k)\breve{x}_k(\xi_{1:K}) \geq h_k(\xi_k) - T_k(\xi_k)\breve{x}_{k-1}(\xi_{1:K}),$$

$$\forall \xi_{1:K} \in \Xi_{1:K}, \ k = 2, \ldots, K, \tag{4.26c}$$

$$\breve{x}_k(\xi_{1:K}) \in X_k, \ \forall \xi_{1:K} \in \Xi_{1:K}, \ k = 1, \ldots, K. \tag{4.26d}$$

The constraints (4.26c) are now disjunctive for each sample path $\xi_{1:K}$. After substituting the expectation with integration the Lagrangian formulation of (4.26) is therefore,

$$\begin{aligned}\min \quad & \sum_{k=1}^{K} \int\limits_{\Xi_{1:K}} P_{1:K}(\xi_{1:K}) c_k^\top(\xi_k) x_k(\xi_{1:k}) d\xi_{1:K} \\ & \qquad + \sum_{k=1}^{K} \int\limits_{\Xi_{1:K}} \mu_k^\top(\xi_{1:K}) \left(\breve{x}_k(\xi_{1:K}) - x_k(\xi_{1:k})\right) d\xi_{1:K} \qquad (4.27\text{a}) \\ \text{s.t.} \quad & W_k(\xi_k)\breve{x}_k(\xi_{1:K}) \geq h_k(\xi_k) - T_k(\xi_k)\breve{x}_{k-1}(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 2, \ldots, K, \\ & \qquad (4.27\text{b}) \\ & \breve{x}_k(\xi_{1:K}) \in X_k,\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 1, \ldots, K, \qquad (4.27\text{c})\end{aligned}$$

where $P_{1:K}(\xi_{1:K}) = P_2^{\xi_1}(\xi_2) P_3^{\xi_{1:2}}(\xi_3) \cdots P_K^{\xi_{1:K-1}}(\xi_K)$.

Notice that the objective function of (4.27) can be rewritten as

$$\begin{aligned}& \sum_{k=1}^{K} \int\limits_{\Xi_{1:K}} \left(P_{1:K}(\xi_{1:K}) c_k(\xi_k) - \mu_k(\xi_{1:K})\right)^\top x_k(\xi_{1:k}) d\xi_{1:K} \\ & \qquad + \sum_{k=1}^{K} \int\limits_{\Xi_{1:K}} \mu_k^\top(\xi_{1:K}) \breve{x}_k(\xi_{1:K}) d\xi_{1:K}. \qquad (4.28)\end{aligned}$$

Therefore, to guarantee the finiteness of problem (4.27), the following must hold for each $x_k(\xi_{1:k})$:

$$\int\limits_{\Xi_{k+1:K}} \left(P_{1:K}(\xi_{1:K}) c_k(\xi_k) - \mu_k(\xi_{1:K})\right) d\xi_{k+1:K} = 0. \qquad (4.29)$$

Otherwise, $x_k(\xi_{1:k})$ can be changed indefinitely to minimize (4.27). Thus, we move (4.29) to initial maximization problem, and the statement of the proposition follows. □

Due to the multiplication of probabilities $P_k^{\xi_{1:k-1}}(\xi_k)$ in (4.20c), it is not trivial to solve (4.20) for a general ambiguity set $\mathcal{P}_k$. In the next section, we start from (4.14) to formulate the Lagrangian dual for the Wasserstein ball ambiguity set.

4.4.2 Lagrangian Dual for Wasserstein Ball

We consider a formulation equivalent to (4.14) by considering non-anticipativity constraints corresponding to $x_k(\xi_{1:k}),\ \alpha_k(\xi_{1:k}),\ \beta_k^s(\xi_{1:k})$.

Proposition IV.4. *The Lagrangian relaxation of the deterministic formulation* (4.14) *using scenario decomposition is given by*

$$\underline{z}^{WLD} := \max \int\limits_{\Xi_{1:K}} \underline{D}^W(\bar{\mu}_{1:K}(\xi_{1:K}), \bar{\nu}_{1:K}(\xi_{1:K}), \bar{\boldsymbol{u}}_{1:K}(\xi_{1:K}), \xi_{1:K}) d\xi_{1:K} \tag{4.30a}$$

$$s.t. \int\limits_{\Xi_{2:K}} \bar{\mu}_1(\xi_{1:K}) d\xi_{2:K} = c_1(\xi_1), \tag{4.30b}$$

$$\int\limits_{\Xi_{k+1:K}} \bar{\mu}_k(\xi_{1:K}) d\xi_{k+1:K} = 0,\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \ldots, K, \tag{4.30c}$$

$$\int\limits_{\Xi_{2:K}} \bar{\nu}_1(\xi_{1:K}) d\xi_{2:K} = \epsilon_2, \tag{4.30d}$$

$$\int\limits_{\Xi_{k+1:K}} \bar{\nu}_k(\xi_{1:K}) d\xi_{k+1:K} = 0,\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \ldots, K-1, \tag{4.30e}$$

$$\int\limits_{\Xi_{2:K}} \bar{u}_1^s(\xi_{1:K}) d\xi_{2:K} = \hat{p}_2^s,\ \forall s \in [N_2], \tag{4.30f}$$

$$\int\limits_{\Xi_{k+1:K}} \bar{u}_k^s(\xi_{1:K}) d\xi_{k+1:K} = 0,$$

$$\forall \xi_{1:k} \in \Xi_{1:k},\ s \in [N_{k+1}],\ k = 2, \ldots, K-1, \tag{4.30g}$$

$$\bar{\mu}_k(\xi_{1:K}) \in \mathbb{R}^{n_k},\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 1, \ldots, K, \tag{4.30h}$$

$$\bar{\nu}_k(\xi_{1:K}) \in \mathbb{R},\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 1, \ldots, K-1, \tag{4.30i}$$

$$\bar{u}_k^s(\xi_{1:K}) \in \mathbb{R},\ \forall \xi_{1:K} \in \Xi_{1:K},\ s \in [N_{k+1}],\ k = 1, \ldots, K-1, \tag{4.30j}$$

where the subproblems are

$$\underline{D}^W(\bar{\mu}_{1:K}(\xi_{1:K}),\bar{\nu}_{1:K}(\xi_{1:K}),\bar{\boldsymbol{u}}_{1:K}(\xi_{1:K}),\xi_{1:K})$$

$$:= \min \quad \sum_{k=1}^{K} \bar{\mu}_k^\top(\xi_{1:K})\breve{x}_k + \sum_{k=1}^{K-1} \bar{\nu}_k(\xi_{1:K})\breve{\alpha}_k + \sum_{k=1}^{K-1}\sum_{s_{k+1}=1}^{N_{k+1}} \bar{u}_k^{s_{k+1}}(\xi_{1:K})\breve{\beta}_k^{s_{k+1}} \tag{4.31a}$$

$$\text{s.t.} \quad \left\|\hat{\xi}_k^s - \xi_k\right\| \breve{\alpha}_{k-1} + \breve{\beta}_{k-1}^s \geq c_k^\top(\xi_k)\breve{x}_k + \epsilon_{k+1}\breve{\alpha}_k + \sum_{s'=1}^{N_{k+1}} \hat{p}_{k+1}^{s'}\breve{\beta}_k^{s'},\ \forall s \in [N_k],\ k = 2,\dots,K-1, \tag{4.31b}$$

$$\left\|\hat{\xi}_K^s - \xi_K\right\| \breve{\alpha}_{K-1} + \breve{\beta}_{K-1}^s \geq c_K^\top(\xi_K)\breve{x}_K,\ \forall s \in [N_K], \tag{4.31c}$$

$$T_k(\xi_k)\breve{x}_{k-1} + W_k(\xi_k)\breve{x}_k \geq h_k(\xi_k), \quad \forall k = 2,\dots,K, \tag{4.31d}$$

$$\breve{x}_k \in X_k, \quad \forall k = 1,\dots,K, \tag{4.31e}$$

$$\breve{\alpha}_k \geq 0, \quad \forall k = 1,\dots,K-1, \tag{4.31f}$$

$$\breve{\beta}_k^s \in \mathbb{R}, \quad \forall s \in [N_{k+1}],\ k = 1,\dots,K-1. \tag{4.31g}$$

Proof. Using non-anticipativity constraints, (4.14) can be formulated as

$$\min \quad c_1^\top(\xi_1)x_1(\xi_1) + \epsilon_2\alpha_1(\xi_1) + \sum_{s_2=1}^{N_2} \hat{p}_2^{s_2}\beta_1^{s_2}(\xi_1) \tag{4.32a}$$

$$\text{s.t.} \quad x_k(\xi_{1:k}) = \breve{x}_k(\xi'_{1:K}),\ \forall\,(\xi_{1:k},\xi'_{1:K}) \text{ such that } \xi_{1:k} = \xi'_{1:k},\ k = 1,\dots,K \tag{4.32b}$$

$$\alpha_k(\xi_{1:k}) = \breve{\alpha}_k(\xi'_{1:K}),\ \forall\,(\xi_{1:k},\xi'_{1:K}) \text{ such that } \xi_{1:k} = \xi'_{1:k},\ k = 1,\dots,K-1 \tag{4.32c}$$

$$\beta_k^s(\xi_{1:k}) = \breve{\beta}_k^s(\xi'_{1:K}),\ \forall\,(\xi_{1:k},\xi'_{1:K}) \text{ such that } \xi_{1:k} = \xi'_{1:k},\quad s \in [N_{k+1}],\ k = 1,\dots,K-1 \tag{4.32d}$$

$$\left\|\hat{\xi}_k^s - \xi_k\right\| \breve{\alpha}_{k-1}(\xi_{1:K}) + \breve{\beta}_{k-1}^s(\xi_{1:K}) \geq c_k^\top(\xi_k)\breve{x}_k(\xi_{1:K}) + \epsilon_{k+1}\breve{\alpha}_k(\xi_{1:K}) + \sum_{s'=1}^{N_{k+1}} \hat{p}_{k+1}^{s'}\breve{\beta}_k^{s'}(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ s \in [N_k],\ k = 2,\dots,K-1, \tag{4.32e}$$

$$\left\|\hat{\xi}_K^s - \xi_K\right\| \breve{\alpha}_{K-1}(\xi_{1:K}) + \breve{\beta}_{K-1}^s(\xi_{1:K}) \geq c_K^\top(\xi_K)\breve{x}_K(\xi_{1:K}),\quad \forall \xi_{1:K} \in \Xi_{1:K},\ s \in [N_K], \tag{4.32f}$$

$$T_k(\xi_k)\breve{x}_{k-1}(\xi_{1:K}) + W_k(\xi_k)\breve{x}_k(\xi_{1:K}) \geq h_k(\xi_k), \quad \forall \xi_{1:K} \in \Xi_{1:K},\ k = 2,\dots,K, \tag{4.32g}$$

$$\breve{x}_k(\xi_{1:K}) \in X_k, \quad \forall \xi_{1:K} \in \Xi_{1:K},\ k = 1,\dots,K, \tag{4.32h}$$

$$\breve{\alpha}_k(\xi_{1:K}) \geq 0, \quad \forall \xi_{1:K} \in \Xi_{1:K},\ k = 1,\dots,K-1, \tag{4.32i}$$

$$\breve{\beta}_k^s(\xi_{1:K}) \in \mathbb{R}, \quad \forall s \in [N_{k+1}],\ \xi_{1:K} \in \Xi_{1:K},\ k = 1,\dots,K-1. \tag{4.32j}$$

Let $\bar{\mu}_k(\xi_{1:K}) \in \mathbb{R}^{n_k}$, $\bar{\nu}_k(\xi_{1:K}) \in \mathbb{R}$, and $\bar{u}_k^{s_{k+1}}(\xi_{1:K}) \in \mathbb{R}$ be the Lagrangian multipliers corresponding to the nonanticipativity constraints (4.32b), (4.32c), and (4.32d). The

Lagrangian formulation is

$$\begin{aligned}\min \quad & c_1^\top(\xi_1)x_1(\xi_1) + \epsilon_2\alpha_1(\xi_1) + \sum_{s=1}^{N_2} \hat{p}_2^s\beta_1^s(\xi_1) \\ & + \sum_{k=1}^{K} \int_{\Xi_{1:K}} \bar{\mu}_k^\top(\xi_{1:K})\left(\breve{x}_k(\xi_{1:K}) - x_k(\xi_{1:k})\right) d\xi_{1:K} \\ & + \sum_{k=1}^{K-1} \int_{\Xi_{1:K}} \bar{\nu}_k(\xi_{1:K})\left(\breve{\alpha}_k(\xi_{1:K}) - \alpha_k(\xi_{1:k})\right) d\xi_{1:K} \\ & + \sum_{k=1}^{K-1} \sum_{s_{k+1}=1}^{N_{k+1}} \int_{\Xi_{1:K}} \bar{u}_k^{s_{k+1}}(\xi_{1:K})\left(\breve{\beta}_k^{s_{k+1}}(\xi_{1:K}) - \beta_k^{s_{k+1}}(\xi_{1:k})\right) d\xi_{1:K} \qquad (4.33a) \\ \text{s.t.} \quad & (4.32e)–(4.32j).\end{aligned}$$

Using similar steps as Propopsition IV.3, the Lagrangian multipliers are subject to (4.30b)–(4.30g) to guarantee the finiteness of (4.33). The Lagrangian dual function is threfore,

$$\begin{aligned}\min \quad & \int_{\Xi_{1:K}} \left(\sum_{k=1}^{K} \bar{\mu}_k^\top(\xi_{1:K})\breve{x}_k(\xi_{1:K}) + \sum_{k=1}^{K-1} \bar{\nu}_k(\xi_{1:K})\breve{\alpha}_k(\xi_{1:K}) \right. \\ & \qquad \left. + \sum_{k=1}^{K-1} \sum_{s_{k+1}=1}^{N_{k+1}} \bar{u}_k^{s_{k+1}}(\xi_{1:K})\breve{\beta}_k^{s_{k+1}}(\xi_{1:K})\right) d\xi_{1:K} \qquad (4.34a) \\ \text{s.t.} \quad & (4.32e)–(4.32j),\end{aligned}$$

which can be decomposed for each $\xi_{1:K}$. □

We demonstrate that problem (4.30) can be transformed to problem (4.20).

Theorem IV.5. *Problem* (4.30) *is equivalent to* (4.20) *when the ambiguity sets are Wasserstein ball ambiguity sets.*

Proof. We eliminate $\breve{\beta}_k^{s_{k+1}}$. Let us define $\breve{r}_k^{s_k}$ as the nonnegative slack of the constraints

(4.31b) and (4.31c). Then,

$$\breve{\beta}^{s_k}_{k-1} = -\left\|\hat{\xi}^{s_k}_k - \xi_k\right\| \breve{\alpha}_{k-1} + c_k^\top(\xi_k)\breve{x}_k + \epsilon_{k+1}\breve{\alpha}_k + \breve{r}^{s_k}_k + \sum_{s_{k+1}=1}^{N_{k+1}} \hat{p}^{s_{k+1}}_{k+1}\breve{\beta}^{s_{k+1}}_k, \quad \forall k = 2, \ldots, K-1, \tag{4.35}$$

$$\breve{\beta}^{s_K}_{K-1} = -\left\|\hat{\xi}^{s_K}_K - \xi_K\right\| \breve{\alpha}_{K-1} + c_K^\top(\xi_K)\breve{x}_K + \breve{r}^{s_K}_K. \tag{4.36}$$

This can be simplified to

$$\begin{aligned}\breve{\beta}^{s_{k+1}}_k &= -\left\|\hat{\xi}^{s_{k+1}}_{k+1} - \xi_{k+1}\right\| \breve{\alpha}_k + c_{k+1}^\top(\xi_{k+1})\breve{x}_{k+1} + \breve{r}^{s_{k+1}}_{k+1} \\ &\quad + \sum_{j=k+2}^{K}\left(\sum_{s_{k+2}=1}^{N_{k+2}} \cdots \sum_{s_j=1}^{N_j} \prod_{i=k+2}^{j} \hat{p}^{s_i}_i \left(\left(\epsilon_j - \left\|\hat{\xi}^{s_j}_j - \xi_j\right\|\right)\breve{\alpha}_{j-1} + c_j^\top(\xi_j)\breve{x}_j + \breve{r}^{s_j}_j\right)\right),\end{aligned} \tag{4.37}$$

$$\begin{aligned}&= -\left\|\hat{\xi}^{s_{k+1}}_{k+1} - \xi_{k+1}\right\| \breve{\alpha}_k + c_{k+1}^\top(\xi_{k+1})\breve{x}_{k+1} + \breve{r}^{s_{k+1}}_{k+1} \\ &\quad + \sum_{j=k+2}^{K}\left(\sum_{s_j=1}^{N_j} \hat{p}^{s_j}_j \left(\left(\epsilon_j - \left\|\hat{\xi}^{s_j}_j - \xi_j\right\|\right)\breve{\alpha}_{j-1} + c_j^\top(\xi_j)\breve{x}_j + \breve{r}^{s_j}_j\right)\right. \\ &\qquad\qquad \left. \times \sum_{s_{k+2}=1}^{N_{k+2}} \cdots \sum_{s_{j-1}=1}^{N_{j-1}} \prod_{i=k+2}^{j-1} \hat{p}^{s_i}_i\right),\end{aligned} \tag{4.38}$$

$$\begin{aligned}&= -\left\|\hat{\xi}^{s_{k+1}}_{k+1} - \xi_{k+1}\right\| \breve{\alpha}_k + \breve{r}^{s_{k+1}}_{k+1} + \sum_{j=k+1}^{K} c_j^\top(\xi_j)\breve{x}_j \\ &\quad + \sum_{j=k+2}^{K} \sum_{s_j=1}^{N_j} \hat{p}^{s_j}_j \left(\left(\epsilon_j - \left\|\hat{\xi}^{s_j}_j - \xi_j\right\|\right)\breve{\alpha}_{j-1} + \breve{r}^{s_j}_j\right),\end{aligned} \tag{4.39}$$

for all $k = 1, \ldots, K-2$. Then,

$$\begin{aligned}
&\sum_{k=1}^{K-1} \sum_{s_{k+1}=1}^{N_{k+1}} \bar{u}_k^{s_{k+1}}(\xi_{1:K}) \breve{\beta}_k^{s_{k+1}} = \\
&\qquad \sum_{k=2}^{K} \left(\sum_{j=1}^{k-1} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}) \right) c_k^\top(\xi_k) \breve{x}_k \\
&\qquad - \sum_{k=1}^{K-1} \sum_{s_{k+1}=1}^{N_{k+1}} \bar{u}_k^{s_{k+1}}(\xi_{1:K}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| \breve{\alpha}_k \\
&\qquad + \sum_{k=2}^{K-1} \left(\sum_{j=1}^{k-1} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}) \right) \sum_{s_{k+1}=1}^{N_{k+1}} \hat{p}_{k+1}^{s_{k+1}} \left(\epsilon_{k+1} - \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| \right) \breve{\alpha}_k \\
&\qquad + \sum_{k=2}^{K} \sum_{s_k=1}^{N_k} \bar{u}_{k-1}^{s_k}(\xi_{1:K}) \breve{r}_k^{s_k} \\
&\qquad + \sum_{k=3}^{K} \left(\sum_{j=1}^{k-2} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}) \right) \sum_{s_k=1}^{N_k} \hat{p}_k^{s_k} \breve{r}_k^{s_k} \qquad (4.40)
\end{aligned}$$

We substitute this to (4.31a). Now, since $\breve{\alpha}_k \geq 0$,

$$\bar{\nu}_1(\xi_{1:K}) - \sum_{s_2=1}^{N_2} \bar{u}_1^{s_2}(\xi_{1:K}) \left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| \geq 0, \qquad (4.41a)$$

$$\begin{aligned}
&\bar{\nu}_k(\xi_{1:K}) + \left(\sum_{j=1}^{k-1} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}) \right) \sum_{s_{k+1}=1}^{N_{k+1}} \hat{p}_{k+1}^{s_{k+1}} \left(\epsilon_{k+1} - \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| \right) \\
&\qquad - \sum_{s_{k+1}=1}^{N_{k+1}} \bar{u}_k^{s_{k+1}}(\xi_{1:K}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| \geq 0, \ \forall k = 2, \ldots, K-1, \qquad (4.41b)
\end{aligned}$$

otherwise, the solution is unbounded. Furthermore, since $\breve{r}_k^{s_k} \geq 0$,

$$\bar{u}_1^{s_2}(\xi_{1:K}) \geq 0, \ \forall s_2 \in [N_2], \qquad (4.42a)$$

$$\bar{u}_k^{s_{k+1}}(\xi_{1:K}) + \left(\sum_{j=1}^{k-1} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}) \right) \hat{p}_{k+1}^{s_{k+1}} \geq 0, \ \forall s_{k+1} \in [N_{k+1}], \ k = 2, \ldots, K-1. \qquad (4.42b)$$

Thus, the subproblems are equivalent to

$$
\begin{aligned}
&\underline{D}^{W'}(\bar{\mu}_{1:K}(\xi_{1:K}),\bar{\boldsymbol{u}}_{1:K}(\xi_{1:K}),\xi_{1:K}) \\
&\qquad := \min \quad \bar{\mu}_1^\top(\xi_{1:K})\breve{x}_1 + \sum_{k=2}^{K}\left(\bar{\mu}_k^\top(\xi_{1:K}) + \left(\sum_{j=1}^{k-1}\sum_{s_{j+1}}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K})\right) c_k^\top(\xi_k)\right)\breve{x}_k
\end{aligned}
\tag{4.43a}
$$

$$
\text{s.t.} \quad (4.31\text{d}),\ (4.31\text{e}),
$$

and problem (4.30) is

$$
\max \quad \int\limits_{\Xi_{1:K}} \underline{D}^{W'}(\bar{\mu}_{1:K}(\xi_{1:K}),\bar{\boldsymbol{u}}_{1:K}(\xi_{1:K}),\xi_{1:K})d\xi_{1:K} \tag{4.44a}
$$

$$
\text{s.t.} \quad (4.30\text{b}),\ (4.30\text{c}),
$$

$$
\int\limits_{\Xi_{2:K}} \sum_{s_2=1}^{N_2} \bar{u}_1^{s_2}(\xi_{1:K}) \left\|\hat{\xi}_2^{s_2} - \xi_2\right\| d\xi_{2:K} \le \epsilon_2, \tag{4.44b}
$$

$$
\begin{aligned}
&\int\limits_{\Xi_{k+1:K}} \sum_{s_{k+1}=1}^{N_{k+1}} \left(\bar{u}_k^{s_{k+1}}(\xi_{1:K}) + \left(\sum_{j=1}^{k-1}\sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K})\right)\hat{p}_{k+1}^{s_{k+1}}\right)\left\|\hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1}\right\| d\xi_{k+1:K} \\
&\quad \le \epsilon_{k+1} \int\limits_{\Xi_{k+1:K}} \left(\sum_{j=1}^{k-1}\sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K})\right) d\xi_{k+1:K},\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2,\ldots,K-1,
\end{aligned}
\tag{4.44c}
$$

$$
\int\limits_{\Xi_{2:K}} \bar{u}_1^{s_2}(\xi_{1:K})d\xi_{2:K} = \hat{p}_2^{s_2},\ \forall s_2 \in [N_2], \tag{4.44d}
$$

$$
\int\limits_{\Xi_{k+1:K}} \bar{u}_k^{s_{k+1}}(\xi_{1:K})d\xi_{k+1:K} = 0,\ \forall \xi_{1:k} \in \Xi_{1:k},\ s_{k+1} \in [N_{k+1}],\ k = 2,\ldots,K-1, \tag{4.44e}
$$

$$
(4.42\text{a}),\ (4.42\text{b}),\ (4.30\text{h}),\ (4.30\text{j}),
$$

where we have eliminated ν_k by using (4.41a) and (4.41b). Let us introduce variables

$v_k(\xi_{1:K}) \in \mathbb{R}$, defined as

$$v_k(\xi_{1:K}) = \sum_{j=1}^{k-1} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 2, \ldots, K-1, \tag{4.45}$$

and variables $\bar{w}_k^{s_{k+1}}(\xi_{1:K})$, defined as

$$\bar{w}_1^{s_2}(\xi_{1:K}) = \bar{u}_1^{s_2}(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ s_2 \in [N_2], \tag{4.46a}$$

$$\bar{w}_k^{s_{k+1}}(\xi_{1:K}) = \bar{u}_k^{s_{k+1}}(\xi_{1:K}) + \hat{p}_{k+1}^{s_{k+1}} v_k(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ s_{k+1} \in [N_{k+1}],\ k = 2, \ldots, K. \tag{4.46b}$$

Then we eliminate $\bar{u}_k^{s_{k+1}}(\xi_{1:K})$ from (4.44b) – (4.44e), (4.42a), (4.42b) and get

$$\int\limits_{\Xi_{2:K}} \sum_{s_2=1}^{N_2} \bar{w}_1^{s_2}(\xi_{1:K}) \left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| d\xi_{2:K} \leq \epsilon_2, \tag{4.47a}$$

$$\int\limits_{\Xi_{k+1:K}} \sum_{s_{k+1}=1}^{N_{k+1}} \bar{w}_k^{s_{k+1}}(\xi_{1:K}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| d\xi_{k+1:K}$$

$$\leq \epsilon_{k+1} \int\limits_{\Xi_{k+1:K}} v_k(\xi_{1:K}) d\xi_{k+1:K},\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \ldots, K-1, \tag{4.47b}$$

$$\int\limits_{\Xi_{2:K}} \bar{w}_1^{s_2}(\xi_{1:K}) d\xi_{2:K} = \hat{p}_2^{s_2},\ \forall s_2 \in [N_2], \tag{4.47c}$$

$$\int\limits_{\Xi_{k+1:K}} \bar{w}_k^{s_{k+1}}(\xi_{1:K}) d\xi_{k+1:K} = \hat{p}_{k+1}^{s_{k+1}} \int\limits_{\Xi_{k+1:K}} v_k(\xi_{1:K}) d\xi_{k+1:K},$$

$$\forall \xi_{1:k} \in \Xi_{1:k},\ s_{k+1} \in [N_{k+1}],\ k = 2, \ldots, K-1, \tag{4.47d}$$

$$\bar{w}_k^{s_{k+1}}(\xi_{1:K}) \geq 0,\ \forall \xi_{1:K} \in \Xi_{1:K},\ s_{k+1} \in [N_{k+1}],\ k = 1, \ldots, K-1, \tag{4.47e}$$

$$v_k(\xi_{1:K}) \in \mathbb{R},\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 2, \ldots, K-1, \tag{4.47f}$$

$$v_2(\xi_{1:K}) = \sum_{s_2=1}^{N_2} \bar{w}_1^{s_2}(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K}. \tag{4.47g}$$

$$v_k(\xi_{1:K}) = \sum_{j=1}^{k-1} \sum_{s_{j+1}=1}^{N_{j+1}} \bar{w}_j^{s_{j+1}}(\xi_{1:K}) - \sum_{j=2}^{k-1} v_j(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 3, \ldots, K-1. \tag{4.47h}$$

The last two equalities (4.47g) and (4.47h) result in

$$v_k(\xi_{1:K}) = \sum_{s_k=1}^{N_k} \bar{w}_{k-1}^{s_k}(\xi_{1:K}),\ \forall \xi_{1:K} \in \Xi_{1:K},\ k = 2, \ldots, K-1, \tag{4.48}$$

which we use to eliminate $v_k(\xi_{1:K})$ from (4.47).

Let us define

$$w_k^{s_{k+1}}(\xi_{1:k+1}) := \int\limits_{\Xi_{k+2:K}} \bar{w}_k^{s_{k+1}}(\xi_{1:K}) d\xi_{k+2:K}, \quad \forall \xi_{1:k+1} \in \Xi_{1:k+1},\ k = 1, \dots, K-2, \tag{4.49}$$

$$w_{K-1}^{s_K}(\xi_{1:K}) := \bar{w}_{K-1}^{s_K}(\xi_{1:K}), \qquad \forall \xi_{1:K} \in \Xi_{1:K}. \tag{4.50}$$

Then, set of constraints (4.47) are equivalently,

$$\int\limits_{\Xi_2} \sum_{s_2=1}^{N_2} w_1^{s_2}(\xi_{1:2}) \left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| d\xi_2 \le \epsilon_2, \tag{4.51a}$$

$$\int\limits_{\Xi_{k+1}} \sum_{s_{k+1}=1}^{N_{k+1}} w_k^{s_{k+1}}(\xi_{1:k+1}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| d\xi_{k+1} \le \epsilon_{k+1} \sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k}),$$
$$\forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \dots, K-1, \tag{4.51b}$$

$$\int\limits_{\Xi_2} w_1^{s_2}(\xi_{1:2}) d\xi_2 = \hat{p}_2^{s_2},\ \forall s_2 \in [N_2], \tag{4.51c}$$

$$\int\limits_{\Xi_{k+1}} w_k^{s_{k+1}}(\xi_{1:k+1}) d\xi_{k+1} = \hat{p}_{k+1}^{s_{k+1}} \sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k}),$$
$$\forall \xi_{1:k} \in \Xi_{1:k},\ s_{k+1} \in [N_{k+1}],\ k = 2, \dots, K-1, \tag{4.51d}$$

$$w_k^{s_{k+1}}(\xi_{1:k+1}) \ge 0,\ \forall \xi_{1:k+1} \in \Xi_{1:k+1},\ s_{k+1} \in [N_{k+1}],\ k = 1, \dots, K-1. \tag{4.51e}$$

These constraints can be classified into the following groups:

- **Set of constraints for the first stage:** (4.51a)**,** (4.51c)

  For notational convenience, we define

$$u_1^{s_2}(\xi_{1:2}) := w_1^{s_2}(\xi_{1:2}), \qquad \forall \xi_{1:2} \in \Xi_{1:2}. \tag{4.52}$$

The set of constraints corresponding to the first stage is

$$\int\limits_{\Xi_2} \sum_{s_2=1}^{N_2} u_1^{s_2}(\xi_{1:2}) \left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| d\xi_2 \leq \epsilon_2, \tag{4.53a}$$

$$\int\limits_{\Xi_2} u_1^{s_2}(\xi_{1:2}) d\xi_2 = \hat{p}_2^{s_2},\ \forall s_2 \in [N_2], \tag{4.53b}$$

$$u_1^{s_2}(\xi_{1:2}) \geq 0,\ \forall \xi_2 \in \Xi_2,\ s_2 \in [N_2]. \tag{4.53c}$$

We introduce variables $P_2^{\xi_1}(\xi_2),\ \forall \xi_2 \in \Xi_2$ and add constraints

$$\sum_{s_2=1}^{N_2} u_1^{s_2}(\xi_{1:2}) = P_2^{\xi_1}(\xi_2),\ \forall \xi_2 \in \Xi_2. \tag{4.54}$$

Then, these set of constraints correspond to a Wasserstein ball ambiguity set $\mathcal{P}_2$.

- **Set of constraints for stages 2 to K:** (4.51b)**,** (4.51d)

  Assume that for a given $\xi_{1:k}$, $\sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k}) > 0$. Then define

$$u_k^{s_{k+1}}(\xi_{1:k+1}) := \frac{w_k^{s_{k+1}}(\xi_{1:k+1})}{\sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k})}, \qquad \forall \xi_{k+1} \in \Xi_{k+1}. \tag{4.55}$$

  As a result, the set of constraints corresponding to stages 2 to K is

$$\int\limits_{\Xi_{k+1}} \sum_{s_{k+1}=1}^{N_{k+1}} u_k^{s_{k+1}}(\xi_{1:k+1}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| d\xi_{k+1} \leq \epsilon_{k+1}, \tag{4.56a}$$

$$\int\limits_{\Xi_{k+1}} u_k^{s_{k+1}}(\xi_{1:k+1}) d\xi_{k+1} = \hat{p}_{k+1}^{s_{k+1}},\ \forall s_{k+1} \in [N_{k+1}], \tag{4.56b}$$

$$u_k^{s_{k+1}}(\xi_{1:k+1}) \geq 0,\ \forall \xi_{k+1} \in \Xi_{k+1},\ s_{k+1} \in [N_{k+1}]. \tag{4.56c}$$

Introducing variables $P_k^{\xi_{1:k-1}}(\xi_k),\ \forall \xi_k \in \Xi_k$ and adding constraints

$$\sum_{s_k=1}^{N_k} u_{k-1}^{s_k}(\xi_{1:k}) = P_k^{\xi_{1:k-1}}(\xi_k),\ \forall \xi_k \in \Xi_k, \tag{4.57}$$

the set of constraints above results in being equivalent to a Wasserstein ambiguity set $\mathcal{P}_k$.

When $\sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k}) = 0$, the solution is trivially $w_k^{s_{k+1}}(\xi_{1:k+1}) = 0,\ \forall \xi_{k+1} \in \Xi_{k+1},\ s_{k+1} \in [N_{k+1}]$.

This analysis gives rise to the following interpretation of the variables:

$$\begin{aligned} u_{k-1}^{s_k}(\xi_{1:k}) &= \mathbb{P}\left(\hat{\xi}_k^{s_k}, \xi_k\right). \\ w_{k-1}^{s_k}(\xi_{1:k}) &= \mathbb{P}\left(\hat{\xi}_k^{s_k}, \xi_{1:k}\right), \\ \bar{w}_{k-1}^{s_k}(\xi_{1:K}) &= \mathbb{P}\left(\hat{\xi}_k^{s_k}, \xi_{1:K}\right). \end{aligned}$$

From (4.48) and using the definition of $v_k(\xi_{1:K})$, we have

$$\begin{aligned} \sum_{j=1}^{k-1} \sum_{s_{j+1}}^{N_{j+1}} \bar{u}_j^{s_{j+1}}(\xi_{1:K}) &= \sum_{s_k=1}^{N_k} \mathbb{P}\left(\hat{\xi}_k^{s_k}, \xi_{1:K}\right) \\ &= \mathbb{P}\left(\xi_{1:K}\right) \\ &= P_2^{\xi_1}(\xi_2) P_3^{\xi_{1:2}}(\xi_3) \cdots P_K^{\xi_{1:K-1}}(\xi_K). \end{aligned}$$

The subproblem is therefore,

$$\min \quad \bar{\mu}_1^\top(\xi_{1:K})\breve{x}_1 + \sum_{k=2}^{K} \left(\bar{\mu}_k^\top(\xi_{1:K}) + \left(P_2^{\xi_1}(\xi_2) P_3^{\xi_{1:2}}(\xi_3) \cdots P_K^{\xi_{1:K-1}}(\xi_K)\right) c_k^\top(\xi_k)\right)\breve{x}_k \tag{4.58a}$$

$$\text{s.t.} \quad (4.31\text{d}),\ (4.31\text{e}).$$

Let us further define

$$\mu_1(\xi_{1:K}) := \bar{\mu}_1(\xi_{1:K}) \tag{4.59a}$$

$$\mu_k(\xi_{1:K}) := \bar{\mu}_k^\top(\xi_{1:K}) + \Big(P_2^{\xi_1}(\xi_2)P_3^{\xi_{1:2}}(\xi_3)\cdots P_K^{\xi_{1:K-1}}(\xi_K)\Big)\, c_k^\top(\xi_k) \tag{4.59b}$$

The subproblem becomes

$$\begin{aligned} \min \quad & \sum_{k=1}^{K} \mu_k^\top(\xi_{1:K})\breve{x}_k && (4.60a)\\ \text{s.t.} \quad & (4.31\text{d}),\ (4.31\text{e}). \end{aligned}$$

The constraints (4.30b) and (4.30c) are therefore,

$$\int\limits_{\Xi_{2:K}} \mu_1(\xi_{1:K})d\xi_{2:K} = c_1(\xi_1), \tag{4.61a}$$

$$\begin{aligned} &\int\limits_{\Xi_{k+1:K}} \mu_k(\xi_{1:K})d\xi_{k+1:K} \\ &= \int\limits_{\Xi_{k+1:K}} P_2^{\xi_1}(\xi_2)P_3^{\xi_{1:2}}(\xi_3)\cdots P_K^{\xi_{1:K-1}}(\xi_K)d\xi_{k+1:K}c_k^\top(\xi_k),\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2,\dots,K. \end{aligned} \tag{4.61b}$$

By combining with the constraints $P_{k+1}^{\xi_{1:k}} \in \mathcal{P}_{k+1}(\xi_{1:k}),\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2,\dots,K-1$, the above formulation is equivalent to the formulation in Proposition IV.3.

□

This equivalent formulation gives a linearization scheme for (4.20):

Corollary IV.6. *Assuming a Wasserstein ball ambiguity set* (4.2), *we can linearize*

(4.20) *by*

$$\max \int\limits_{\Xi_{1:K}} D'(\mu_{1:K}(\xi_{1:K}), \xi_{1:K}) d\xi_{1:K} \tag{4.62a}$$

$$s.t. \int\limits_{\Xi_{2:K}} \mu_1(\xi_{1:K}) d\xi_{2:K} = c_1(\xi_1), \tag{4.62b}$$

$$\int\limits_{\Xi_{k+1:K}} \mu_k(\xi_{1:K}) d\xi_{k+1:K} = c_k(\xi_k) P_{1:k}(\xi_{1:k}),\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \dots, K, \tag{4.62c}$$

$$\int\limits_{\Xi_2} \sum_{s_2=1}^{N_2} w_1^{s_2}(\xi_{1:2}) \left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| d\xi_2 \le \epsilon_2, \tag{4.62d}$$

$$\int\limits_{\Xi_{k+1}} \sum_{s_{k+1}=1}^{N_{k+1}} w_k^{s_{k+1}}(\xi_{1:k+1}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| d\xi_{k+1} \le \epsilon_{k+1} P_{1:k}(\xi_{1:k}),$$

$$\forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \dots, K-1, \tag{4.62e}$$

$$\int\limits_{\Xi_2} w_1^{s_2}(\xi_{1:2}) d\xi_2 = \hat{p}_2^{s_2},\ \forall s_2 \in [N_2], \tag{4.62f}$$

$$\int\limits_{\Xi_{k+1}} w_k^{s_{k+1}}(\xi_{1:k+1}) d\xi_{k+1} = \hat{p}_{k+1}^{s_{k+1}} P_{1:k}(\xi_{1:k}),$$

$$\forall \xi_{1:k} \in \Xi_{1:k},\ s_{k+1} \in [N_{k+1}],\ k = 2, \dots, K-1, \tag{4.62g}$$

$$\sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k}) = P_{1:k}(\xi_{1:k}),\ \forall \xi_{1:k} \in \Xi_{1:k},\ k = 2, \dots, K, \tag{4.62h}$$

$$w_k^{s_{k+1}}(\xi_{1:k+1}) \ge 0,\ \forall \xi_{1:k+1} \in \Xi_{1:k+1},\ s_{k+1} \in [N_{k+1}],\ k = 1, \dots, K-1, \tag{4.62i}$$

where

$$D'(\mu_{1:K}(\xi_{1:K}), \xi_{1:K}) := \min \quad \sum_{k=1}^{K} \mu_k^\top(\xi_{1:K}) x_k \tag{4.63a}$$

$$s.t. \quad W_k(\xi_k) x_k \ge h_k(\xi_k) - T_k(\xi_k) x_{k-1},\ \forall k = 2, \dots, K, \tag{4.63b}$$

$$x_k \in X_k,\ \forall k = 1, \dots K. \tag{4.63c}$$

Note that this formulation is still computationally challenging due to the infinite number of variables and constraints that exist in the problem. In the next section, we describe the algorithmic approach by using a sample average approximation (SAA).

4.5 Algorithms

In this section, we discuss the algorithmic process for dual decomposition. We consider an SAA approach, where $\tilde{N}_k$ i.i.d. random samples $(\xi_k^1, \xi_k^2, \ldots, \xi_k^{\tilde{N}_k}) := \tilde{\Xi}_k^{\tilde{N}_k} \subset \Xi_k$ are used instead. Furthermore, we denote a sample path as $\xi_{1:K} := (\xi_1^{i_1}, \ldots, \xi_K^{i_K}) \in \tilde{\Xi}_{1:K}^{\tilde{N}_{1:K}} := \tilde{\Xi}_1^{\tilde{N}_1} \times \ldots \times \tilde{\Xi}_K^{\tilde{N}_K}$.

The SAA reformulation of (4.62) given by

$$\underline{z}^{LD} =$$

$$\max \quad \sum_{\xi_{1:K} \in \tilde{\Xi}_{1:K}^{\tilde{N}_{1:K}}} D'(\mu_{1:K}(\xi_{1:K}), \xi_{1:K}) \tag{4.64a}$$

$$\text{s.t.} \quad \sum_{\xi_{2:K} \in \tilde{\Xi}_{2:K}^{\tilde{N}_{2:K}}} \mu_1(\xi_{1:K}) = c_1(\xi_1), \tag{4.64b}$$

$$\sum_{\xi_{k+1:K} \in \tilde{\Xi}_{k+1:K}^{\tilde{N}_{k+1:K}}} \mu_k(\xi_{1:K}) - c_k(\xi_k) P_{1:k}(\xi_{1:k}) = 0, \ \forall \xi_{1:k} \in \tilde{\Xi}_{1:k}^{\tilde{N}_{1:k}}, \ k = 2, \ldots, K, \tag{4.64c}$$

$$\sum_{\xi_2 \in \tilde{\Xi}_2^{\tilde{N}_2}} \sum_{s_2=1}^{N_2} w_1^{s_2}(\xi_{1:2}) \left\| \hat{\xi}_2^{s_2} - \xi_2 \right\| d\xi_2 \leq \epsilon_2, \tag{4.64d}$$

$$\sum_{\xi_{k+1} \in \tilde{\Xi}_{k+1}^{\tilde{N}_{k+1}}} \sum_{s_{k+1}=1}^{N_{k+1}} w_k^{s_{k+1}}(\xi_{1:k+1}) \left\| \hat{\xi}_{k+1}^{s_{k+1}} - \xi_{k+1} \right\| - \epsilon_{k+1} P_{1:k}(\xi_{1:k}) \leq 0,$$
$$\forall \xi_{1:k} \in \tilde{\Xi}_{1:k}^{\tilde{N}_{1:k}}, \ k = 2, \ldots, K-1, \tag{4.64e}$$

$$\sum_{\xi_2 \in \tilde{\Xi}_2^{\tilde{N}_2}} w_1^{s_2}(\xi_{1:2}) = \hat{p}_2^{s_2}, \ \forall s_2 \in [N_2], \tag{4.64f}$$

$$\sum_{\xi_{k+1} \in \tilde{\Xi}_{k+1}^{\tilde{N}_{k+1}}} w_k^{s_{k+1}}(\xi_{1:k+1}) - \hat{p}_{k+1}^{s_{k+1}} P_{1:k}(\xi_{1:k}) = 0,$$
$$\forall \xi_{1:k} \in \tilde{\Xi}_{1:k}^{\tilde{N}_{1:k}}, \ s_{k+1} \in [N_{k+1}], \ k = 2, \ldots, K-1, \tag{4.64g}$$

$$\sum_{s_k=1}^{N_k} w_{k-1}^{s_k}(\xi_{1:k}) - P_{1:k}(\xi_{1:k}) = 0, \ \forall \xi_{1:k} \in \tilde{\Xi}_{1:k}^{\tilde{N}_{1:k}}, \ k = 2, \ldots, K, \tag{4.64h}$$

$$w_k^{s_{k+1}}(\xi_{1:k+1}) \geq 0, \ \forall \xi_{1:k+1} \in \tilde{\Xi}_{1:k+1}^{\tilde{N}_{1:k+1}}, \ s_{k+1} \in [N_{k+1}], \ k = 1, \ldots, K-1. \tag{4.64i}$$

This can be solved by using algorithms such as the proximal method (*Kim and Dandurand*, 2018), or the trust-region method (*Kim et al.*, 2019).

We follow *Carøe and Schultz* (1999) for the branch-and-bound algorithm, which is a deterministic algorithm for obtaining an optimal solution. The dual decomposition gives a lower bound to the optimal objective value, and the scenario solutions $\hat{x}(\xi_{1:K})$ do not satisfy the non-anticipativity constraints unless the duality gap is zero. In the following, we let $\mathcal{I}$ be the list of incumbent problems with z_i being the lower bound associated with problem $I_i \in \mathcal{I}$. The algorithm is given in Algorithm 5.

Algorithm 5 *Carøe and Schultz* (1999) branch-and-bound algorithm

```
Initialize: $\mathcal{I} = \{I_1\}$, $\bar{z} = \infty$, $z_1 = -\infty$
repeat
  Select and delete problem $I_i$ from $\mathcal{I}$, and obtain the Lagrangian relaxation $\underline{z}_i^{LD}$.
  if $\underline{z}_i^{LD} < \bar{z}$ then
    if All scenario solutions are identical then
      Calculate the objective $\hat{z}_i$.
      Update $\bar{z} = \min\{\bar{z}, \hat{z}_i\}$
      Eliminate all problems in $\mathcal{I}$ with $z_i \geq \bar{z}$.
    else
      Get the average value $\bar{x}$ and use heuristics to obtain feasible solution $\bar{x}^R$.
      Calculate the objective $\hat{z}_i$.
      Update $\bar{z} = \min\{\bar{z}, \hat{z}_i\}$
      Eliminate all problems in $\mathcal{I}$ with $z_i \geq \bar{z}$.
      Select an inconsistent variable $x$.
      Create two new problems from $I_i$ with the associated lower bound $\underline{z}_i^{LD}$,
      and an additional constraint $x \leq \lfloor \bar{x} \rfloor$ or $x \geq \lfloor \bar{x} \rfloor + 1$. Add to $\mathcal{I}$.
    end if
  end if
until $\mathcal{I} = \emptyset$
```

4.6 Computational Study on Transmission Expansion Problem with Hydro Storage

The transmission Expansion Planning (TEP) problem aims to improve and update the electricity transmission infrastructure to adapt to the changes of load and generation in power systems by minimizing the cost of expanding existing transmission circuits for future operation. Recently in February 2021, there has been a massive electricity generation failure in Texas, caused by a series of severe winter storms. This was in the tail scenario where the demand for electricity was high, but multiple generation facilities had failed due to frozen power equipment at the same time (*Penney*, 2021). Another cause of the failure was because the power grid in Texas is isolated from the the other two major national grids. Combining with the recent trend of increasing risk in disasters caused by climate change (*NOAA National Centers for Environmental Information (NCEI)*, 2021), we are interested in a risk-aware planning for the transmission expansion problem.

We refer to *Romero et al.* (2002) for the DC model formulation of the TEP problem. In addition, we consider the pumped hydroelectric energy storage system which is able to store the excess energy and release it when it becomes necessary. This adds an extra set of continuous state variables for the multistage stochastic model, which makes it difficult to solve using SDDP or SDDiP methods.

4.6.1 Formulation

4.6.1.1 Notations

We consider an electric grid with n buses and define E as the set of all right-of-ways connecting buses. Let S be the node-branch incidence matrix with dimension $n \times |E|$. Let T be the number of planning stages. For each right-of-way $(i, j) \in E$, let n_{ij}^0 denote the initial number of lines between bus i and bus j and $\bar{n}_{ij}$ be the maximum number of lines allowed between bus i and bus j. For each line between bus i and bus j, we denote the susceptance of the line by γ_{ij}, the cost to add a new line by c_{ij} and the maximum power flow by $\bar{f}_{ij}$. We use d to represent the discount factor per quarter-year throughout the planning horizon. At some buses, there are hydroelectric reservoirs that are able to generate electricity or store water by consuming the power in the grid using a pump. For each reservoir location i, the efficiency of the hydropower generation is given by η_{Gi}, the efficiency of the pumping process is given by η_{Pi}, and the water that remains after evaporation per unit stage is given by η_{Ri}^t, which may change over time. The capacity of the reservoir i is denoted as RC_i and the initial reservoir level is denoted as RL_i.

For simplicity, we treat the value of maximum power generation G_i as deterministic and fixed, while the load $D_i(\omega_t)$ is uncertain and changes over time.

Let $x^t = (x_{ijk}^t, (i,j) \in E, k = 1, \ldots, \bar{n}_{ij})^\top$ be a binary decision vector such that $x_{ijk}^t = 1$ if we decide to construct the k-th line in (i,j) right-of-way in stage t and $x_{ijk}^t = 0$ otherwise. Let $g^t \in \mathbb{R}_+^n, \theta^t \in \mathbb{R}^n, DC^t \in \mathbb{R}_+^n$ be recourse decision vectors, each of dimension n, representing the power generation, voltage angle, and load curtailment at each bus in stage t, respectively; $f^t \in \mathbb{R}^{|E|}$ is the vector of maximum power flow on each of right-of-ways.

We let $y^t = (y_i^t, i = 1, \ldots, n)^\top$ be a decision vector corresponding to the reservoir level at the end of stage t at bus i. The unit is the same as the one for the power flow. At the beginning of the stage, the level of the reservoir i is $\eta_{Ri}^t y_i^{t-1}$. The decision maker has the choice to process b_{Gi} of water and generate electricity, or consume b_{Pi} of power to store water in the reservoir.

We present a summary of notations below.

Parameters

n:	number of buses in a given transmission network
E:	set of all right-of-ways connecting buses
T:	number of planning stages, i.e., the length of the planning
S:	node-branch incidence binary matrix with size $n \times \|E\|$
c_{ij}:	cost of a line added to the $i-j$ right-of-way (\$)
γ_{ij}:	susceptance of the line between buses i and j
n_{ij}^{0}:	initial number of lines between buses i and j
$\bar{n}_{ij}$:	maximum allowable number of lines that can be added to the $i-j$ right-of-way
p_D^t:	unit penalty for load curtailment in stage t
$\bar{f}_{ij}$:	maximum power flow on $i-j$ right-of-way per line
d:	annual discount factor
η_{Gi}:	efficiency of hydropower generation at bus i
η_{Pi}:	efficiency of pumping process at bus i
η_{Ri}^t:	rate of change of the water reservoir at bus i in stage t
RC_i:	maximum reservoir capacity at bus i
RL_i:	initial reservoir level at bus i
G_i:	amount of maximum power generation at bus i
$D^t(\omega_t)$ $= \left(D_i^t(\omega_t),\ i = 1, \ldots, n\right)^\top$:	vector of the amount of stochastic load $D_i^t(\omega_t)$ at bus i in stage t with event ω_t

Decision Variables

$x_{ijk}^t \in \{0,1\}$: binary variable indicating whether or not to install the k^{th} line of the $i-j$ right-of-way in stage t, such that $x_{ijk}^t = 1$ if yes and $x_{ijk}^t = 0$ otherwise.

$y_i^t \geq 0$: reservoir level at the end of stage t at bus i

Recourse Variables

$g_i^t \geq 0$: power generation at bus i in stage t

f_{ijk}^t: power flow on the k^{th} line of the $i-j$ right-of-way in stage t

$-\frac{\pi}{2} \leq \theta_i^t \leq \frac{\pi}{2}$: voltage angle at bus i in stage t

$DC_i^t \geq 0$: load curtailment at bus i in stage t

$b_{Gi}^t \geq 0$: power generation from reservoir at bus i

$b_{Pi}^t \geq 0$: pumping quantity at bus i

4.6.1.2 Distributionally Robust Multistage Problem Formulation

The goal is to minimize the present value of the investment while minimizing the expected penalty of load curtailment. The problem can be formulated as follows.

$$\min \sum_{(i,j)\in E} \sum_{k=n_{ij}^0}^{\bar{n}_{ij}} c_{ij} x_{ijk}^1 + \max_{P_2 \in \mathcal{P}_2} \mathbb{E}_{\xi_2} \left[Q_2(x^1, y^1, \xi_2) \right] \tag{4.65a}$$

$$\text{s.t. } x_{ijk}^1 = 1 \qquad \forall (i,j) \in E,\ k = 1, \ldots, n_{ij}^0 \tag{4.65b}$$

$$x_{ijk}^1 \leq x_{ij,k-1}^1 \qquad \forall (i,j) \in E,\ k = n_{ij}^0 + 1, \ldots, \bar{n}_{ij} \tag{4.65c}$$

$$x_{ijk}^1 \text{ binary} \qquad \forall (i,j) \in E,\ k = 1, \ldots, \bar{n}_{ij}, \tag{4.65d}$$

$$y_i^1 = RL_i \qquad \forall i = 1, \ldots, n. \tag{4.65e}$$

In above formulation, the objective function (4.65a) minimizes the present value of initial capacity expansion investment plus the expectation of future expenditure Q_2 starting in stage 2. Constraints (4.65b) set up the initial transmission lines that we have. Constraints (4.65c) enforce that we plan the construction of lines from lower index to higher index on each right-of-way to avoid symmetric solutions. Constraints (4.65d) enforce x^1 being binary decision variables. Constraints (4.65e) initializes the water level of the reservoir.

For all $t = 2, \dots, T-1$, we have

$$
\begin{aligned}
Q_t(x^{t-1}, & y^{t-1}, \xi_t) = \\
& \min\; d^{t-1} \left(\sum_{(i,j)\in E} \sum_{k=1}^{\bar{n}_{ij}} c_{ij}(x_{ijk}^t - x_{ijk}^{t-1}) + \sum_{i=1}^{n} p_D^t DC_i^t \right) \\
& \qquad\qquad\qquad + \max_{P_{t+1}\in\mathcal{P}_{t+1}} \mathbb{E}_{\xi_{t+1}} \left[Q_{t+1}(x^t, y^t, \xi_{t+1}) \right] && \text{(4.66a)} \\
& \text{s.t. } Sf^t + g^t + DC^t + \eta_G \bullet b_G^t - b_P^t = D^t(\omega_t) && \text{(4.66b)} \\
& \quad \sum_{k=1}^{n_{ij}^0} f_{ijk}^t - \gamma_{ij} n_{ij}^0 (\theta_i^t - \theta_j^t) = 0 \qquad \forall (i,j) \in E && \text{(4.66c)} \\
& \quad f_{ijk}^t - \gamma_{ij}(\theta_i^t - \theta_j^t) \le M(1 - x_{ijk}^{t-1}) \qquad \forall (i,j) \in E, k = n_{ij}^0 + 1, \dots, \bar{n}_{ij} && \text{(4.66d)} \\
& \quad f_{ijk}^t - \gamma_{ij}(\theta_i^t - \theta_j^t) \ge -M(1 - x_{ijk}^{t-1}) \qquad \forall (i,j) \in E, k = n_{ij}^0 + 1, \dots, \bar{n}_{ij} && \text{(4.66e)} \\
& \quad f_{ijk}^t \le \bar{f}_{ij} x_{ijk}^{t-1} \qquad \forall (i,j) \in E,\ k = 1, \dots, n_{ij}^t && \text{(4.66f)} \\
& \quad -f_{ijk}^t \le \bar{f}_{ij} x_{ijk}^{t-1} \qquad \forall (i,j) \in E,\ k = 1, \dots, n_{ij}^t && \text{(4.66g)} \\
& \quad 0 \le g^t \le G && \text{(4.66h)} \\
& \quad 0 \le DC^t \le D^t(\omega_t) && \text{(4.66i)} \\
& \quad -\frac{\pi}{2} \le \theta_i^t \le \frac{\pi}{2} \qquad \forall i = 1, \dots, n && \text{(4.66j)} \\
& \quad x_{ijk}^t \ge x_{ijk}^{t-1} \qquad \forall (i,j) \in E,\ k = 1, \dots, \bar{n}_{ij} && \text{(4.66k)} \\
& \quad x_{ijk}^t \le x_{ij,k-1}^t \qquad \forall (i,j) \in E,\ k = 2, \dots, \bar{n}_{ij} && \text{(4.66l)} \\
& \quad x_{ijk}^t \text{ binary} \qquad \forall (i,j) \in E,\ k = 1, \dots, \bar{n}_{ij}, && \text{(4.66m)} \\
& \quad y_i^t = \eta_{Ri}^t y_i^{t-1} + \eta_{Pi} b_{Pi}^t - b_{Gi}^t \qquad \forall i = 1, \dots, n && \text{(4.66n)} \\
& \quad 0 \le y_i^t \le RC_i \qquad \forall i = 1, \dots, n && \text{(4.66o)} \\
& \quad 0 \le b_{Gi}^t \le \eta_{Ri}^t y_i^{t-1} \qquad \forall i = 1, \dots, n && \text{(4.66p)} \\
& \quad b_{Pi}^t \ge 0 \qquad \forall i = 1, \dots, n && \text{(4.66q)}
\end{aligned}
$$

Here, $\bullet$ is an element-wise multiplication of two vectors.

In above formulations, $Q_t(x^{t-1}, y^{t-1}, \xi_t)$ computes the minimum discounted construction and operational cost in stage t given the realization $\xi_t = D^t(\omega_t)$ accordingly to the objective function (4.66a). Constraints (4.66b) ensure flow balance in the DC power flow system, which model Kirchhoff's current law; constraints (4.66c) provide an expression of Ohm's law for the equivalent DC network for original network (without any expansion); constraints (4.66d) and (4.66e) express Ohm's law for the expanded DC network line by line and they require a sufficient large "big M" coefficient to ensure feasibility when $x_{ijk}^{t-1} = 0$; constraints (4.66f) and (4.66g) ensure power flow limits on transmission lines and transformers; constraints (4.66h) and (4.66i) provide power generation and demand limits, respectively; constraints (4.66k) link the expansion decision x^t with x^{t-1} such that if the line was used in stage $t-1$, then it should also be used in stage t; constraints (4.66l) and (4.66m) are analogy to constraints (4.65c) and (4.65d). Constraints (4.66n) dictate the change of reservoir levels during the stage; constraints (4.66o) ensure reservoir levels do not exceed the capacity; constraints (4.66p) ensure the power generation cannot exceed the available quantity at the initial part of the stage; constraints (4.66q) ensure non-negativity of the pump-back power.

Finally, for $t = T$, we have a similar formulation as model (4.66) except that we do not need to plan the expansion. Therefore, for $t = T$, we omit the expansion cost in objective function (4.66a) and constraints (4.66k)–(4.66m).

4.6.2 Numerical Instances

The Garver 6-bus system is a small size power system containing 6 buses and 15 transmission right-of-ways that can be added. The initial topology of the network is given in Figure 4.1 and the initial detailed data are given in Tables 4.1 and 4.2. The p.u. for reactance data considers a 100MW base.

Table 4.1: Garver 6-bus bus data

BusID	GenMax	Load
1	150	80
2	0	240
3	360	40
4	0	160
5	0	240
6	600	0

Table 4.2: Garver 6-bus branch data

From	To	n_{ij}^0	Reactance	$\bar{f}_{ij}$	Cost
1	2	1	0.4	100	40
1	3	0	0.38	100	38
1	4	1	0.6	80	60
1	5	1	0.2	100	20
1	6	0	0.68	70	68
2	3	1	0.2	100	20
2	4	1	0.4	100	40
2	5	0	0.31	100	31
2	6	0	0.3	100	30
3	4	0	0.59	82	59
3	5	1	0.2	100	20
3	6	0	0.48	100	48
4	5	0	0.63	75	63
4	6	0	0.3	100	30
5	6	0	0.61	78	61

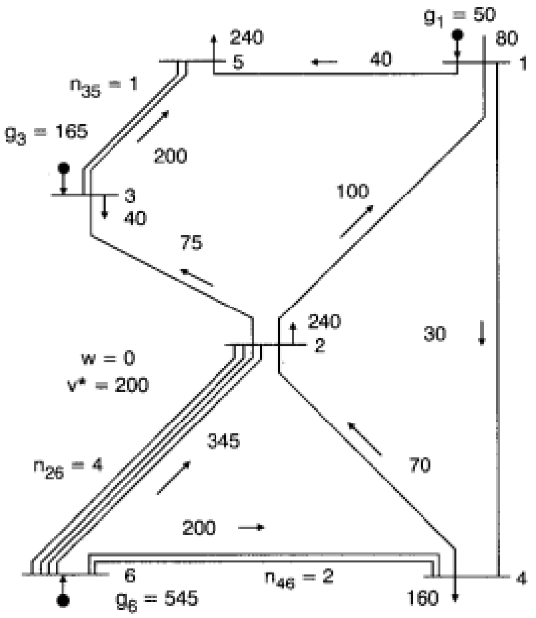


Figure 4.1: The topology of Garver 6-bus system.

4.6.2.1 Parameter Settings

The load takes values between 0.5 and 1.5 times the value given in Table 4.1 in the second stage. Then, the average load increases by 10% as t increases. We set the annual discount rate at 5%, and choose the penalty of load curtailment $p_D^t = \$10^4/MWh$. The hydroelectric reservoir only exists at bus 3, with the initial level being 50MW and the capacity being 200MW. The values η_{Gi}, η_{Pi}, η_{Ri} are set to 0.9.

4.6.2.2 Heuristics in the Branch-and-bound method

For the heuristic solution providing an upper bound of the optimal objective value, the binary variables x_{ijk}^t corresponding to installing the lines are averaged and rounded to the closest integer. However, we prioritize satisfying the constraints requiring $x_{ijk}^t = 1$ if $x_{ijk}^{t-1} = 1$, and $x_{ijk}^t = 0$ if $x_{ijk-1}^t = 0$. The reservoir level y_i are substituted with the maximum value among all the subproblem solutions $\hat{y}_i(\xi_{1:T})$.

The branching policy is a depth-first search, prioritizing variables in the earlier stage. Additionally, increasing the lines ($x_{ijk}^t = 1$) is prioritized over the other ($x_{ijk}^t =$

0) because the penalty for load curtailment is large.

4.6.3 Results

4.6.3.1 Solution Comparison

We focus on the case where there are 3 stages and 2 samples in each stage. We assume that the total load values are random and assume that individual loads in each bus have the same ratio as the standard load quantity in Table 4.1. Between the first and the second stage, the samples of the load values are either 1.0 or 1.2 times of the standard value. Similarly, between the second and the third stage, the samples of the load values are either 1.0 or 1.2 times the previous stage. Therefore, we have four trajectories with the multipliers: $[(1.0, 1.0), (1.0, 1.2), (1.2, 1.2), (1.2, 1.44)]$. Notice that the scenarios are not stage independent. We vary the value of $\epsilon \in \{1.0, 10.0, 50.0, 100.0\}$ and compare the results. In Table 4.3, we present the objective value and the anticipated probability of each trajectory. As expected, the DM anticipates a higher probability for the worst-case trajectory and is required to adjust the decisions to install more lines.

Table 4.3: Cost and probabilities for each sample trajectories

ϵ	$(1.0, 1.0)$		$(1.0, 1.2)$		$(1.2, 1.2)$		$(1.2, 1.44)$		Total
	Obj.	Prob.	Obj.	Prob.	Obj.	Prob.	Obj.	Prob.	Obj.
1.0	160.0	24.7%	160.0	24.7%	245.5	25.0%	245.5	25.6%	203.3
10.0	190.0	21.7%	190.0	21.7%	218.5	28.3%	218.5	28.3%	206.1
50.0	190.0	8.6%	190.0	8.6%	218.5	41.4%	218.5	41.4%	213.6
100.0	209.0	0.0%	209.0	0.0%	218.5	0.0%	218.5	100.0%	218.5

4.6.3.2 Computation Time

We compare the computation time for three sets of instances. We used 2.20GHz, 2201 Mhz, Intel(R) Xeon(R) CPU E5-2630 v4 with 10 Core(s), 20 Logical Processors on Windows Server 2012 R2, and implemented the algorithm with Julia 1.5 and

Gurobi 9.1 using package DualDecomposition.jl (*Kim et al.*, 2021). The first one is the same setting as in the previous section, changing the value of ϵ. The second test is changing the number of scenarios between $\{2, 3, 4\}$ for a 3-stage problem with $\epsilon = 10.0$. Finally, we change the number of stages between $\{2, 3, 4, 5\}$ with 2 scenarios each and $\epsilon = 10.0$. We report the computational time in Tables 4.4–4.6. When the time limit is reached, we report the optimality gap when the solution was found.

Table 4.4: Computation time for different ϵ values

ϵ	Time (s)	Gap (%)
1.0	172.2	0.0
10.0	90.8	0.0
50.0	214.6	0.0
100.0	3141.5	0.0

Table 4.5: Computation time for different scenario numbers

Scenario per stage	Time (s)	Gap (%)
2	90.8	0.0
3	825.6	0.0
4	2456.6	0.0
5	3600.0	6.1

Table 4.6: Computation time for different stage numbers

Stages	Time (s)	Gap (%)
2	31.5	0.0
3	90.8	0.0
4	1926.2	0.0
5	3600.0	–

Over the three sets of instances, the increase of computation time for different stage numbers is the fastest. There are 16 scenarios for a 5-stage 2-scenario problem, and there are also 16 scenarios for a 3-stage 4-scenario problem. The former took a longer time as the subproblems have more variables and constraints for longer stages. Moreover, the number of scenarios increases exponentially. We also observe that it generally takes a longer period of time when the radius ϵ increases as there were more variables to branch while running the algorithm.

4.7 Concluding Remarks

In this chapter, we formulated the dual decomposition method for multistage distributionally robust mixed-integer programming using a Wasserstein-based ambiguity set. We implemented a branch-and-bound algorithm combined with dual decomposition to solve a transmission expansion problem with hydro storage. In the future, we would like to investigate methods to further speed up the dual decomposition algorithm using the multistage structure of the problem.

CHAPTER V

Conclusion

In this dissertation, we focused on three different approaches to combining sequential decision making with distributionally robust optimization. In Chapter II, we proposed DRPOMDP and investigated the conditions where useful properties of POMDPs, such as the convexity of the value function, can also be used for the distributionally robust case. We adapted the HSVI method using the convex property of the value function to efficiently solve the infinite horizon problem. In Chapter III, we proved new theoretical guarantees on the Wasserstein distance of true and empirical distributions that are constructed from previously collected data. We then applied the theoretical bound to the regret-based reinforcement learning problem and empirically observed the advantage of using the Wasserstein distance based ambiguity set over the total variational distance. In Chapter IV, we adapted the two-stage distributionally robust MIP to the multistage stochastic MIP, extending the applicability of the dual decomposition algorithm. We discovered the multistage variant of the Wasserstein distance based ambiguity set, and implemented a branch-and-bound algorithm to solve the problem to optimality.

In the future research, we plan to improve the efficiency of the algorithms where problems with similar structures are solved repeatedly, and where data is provided iterative and online. We anticipate real-world applications of these algorithms in a large-scale, complex systems in energy infrastructure planning.

APPENDICES

APPENDIX A

Distributionally Robust Partially Observable Markov Decision Processes

A.1 Relaxation of a-rectangularity

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

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

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

$$\tilde{\mathcal{D}}_s = \left\{ \tilde{\mu}_s \begin{pmatrix} \boldsymbol{p}_s \\ \boldsymbol{r}_s \\ \tilde{\boldsymbol{u}}_s \end{pmatrix} \middle| \begin{array}{l} \mathbb{E}_{(\boldsymbol{p}_s,\boldsymbol{r}_s,\tilde{\boldsymbol{u}}_s)\sim\tilde{\mu}_s}\left[F_s\boldsymbol{p}_s + G_s\boldsymbol{r}_s + H_s\tilde{\boldsymbol{u}}_s\right] = \boldsymbol{c}_s, \\ \tilde{\mu}_s\left(\mathcal{X}_s\right) = 1 \end{array} \right\}, \tag{A.2}$$

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

$$\mathcal{X}_s = \left\{ \begin{pmatrix} \boldsymbol{p}_s \\ \boldsymbol{r}_s \\ \tilde{\boldsymbol{u}}_s \end{pmatrix} \in \begin{array}{c} \mathbb{R}^{|\mathcal{A}|\times|\mathcal{S}|\times|\mathcal{Z}|} \\ \mathbb{R}^{|\mathcal{A}|} \\ \mathbb{R}^{L} \end{array} \middle| \; B_s\boldsymbol{p}_s + C_s r_s + E_s\tilde{\boldsymbol{u}}_s \preceq_{K_s} \boldsymbol{d}_s \right\}. \tag{A.3}$$

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

The value function is also convex in the form (2.10), since for $t < T$,

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

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

A.2 General Ambiguity Set

In this section, we provide a general form of the ambiguity set where the mean values are on an affine manifold, and the supports are conic representable. For all

$a \in \mathcal{A}$ and $s \in \mathcal{S}$, we define a non-empty ambiguity set

$$\tilde{\mathcal{D}}_{as} = \left\{ \tilde{\mu}_{as} \begin{pmatrix} \boldsymbol{p}_{as} \\ r_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \middle| \begin{array}{l} \mathbb{E}_{(\boldsymbol{p}_{as}, r_{as}, \tilde{\boldsymbol{u}}_{as}) \sim \tilde{\mu}_{as}} \left[F_{as}\boldsymbol{p}_{as} + G_{as} r_{as} + H_{as}\tilde{\boldsymbol{u}}_{as} \right] = \boldsymbol{c}_{as}, \\ \tilde{\mu}_{as} \left(\mathcal{X}_{as} \right) = 1 \end{array} \right\}, \quad \text{(A.4)}$$

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

$$\mathcal{X}_{as} = \left\{ \begin{pmatrix} \boldsymbol{p}_{as} \\ r_{as} \\ \tilde{\boldsymbol{u}}_{as} \end{pmatrix} \in \begin{array}{c} \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Z}|} \\ \mathbb{R} \\ \mathbb{R}^L \end{array} \middle| \; B_{as}\boldsymbol{p}_{as} + C_{as} r_{as} + E_{as}\tilde{\boldsymbol{u}}_{as} \preceq_{K_{as}} \boldsymbol{d}_{as} \right\}. \quad \text{(A.5)}$$

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

A.3 Proofs of Theorems II.3 and II.4

First, we provide a detailed proof for Theorem II.3 below.

Proof. We show the result by induction. When $t = T$, $V^T(\boldsymbol{b}) = 0$ satisfies (2.10). For

$t < T$, the inner problem $Q^t(\boldsymbol{b}, a)$ described in (2.7) becomes

$$\begin{aligned}
\min_{\tilde{\mu}_a \in \mathcal{P}\left(\tilde{\mathcal{X}}_a\right)} \quad & \mathbb{E}_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \sim \tilde{\mu}_a} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1} \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) \right] && \text{(A.6a)} \\
\text{s.t.} \quad & \mathbb{E}_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \sim \tilde{\mu}_a} \left[\tilde{\boldsymbol{u}}_{as} \right] = \boldsymbol{c}_{as}, && \forall s \in \mathcal{S} \qquad \text{(A.6b)} \\
& \mathbb{E}_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \sim \tilde{\mu}_a} \left[I \left((\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as}) \in \tilde{\mathcal{X}}_{as} \right) \right] = 1, && \forall s \in \mathcal{S} \qquad \text{(A.6c)}
\end{aligned}$$

for all $a \in \mathcal{A}$. Here $I(\cdot)$ is an indicator function, such that if event $\cdot$ is true, it returns value 1 and 0 otherwise. Associating the dual variables $\boldsymbol{\rho}_{as}$ and ω_{as} with constraints (A.6b) and (A.6c), respectively, we formulate the dual of (A.6) as

$$\begin{aligned}
\max_{\boldsymbol{\rho}_a, \boldsymbol{\omega}_a} \quad & \sum_{s \in \mathcal{S}} \boldsymbol{c}_{as}^\top \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} \omega_{as} && \text{(A.7a)} \\
\text{s.t.} \quad & \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} \omega_{as} && \text{(A.7b)} \\
& \quad \leq \sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1} \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) \qquad \forall (\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a) \in \tilde{\mathcal{X}}_a \\
& \boldsymbol{\rho}_{as} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Z}|}, \ \omega_{as} \in \mathbb{R} && \forall s \in \mathcal{S}. \qquad \text{(A.7c)}
\end{aligned}$$

Constraints (A.7b) are further equivalent to the following inequality with a minimization problem on the right-hand side (RHS).

$$\begin{aligned}
& \sum_{s \in \mathcal{S}} \omega_{as} \leq && \text{(A.8a)} \\
& \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad \sum_{s \in \mathcal{S}} b_s \left(r_{as} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} V^{t+1} \left(\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z) \right) \right) - \sum_{s \in \mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} \\
& \quad \text{s.t.} \quad \tilde{\boldsymbol{u}}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as} && \forall s \in \mathcal{S} \qquad \text{(A.8b)} \\
& \quad \tilde{\boldsymbol{u}}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as} && \forall s \in \mathcal{S} \qquad \text{(A.8c)} \\
& \quad \mathbf{1}^\top \boldsymbol{p}_{as} = 1 && \forall s \in \mathcal{S} \qquad \text{(A.8d)} \\
& \quad \boldsymbol{p}_{as} \geq 0 && \forall s \in \mathcal{S}. \qquad \text{(A.8e)}
\end{aligned}$$

Substituting (2.10) for V^{t+1} and (2.1) for $\boldsymbol{f}(\boldsymbol{b}, a, \boldsymbol{p}_a, z)$, we obtain

$$\begin{aligned}\text{RHS of (A.8)} = \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad & \sum_{s\in\mathcal{S}} b_s r_{as} + \beta \sum_{z\in\mathcal{Z}} \max_{\boldsymbol{\alpha}_{az}\in\Lambda^{t+1}} \left[\boldsymbol{\alpha}_{az}^\top \sum_{s\in\mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s\right] - \sum_{s\in\mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} \\ \text{s.t.} \quad & \text{(A.8b)–(A.8e).}\end{aligned} \tag{A.9}$$

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

$$\begin{aligned}\text{(A.9)} = \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad & \max_{\substack{\boldsymbol{\alpha}_{az}\in\text{Conv}(\Lambda^{t+1}) \\ \forall z\in\mathcal{Z}}} \left[\sum_{s\in\mathcal{S}} b_s r_{as} + \beta \sum_{z\in\mathcal{Z}} \boldsymbol{\alpha}_{az}^\top \sum_{s\in\mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s - \sum_{s\in\mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as}\right] \\ \text{s.t.} \quad & \text{(A.8b)–(A.8e)}\end{aligned} \tag{A.10}$$

The expression in the bracket is convex (linear) in $(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)$ for fixed $\boldsymbol{\alpha}_{az},\ z \in \mathcal{Z}$, and concave (affine) in $\boldsymbol{\alpha}_{az},\ z \in \mathcal{Z}$ given fixed values of $(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)$. Moreover, (A.8b)–(A.8e) and $\text{Conv}\left(\Lambda^{t+1}\right)$ are convex sets. The minimax theorem (see, e.g., *Osogami* (2015), *Du and Pardalos* (2013)) ensures that the problem is equivalent to

$$\begin{aligned}\text{(A.10)} = \max_{\substack{\boldsymbol{\alpha}_{az}\in\text{Conv}(\Lambda^{t+1}) \\ \forall z\in\mathcal{Z}}} \min_{(\boldsymbol{p}_a, \tilde{\boldsymbol{u}}_a)} \quad & \sum_{s\in\mathcal{S}} b_s r_{as} + \beta \sum_{z\in\mathcal{Z}} \boldsymbol{\alpha}_{az}^\top \sum_{s\in\mathcal{S}} \boldsymbol{J}_z \boldsymbol{p}_{as} b_s - \sum_{s\in\mathcal{S}} \tilde{\boldsymbol{u}}_{as}^\top \boldsymbol{\rho}_{as} \\ \text{s.t.} \quad & \text{(A.8b)–(A.8e)}\end{aligned} \tag{A.11}$$

We take the dual of the inner minimization by associating dual variables $\boldsymbol{\kappa}_{as}^1$, $\boldsymbol{\kappa}_{as}^2$, σ_{as} with constraints (A.8b)–(A.8d), respectively. We thus have the following equivalence:

$$
\begin{aligned}
\text{(A.11)} = \max_{\substack{\boldsymbol{\alpha}_{az} \in \text{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \max_{\boldsymbol{\kappa}_a^1, \boldsymbol{\kappa}_a^2, \boldsymbol{\sigma}_a} \quad & \sum_{s \in \mathcal{S}} b_s r_{as} + \sum_{s \in \mathcal{S}} \left(-\bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^1 + \bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^2 + \sigma_{as} \right) && \text{(A.12a)} \\
\text{s.t.} \quad & \beta b_s \sum_{z \in \mathcal{Z}} \boldsymbol{J}_z{}^\top \boldsymbol{\alpha}_{az} + \boldsymbol{\kappa}_{as}^1 - \boldsymbol{\kappa}_{as}^2 - \mathbf{1}\sigma_{as} \geq 0, \quad \forall s \in \mathcal{S} && \text{(A.12b)} \\
& \boldsymbol{\kappa}_{as}^1 + \boldsymbol{\kappa}_{as}^2 + \boldsymbol{\rho}_{as} = 0, \quad \forall s \in \mathcal{S} && \text{(A.12c)} \\
& \boldsymbol{\kappa}_{as}^1, \boldsymbol{\kappa}_{as}^2 \in \mathbb{R}_+^{|\mathcal{S}| \times |\mathcal{Z}|}, \sigma_{as} \in \mathbb{R}, \quad \forall s \in \mathcal{S}, && \text{(A.12d)}
\end{aligned}
$$

Due to (A.8), we substitute $\sum_{s \in \mathcal{S}} \omega_{as}$ in the objective function (A.7a) with (A.12). As a result, the value function (2.5) is equivalent to

$$
\begin{aligned}
V^t(\boldsymbol{b}) = \max_{a \in \mathcal{A}} \max_{\substack{\boldsymbol{\alpha}_{az} \in \text{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} & && \text{(A.13a)} \\
\max_{\boldsymbol{\rho}_a, \boldsymbol{\kappa}_a^1, \boldsymbol{\kappa}_a^2, \boldsymbol{\sigma}_a} \quad & \sum_{s \in \mathcal{S}} \boldsymbol{c}_{as}^\top \boldsymbol{\rho}_{as} + \sum_{s \in \mathcal{S}} b_s r_{as} + \sum_{s \in \mathcal{S}} \left(-\bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^1 + \bar{p}_{as}^\top \boldsymbol{\kappa}_{as}^2 + \sigma_{as} \right) && \\
\text{s.t.} \quad & \text{(A.12b)–(A.12d)} && \\
& \boldsymbol{\rho}_{as} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{Z}|} \ \forall s \in \mathcal{S}, && \text{(A.13b)}
\end{aligned}
$$

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

$$
V^t(\boldsymbol{b}) = \max_{a \in \mathcal{A}} \max_{\substack{\boldsymbol{\alpha}_{az} \in \text{Conv}(\Lambda^{t+1}) \\ \forall z \in \mathcal{Z}}} \sum_{s \in \mathcal{S}} b_s \times \Xi(a, \boldsymbol{\alpha}_{az} \ \forall z \in \mathcal{Z}, s), \tag{A.14}
$$

where

$$
\begin{aligned}
\Xi(a, \boldsymbol{\alpha}_{az} \ \forall z \in \mathcal{Z}, s) = \min_{(\boldsymbol{p}_{as}, \tilde{\boldsymbol{u}}_{as})} \quad & \beta \sum_{z \in \mathcal{Z}} \boldsymbol{\alpha}_{az}^\top \boldsymbol{J}_z \boldsymbol{p}_{as} + r_{as} && \text{(A.15a)} \\
\text{s.t.} \quad & \boldsymbol{c}_{as} \geq \boldsymbol{p}_{as} - \bar{\boldsymbol{p}}_{as} && \text{(A.15b)} \\
& \boldsymbol{c}_{as} \geq \bar{\boldsymbol{p}}_{as} - \boldsymbol{p}_{as} && \text{(A.15c)} \\
& \mathbf{1}^\top \boldsymbol{p}_{as} = 1 && \text{(A.15d)} \\
& \boldsymbol{p}_{as} \geq 0. && \text{(A.15e)}
\end{aligned}
$$

Defining set Λ^t as

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

it follows that the above value function in (A.14) is of the form (2.10). Furthermore, by induction, this is true for all t. This completes the proof. □

The proof of Theorem II.4 is given as follows.

Proof. Consider two arbitrary value functions V_1 and V_2. Given belief state $\boldsymbol{b}$, let

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

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

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

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

$$\begin{aligned}
&0 \leq \mathcal{L}V_1(\boldsymbol{b}) - \mathcal{L}V_2(\boldsymbol{b}) \\
&\quad = \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star}) \sim \mu_{a_1^\star,1}^\star} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{a_1^\star s} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right] \\
&\qquad - \mathbb{E}_{(\boldsymbol{p}_{a_2^\star}, \boldsymbol{r}_{a_2^\star}) \sim \mu_{a_2^\star,2}^\star} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{a_2^\star s} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_2^\star s} V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_2^\star, \boldsymbol{p}_{a_2^\star}, z) \right) \right) \right] \\
&\quad \leq \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star}) \sim \mu_{a_1^\star,2}^\star} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{a_1^\star s} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right] \\
&\qquad - \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star}) \sim \mu_{a_1^\star,2}^\star} \left[\sum_{s \in \mathcal{S}} b_s \left(r_{a_1^\star s} + \beta \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right] \\
&\quad = \beta \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star}) \sim \mu_{a_1^\star,2}^\star} \left[\sum_{s \in \mathcal{S}} b_s \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \times \left(V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, z, \boldsymbol{p}_{a_1^\star}) \right) - V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) \right) \right].
\end{aligned} \tag{A.16}$$

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

$$\begin{aligned}
\text{(A.16)} \leq \beta \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star}) \sim \mu^\star_{a_1^\star,2}} & \left[\sum_{s \in \mathcal{S}} b_s \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \times \left| V_1 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, \boldsymbol{p}_{a_1^\star}, z) \right) - V_2 \left(\boldsymbol{f}(\boldsymbol{b}, a_1^\star, z, \boldsymbol{p}_{a_1^\star}) \right) \right| \right] \\
\leq \beta \mathbb{E}_{(\boldsymbol{p}_{a_1^\star}, \boldsymbol{r}_{a_1^\star}) \sim \mu^\star_{a_1^\star,2}} & \left[\sum_{s \in \mathcal{S}} b_s \sum_{z \in \mathcal{Z}} \mathbf{1}^\top \boldsymbol{J}_z \boldsymbol{p}_{a_1^\star s} \sup_{\boldsymbol{b}' \in \Delta(\mathcal{S})} \left| V_1(\boldsymbol{b}') - V_2(\boldsymbol{b}') \right| \right] \\
= \beta \sup_{\boldsymbol{b}' \in \Delta(\mathcal{S})} & \left| V_1(\boldsymbol{b}') - V_2(\boldsymbol{b}') \right| .
\end{aligned}$$

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

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

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

and therefore,

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

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

BIBLIOGRAPHY

BIBLIOGRAPHY

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

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

Agrawal, R. (2020), Finite-sample concentration of the multinomial in relative entropy, *IEEE Transactions on Information Theory.*

Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath (1999), Coherent measures of risk, *Mathematical finance*, *9*(3), 203–228.

Azar, M. G., I. Osband, and R. Munos (2017), Minimax regret bounds for reinforcement learning, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 263–272, JMLR. org.

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

Birge, J. R., and F. Louveaux (2011), *Introduction to stochastic programming*, Springer Science & Business Media.

Bolley, F., and C. Villani (2005), Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities, *Annales de la Faculté des sciences de Toulouse: Mathématiques*, *14*(3), 331–352.

Boyd, S., S. P. Boyd, and L. Vandenberghe (2004), *Convex optimization*, Cambridge university press.

Carøe, C. C., and R. Schultz (1999), Dual decomposition in stochastic integer programming, *Operations Research Letters*, *24*(1-2), 37–45.

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

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

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

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

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

Duque, D., and D. P. Morton (2020), Distributionally robust stochastic dual dynamic programming, *SIAM Journal on Optimization*, *30*(4), 2841–2865.

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

Filippi, S., O. Cappe, and A. Garivier (2010), Optimism in reinforcement learning and kullback-leibler divergence, *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, doi:10.1109/allerton.2010.5706896.

Fournier, N., and A. Guillin (2015), On the rate of convergence in wasserstein distance of the empirical measure, *Probability Theory and Related Fields*, *162*(3-4), 707–738.

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

Gassmann, H. I. (1990), Mslip: A computer code for the multistage stochastic linear programming problem, *Mathematical Programming*, *47*(1), 407–423.

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

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

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

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

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

Jagtenberg, C., S. Bhulai, and R. van der Mei (2017), Optimal ambulance dispatching, in *Markov Decision Processes in Practice*, pp. 269–291, Springer.

Jaksch, T., R. Ortner, and P. Auer (2010), Near-optimal regret bounds for reinforcement learning, *Journal of Machine Learning Research*, *11*(Apr), 1563–1600.

Ji, R., and M. Lejeune (2018), Data-driven optimization of reward-risk ratio measures, *Available at SSRN 2707122*.

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

Kim, K. (2020), Dual decomposition of two-stage distributionally robust mixed-integer programming under the Wasserstein ambiguity set, Preprint manuscript.

Kim, K., and B. Dandurand (2018), Scalable branching on dual decomposition of stochastic mixed-integer programming problems, *To appear in Mathematical Programming Computation*.

Kim, K., C. G. Petra, and V. M. Zavala (2019), An asynchronous bundle-trust-region method for dual decomposition of stochastic mixed-integer programming, *SIAM Journal on Optimization*, *29*(1), 318–342.

Kim, K., H. Nakao, Y. Kim, M. Schanen, and W. Zhang (2021), DualDecomposition.jl: Parallel Dual Decomposition in Julia, doi:10.5281/zenodo.4574769.

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

Lattimore, T., and C. Szepesvári (2020), *Bandit algorithms*, Cambridge University Press.

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

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

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

NOAA National Centers for Environmental Information (NCEI) (2021), U.s. billion-dollar weather and climate disasters.

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

Papadimitriou, C. H., and J. N. Tsitsiklis (1987), The complexity of markov decision processes, *Mathematics of operations research*, *12*(3), 441–450.

Penney, V. (2021), How texas' power generation failed during the storm, in charts, *The New York Times*.

Pereira, M. V., and L. M. Pinto (1991), Multi-stage stochastic optimization applied to energy planning, *Mathematical programming*, *52*(1-3), 359–375.

Philpott, A. B., V. L. de Matos, and L. Kapelevich (2018), Distributionally robust sddp, *Computational Management Science*, *15*(3-4), 431–454.

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

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

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

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

Rockafellar, R. T., and S. Uryasev (2002), Conditional value-at-risk for general loss distributions, *Journal of banking & finance*, *26*(7), 1443–1471.

Romero, R., A. Monticelli, A. Garcia, and S. Haffner (2002), Test systems and mathematical models for transmission network expansion planning, *IEE Proceedings-Generation, Transmission and Distribution*, *149*(1), 27–36.

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

Shani, G., J. Pineau, and R. Kaplow (2013), A survey of point-based pomdp solvers, *Autonomous Agents and Multi-Agent Systems*, *27*(1), 1–51.

Shapiro, A., D. Dentcheva, and A. Ruszczyński (2009), *Lectures on stochastic programming: modeling and theory*, SIAM.

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

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

Tewari, A., and P. L. Bartlett (2008), Optimistic linear programming gives logarithmic regret for irreducible mdps, in *Advances in Neural Information Processing Systems*, pp. 1505–1512.

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

Villani, C. (2008), *Optimal transport: old and new*, vol. 338, Springer Science & Business Media.

Weissman, T., E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger (2003), Inequalities for the L1 deviation of the empirical distribution, *Hewlett-Packard Labs, Tech. Rep.*

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

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

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

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

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

Yu, X., and S. Shen (2020), Multistage distributionally robust mixed-integer programming with decision-dependent moment-based ambiguity sets, *Mathematical Programming*, pp. 1–40.

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

Zou, J., S. Ahmed, and X. A. Sun (2019), Stochastic dual dynamic integer programming, *Mathematical Programming*, *175*(1), 461–502.

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