Robust Markov decision processes with data-driven, distance-based ambiguity sets

Sivaramakrishnan Ramani

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Archis Ghate, Chair

Chiwei Yan

Chaoyue Zhao

Program Authorized to Offer Degree:
Industrial and Systems Engineering

University of Washington

Abstract

Robust Markov decision processes with data-driven, distance-based ambiguity sets

Sivaramakrishnan Ramani

Chair of the Supervisory Committee:
Professor Archis Ghate
Industrial & Systems Engineering

We consider finite- and infinite-horizon Markov decision processes (MDPs), with unknown state-transition probabilities. These transition probabilities are assumed to belong to certain ambiguity sets, and the goal is to maximize the worst-case expected total discounted reward over all probabilities from these sets. Specifically, the ambiguity set is a ball — it includes all probability distributions within a certain distance from the empirical distribution constructed using historical, independent observations of state transitions. We therefore call these problems robust MDPs (RMDPs) with data-driven, distance-based ambiguity sets.

The literature on data-driven robust optimization mentions (i) robust value convergence with respect to sample size, (ii) out-of-sample value convergence with respect to sample size, (iii) probabilistic performance guarantees on out-of-sample values, and (iv) a probabilistic convergence rate, as desirable properties. The research objective of this dissertation is to establish essentially a minimal set of conditions under which RMDPs with data-driven, distance-based ambiguity sets exhibit these four properties.

We first achieve this for the (s, a)-rectangular RMDPs ((s, a)-RMDPs) studied in Chapter 2. There, the ambiguity set for the whole MDP equals a Cartesian product of ambiguity sets for individual state-action pairs. We establish robust and out-of-sample value convergence under a generalized Pinsker's inequality, if the radii of the ambiguity balls approach zero as the sample-size diverges to infinity. This inequality links convergence of probability

distributions with respect to the distance function, to their topological convergence in a Euclidean space. We also establish that, for finite sample-sizes, the optimal value of the RMDP provides a lower bound on the value of the robust optimal policy in the true MDP, with a high probability. This probabilistic performance guarantee relies on a certain concentration inequality. Under these generalized Pinsker and concentration inequalities, we also derive a probabilistic rate of convergence for the robust and out-of-sample values to the true optimal value. These two inequalities hold for several well-known distance functions including total variation, Burg, Hellinger, and Wasserstein. We present computational results on a generic MDP and a machine repair example using total variation, Burg, and Wasserstein distances. These results illustrate that the generality of our framework provides broad choices when attempting to trade-off conservativeness of robust optimal policies against their out-of-sample performance by tuning ambiguity ball radii.

In Chapter 3, we extend results from Chapter 2 to a so-called s-rectangular framework. In this more general context, the ambiguity set for the MDP is a Cartesian product of ambiguity sets for individual states. In that chapter, we introduce a family of distance-based s-rectangular RMDPs (s-RMDPs) indexed with $\rho \in [1, \infty]$. In each state, the ambiguity set of transition probability mass functions (pmfs) across different actions equals a sublevel set of the ℓ_ρ-norm of a vector of distances from reference pmfs. Setting $\rho = \infty$ in this family recovers (s, a)-RMDPs. For any s-RMDP from this family, there is an (s, a)-RMDP whose robust optimal value is at least as good; and vice versa. This occurs because the s- and (s, a)-RMDPs can employ different ambiguity set radii, casting a nuanced doubt on the anecdotal belief that (s, a)-RMDPs are more conservative than s-RMDPs. More strongly, if the distance function is lower semicontinuous and convex, then, for any s-RMDP, there exists an (s, a)-RMDP with an identical robust optimal value. This suggests that appropriate caution should be exercised before interpreting too broadly any anecdotal claims that (s, a)-RMDPs are more conservative than s-rectangular ones. We also study data-driven versions of our s-RMDPs,

where the reference pmf equals the empirical pmf constructed from state transition samples. We prove that the robust optimal values, and the out-of-sample values of robust optimal policies both converge to the true optimal, asymptotically with sample sizes, if the distance function satisfies the generalized Pinsker's inequality introduced in Chapter 2. The robust optimal value also provides a probabilistic lower bound on the out-of-sample value of a robust optimal policy, when the distance function satisfies the concentration inequality. This finite-sample guarantee admits a surprising conclusion — (s, a)-RMDPs are the least conservative among all s-RMDPs within our family. Like in Chapter 2, under these generalized Pinsker and concentration inequalities, we also derive a probabilistic rate of convergence for the robust and out-of-sample values to the true optimal value. Though similar asymptotic and finite-sample results were developed for (s, a)-RMDPs in Chapter 2, the proof techniques in this chapter are different and more sophisticated. These more involved proofs are needed because the structure of s-RMDPs introduces new analytical hurdles in our attempt to establish the sufficiency of generalized Pinsker and concentration inequalities. For example, it is no longer adequate to limit attention to deterministic policies — randomization may be needed for optimality. We also present computational experiments on a machine repair example using the total variation distance and $\rho = 1$. The results of those experiments validate the claims established in that chapter.

Finally, in Chapter 4, we develop a data-driven, distance-based RMDP framework on separable complete metric spaces. We extend our asymptotic and finite-sample results to that setup. Unlike our first two studies, this more general endeavor relies on measure-theoretic concepts from minimax control.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1
INTRODUCTION

1.1 Background

Markov decision processes (MDPs) provide a systematic approach for modeling sequential decision problems under uncertain environments [50]. In MDPs, a decision-maker observes the state of a system at discrete time- points referred to as stages. Actions are taken at every stage after observing the state. Upon taking a particular action, the system stochastically transitions to a new state and collects an immediate reward (or incurs an immediate cost). It is assumed that the stochastic evolution of the system is governed by a Markov process whose transition probabilities are determined by the current state of the system and the action taken. The goal in the MDP is to choose a sequence of actions (typically, dependent on the state) so as to optimize a certain performance metric of interest.

The decision-maker often does not know the "true" transition probabilities which govern the evolution of the Markov process. Robust Markov decision processes (RMDPs) address this challenge via a worst-case approach. The decision-maker assumes that the unknown transition probabilities belong to certain ambiguity sets. The decision-maker wishes to maximize the smallest expected total discounted reward over all possible transition probabilities from these ambiguity sets. While some RMDP problems can be notoriously difficult [67, Table 1], certain classes of RMDP are more amenable to computation. Two common such classes are (s, a)-*rectangular* RMDPs [32, 46] and s-*rectangular* RMDPs [67]. In (s, a)-rectangular RMDPs, the ambiguity set equals a Cartesian product of ambiguity sets across state-action pairs. Ambiguity sets with this Cartesian product structure are termed (s, a)-rectangular. The other common class of RMDPs, namely the s-rectangular RMDPs, is a generalization of (s, a)-rectangular RMDPs. There, the ambiguity sets can be decomposed

as a Cartesian product of ambiguity sets across states (rather than state-action pairs as in (s, a)-rectangularity RMDPs).

The concept of (s, a)-rectangular RMDPs goes back to the early 1970s [54], and they have been applied to problems in healthcare [72]; portfolio optimization [23]; revenue management [53]; inventory control [37]; aircraft navigation [38]; and power systems [31]. RMDPs with ambiguity sets other than (s, a)- and s-rectangular ones have also been explored in the literature. A generalization of (s, a)-rectangularity, termed k-rectangularity, was developed in [42]. A robust counterpart of Bellman's backward recursion under k-rectangularity was developed for finite-horizon RMDPs. Another recent work [11] studied RMDPs under a broad class of ambiguity sets, both under the distance-based and moment-based frameworks. Moment-based ambiguity sets were employed in [45] to model a distributionally robust partially observable MDP. A so-called factor uncertainty model, quite different from other papers on RMDPs, was recently introduced in [26]. It was motivated by applications in healthcare, wherein the typical multiplicative separability of ambiguity sets across states or state-action pairs may not hold because the uncertainty is determined by shared underlying factors.

Our work fits within the broad area of distributionally robust optimization (DRO), where the decision-maker does not know the distribution of an uncertain parameter in a stochastic optimization problem [51]. The decision-maker then plans against the worst possible distribution within a family. This family is called the ambiguity set. DRO goes back to the late 1950s [55] and has recently emerged as a popular methodology for data-driven decision making. The literature on DRO includes two major approaches to model ambiguity sets: moment-based and distance-based. In the moment-based framework, the ambiguity set includes all distributions with some moment restrictions. Ambiguity sets in the distance-based approach include all distributions that are, in some sense, close to a reference distribution. We refer the reader to [13, 24, 55, 59, 68] and references therein for work on the moment-based approach. For distance-based methodology, refer to [6, 7, 17, 18, 21, 33, 34] and literature reviews therein. A generic data-driven distance-based framework using the Wasserstein metric was developed in [19]. That paper provided an asymptotic convergence analysis and a

finite-sample probabilistic guarantee for the specific context at hand. A similar approach was applied to two-stage stochastic programs in [73]. Ambiguity sets based on goodness-of-fit tests were proposed in [10] for robust Sample Average Approximation (SAA). A rigorous meta-model for finding an optimal framework that uses data to arrive at decisions in stochastic problems was formulated in [64]. The meta-model was shown to have a unique solution: the distributionally robust framework with ambiguity sets defined via the Burg distance. This idea was further generalized along several directions in [60].

1.2 Organization of the dissertation

Chapter 2 studies finite- and infinite-horizon (s, a)-rectangular RMDPs under data-driven distance-based ambiguity sets. We prove the convergence of robust and out-of-sample values to the true optimal value. We derive probabilistic performance guarantee on out-of-sample values. We also establish a probabilistic rate of convergence for the robust and the out-of-sample values to the true optimal value. We conclude the chapter by presenting results from computational experiments conducted using various distance functions. Chapter 3 studies distance-based RMDPs under s-rectangular ambiguity sets. We consider a family of s-rectangular RMDPs indexed with $\rho \in [1, \infty]$. In this family, the ambiguity set corresponding to each state is characterized as a sublevel set of the ℓ_ρ-norm of a vector of distances from reference pmfs. This construction allows an explicit connection between distance-based (s, a)-rectangular RMDP and distance-based s-rectangular RMDP from our family. By exploiting this connection, we undertake a rigorous study of the relative conservatiness properties between distance-based (s, a)- versus distance-based s-rectangular RMDPs. Apart from analysing this relative conservativeness behavior, we also study data-driven versions of s-rectangular RMDPs from this family. Like in Chapter 2, we show the asymptotic convergence of the robust and out-of-sample values and establish the finite-sample properties for this class of s-rectangular RMDPs. Though these results are a counterpart of the (s, a)-rectangular case, the proof techniques used to establish these results are different. This is because the structural properties of s-rectangular RMDPs are different from (s, a)-rectangular

RMDPs, and these differences require a different approach to establish the results. We wrap up the chapter by presenting a numerical study conducted on this class of s-rectangular RMDP wherein the ambiguity set is constructed using the Total Variation distance. Chapter 4 develops an RMDP framework on separable complete metric spaces. Asymptotic and finite-sample results akin to our first two studies are established using a measure-theoretic approach. We also present applications which fit into the modeling framework described in that chapter.

Chapter 2

DATA-DRIVEN ROBUST MARKOV DECISION PROCESSES USING (S, A)-RECTANGULAR AMBIGUITY SETS

2.1 Introduction

A finite-horizon Markov decision process (MDP) is described as follows [50]. A system is in state $i \in S \stackrel{\text{def}}{=} \{1, \cdots, n\}$ at the beginning of stage t. After observing this state, a decision-maker chooses an action $a \in A \stackrel{\text{def}}{=} \{1, \cdots, m\}$. The system then transitions into a state $j \in S$ with probability $p_t(j|i, a)$ and the decision-maker earns a reward $r_t(j|i, a)$. Future rewards are discounted by a factor $0 \leq \alpha \leq 1$. This continues until the end of stage T. We assume, to keep notation at a minimum, that the terminal reward at the end of stage T is 0. The initial state is drawn from a probability mass function (pmf) f over S. The decision-maker wishes to choose actions to maximize the expected total discounted reward over T stages.

A (deterministic Markovian) policy $\pi = (\pi_1, \cdots, \pi_T)$ is a tuple of mappings such that $\pi_t(i) \in A$ is the action prescribed in state $i \in S$ in stage t. Its value equals

$$J^{\pi}_{P^{\pi}(1:T)} \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(i) J^{\pi}_{P^{\pi}(1:T)}(i), \tag{2.1}$$

where

$$J^{\pi}_{P^{\pi}(1:T)}(i) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{P^{\pi}(1:T)} \left[\sum_{t=1}^{T} \alpha^{t-1} r_t(\mathbf{s}_{t+1}|\mathbf{s}_t, \pi_t(\mathbf{s}_t)) \Big| \mathbf{s}_1 = i \right], \text{ for } i = 1, \ldots, n. \tag{2.2}$$

The subscript $P^{\pi}(1:T) \stackrel{\text{def}}{=} (P^{\pi}_1, \cdots, P^{\pi}_T)$ is a T-tuple of $n \times n$ matrices P^{π}_t. The row i-column j entry of P^{π}_t is $p_t(j|i, \pi_t(i))$, for $i, j \in S$. This subscript emphasizes that the expectation is with respect to the random trajectory $(\mathbf{s}_2, \cdots, \mathbf{s}_{T+1})$ induced by $P^{\pi}(1:T)$. The conditioning

$\mathbf{s}_1 = i$ signifies that the initial state is i. The finite set of all deterministic Markovian policies is denoted by Π. The decision-maker can maximize the expected total discounted reward by solving the problem

$$J^* \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \; J^{\pi}_{P^{\pi}(1:T)}. \tag{2.3}$$

Problems of this form will be called stochastic optimization problems. A policy that attains the maximum in (2.3) is found via Bellman's backward recursion [50].

The "true" transition probabilities in $P^{\pi}(1:T)$ are often unknown in practice. In this chapter, we use RMDPs under (s,a)-rectangular ambiguity sets to address the ambiguity present in transition probabilities. These RMDPs are intuitive to underatnd from an adversarial viewpoint. In (s,a)-rectangular RMDPs, the decision-maker chooses a policy $\pi \in \Pi$. A hypothetical adversary observes π and then chooses a $P^{\pi}(1:T)$ from an ambiguity set $\mathcal{P}^{\pi}(1:T)$ so as to minimize $J^{\pi}_{P^{\pi}(1:T)}$. Per the (s,a)-rectangularity assumption, the ambiguity set $\mathcal{P}^{\pi}(1:T)$ is such that the adversary's choice of transition probabilities for a state-action pair $(i, \pi_t(i))$ in stage t does not restrict its choices in the future or in other state-action pairs $(j, \pi_t(j))$. This is a type of multiplicative separability assumption and can be expressed as $\mathcal{P}^{\pi}(1:T) = \bigtimes_{t=1}^{T} \bigtimes_{i \in S} \mathcal{P}_t(i, \pi_t(i))$. Here, $\mathcal{P}_t(i, \pi_t(i)) \subseteq \Delta^n$ is an ambiguity set of transition pmfs $p_t(\cdot|i, \pi_t(i)) \stackrel{\text{def}}{=} (p_t(1|i, \pi_t(i)), \cdots, p_t(n|i, \pi_t(i)))$ corresponding to the state-action pair $(i, \pi_t(i))$ at stage t. The notation Δ^n represents the probability simplex in $\mathbb{R}^n$. Like the non-robust MDPs, (s,a)-rectangular RMDPs also admits an optimal policy that is deterministic and Markovian [32].

Since this entire chapter is about (s,a)-rectangular RMDPs, in the rest of the chapter, we henceforth refer to (s,a)-rectangularity simply as rectangularity.

2.2 Contributions of this chapter

We consider rectangular RMDPs with ambiguity sets

$$\mathcal{P}^N_t(i,a;\epsilon) \stackrel{\text{def}}{=} \left\{ p \in \Delta^n | d(p, \hat{p}^N_t(\cdot|i,a)) \leq \epsilon \right\}, \; \forall (i,a) \in S \times A, \text{ and } t = 1, \ldots, T, \tag{2.4}$$

where $0 \leq \epsilon \leq \infty$. The superscript N and argument ϵ are included in this notation for emphasis. Here, $\hat{p}_t^N(\cdot|i,a) \in \Delta^n$ is the empirical (frequency) estimate of $p_t(\cdot|i,a)$, calculated based on N independent, historical observations of the next state reached upon choosing action $a \in A$ in state i in stage t. The function $d : \Delta^n \times \Delta^n \to \{\mathbb{R}_+ \cup \infty\}$ is a nonnegative extended real-valued distance function (not necessarily a metric) that quantifies a sense of closeness between distributions. Thus, ambiguity sets are d-balls centered at empirical estimates of the unknown transition pmfs.

We study finite-horizon problems in Section 2.4 and infinite-horizon problems in Section 2.5, under two assumptions. Assumption 1 requires that the distance function "metrizes" the usual topology of componentwise convergence in $\mathbb{R}^n$. Lemma 1 establishes a natural sufficient condition, which we call the generalized Pinsker's inequality, for Assumption 1. This condition is stated in Assumption 2, and it requires that a certain transformation of the distance between two pmfs provides an upper bound on the ℓ_1 distance between them. Assumption 3 is a concentration inequality [43], which provides a probabilistic guarantee that the true pmf of a random variable with a finite support belongs to the ambiguity ball centered at the empirical pmf. Our main contributions are listed below.

- We prove that the optimal values of the RMDPs converge almost surely to the true optimal value as the sample-size $N \to \infty$. This is termed *robust value convergence.* We then prove that the values (evaluated under the true transition pmfs) of policies that solve the RMDPs for large sample-sizes N almost surely equal the true optimal value. This is referred to as *out-of-sample value lock-in.* Both these results are proved under Assumption 1.

- For finite sample-sizes, we show that the value of the robust optimal policy evaluated under the true transition pmfs is bounded below, with a high probability, by the optimal value of the RMDP. This is called *probabilistic performance guarantee on out-of-sample values* and is proved under Assumption 3.

- When a slight strengthening of Assumption 2 holds, we derive a probabilistic rate of convergence for the robust and out-of-sample values to the true optimal value in the infinite-horizon setting. A similar rate can also be derived for finite-horizon MDP but we omit it for brevity.

- In Section 2.6, we show that Assumption 1 and Assumption 3 hold for many well-known distance functions such as Total Variation (TV), Burg, Hellinger, and Wasserstein. Assumption 1 also holds for distance functions such as Kullback-Leibler (KL), χ^2, and modified χ^2, wherein Assumption 3 holds "approximately" for large sample-sizes. In Section 2.6, we also derive concrete formulas for ambiguity radii that meet our probabilistic performance guarantees for several distances.

We remark as an aside that all results in this chapter can be adapted to an optimistic uncertainty model, where the adversary is a max-player instead of a min-player.

Our RMDP can be solved via a robust counterpart of the value iteration algorithm [32, page 267]. An important step in that process is the solution of the inner-optimization problem. Tailored efficient solution methods for the inner-optimization problem within a robust counterpart of Bellman's recursion are available for the distances listed in Section 2.6. These are briefly outlined in Section 2.7. Using these solution methods, we conduct computational experiments with the TV, Burg, and Wasserstein distances on a randomly generated MDP instance and a machine repair example commonly studied in the literature. Those results are presented in Section 2.8. We also construct counterexamples to demonstrate that Assumption 1 and Assumption 3 cannot be dropped from the hypotheses of our asymptotic and finite-sample results, respectively.

Esfahani and Kuhn [19], Bertsimas et al. [10], Van Parys et al. [64], and Sutter et al. [60] discussed (i) robust value convergence, (ii) out-of-sample value convergence, (iii) probabilistic performance guarantees on out-of-sample values, and (iv) a probabilistic convergence rate, as desirable properties that a data-driven robust optimization framework should posses. Our work establishes essentially a minimal set of conditions under which rectangular RMDPs

with data-driven distance-based ambiguity sets exhibit these four properties. To the best of our knowledge, we are not aware of any published work that achieves this for any class of RMDPs, as described in the literature review next.

2.3 Related literature

Three papers have considered (s, a)-rectangular RMDPs with ambiguity sets defined via the Wasserstein metric. Yang [69] studied finite-horizon MDPs with finite state- and action-spaces, where both rewards and transition probabilities were unknown. Derman and Mannor [14] investigated finite-state, finite-action, infinite-horizon problems. Yang [70] focused on infinite-horizon problems with continuous state- and action-spaces, and relied on measure-theoretic assumptions from earlier work on minimax control [25]. Ambiguity sets based on Kullback-Leibler, modified χ^2, and TV distances were studied for (s, a)-rectangular RMDPs in [32, 46]. Finite- and infinite-horizon (s, a)-rectangular RMDPs in separable metric spaces were studied in [62] using ambiguity sets based on the TV distance. This was extended to average cost RMDPs in [63]. A modified policy iteration algorithm was proposed for solving (s, a)-rectangular RMDPs in [36]. None of these papers prove asymptotic value convergence with increasing sample-sizes. Some do not provide a probabilistic guarantee on out-of-sample performance.

Data-driven RMDPs under other notions of rectangularity have also been explored in the literature. The work in [67] employed observation histories to build a non-rectangular ambiguity set. Asymptotic behavior of the robust problem with respect to the length of the observation history was analyzed, though no finite-sample guarantees were derived. The work in [66] extended [67] to derive a finite-sample guarantee for infinite-horizon problems. They used finite-length observation histories to construct non-rectangular ambiguity sets using the Wasserstein metric. However, the paper did not prove any asymptotic convergence result with respect to the length of the observation history.

2.4 Finite-horizon problems

We will assume throughout this section, merely for notational convenience, that, for every $(i, a) \in S \times A$ and every stage $t = 1, \ldots, T$, the true transition pmf has full support S; that is, there is a positive probability of reaching any state in S in stage $t + 1$. All our proofs go through even when this does not hold, albeit with tedious notation wherein various subscripts, superscripts, and cardinalities are indexed by specific state-action pairs. Indeed, we are able to computationally tackle this more general situation in our machine repair example in Section 2.8.3. Following the style of notation in (2.4), we write the ambiguity set corresponding to policy π as

$$\mathcal{P}^{N,\pi}(\epsilon; (1:T)) = \bigtimes_{t=1}^{T} \bigtimes_{i \in S} \mathcal{P}_t^N(i, \pi_t(i); \epsilon). \tag{2.5}$$

The data-driven RMDP problem can then be formulated as

$$\tilde{J}^N(\epsilon) \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \; \tilde{J}^{N,\pi}(\epsilon), \text{ where } \tilde{J}^{N,\pi}(\epsilon) \stackrel{\text{def}}{=} \inf_{P^\pi(1:T) \in \mathcal{P}^{N,\pi}(\epsilon;(1:T))} J^\pi_{P^\pi(1:T)}. \tag{2.6}$$

To avoid trivialities throughout this section, we work under the assumption that the ambiguity balls $\mathcal{P}_t^N(i, a; \epsilon)$ are nonempty with probability 1 (wp 1), for all states i, actions a, stages t, sample-sizes N, and $0 \leq \epsilon \leq \infty$. This ensures that the optimal value of each inner-optimization problem encountered in this section is finite, wp 1. Example 1 demonstrates that this nonemptiness assumption is not vacuous.

Example 1. Consider the distance $d(p, q)$ where $d(p, q) = 0$, if, for each j, p_j, q_j are both irrational; and 1 otherwise. Let $q \in \Delta^n$ and $\hat{q}^N$ be its empirical estimate. Then, $d(p, \hat{q}^N) = 1$, for all $p \in \Delta^n$, as each component of $\hat{q}^N$ is a ratio of counts. Thus, the ambiguity ball is empty, for all $0 \leq \epsilon < 1$.

A sufficient condition for the nonemptiness assumption is that $d(q, q) = 0$, for every $q \in \Delta^n$. This guarantees that the center of the ambiguity ball belongs to the ambiguity ball. This condition holds for all distances listed in Section 2.6.

2.4.1 Value convergence to true optimal value

The shorthand $\xrightarrow{N,\text{wp1}}$ denotes limits that hold with probability 1 as $N \to \infty$. This is also called almost sure convergence. The strong law of large numbers (SLLN) assures us that empirical estimates of the true transition probabilities converge wp 1 to the true transition probabilities as $N \to \infty$. Imagine, for the moment, that the decision-maker implements an SAA approach [1] whereby, stochastic optimization problem (2.3) for the true MDP is solved approximately by replacing $P^\pi(1:T)$ with its empirical estimate $\hat{P}^{N,\pi}(1:T)$. Suppose $J^{*,N}$ denotes the resulting optimal value. It is possible to show that $J^{*,N} \xrightarrow{N,\text{wp1}} J^*$. One way to prove this is via combining the continuity of value functions in transition probabilities (see Lemma 3) with the fact that the set of deterministic Markovian policies is finite. However, one motivation for pursuing data-driven robust optimization instead of SAA is that SAA solutions can be too sensitive to the sampled training data [10]. Nevertheless, it is natural to ask whether the data-driven robust optimal values $\tilde{J}^N(\epsilon)$ converge to the optimal value J^* of the true MDP defined in (2.3), akin to SAA. This would not be true if the radius ϵ of the ambiguity balls did not depend on N. It is perhaps intuitive, however, that such a convergence result would hold if the radius vanished to 0 as N increased. This intuition is rooted in the belief that the simultaneous convergence of the empirical estimates to the true transition probabilities along with the shrinking radii of the ambiguity balls, should render the data-driven RMDP (2.6) to increasingly well-approximate the true stochastic problem (2.3). While this intuition is largely correct, one should be careful about at least two interrelated issues. The first is that the centers of these balls are at the empirical estimates and hence dependent on N. That is, the balls are not only shrinking but also moving with N. The second is topological. In particular, the question is whether or not the distance function d, although not a metric in itself, in some sense "metrizes" the natural topology of componentwise convergence in $\Delta^n \subset \mathbb{R}^n$. Assumption 1 addresses this, thereby leading to an appropriate convergence result in Theorem 1.

Assumption 1. Suppose $p \in \Delta^n$ is any pmf. Suppose $\hat{p}^N \in \Delta^n$ is a sequence of its

empirical estimates. Suppose $p^N \in \Delta^n$ is any sequence such that $d(p^N, \hat{p}^N) \xrightarrow{N,\text{wp1}} 0$. Then, $|p_j^N - \hat{p}_j^N| \xrightarrow{N,\text{wp1}} 0$, for each $j = 1, \cdots, n$.

This requires that if the distance function deems two pmfs to be close, then they are close (in the usual metric in $\mathbb{R}^n$). Example 2 demonstrates that this assumption is not vacuous.

Example 2. For any $p, q \in \Delta^n$, consider the distance $d(p, q)$ defined as

$$d(p,q) = \begin{cases} 0, \text{ if, for each } j,\ p_j, q_j \text{ are either both rational or both irrational} \\ 1, \text{ otherwise.} \end{cases}$$

Suppose that $p = (1/2, 1/2) \in \Delta^2$ is the pmf of a Bernoulli random variable that takes values ± 1. Thus, $\hat{p}^N = (k/N, (N-k)/N)$, where k is the number of occurrences of $+1$ among N independent trials. Now consider the constant sequence $p^N = (2/3, 1/3) \in \Delta^2$. Then, $d(p^N, \hat{p}^N) = 0$, for all N. This is because both k/N and $2/3$ are rational, and both $(N-k)/N$ and $1/3$ are rational. Moreover, both k/N and $(N-k)/N$ converge to $1/2$, wp 1, by the SLLN. Thus, $\lim_{N\to\infty} \left|p_1^N - \hat{p}_1^N\right| = \lim_{N\to\infty} \left|\frac{2}{3} - \frac{k}{N}\right| = \frac{1}{6}$, wp 1. Similarly, $\lim_{N\to\infty} \left|p_2^N - \hat{p}_2^N\right| = \lim_{N\to\infty} \left|\frac{1}{3} - \frac{N-k}{N}\right| = \frac{1}{6}$, wp 1. Thus, Assumption 1 does not hold.

One natural way to ascertain that Assumption 1 holds, is to instead establish the following sufficient condition.

Assumption 2. There exists a continuous function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$ such that $\|p - q\|_1 \leq \psi(d(p, q))$, for all $p, q \in \Delta^n$.

For the Kullback-Leibler distance, this bound is called Pinsker's inequality. Motivated by that, we call this assumption the generalized Pinsker's inequality.

Lemma 1. Suppose Assumption 2 holds. Then Assumption 1 holds.

Proof. Recalling notation from Assumption 1, we have, $0 \leq |p_j^N - \hat{p}_j^N| \leq \left\|p^N - \hat{p}^N\right\|_1 \leq \psi(d(p^N, \hat{p}^N))$. Here, the last inequality follows by Assumption 2. Assumption 1 then follows since $\lim_{N\to\infty} \psi(d(p^N, \hat{p}^N)) = \psi\left(\lim_{N\to\infty} d(p^N, \hat{p}^N)\right) = \psi(0) = 0$. The first equality follows by continuity of $\psi(\cdot)$ and the last equality by Assumption 2. □

Section 2.6 demonstrates that this generalized Pinsker's inequality holds for several distances commonly studied in the literature.

Theorem 1. Suppose the radii of the ambiguity balls are dependent on N such that $\lim_{N\to\infty} \epsilon^N = 0$. Suppose Assumption 1 holds. Then, $\tilde{J}^N(\epsilon^N) \xrightarrow{N,\text{wp1}} J^*$.

We first present a couple of technical lemmas which will be used in proving this theorem.

Let $\Delta^{n^2T} \stackrel{\text{def}}{=} \bigtimes_{t=1}^{T} \bigtimes_{i=1}^{n} \Delta^n$. We embed tuples $P^\pi(1:T) = (P_1^\pi, \cdots, P_T^\pi) \in \Delta^{n^2T}$ of $n \times n$ transition matrices P_t^π into $\mathbb{R}^{n^2T}$ with its natural topology. Thus, a sequence $P^{N,\pi}(1:T) = (P_1^{N,\pi}, \cdots, P_T^{N,\pi})$ converges to $P^\pi(1:T)$ if, and only if, each component sequence $p_t^N(j|i,\pi_t(i))$ converges to the corresponding component $p_t(j|i,\pi_t(i))$. Moreover, the limit $P^\pi(1:T)$ belongs to Δ^{n^2T}.

Lemma 2. Suppose Assumption 1 holds. Suppose the radii of ambiguity sets are such that $\lim_{N\to\infty} \epsilon^N = 0$. Let $P^{N,\pi}(1:T) \in \mathcal{P}^{N,\pi}(\epsilon^N;(1:T))$ be any sequence. Denote the true transition probabilities for policy π by $P^\pi(1:T)$. Then, $P^{N,\pi}(1:T) \xrightarrow{N,\text{wp1}} P^\pi(1:T)$.

Proof. Consider any fixed states $i, j \in S$ and stage t. It suffices to establish that $|p_t^N(j|i,\pi_t(i)) - p_t(j|i,\pi_t(i))| \xrightarrow{N,\text{wp1}} 0$. The triangle inequality yields

$$\begin{aligned} 0 &\le |p_t^N(j|i,\pi_t(i)) - p_t(j|i,\pi_t(i))| \\ &\le |p_t^N(j|i,\pi_t(i)) - \hat{p}_t^N(j|i,\pi_t(i))| + |\hat{p}_t^N(j|i,\pi_t(i)) - p_t(j|i,\pi_t(i))|. \end{aligned}$$

We show that each term above converges to 0, wp 1, as $N \to \infty$. By the SLLN, $|\hat{p}_t^N(j|i,\pi_t(i)) - p_t(j|i,\pi_t(i))| \xrightarrow{N,\text{wp1}} 0$. Moreover, $p_t^N(\cdot|i,\pi_t(i)) \in \mathcal{P}_t^N(i,\pi_t(i);\epsilon^N)$ implies that $d(p_t^N(\cdot|i,\pi_t(i)), \hat{p}_t^N(\cdot|i,\pi_t(i))) \le \epsilon^N$. Since $\lim_{N\to\infty} \epsilon^N = 0$, we have that $d(p_t^N(\cdot|i,\pi_t(i)), \hat{p}_t^N(\cdot|i,\pi_t(i))) \xrightarrow{N,\text{wp1}} 0$. Then, Assumption 1 implies $|p_t^N(j|i,\pi_t(i)) - \hat{p}_t^N(j|i,\pi_t(i))| \xrightarrow{N,\text{wp1}} 0$. □

Lemma 3. Suppose $\pi \in \Pi$ is a fixed policy. Suppose $\lim_{N\to\infty} P^{N,\pi}(1:T) = P^{\pi}(1:T)$ in Δ^{n^2T}, wp 1. Then, $J^{\pi}_{P^{N,\pi}(1:T)} \xrightarrow{N,\text{wp1}} J^{\pi}_{P^{\pi}(1:T)}$.

Proof. From the definition of expectation in (2.2), $J^{\pi}_{P^{\pi}(1:T)}(s_1)$ equals

$$\sum_{s_2\in S}\cdots\sum_{s_{T+1}\in S}\mathbb{P}\left(\mathbf{s}_2 = s_2,\cdots,\mathbf{s}_{T+1} = s_{T+1}|\mathbf{s}_1 = s_1,\pi\right)\left[\sum_{t=1}^{T}\alpha^{t-1}r_t(s_{t+1}|s_t,\pi_t(s_t))\right]$$
$$=\sum_{s_2\in S}\cdots\sum_{s_{T+1}\in S}\left[\prod_{t=1}^{T}p_t(s_{t+1}|s_t,\pi_t(s_t))\right]\left[\sum_{t=1}^{T}\alpha^{t-1}r_t(s_{t+1}|s_t,\pi_t(s_t))\right].$$

The last equality follows by the Markov property and the definition of transition probabilities. Thus, $J^{\pi}_{P^{\pi}(1:T)}(s_1)$ is a polynomial function of the components of $P^{\pi}(1:T)$ and hence continuous over $\Delta^{n^2T} \subset \mathbb{R}^{n^2T}$. Thus, $J^{\pi}_{P^{\pi}(1:T)}$ is also continuous. The conclusion then follows by the continuous mapping theorem. □

Proof of Theorem 1. Fix any $\pi \in \Pi$ and let $P^{\pi}(1:T)$ denote the corresponding tuple of true transition probability matrices. We will show that

$$\tilde{J}^{N,\pi}(\epsilon^N) \xrightarrow{N,\text{wp1}} J^{\pi}_{P^{\pi}(1:T)}. \tag{2.7}$$

The result will then follow because, wp 1,

$$\lim_{N\to\infty}\tilde{J}^{N}(\epsilon^N) = \lim_{N\to\infty}\max_{\pi\in\Pi}\tilde{J}^{N,\pi}(\epsilon^N) \overset{(a)}{=} \max_{\pi\in\Pi}\lim_{N\to\infty}\tilde{J}^{N,\pi}(\epsilon^N) = \max_{\pi\in\Pi}J^{\pi}_{P^{\pi}(1:T)} = J^{*}.$$

The interchange of limit and maximum in "(a)" is allowed because Π is a finite set.

To establish (2.7), fix any $\delta > 0$. We will prove that, wp 1, $\exists\ N^*$ large enough such that $\left|\tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{\pi}(1:T)}\right| < \delta$ for all $N \geq N^*$. Notice that there exists a $P^{N,\pi}(1:T) \in \mathcal{P}^{N,\pi}(\epsilon^N;(1:T))$ such that $\tilde{J}^{N,\pi}(\epsilon^N) < J^{\pi}_{P^{N,\pi}(1:T)} + \frac{\delta}{2}$ and $\tilde{J}^{N,\pi}(\epsilon^N) + \frac{\delta}{2} \geq J^{\pi}_{P^{N,\pi}(1:T)}$, for every $N \geq 1$. This holds by the definition of infimum. That is, $\left|\tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{N,\pi}(1:T)}\right| \leq \frac{\delta}{2}$. Thus,

for each $N \geq 1$, the triangle inequality yields

$$\begin{aligned}\left|\tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{\pi}(1:T)}\right| &\leq \left|\tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{N,\pi}(1:T)}\right| + \left|J^{\pi}_{P^{N,\pi}(1:T)} - J^{\pi}_{P^{\pi}(1:T)}\right| \\ &\leq \frac{\delta}{2} + \left|J^{\pi}_{P^{N,\pi}(1:T)} - J^{\pi}_{P^{\pi}(1:T)}\right|. \end{aligned} \tag{2.8}$$

From Lemma 2, we know that $P^{N,\pi}(1:T) \xrightarrow{N,\text{wp1}} P^{\pi}(1:T)$. Lemma 3 then implies that $J^{\pi}_{P^{N,\pi}(1:T)} \xrightarrow{N,\text{wp1}} J^{\pi}_{P^{\pi}(1:T)}$. Thus, wp 1, $\exists\ N^*$ such that $\left|J^{\pi}_{P^{N,\pi}(1:T)} - J^{\pi}_{P^{\pi}(1:T)}\right| < \frac{\delta}{2}$ for all $N \geq N^*$. Utilizing this in (2.8) establishes (2.7). □

The next example demonstrates that we cannot drop Assumption 1 from the hypothesis of this theorem.

Example 3. Consider a single-stage MDP with $S = \{1, 2\}$; $A = \{a\}$; all true $p(\cdot|\cdot, \cdot) = 0.5$; rewards $r(1|\cdot, a) = 2$ and $r(2|\cdot, a) = 0$; and initial state pmf $f(1) = f(2) = 0.5$. Only one policy is available — $\pi(1) = a, \pi(2) = a$ — and the optimal value $J^* = 1$. Suppose N independent observations of the next state reached upon choosing action a are available for each state. The corresponding empirical estimates would be ratios of counts and thus rational. Assume that the ambiguity balls $\mathcal{P}^N(1, a; \epsilon^N)$ and $\mathcal{P}^N(2, a; \epsilon^N)$ are defined via the rational/irrational distance described in Example 2 with radius $\epsilon^N > 0$. Recall from Example 2 that Assumption 1 does not hold for this distance. The degenerate transition probabilities $p(1|1, a) = 0, p(2|1, a) = 1$ and $p(1|2, a) = 0, p(2|2, a) = 1$ belong to these ambiguity balls. In fact, since the corresponding value function is 0 and we know that the value function in this case is nonnegative, these degenerate transition pmfs are, in fact, the worst-case pmfs. In short, $\tilde{J}^N(\epsilon^N) = 0$, for every N. Consequently, $0 = \lim_{N\to\infty} \tilde{J}^N(\epsilon^N) \neq J^* = 1$. Thus, the conclusion of Theorem 1 fails.

We next establish an important consequence of the above theorem. To state and prove it formally, we need additional notation. Let $\hat{\pi}^N$ denote an optimal policy that solves the data-driven RMDP problem (2.6) and $P^{\hat{\pi}^N}(1:T)$ be the corresponding T-tuple of true

transition probability matrices of size $n \times n$ each. Let $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ denote the value of this policy in the true MDP. This is often called the out-of-sample value of $\hat{\pi}^N$. This is because we are assessing the performance of $\hat{\pi}^N$ on the true transition probabilities, whereas $\hat{\pi}^N$ was computed only based on a (training) sample drawn from these true transition probabilities.

Theorem 2. Suppose the radii of the ambiguity balls are dependent on N such that $\lim_{N\to\infty} \epsilon^N = 0$. Suppose Assumption 1 holds. Then, the robust optimal policy $\hat{\pi}^N$ solves (2.3) for all sufficiently large N, wp 1. Consequently, $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)} = J^*$, for all sufficiently large N, wp 1.

Proof. Let $\bar{\pi} \in \Pi$ be a policy that occurs infinitely often wp 1 in the sequence $\{\hat{\pi}^N\} \in \Pi$. Such a policy exists since Π is finite. Let $N^{\bar{\pi}}_k$ denote the subsequence that indexes all occurrences of $\bar{\pi}$ in $\{\hat{\pi}^N\}$. That is, $\hat{\pi}^{N^{\bar{\pi}}_k} = \bar{\pi}$, for all k, wp 1. Consequently, by the triangle inequality, we have,

$$\begin{aligned} 0 \leq \left| J^{\bar{\pi}}_{P^{\bar{\pi}}(1:T)} - J^* \right| &\leq \left| J^{\bar{\pi}}_{P^{\bar{\pi}}(1:T)} - \tilde{J}^{N^{\bar{\pi}}_k, \hat{\pi}^{N^{\bar{\pi}}_k}}(\epsilon^{N^{\bar{\pi}}_k}) \right| + \left| \tilde{J}^{N^{\bar{\pi}}_k, \hat{\pi}^{N^{\bar{\pi}}_k}}(\epsilon^{N^{\bar{\pi}}_k}) - J^* \right| \\ &= \left| J^{\bar{\pi}}_{P^{\bar{\pi}}(1:T)} - \tilde{J}^{N^{\bar{\pi}}_k, \bar{\pi}}(\epsilon^{N^{\bar{\pi}}_k}) \right| + \left| \tilde{J}^{N^{\bar{\pi}}_k, \hat{\pi}^{N^{\bar{\pi}}_k}}(\epsilon^{N^{\bar{\pi}}_k}) - J^* \right|, \end{aligned}$$

for all k, wp 1. Here, the last equality was obtained simply by replacing $\hat{\pi}^{N^{\bar{\pi}}_k}$ with $\bar{\pi}$ in the first absolute value expression. From Theorem 1 and its proof, we know that each one of these absolute value terms in the upper bound converges to 0, wp 1, as $k \to \infty$. Thus, we must have that $J^{\bar{\pi}}_{P^{\bar{\pi}}(1:T)} = J^*$, thereby, showing that $\bar{\pi}$ is optimal to (2.3). Moreover, since Π is finite, the sequence $\{\hat{\pi}^N\}$ eventually only includes policies that appear infinitely often, wp 1. All such policies must be optimal to (2.3) as established above. This proves the first claim in the theorem. The second claim then follows immediately because J^* is the optimal objective value in problem (2.3). □

2.4.2 Probabilistic guarantee on the performance of robust optimal policy

Although Theorem 2 shows that $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ converges to the true optimal value, the decision-maker would additionally want $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ to be sufficiently high with a large enough probability (with respect to the uncertainty in the sampled training data) for finite sample-sizes. In particular, suppose the decision-maker wants to have a probabilistic confidence of $1-\gamma$, for some $\gamma \in (0,1)$, that $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ is sufficiently high. In other words, the decision-maker wants to obtain a lower bound on $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ that holds with probability at least $1-\gamma$. Two questions arise in this context: (i) what is the numerical value of such a lower bound? and (ii) what radius for the ambiguity balls should the decision-maker utilize to achieve a $1-\gamma$ confidence for that lower bound? These two questions are answered in this section under the following assumption.

Assumption 3. Fix an integer $N \geq 1$. Then, for every $\beta \in (0,1)$, there exists an $0 < \epsilon^N_\beta < \sup_{p,q\in\Delta^n} d(p,q)$ with the following property: for every pmf $q \in \Delta^n$ and its empirical estimate $\hat{q}^N$, we have,

$$\mathbb{P}[d(q,\hat{q}^N) \leq \epsilon^N_\beta] \geq 1-\beta. \tag{2.9}$$

Inequalities of the form (2.9) are called concentration inequalities [43]. The probability on the left hand side of (2.9) is induced by the uncertainty in $\hat{q}^N$, that is, in the sampled training data. The positive real number ϵ^N_β typically depends on n, but this is suppressed in the notation for brevity. If $\sup_{p,q\in\Delta^n} d(p,q) = \epsilon \leq \infty$, then $\mathbb{P}\left[d(q,\hat{q}^N) \leq \epsilon\right] = 1$. Thus, this supremum trivially satisfies the concentration inequality (2.9). The requirement that ϵ^N_β be strictly smaller than the supremum, excludes this trivial situation. Several distances listed in Section 2.6 satisfy this assumption. In fact, we derive concrete expressions for ϵ^N_β for several distances there. Those formulas reveal the additional property, not explicitly included in the above assumption, that $\lim_{N\to\infty} \epsilon^N_\beta = 0$. Consequently, these radii satisfy the hypothesis of Theorem 1 and Theorem 2, thereby guaranteeing almost sure convergence of $\tilde{J}^N(\epsilon^N_\beta)$ and $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ to J^*. The next example demonstrates that Assumption 3 is not vacuous.

Example 4. Consider the distance

$$d(p,q) = \begin{cases} 0, \text{ if } p = q \\ 1, \text{ otherwise} \end{cases}, \quad \forall p, q \in \Delta^n.$$

Note that $\sup_{p,q \in \Delta^n} d(p,q) = 1$. Now suppose the true pmf $q = (1/2, 1/2) \in \Delta^2$. Then, for any $N \geq 1$ and any $0 < \epsilon < 1$, we have,

$$\mathbb{P}[d(q, \hat{q}^N) \leq \epsilon] = \mathbb{P}[d(q, \hat{q}^N) = 0] = \mathbb{P}[\hat{q}^N = q] = \begin{cases} \binom{N}{N/2} \left(\frac{1}{2}\right)^N, \text{ if } N \text{ even}, \\ 0, \text{ if } N \text{ odd}. \end{cases}$$

The binomial coefficient $\binom{N}{N/2}$ can be upper bounded as $\binom{N}{N/2} = \binom{N-1}{N/2} + \binom{N-1}{(N/2)-1} \leq \sum_{k=0}^{N-1} \binom{N-1}{k} = 2^{N-1}$. Here, the first equality holds by Pascal's identity and the last equality holds by the binomial theorem. Hence, $\mathbb{P}[d(q, \hat{q}^N) \leq \epsilon]$ is bounded above by $1/2$, and inequality (2.9) in Assumption 3 fails for every $\beta \in (0, 0.5)$.

Now recall that $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)}$ denotes the value of policy $\hat{\pi}^N$ in the MDP with true transition probabilities $P^{\hat{\pi}^N}(1:T)$ — that is, the out-of-sample value of $\hat{\pi}^N$. This quantity is unknown to the decision-maker since the true transition probabilities are unknown. Also recall that $\tilde{J}^N(\epsilon^N_\beta)$ is the optimal value of the data-driven robust MDP (2.6). The decision-maker can compute this quantity. As previewed at the beginning of this section, our next result establishes that the latter (known value) provides a lower bound on the former (unknown value) with arbitrarily high probability.

Theorem 3. Fix the sample-size $N \geq 1$. Suppose Assumption 3 holds. Consider any fixed $\gamma \in (0,1)$, and let $0 < \beta(\gamma) = 1 - (1-\gamma)^{1/nmT} < 1$. Let $0 < \epsilon^N_{\beta(\gamma)} < \sup_{p,q \in \Delta^n} d(p,q)$ be as in Assumption 3. Then,

$$\mathbb{P}\left[J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)} \geq \tilde{J}^N\left(\epsilon^N_{\beta(\gamma)}\right)\right] \geq 1 - \gamma. \tag{2.10}$$

Proof. We know that $\tilde{J}^{N,\pi}(\epsilon^N_{\beta(\gamma)}) < \infty$, wp 1, for every $\pi \in \Pi$. This is because the ambiguity balls are nonempty by our assumption. Since $\hat{\pi}^N$ is optimal to (2.6), we have,

$$\tilde{J}^N(\epsilon^N_{\beta(\gamma)}) = \tilde{J}^{N,\hat{\pi}^N}(\epsilon^N_{\beta(\gamma)}) = \inf_{Q^{\hat{\pi}^N}(1:T) \in \mathcal{P}^{N,\hat{\pi}^N}(\epsilon^N_{\beta(\gamma)})} J^{\hat{\pi}^N}_{Q^{\hat{\pi}^N}(1:T)} \leq J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)},$$

if the tuple of true transition probability matrices $P^{\hat{\pi}^N}(1:T) \in \mathcal{P}^{N,\hat{\pi}^N}(\epsilon^N_{\beta(\gamma)}; (1:T))$. By rectangularity, this latter event occurs if, and only if, the true transition pmf $p_t(\cdot|i, \hat{\pi}^N_t(i))$ belongs to the ambiguity ball $\mathcal{P}^N_t(i, \hat{\pi}^N_t(i); \epsilon^N_{\beta(\gamma)})$, for every state $i = 1, \ldots, n$ and every stage $t = 1, \ldots, T$. That is, if, and only if,

$$d(p_t(\cdot|i, \hat{\pi}^N_t(i)), \hat{p}^N_t(\cdot|i, \hat{\pi}^N_t(i))) \leq \epsilon^N_{\beta(\gamma)},$$

for every state $i = 1, \ldots, n$ and every stage $t = 1, \ldots, T$. This discussion shows that

$$\begin{aligned}
&\mathbb{P}\left[J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}(1:T)} \geq \tilde{J}^N(\epsilon^N_{\beta(\gamma)})\right] \\
&\geq \mathbb{P}\left[d(p_t(\cdot|i, \hat{\pi}^N_t(i)), \hat{p}^N_t(\cdot|i, \hat{\pi}^N_t(i))) \leq \epsilon^N_{\beta(\gamma)},\ i = 1, \ldots, n,\ t = 1, \ldots, T\right] \\
&\geq \mathbb{P}\left[d(p_t(\cdot|i, a), \hat{p}^N_t(\cdot|i, a)) \leq \epsilon^N_{\beta(\gamma)},\ i = 1, \ldots, n,\ a = 1, \ldots, m,\ t = 1, \ldots, T\right] \\
&\overset{(a)}{=} \prod_{t=1}^{T} \prod_{i=1}^{n} \prod_{a=1}^{m} \mathbb{P}\left[d(p_t(\cdot|i, a), \hat{p}^N_t(\cdot|i, a)) \leq \epsilon^N_{\beta(\gamma)}\right] \\
&\overset{(b)}{\geq} \prod_{t=1}^{T} \prod_{i=1}^{n} \prod_{a=1}^{m} (1 - \beta(\gamma)) = (1 - \beta(\gamma))^{nmT} \overset{(c)}{=} 1 - \gamma.
\end{aligned}$$

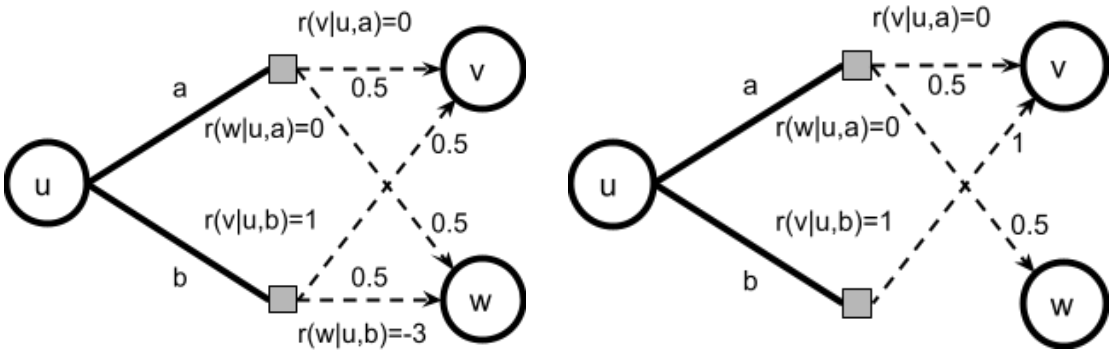
Equality "(a)" holds because (training) sampled observations of the next state reached are independent across states i, actions a, and stages t. Inequality "(b)" follows from (2.9). Equality "(c)" holds by definition of $\beta(\gamma)$. This completes the proof. □

We emphasize again that, in Section 2.6, we derive explicit formulas for ambiguity ball radii

ϵ_β^N compatible with (2.9), for several distances. These can be utilized in Theorem 3 to ensure the requisite $1-\gamma$ confidence promised by (4.19). The ambiguity ball radius prescribed by this theorem depends on n, m, and T. The dependence on n is inevitable because that is the dimension of the pmfs involved. The number of actions m appears in the formula for the radius because the pmf of the next state depends on the action chosen. The dependence on T is induced by the multi-stage nonstationary nature of the problem and seems impossible to avoid.

The next example demonstrates that we cannot drop Assumption 3 from the hypothesis of the theorem.

Example 5. Consider the single-stage MDP depicted in Figure 2.1a with action-space $A = \{a, b\}$; true $p(\cdot|\cdot,\cdot) = 0.5$; rewards are displayed next to the dotted transition arrows. Assume that the decision-maker knows the true transition probabilities for the state-action pair (u, a), but does not know them for the pair (u, b). The decision-maker has a single sample ($N = 1$) of the next state reached upon choosing action b in state u. Suppose this observed state was v. Therefore, the empirical pmf is simply $\hat{p}^1(v|u, b) = 1$ and $\hat{p}^1(w|u, b) = 0$. This event occurs with probability 0.5. Suppose the decision-maker utilizes the 0/1 distance from Example 4. Regardless of the radius $0 < \epsilon < 1$ of the ambiguity ball, the ambiguity ball includes only the empirical pmf. The worst-case pmf is therefore equal to this empirical pmf as shown in Figure 2.1b. There are only 2 policies — $\pi_1(u) = a$ and $\pi_2(u) = b$. Since there is no ambiguity about transition probabilities $p(\cdot|u, a)$, $\tilde{J}^{N,\pi_1} = J^{\pi_1} = 0.5 \times 0 + 0.5 \times 0 = 0$. Hence, $\tilde{J}(N, \epsilon) = \max\{\tilde{J}^{N,\pi_1}, \tilde{J}^{N,\pi_2}(\epsilon)\} = \max\{0, 1\} = 1$. Thus, it is optimal in the data-driven robust problem to choose action b in state u. The value of this policy in the true problem is $J^{\pi_2} = 0.5 \times 1 + 0.5 \times (-3) = -1$. This shows that $\mathbb{P}(J^{\pi_2} < \tilde{J}^{N,\pi_2}(\epsilon)) \geq 0.5$. Thus, (say) if $\gamma = 0.1$, the conclusion of Theorem 3 fails.



(a) Simple 1-stage MDP. (b) Worst-case pmf.

Figure 2.1: A schematic of Example 5, which demonstrates that we cannot drop Assumption 3 from the hypothesis of Theorem 3.

2.5 Stationary infinite-horizon problems

In a stationary infinite-horizon MDP over stages $t = 1, 2, \cdots$, the transition probabilities $p(j|i, a)$ and rewards $r(j|i, a)$ do not vary over stages, and the discount factor α is strictly less than 1 [50]. The decision-maker wishes to maximize the expected total discounted reward over an infinite horizon.

A (stationary, deterministic Markovian) policy π is a mapping that prescribes action $\pi(i) \in A$ to state $i \in S$ in every stage. Its value equals

$$J^{\pi}_{P^{\pi}} \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(i) J^{\pi}_{P^{\pi}}(i), \tag{2.11}$$

where

$$J^{\pi}_{P^{\pi}}(i) \stackrel{\text{def}}{=} \lim_{T\to\infty} \left\{ \mathop{\mathbb{E}}_{P^{\pi}} \left[\sum_{t=1}^{T} \alpha^{t-1} r(\mathbf{s}_{t+1}|\mathbf{s}_t, \pi(\mathbf{s}_t)) \Big| \mathbf{s}_1 = i \right] \right\}, \text{ for } i = 1, \ldots, n. \tag{2.12}$$

The subscript P^{π} is matrix of size $n \times n$ with (i, j)th entry $p(j|i, \pi(i))$, for $i, j \in S$.

This subscript emphasizes that the expectation is with respect to the random trajectory $(\mathbf{s}_2, \mathbf{s}_3, \cdots)$ induced by P^π. The finite set of all stationary, deterministic Markovian policies is denoted by Π. The decision-maker can maximize expected total discounted reward by solving the problem

$$J^* \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \; J^\pi_{P^\pi}. \tag{2.13}$$

A policy that attains the maximum in this problem can be found by solving Bellman's equations via value iteration or policy iteration [50].

A robust counterpart of this stationary infinite-horizon MDP was studied under the rectangularity assumption in [32, 46]. There, since the transition probabilities out of state-action pair (s, a) do not vary across stages, the decision-maker employs stage-invariant ambiguity sets $\mathcal{P}(s, a) \subseteq \Delta^n$. However, in the adversarial interpretation of this approach, the adversary is allowed to choose different transition probabilities in the same state-action pair in different stages. This is called a "dynamic model" of ambiguity [32]. Per this model, the ambiguity set corresponding to policy π is expressed as $\mathcal{P}^\pi_\infty = \bigtimes_{t=1}^{\infty} \bigtimes_{i \in S} \mathcal{P}(i, \pi(i))$. The decision-maker can maximize the worst-case expected total discounted reward by solving the problem

$$\tilde{J} \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \; \tilde{J}^\pi, \text{ where } \tilde{J}^\pi \stackrel{\text{def}}{=} \inf_{P^\pi_\infty \in \mathcal{P}^\pi_\infty} J^\pi_{P^\pi_\infty}. \tag{2.14}$$

Here, $P^\pi_\infty \stackrel{\text{def}}{=} (P^\pi_1, P^\pi_2, \cdots)$ denotes an infinite tuple of $n \times n$ transition probability matrices such that the pmf $p_t(\cdot|i, \pi(i))$ in the ith row of P^π_t belongs to $\mathcal{P}(i, \pi(i)) \subseteq \Delta^n$. Moreover, with a slight abuse of notation, $J^\pi_{P^\pi_\infty}$ represents the expected total discounted reward induced by the infinite tuple P^π_∞ of transition probability matrices. A policy that maximizes $\tilde{J}^\pi$ over Π can be found via a robust counterpart of value iteration or policy iteration [32].

In this section, we focus on rectangular stationary infinite-horizon RMDPs where the decision-maker employs data-driven ambiguity sets

$$\mathcal{P}^N(i, a; \epsilon) \stackrel{\text{def}}{=} \left\{ p \in \Delta^n | d(p, \hat{p}^N(\cdot|i, a)) \leq \epsilon \right\}, \; \forall (i, a) \in S \times A, \tag{2.15}$$

where $0 \leq \epsilon \leq \infty$. Here, $\hat{p}^N(\cdot|i,a)$ is the empirical estimate of the true transition pmf upon choosing action $a \in A$ in state $i \in S$. Following the style of notation in Section 2.4, we write the data-driven ambiguity set corresponding to policy π as

$$\mathcal{P}_{\infty}^{N,\pi}(\epsilon) = \bigtimes_{t=1}^{\infty} \bigtimes_{i \in S} \mathcal{P}^N(i, \pi(i); \epsilon). \tag{2.16}$$

We then rewrite the corresponding problem (2.14) for emphasis and ease of reference as

$$\tilde{J}^N(\epsilon) \stackrel{\text{def}}{=} \max_{\pi \in \Pi} \ \tilde{J}^{N,\pi}(\epsilon), \text{ where } \tilde{J}^{N,\pi}(\epsilon) \stackrel{\text{def}}{=} \inf_{P_{\infty}^{\pi} \in \mathcal{P}_{\infty}^{N,\pi}(\epsilon)} J_{P_{\infty}^{\pi}}^{\pi}. \tag{2.17}$$

We call this the data-driven stationary infinite-horizon RMDP.

As in Section 2.4, we will assume throughout that, for every $(i,a) \in S \times A$, the true transition pmf has full support S. Again, this is done merely for notational convenience. Finally, similar to Section 2.4, we work under the assumption that ambiguity balls $\mathcal{P}^N(i,a;\epsilon)$ are nonempty, wp 1, for all states i, actions a, and sample-sizes N. This ensures that the optimal value of each inner-optimization problem encountered in this section is finite, wp 1.

2.5.1 Value convergence to the true optimal value

This section extends Theorem 1 and Theorem 2 to the stationary infinite-horizon case.

Theorem 4. Suppose the radii of the ambiguity balls are dependent on the sample-size N such that $\lim_{N \to \infty} \epsilon^N = 0$. Suppose Assumption 1 holds. Then,

$$\tilde{J}^N(\epsilon^N) \xrightarrow{N,\text{wp1}} J^*.$$

Proof. Fix any $\pi \in \Pi$ and let P^{π} denote the corresponding true transition probability matrix of size $n \times n$. As done in the proof of Theorem 1, it is enough to show that

$$\tilde{J}^{N,\pi}(\epsilon^N) \xrightarrow{N,\text{wp1}} J_{P^{\pi}}^{\pi}. \tag{2.18}$$

Fix any $T \in \{1, 2, \cdots\}$. By the rectangularity property, that is, by the Cartesian product structure of $\mathcal{P}_\infty^{N,\pi}(\epsilon^N)$, it can be decomposed as $\mathcal{P}_\infty^{N,\pi}(\epsilon^N; (1:T)) \times \mathcal{P}_\infty^{N,\pi}(\epsilon^N; (T+1:\infty))$. Here,

$$\mathcal{P}_\infty^{N,\pi}(\epsilon^N; (1:T)) \stackrel{\text{def}}{=} \bigtimes_{t=1}^{T} \bigtimes_{i \in S} \mathcal{P}^N(i, \pi(i); \epsilon^N)$$
$$\mathcal{P}_\infty^{N,\pi}(\epsilon^N; (T+1:\infty)) \stackrel{\text{def}}{=} \bigtimes_{t=T+1}^{\infty} \bigtimes_{i \in S} \mathcal{P}^N(i, \pi(i); \epsilon^N).$$

Consequently, any infinite tuple $Q_\infty^\pi = (Q_1^\pi, Q_2^\pi, \cdots)$ of transition probability matrices within $\mathcal{P}_\infty^{N,\pi}(\epsilon^N)$ can be decomposed as $(\underbrace{Q_1^\pi, Q_2^\pi, \cdots, Q_T^\pi}_{Q^\pi(1:T)}; \underbrace{Q_{T+1}^\pi, Q_{T+2}^\pi, \cdots}_{Q^\pi(T+1:\infty)})$, where $Q^\pi(1:T) \in \mathcal{P}_\infty^{N,\pi}(\epsilon^N; (1:T))$ and $Q^\pi(T+1:\infty) \in \mathcal{P}_\infty^{N,\pi}(\epsilon^N; (T+1:\infty))$. Therefore,

$$\tilde{J}^{N,\pi}(\epsilon^N) = \inf_{Q_\infty^\pi \in \mathcal{P}_\infty^{N,\pi}(\epsilon^N)} \sum_{i=1}^{n} f(i) \Bigg\{ \mathop{\mathbb{E}}_{Q_\infty^\pi} \Bigg[\sum_{t=1}^{T} \alpha^{t-1} r(\mathbf{s}_{t+1} | \mathbf{s}_t, \pi(\mathbf{s}_t)) + \sum_{t=T+1}^{\infty} \alpha^{t-1} r(\mathbf{s}_{t+1} | \mathbf{s}_t, \pi(\mathbf{s}_t)) \Bigg| \mathbf{s}_1 = i \Bigg] \Bigg\}.$$

Suppose $C > 0$ is some constant such that $|r(j|i,a)| \leq C$, for all $i, j \in S$ and $a \in A$. Such a constant exists because S and A are finite sets. Using this in the above expression for $\tilde{J}^{N,\pi}(\epsilon^N)$ and noticing that the stochastic state evolution $(\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{T+1})$ does not depend on $Q^\pi(T+1:\infty)$, we have,

$$\frac{-\alpha^T C}{1-\alpha} + \inf_{Q^\pi(1:T) \in \mathcal{P}^{N,\pi}(\epsilon^N; (1:T))} \sum_{i=1}^{n} f(i) \mathop{\mathbb{E}}_{Q^\pi(1:T)} \left[\left(\sum_{t=1}^{T} \alpha^{t-1} r(\mathbf{s}_{t+1} | \mathbf{s}_t, \pi(\mathbf{s}_t)) \right) \middle| \mathbf{s}_1 = i \right]$$
$$\leq \tilde{J}^{N,\pi}(\epsilon^N) \leq$$
$$\frac{\alpha^T C}{1-\alpha} + \inf_{Q^\pi(1:T) \in \mathcal{P}^{N,\pi}(\epsilon^N; (1:T))} \sum_{i=1}^{n} f(i) \mathop{\mathbb{E}}_{Q^\pi(1:T)} \left[\left(\sum_{t=1}^{T} \alpha^{t-1} r(\mathbf{s}_{t+1} | \mathbf{s}_t, \pi(\mathbf{s}_t)) \right) \middle| \mathbf{s}_1 = i \right].$$

Note that $\inf_{Q^\pi(1:T)\in\mathcal{P}^{N,\pi}(\epsilon^N;(1:T))} \sum_{i=1}^{n} f(i) \mathop{\mathbb{E}}_{Q^\pi(1:T)} \left[\left(\sum_{t=1}^{T} \alpha^{t-1} r(\mathbf{s}_{t+1}|\mathbf{s}_t, \pi(\mathbf{s}_t)) \right) \middle| \mathbf{s}_1 = i \right]$ is the value of executing policy π in every stage of a T-stage data-driven RMDP as in Section 2.4. We use the notation $\tilde{J}^{N,\pi}(T;\epsilon^N)$ to represent this value. Similarly, let $J^\pi_{P^\pi}(T)$ denote the value of implementing policy π in every stage of a T-stage MDP with a true transition probability matrix P^π of size $n \times n$. From Theorem 1, we know that $\tilde{J}^{N,\pi}(T;\epsilon^N) \xrightarrow{N,\text{wp1}} J^\pi_{P^\pi}(T)$. Thus, utilizing this limit in the above lower and upper bounds on $\tilde{J}^{N,\pi}(\epsilon^N)$ yields

$$\frac{-\alpha^T C}{1-\alpha} + J^\pi_{P^\pi}(T) \leq \liminf_{N\to\infty} \tilde{J}^{N,\pi}(\epsilon^N) \leq \limsup_{N\to\infty} \tilde{J}^{N,\pi}(\epsilon^N) \leq \frac{\alpha^T C}{1-\alpha} + J^\pi_{P^\pi}(T).$$

The above inequalities hold for every $T \in \{1, 2, \cdots\}$. Also, $\frac{\alpha^T C}{1-\alpha}$ converges to 0 as $T \to \infty$. We further note from (2.11)-(2.12) that $\lim_{T\to\infty} J^\pi_{P^\pi}(T) = J^\pi_{P^\pi}$ by definition. Passing to the limit, we get that, wp 1, $\liminf_{N\to\infty} \tilde{J}^{N,\pi}(\epsilon^N) = J^\pi(\epsilon^N)$ and $\limsup_{N\to\infty} \tilde{J}^{N,\pi}(\epsilon^N) = J^\pi(\epsilon^N)$. Hence, $\tilde{J}^{N,\pi}(\epsilon^N) \xrightarrow{N,\text{wp1}} J^\pi(\epsilon^N)$. □

The next example demonstrates that we cannot drop Assumption 1 from the hypothesis of this theorem.

Example 6. This is a modification of Example 3. Consider a 2-state, 1-action, stationary infinite-horizon MDP with $S = \{1, 2\}$; $A = \{a\}$; all true $p(\cdot|\cdot,\cdot) = 0.5$; rewards $r(1|\cdot, a) = 2$ and $r(2|\cdot, a) = 0$; discount factor $\alpha \in (0, 1)$; and initial state pmf $f(1) = f(2) = 0.5$. It is easy to calculate that the optimal value $J^* = 1/(1-\alpha)$. Now suppose that the decision-maker employs the distance described in Example 2. Recall that Assumption 1 does not hold for this distance. It is straightforward to show, via an argument identical to Example 3, that $0 = \lim_{N\to\infty} \tilde{J}^N(\epsilon^N) \neq J^* = 1/(1-\alpha)$. Thus, the conclusion of Theorem 4 fails.

The next result is a counterpart of Theorem 2. The proof is identical to the proof of Theorem 2, and hence omitted. Let $\hat{\pi}^N$ denote an optimal policy that solves the data-driven stationary infinite-horizon RMDP problem (2.17) and $P^{\hat{\pi}^N}$ be the corresponding true transition probability matrix. Let $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ denote the value of this policy in the true MDP.

Theorem 5. Suppose the radii of the ambiguity balls are dependent on the sample-size N such that $\lim_{N\to\infty} \epsilon^N = 0$. Suppose Assumption 1 holds. Then, the robust optimal policy $\hat{\pi}^N$ solves (2.13) for all sufficiently large N, wp 1. Consequently, $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} = J^*$, for all sufficiently large N, wp 1.

2.5.2 Probabilistic guarantee on the performance of robust optimal policy

This section extends Theorem 3 to the stationary infinite-horizon case.

Theorem 6. Fix the sample-size $N \geq 1$. Suppose Assumption 3 holds. Consider any fixed $\gamma \in (0,1)$, and let $0 < \beta(\gamma) = 1 - (1-\gamma)^{1/nm} < 1$. Let $0 < \epsilon^N_{\beta(\gamma)} < \sup_{p,q\in\Delta^n} d(p,q)$ be as in Assumption 3. Then,

$$\mathbb{P}\left[J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \geq \tilde{J}^N(\epsilon^N_{\beta(\gamma)})\right] \geq 1-\gamma. \tag{2.19}$$

Proof. Define Bellman's evaluation operator $\Phi_{\hat{\pi}^N} : \mathbb{R}^n \to \mathbb{R}^n$ as

$$[\Phi_{\hat{\pi}^N} V](i) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{p(\cdot|i,\hat{\pi}^N(i))} \left[r(\mathbf{s}|i,\hat{\pi}^N(i)) + \alpha V(\mathbf{s})\right], \text{ for } i = 1,\ldots,n. \tag{2.20}$$

Here, the subscript emphasizes that the expectation is taken with respect to the true transition pmf $p(\cdot|i,\hat{\pi}^N(i))$ in state i for policy $\hat{\pi}^N$. A standard monotonicity and successive approximation argument [50, Chapter 6] in the theory of stationary infinite-horizon MDPs provides that

$$\begin{aligned}\left\{[\Phi_{\hat{\pi}^N}\tilde{J}^N(\epsilon^N_{\beta(\gamma)})](i) \geq \tilde{J}^N(i;\epsilon^N_{\beta(\gamma)}),\ i=1,\ldots,n\right\}\\ \subseteq \left\{J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}(i) \geq \tilde{J}^N(i;\epsilon^N_{\beta(\gamma)}),\ i=1,\ldots,n\right\}.\end{aligned} \tag{2.21}$$

A similar property was also employed in [70] in a continuous state, continuous action, data-driven robust stochastic control problem with ambiguity sets defined using the Wasserstein metric. This yields,

$$\mathbb{P}\left[J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \geq \tilde{J}^N(\epsilon^N_{\beta(\gamma)})\right] = \mathbb{P}\left[\sum_{i=1}^n f(i) J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}(i) \geq \sum_{i=1}^n f(i)\tilde{J}^N(i;\epsilon^N_{\beta(\gamma)})\right]$$

$$\overset{(a)}{\geq} \mathbb{P}\left[[\Phi_{\hat{\pi}^N}\tilde{J}^N(\epsilon^N_{\beta(\gamma)})](i) \geq \tilde{J}^N(i;\epsilon^N_{\beta(\gamma)}),\ \forall i \in S\right]$$

$$\overset{(b)}{=} \mathbb{P}\left[\underset{p(\cdot|i,\hat{\pi}^N(i))}{\mathbb{E}}\left[r(\mathbf{s}|i,\hat{\pi}^N(i)) + \alpha\tilde{J}^N(\mathbf{s};\epsilon^N_{\beta(\gamma)})\right]\right.$$

$$\left. \geq \inf_{q(\cdot|i,\hat{\pi}^N(i))\in\mathcal{P}^N(i,\hat{\pi}^N(i);\epsilon^N_{\beta(\gamma)})}\left(\underset{q(\cdot|i,\hat{\pi}^N(i))}{\mathbb{E}}\left[r(\mathbf{s}|i,\hat{\pi}^N(i)) + \alpha\tilde{J}^N(\mathbf{s};\epsilon^N_{\beta(\gamma)})\right]\right),\ \forall i \in S\right]$$

$$\geq \mathbb{P}\left[p(\cdot|i,\hat{\pi}^N(i)) \in \mathcal{P}^N(i,\hat{\pi}^N(i);\epsilon^N_{\beta(\gamma)}),\ \forall i \in S\right]$$

$$= \mathbb{P}\left[d(p(\cdot|i,\hat{\pi}^N(i)),\hat{p}^N(\cdot|i,\hat{\pi}^N(i))) \leq \epsilon^N_{\beta(\gamma)},\ \forall i \in S\right]$$

$$\geq \mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)) \leq \epsilon^N_{\beta(\gamma)},\ \forall i \in S,\ \forall a \in A\right]$$

$$= \prod_{i=1}^n\prod_{a=1}^m \mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)) \leq \epsilon^N_{\beta(\gamma)}\right] \geq \prod_{i=1}^n\prod_{a=1}^m (1-\beta(\gamma)) = (1-\beta(\gamma))^{nm} = 1-\gamma.$$

Inequality "(a)" follows from (2.21). Equality "(b)" follows from (2.20) and the fact that actions $\hat{\pi}^N(i)$ attain the maxima in robust Bellman's equations of optimality (Corollary 3.1 in [32]). In particular, the latter yields that for all $i = 1, \ldots, n$

$$\tilde{J}^N(i;\epsilon^N_{\beta(\gamma)}) = \inf_{q(\cdot|i,\hat{\pi}^N(i))\in\mathcal{P}^N(i,\hat{\pi}^N(i);\epsilon^N_{\beta(\gamma)})}\left(\underset{q(\cdot|i,\hat{\pi}^N(i))}{\mathbb{E}}\left[r(\mathbf{s}|i,\hat{\pi}^N(i)) + \alpha\tilde{J}^N(\mathbf{s};\epsilon^N_{\beta(\gamma)})\right]\right).$$

The remaining inequalities and equalities are derived similar to that in the proof of Theorem 3. □

The radius of the ambiguity set depends only on n and m in this theorem, unlike the finite horizon case in Theorem 3 where it also depends on T. As stated in Section 2.4.2, the dependence on T was induced by the multi-stage nonstationary nature of the problem there. Here, although the problem is infinite-horizon and hence multi-stage, it is stationary. This

stationarity renders parts of the proof similar to a single-stage problem. This manifests itself, for instance, in (2.21) and equality "(b)".

The next example demonstrates that we cannot drop Assumption 3 from the hypothesis of this theorem.

Example 7. This is a modification of Example 5. Consider a stationary infinite-horizon MDP with $S = \{1, 2\}$; $A = \{a, b\}$. We assume that both the actions are allowed in state 1 while only action a is allowed in state 2. All true transition probabilities $p(\cdot|\cdot, \cdot) = 0.5$; rewards $r(\cdot|\cdot, a) = 0, r(1|1, b) = 1, r(2|1, b) = -3$; discount factor $\alpha \in (0, 1)$; and initial state pmf $f(1) = f(2) = 0.5$. There are only two policies — $\pi_1(1) = a, \pi_1(2) = a$ and $\pi_2(1) = b, \pi_2(2) = a$. It is easy to see that $J^{\pi_1} = 0$. For the policy π_2, using the policy evaluation equations [50, Equation (6.1.9)] for the true MDP gives us $J^{\pi_2}(1) = \frac{-1+\alpha/2}{1-\alpha}$ and $J^{\pi_2}(2) = \frac{-\alpha/2}{1-\alpha}$. Consequently, $J^{\pi_2} = 0.5(J^{\pi_2}(1) + J^{\pi_2}(2)) = \frac{-1}{2(1-\alpha)}$. Assume that the decision-maker knows the true transition probabilities for the state-action pairs $(1, a)$, $(2, a)$ but does not know them for the pair $(1, b)$. The decision-maker has a single sample ($N = 1$) of the next state reached upon choosing action b in state 1. Suppose this observed state was 1 and the decision-maker utilizes the 0/1 distance from Example 4 where Assumption 3 does not hold. Using an argument identical to Example 5, we can show that the worst-case pmf equals empirical pmf, $\hat{p}^1(1|1, b) = 1$ and $\hat{p}^1(2|1, b) = 0$, for any radius $0 < \epsilon < 1$. Further, $\tilde{J}^{N,\pi_1} = 0$. Using the policy evaluation equations [32, Equation (29)] for the data-driven RMDP, we get $\tilde{J}^{N,\pi_2}(1; \epsilon) = \frac{1}{1-\alpha}$ and $\tilde{J}^{N,\pi_2}(2; \epsilon) = \frac{\alpha}{(1-\alpha)(2-\alpha)}$. Hence, $\tilde{J}^{N,\pi_2} = \frac{1}{(1-\alpha)(2-\alpha)}$ which implies that $\tilde{J}^N(\epsilon) = \tilde{J}^{N,\pi_2} = \frac{1}{(1-\alpha)(2-\alpha)}$. This shows that $\mathbb{P}(J^{\pi_2} < \tilde{J}^{N,\pi_2}(\epsilon)) \geq 0.5$. Thus, (say) if $\gamma = 0.1$, the conclusion of Theorem 6 fails.

2.5.3 Rate of convergence

Theorem 4 and Theorem 5 are asymptotic convergence results; they do not provide insight into the rate of convergence. Specifically, it is not clear how close $\tilde{J}^N(\epsilon^N)$ or $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ is to J^*, as a function of N. Theorem 7 at the end of this section addresses this issue. It relies on a

few intemediate bounds that are derived next. Recall that the true transition pmf for any state $i \in S$ and action $a \in A$ is denoted by $p(\cdot|i,a)$. Similarly, the matrix of true transition probabilities induced by any policy $\pi \in \Pi$ is denoted by P^π. Furthermore, there is a constant $C > 0$ such that $\max_{\substack{i,j\in S\\ a\in A}} |r(j|i,a)| < C$, because the state- and action-spaces are finite. The next result, in varying forms, is known as the "Simulation Lemma" in the reinforcement learning literature (see [52, Lemma A.1]). The lemma estimates how a difference between two transition matrices translates into a difference between the corresponding values of a policy.

Lemma 4. [52, Lemma A.1] Let P_1^π and P_2^π be any two transition matrices induced by a policy $\pi \in \Pi$. Then,

$$\left| J^\pi_{P_1^\pi} - J^\pi_{P_2^\pi} \right| \leq \frac{2C}{(1-\alpha)^2} \max_{i\in S} d_{\mathrm{TV}}(p_1(\cdot|i,\pi(i)), p_2(\cdot|i,\pi(i))). \tag{2.22}$$

Here, $d_{\mathrm{TV}}(p,q) \stackrel{\mathrm{def}}{=} \frac{1}{2}\|p-q\|_1$ is the Total Variation (TV) distance between pmfs p and q [22].

Lemma 5. Suppose that the distance function d satisfies the generalized Pinsker inequality stated in Assumption 2, and that the function ψ introduced there is nondecreasing. Fix the sample-size $N \geq 1$. Suppose the decision-maker employs ambiguity balls of radius ϵ^N. Then,

$$\left| \tilde{J}^N(\epsilon^N) - J^* \right| \leq \frac{\left(\psi(\epsilon^N) + \max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)))\right) C}{(1-\alpha)^2}, \text{ and} \tag{2.23}$$

$$J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \leq \frac{2\left(\psi(\epsilon^N) + \max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)))\right) C}{(1-\alpha)^2}. \tag{2.24}$$

Proof. Consider any stationary policy $\pi \in \Pi$ and fix any $\delta > 0$. By definition of the infimum, there exists a feasible solution to the inner optimization problem whose value is within δ of the optimal value $\tilde{J}^{N,\pi}(\epsilon^N)$. In fact, by [32, Lemma 3.3], such a value can be attained by a

stationary solution of the form $P_\infty^{N,\pi,\delta} = (P^{N,\pi,\delta}, P^{N,\pi,\delta}, \ldots)$. We thus denote this value by $J^\pi_{P^{N,\pi,\delta}}$ and note that $\left|J^\pi_{P^{N,\pi,\delta}} - \tilde{J}^{N,\pi}(\epsilon^N)\right| \leq \delta$. Hence,

$$
\begin{aligned}
\left|\tilde{J}^N(\epsilon^N) - J^*\right| &= \left|\max_{\pi\in\Pi} \tilde{J}^{N,\pi}(\epsilon^N) - \max_{\pi\in\Pi} J^\pi_{P^\pi}\right| \\
&\leq \max_{\pi\in\Pi} \left|\tilde{J}^{N,\pi}(\epsilon^N) - J^\pi_{P^\pi}\right| \\
&\leq \max_{\pi\in\Pi} |J^\pi_{P^{N,\pi,\delta}} - J^\pi_{P^\pi}| + \delta \\
&\overset{(a)}{\leq} \max_{\pi\in\Pi} \left\{ \frac{2C \max\limits_{i\in S} \; d_{\mathrm{TV}}(p(\cdot|i,\pi(i)), p^{N,\pi,\delta}(\cdot|i,\pi(i)))}{(1-\alpha)^2} \right\} + \delta \\
&\overset{(b)}{\leq} \frac{2C}{(1-\alpha)^2} \left(\max_{\pi\in\Pi, i\in S} \left\{ d_{\mathrm{TV}}(p(\cdot|i,\pi(i)), \hat{p}^N(\cdot|i,\pi(i))) + \right.\right. \\
&\qquad\qquad \left.\left. d_{\mathrm{TV}}(\hat{p}^N(\cdot|i,\pi(i)), p^{N,\pi,\delta}(\cdot|i,\pi(i))) \right\} \right) + \delta \\
&\overset{(c)}{\leq} \frac{C}{(1-\alpha)^2} \left(\max_{\pi\in\Pi, i\in S} \left\{ \psi(d(p(\cdot|i,\pi(i)), \hat{p}^N(\cdot|i,\pi(i)))) + \right.\right. \\
&\qquad\qquad \left.\left. \psi(d(\hat{p}^N(\cdot|i,\pi(i)), p^{N,\pi,\delta}(\cdot|i,\pi(i)))) \right\} \right) + \delta \\
&\leq \frac{C}{(1-\alpha)^2} \left(\max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a))) + \right. \\
&\qquad\qquad \left. \max_{\pi\in\Pi, i\in S} \psi(d(\hat{p}^N(\cdot|i,\pi(i)), p^{N,\pi,\delta}(\cdot|i,\pi(i)))) \right) + \delta \\
&\overset{(d)}{\leq} \frac{C}{(1-\alpha)^2} \left(\max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a))) + \psi(\epsilon^N) \right) + \delta.
\end{aligned}
$$

The inequality in "(a)" follows from (2.22) in Lemma 4. Step "(b)" is the triangle inequality. The inequality in "(c)" follows from the generalized Pinsker's inequality (Assumption 2) along with the fact that $d_{\mathrm{TV}}(p,q) = \frac{1}{2}\|p-q\|_1$. For the final inequality in "(d)", first note that $p^{N,\pi,\delta}(\cdot|i,\pi(i)) \in \mathcal{P}^N(i,\pi(i);\epsilon^N)\ \forall \pi \in \Pi, i \in S$. Hence, $d(\hat{p}^N(\cdot|i,\pi(i)), p^{N,\pi,\delta}(\cdot|i,\pi(i))) \leq \epsilon^N$, $\forall \pi \in \Pi, i \in S$. Since ψ is nondecreasing, the inequality in "(d)" follows. Finally, since

$\delta > 0$ is arbitrary, the bound in (2.23) follows.

For the second claim, we prove that $J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \leq 2 \max_{\pi \in \Pi} \left| \tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{\pi}} \right|$. The bound in (2.24) then follows by repeating the argument in the proof of the first claim. Note that

$$J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} = \max_{\pi \in \Pi} J^{\pi}_{P^{\pi}} - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \leq \left| \max_{\pi \in \Pi} J^{\pi}_{P^{\pi}} - \max_{\pi \in \Pi} \tilde{J}^{N,\pi}(\epsilon^N) \right| + \left| \max_{\pi \in \Pi} \tilde{J}^{N,\pi}(\epsilon^N) - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \right|$$
$$\leq \max_{\pi \in \Pi} \left| J^{\pi}_{P^{\pi}} - \tilde{J}^{N,\pi}(\epsilon^N) \right| + \left| \max_{\pi \in \Pi} \tilde{J}^{N,\pi}(\epsilon^N) - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \right| \leq 2 \max_{\pi \in \Pi} \left| \tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{\pi}} \right|.$$

The last inequality follows by definition of $\hat{\pi}^N$ since

$$\left| \max_{\pi \in \Pi} \tilde{J}^{N,\pi}(\epsilon^N) - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \right| = \left| \tilde{J}^{N,\hat{\pi}^N}(\epsilon^N) - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \right| \leq \max_{\pi \in \Pi} \left| \tilde{J}^{N,\pi}(\epsilon^N) - J^{\pi}_{P^{\pi}} \right|.$$

This completes the proof. □

We are now ready to state and prove the main result of this section — the rate of convergence in terms of the radii of the ambiguity balls.

Theorem 7. Suppose that the distance function d satisfies the generalized Pinsker inequality stated in Assumption 2, and that the function ψ introduced there is nondecreasing. Fix the sample-size $N \geq 1$. Suppose the decision-maker employs ambiguity balls of radius ϵ^N. Let $\beta \in [0, 1]$ be such that

$$\Pr[d(p(\cdot|i, a), \hat{p}^N(|i, a)) \leq \epsilon^N] \geq 1 - \beta, \ \forall i \in S, a \in A. \tag{2.25}$$

Then, with probability at least $(1 - \beta)^{nm}$, we have,

$$\left| \tilde{J}^N(\epsilon^N) - J^* \right| \leq \frac{2C\psi(\epsilon^N)}{(1 - \alpha)^2} \quad \text{and} \quad J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \leq \frac{4C\psi(\epsilon^N)}{(1 - \alpha)^2}. \tag{2.26}$$

Proof. Observe, from bounds (2.23) and (2.24) in Lemma 5, that the two inequalities in (2.26) hold if the event

$$\left[\max_{i\in S,a\in A}\psi(d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)))\leq\psi(\epsilon^N)\right]$$

occurs. The conclusion of the theorem then follows because

$$\begin{aligned}
&\mathbb{P}\left[\max_{i\in S,a\in A}\psi(d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)))\leq\psi(\epsilon^N)\right]\\
&=\mathbb{P}\left[\psi(d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)))\leq\psi(\epsilon^N),\quad\forall i\in S,a\in A\right]\\
&\geq\mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a))\leq\epsilon^N,\quad\forall i\in S,a\in A\right]\\
&=\prod_{i\in S}\prod_{a\in A}\mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a))\leq\epsilon^N\right]\geq(1-\beta)^{nm}.
\end{aligned}$$

The first inequality holds because ψ is nondecreasing. The second equality holds because samples are independent across S and A. The final inequality follows from (2.25) because $|S|=n$ and $|A|=m$. □

The bounds in (2.26) demonstrate that the optimality errors $\left|\tilde{J}^N(\epsilon^N)-J^*\right|$ and $J^*-J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ can be made smaller than any $\delta>0$ by choosing ϵ^N to be sufficiently small, as all other terms in the bounds are known. The additional requirement that ψ be nondecreasing, is mild — it holds for all distances we list in Section 2.6. An appropriate value of β in (2.25) can be derived from concentration inequalities similar to Section 2.6. We expect these values of β to be smaller for larger values of ϵ^N and correspondingly larger values of $\psi(\epsilon^N)$. In other words, a higher probabilistic confidence is associated with larger right hand sides in the two inequalities in (2.26). This is intuitive, since these larger upper bounds in (2.26) mean that $\tilde{J}^N(\epsilon^N)$ and $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ could be farther away from J^*. Moreover, rougly speaking, everything else being equals, the probabilistic confidence $(1-\beta)^{nm}$ deteriorates as the number of states n or the number of actions m increases.

A similar convergence rate for finite-horizon problems in Section 2.4 can also be derived. We omitted it there for brevity and to avoid repetition.

2.6 Distances that satisfy Assumption 1 and Assumption 3

This section lists several well-known distance functions that can be utilized to construct data-driven ambiguity sets. Distances that are also metrics are identified as such. We begin with distances that satisfy Assumption 2 and Assumption 3 (recall that Assumption 2 is a sufficient condition for Assumption 1). For each of these distances, we also provide a formula for the ambiguity ball radius ϵ_β^N that can be utilized in Theorem 3 and Theorem 6, for any $\beta \in (0,1)$. Although these formulas can be pieced together from existing results in the literature, we were unable to find them all in one place and in the precise format we need. We therefore derive them briefly for completeness, and cast them in appropriate form within our context. We then list distances that satisfy Assumption 2, but that do not have an associated concentration inequality, in the literature known to us, of the form (2.9) suitable for Assumption 3. For those distances, we provide an approximate counterpart of (2.9).

2.6.1 Total Variation (TV)

This distance (metric) is $d_{\mathrm{TV}}(p,q) \stackrel{\text{def}}{=} \frac{1}{2}\|p-q\|_1$ [22]. It is easy to see that $d_{\mathrm{TV}}(q,q)=0$, for all $q \in \Delta^n$. Assumption 2 holds with $\psi(z)=2z$, for $z \in \mathbb{R}_+$.

Lemma 6. The TV metric satisfies the concentration inequality

$$\mathbb{P}\left[d_{\mathrm{TV}}(q,\hat{q}^N) \le \epsilon\right] \ge 1 - 2\exp\left(-\frac{N}{2}\left(2\epsilon - \sqrt{\frac{n}{N}}\right)^2\right), \quad \forall \epsilon \ge \frac{1}{2}\left(\sqrt{\frac{n}{N}} + \sqrt{\frac{2\ln(2)}{N}}\right). \tag{2.27}$$

Proof. Consider any $\epsilon \ge \frac{1}{2}\sqrt{\frac{n}{N}}$. We have,

$$\begin{aligned}\mathbb{P}\left[d_{\mathrm{TV}}(q,\hat{q}^N) \le \epsilon\right] &= \mathbb{P}\left[\left\|q-\hat{q}^N\right\|_1 \le 2\epsilon\right] \\ &= \mathbb{P}\left[\left\|q-\hat{q}^N\right\|_1 - \mathbb{E}\left(\left\|q-\hat{q}^N\right\|_1\right) \le 2\epsilon - \mathbb{E}\left(\left\|q-\hat{q}^N\right\|_1\right)\right]\end{aligned}$$

$$\begin{aligned}
&\overset{(a)}{\geq} \mathbb{P}\left[\left\|q-\hat{q}^N\right\|_1 - \mathbb{E}\left(\left\|q-\hat{q}^N\right\|_1\right) \leq 2\epsilon - \sqrt{\frac{n}{N}}\right] \\
&\geq 1 - \mathbb{P}\left[\left|\left\|q-\hat{q}^N\right\|_1 - \mathbb{E}\left(\left\|q-\hat{q}^N\right\|_1\right)\right| > 2\epsilon - \sqrt{\frac{n}{N}}\right] \\
&\overset{(b)}{\geq} 1 - 2\exp\left(-\frac{N}{2}\left(2\epsilon - \sqrt{\frac{n}{N}}\right)^2\right).
\end{aligned}$$

Here, "(a)" follows from Lemma 5 in [8], which establishes that $\mathbb{E}\left(\left\|q-\hat{q}^N\right\|_1\right) \leq \sqrt{n/N}$ and "(b)" holds by [8, Inequality (17)], which in turn follows by McDiarmid's inequality [44]. The additional restriction of $\epsilon \geq \frac{1}{2}\left(\sqrt{\frac{n}{N}} + \sqrt{\frac{2\ln(2)}{N}}\right)$ is imposed in (2.27) to ensure that the probability lower bound is nonnegative. □

It is easy to verify that, by equating $2\exp\left(-\frac{N}{2}\left(2\epsilon - \sqrt{\frac{n}{N}}\right)^2\right)$ to $0 < \beta < 1$, we obtain the ambiguity ball radius

$$\epsilon_\beta^N(\mathrm{TV}) = \frac{1}{2}\left(\sqrt{\frac{n}{N}} + \sqrt{\frac{2\ln(2/\beta)}{N}}\right). \tag{2.28}$$

It is known that $\sup_{p,q\in\Delta^n} d_{\mathrm{TV}}(p,q) = 1$ [22]. Thus, $\epsilon_\beta^N(\mathrm{TV}) < 1$ as required in Assumption 3, for all sufficiently large N. As an aside, although this is not a requirement in Assumption 3, we note that $\lim_{N\to\infty} \epsilon_\beta^N(\mathrm{TV}) = 0$. This means that this ambiguity ball radius satisfies the hypothesis of Theorem 1 and Theorem 4.

2.6.2 Burg

This distance is $d_{\mathrm{Burg}}(p,q) = \sum_{i=1}^n q_i \ln\left(\frac{q_i}{p_i}\right)$. The convention in the literature is to interpret $0\ln(0/x) = 0$, for all real numbers x [22, page 422]. This helps to see that $d(q,q) = 0$, for all $q \in \Delta^n$. Our definition of the Burg distance here is consistent with [7, Table 2]. Note that $d_{\mathrm{Burg}}(p,q) = d_{\mathrm{KL}}(q,p)$ where d_{KL} denotes the Kullback-Leibler distance discussed in Section 2.6.5. From [22, page 429], we have $2d_{\mathrm{TV}}(p,q) = \|p-q\|_1 \leq \sqrt{2d_{\mathrm{Burg}}(p,q)}$. Thus,

Assumption 2 holds with $\psi(z) = \sqrt{2z}$, for $z \in \mathbb{R}_+$. Inequality (1) in [43] implies that

$$\mathbb{P}\left[d_{\text{Burg}}(q, \hat{q}^N) \leq \epsilon\right] \geq 1 - \binom{N+n-1}{n-1} \exp(-N\epsilon), \quad \forall \epsilon \geq \frac{1}{N} \ln \binom{N+n-1}{n-1}. \tag{2.29}$$

The restriction $\epsilon \geq \frac{1}{N} \ln \binom{N+n-1}{n-1}$ is imposed to ensure that the probability lower bound is nonnegative. By equating $\binom{N+n-1}{n-1} \exp(-N\epsilon)$ to $0 < \beta < 1$, we obtain the ambiguity ball radius

$$\epsilon_\beta^N(\text{Burg}) = \frac{1}{N} \ln \left\{ \binom{N+n-1}{n-1} / \beta \right\}. \tag{2.30}$$

It is known that $\sup_{p,q \in \Delta^n} d_{\text{Burg}}(p, q) = \infty$ [22]. Thus, $\epsilon_\beta^N(\text{Burg}) < \sup_{p,q \in \Delta^n} d_{\text{Burg}}(p, q)$ as required in Assumption 3, for all N. Again, it is possible to show, after some algebraic simplification, that $\lim_{N \to \infty} \epsilon_\beta^N(\text{Burg}) = 0$. So this ambiguity ball radius satisfies the hypothesis of Theorem 1 and Theorem 4.

2.6.3 Hellinger

This distance is given by $d_{\text{H}}(p, q) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2$ [7, Table 2]. It is easy to see that $d(q, q) = 0$, for all $q \in \Delta^n$. Inequality (8) in [22] implies that $\|p - q\|_1 = 2 d_{\text{TV}}(p, q) \leq 2 \left[d_{\text{H}}(p, q)\right]^{1/2}$. Thus, Assumption 2 holds with $\psi(z) = 2\sqrt{z}$, for $z \in \mathbb{R}_+$.

Lemma 7. The Hellinger distance satisfies the concentration inequality

$$\mathbb{P}\left[d_{\text{H}}(q, \hat{q}^N) \leq \epsilon\right] \geq 1 - 2 \exp\left(-\frac{N}{2}\left(\epsilon - \sqrt{\frac{n}{N}}\right)^2\right), \quad \forall \epsilon \geq \sqrt{\frac{n}{N}} + \sqrt{\frac{2\ln(2)}{N}}. \tag{2.31}$$

Proof. Inequality (8) in [22] states that $\frac{d_{\text{H}}(p,q)}{2} \leq d_{\text{TV}}(p, q)$. Thus, we have,

$$\mathbb{P}\left[d_{\mathrm{H}}(q,\hat{q}^N)\leq\epsilon\right]=\mathbb{P}\left[\frac{d_{\mathrm{H}}(q,\hat{q}^N)}{2}\leq\frac{\epsilon}{2}\right]\geq\mathbb{P}\left[d_{\mathrm{TV}}(q,\hat{q}^N)\leq\frac{\epsilon}{2}\right]$$
$$\geq 1-2\exp\left(-\frac{N}{2}\left(\epsilon-\sqrt{\frac{n}{N}}\right)^2\right),\quad\forall\epsilon\geq\sqrt{\frac{n}{N}}+\sqrt{\frac{2\ln(2)}{N}},$$

where the last inequality follows from (2.27). □

By equating $2\exp\left(-\frac{N}{2}\left(\epsilon-\sqrt{\frac{n}{N}}\right)^2\right)$ to $0<\beta<1$, we obtain the ambiguity ball radius

$$\epsilon_\beta^N(\mathrm{H})=\sqrt{\frac{n}{N}}+\sqrt{\frac{2\ln(2/\beta)}{N}}. \tag{2.32}$$

It is known that $\sup_{p,q\in\Delta^n} d_{\mathrm{H}}(p,q)=2$ [22]. Thus, $\epsilon_\beta^N(\mathrm{H})<2$ as required in Assumption 3, for all sufficiently large N. Again, note that $\lim_{N\to\infty}\epsilon_\beta^N(\mathrm{H})=0$. So this ambiguity ball radius satisfies the hypothesis of Theorem 1 and Theorem 4.

2.6.4 Wasserstein

The Wasserstein distance (metric) between $p,q\in\Delta^n$ equals the optimal cost of a transportation problem [22, page 424]. This problem includes n nodes with supplies $p_1,p_2,\cdots,p_n$ and n other nodes with demands $q_1,q_2,\cdots,q_n$. The unit cost of shipping from node $i=1,\ldots,n$ to node $j=1,\ldots,n$ is $|i-j|$. The goal is to find a minimum cost transportation plan to meet this demand. Let x_{ij} denote the amount transported from node i to node j. Then, the Wasserstein distance (or 1-Wasserstein distance) [61, page 424] is given by

$$d_{\mathrm{W}}(p,q)\stackrel{\mathrm{def}}{=}\min_x\left\{\sum_{i=1}^n\sum_{j=1}^n|i-j|x_{ij}\;\middle|\;\sum_{i=1}^n x_{ij}=q_j,\ j=1,\ldots,n;\right.$$
$$\left.\sum_{j=1}^n x_{ij}=p_i,\ i=1,\ldots,n;\ x_{ij}\geq 0,\ i=1,\ldots,n,\ j=1,\ldots,n\right\}. \tag{2.33}$$

When $p = q$, the optimal transportation plan involves shipping amount $p_i = q_i$ from node i to node $j = i$ at a total cost of 0. Thus, $d_{\mathrm{W}}(q, q) = 0$, for all $q \in \Delta^n$. Inequality (7) in [22] implies that $\|p - q\|_1 = 2d_{\mathrm{TV}}(p, q) \leq 2d_{\mathrm{W}}(p, q)$. Thus, Assumption 2 is satisfied with $\psi(z) = 2z$, for $z \in \mathbb{R}_+$. Theorem 2 in [20] implies that

$$\mathbb{P}\left[d_{\mathrm{W}}(q, \hat{q}^N) \leq \epsilon\right] \geq 1 - \max\{1, c_1\} \exp(-c_2 N \epsilon^2), \quad \forall \epsilon \geq \sqrt{\frac{\ln(\max\{1, c_1\})}{c_2 N}}. \tag{2.34}$$

Here, c_1, c_2 are positive constants that may depend on n. We have used $\max\{1, c_1\}$ instead of just c_1 here to ensure that the logarithm under the square root sign is nonnegative. The restriction $\epsilon \geq \sqrt{\frac{\ln(\max\{1,c_1\})}{c_2 N}}$ is imposed to ensure that the probability lower bound is nonnegative. By equating $\max\{1, c_1\} \exp(-c_2 N \epsilon^2)$ to $0 < \beta < 1$, we obtain the ambiguity ball radius

$$\epsilon_\beta^N(\mathrm{W}) = \sqrt{\frac{\ln(\max\{1, c_1\}/\beta)}{c_2 N}}. \tag{2.35}$$

It is known that $\sup_{p,q \in \Delta^n} d_{\mathrm{W}}(p, q) = n$ [22]. Thus, $\epsilon_\beta^N(\mathrm{W}) < n$ as required in Assumption 3 for all sufficiently large N. Again, note that $\lim_{N \to \infty} \epsilon_\beta^N(\mathrm{W}) = 0$. So this ambiguity ball radius satisfies the hypothesis of Theorem 1 and Theorem 4. One drawback of this formula is that there does not appear to be any known way to compute the positive constants c_1, c_2. Thus, this ambiguity ball radius cannot be used as-is in practice. Though we verified Assumption 2 and Assumption 3 for the 1-Wasserstein distance, they also hold for the ρ-Wasserstein distance with $\rho \in [1, \infty)$.

2.6.5 Kullback-Leibler

As briefly alluded to in Section 2.6.2, this distance is given by $d_{\mathrm{KL}}(p, q) = \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right)$ [7, Table 2]. The convention in the literature is to interpret $0 \ln(0/x) = 0$, for all real numbers x; and $x \ln(x/0) = \infty$, for all real nonzero x [22, page 422]. The former interpretation helps

to confirm that $d(q,q) = 0$, for all $q \in \Delta^n$. From [22, page 429], we have, $\|p-q\|_1 = 2d_{\mathrm{TV}}(p,q) \leq \sqrt{2d_{\mathrm{KL}}(p,q)}$. Thus, Assumption 2 holds with $\psi(z) = \sqrt{2z}$, for $z \in \mathbb{R}_+$. In this thesis, we will only work with pmfs $p \in \Delta^n$ such that $p_i > 0$, for all $i = 1, \ldots, n$, owing to the full support assumption. Thus, if $\hat{q}_i^N = 0$ for any $i = 1, \ldots, n$, then $d_{\mathrm{KL}}(p, \hat{q}^N) = \infty$, for all $p \neq \hat{q}^N$. Thus, for any $\epsilon > 0$, $\hat{q}^N$ is the only pmf in the ambiguity ball in this scenario. We are not aware of a concentration inequality for this Kullback-Leibler distance. Thus, we are unable to ascertain whether or not Assumption 3 holds. Nevertheless, it is known that the statistic $2Nd_{\mathrm{KL}}(q, \hat{q}^N)$ converges in distribution to a χ^2 distribution with $n-1$ degrees of freedom [47, Theorem 3.1]. Thus, for large N, this yields,

$$\mathbb{P}\left[d_{\mathrm{KL}}(q, \hat{q}^N) \leq \epsilon\right] = \mathbb{P}\left[2Nd_{\mathrm{KL}}(q, \hat{q}^N) \leq 2N\epsilon\right] \approx 1 - \bar{F}_{\chi^2_{n-1}}(2N\epsilon), \tag{2.36}$$

where $\bar{F}_{\chi^2_{n-1}}(x)$ denotes the probability that a χ^2 random variable with $n-1$ degrees of freedom takes a value strictly greater than $x \in \mathbb{R}$. Thus, equating $\bar{F}_{\chi^2_{n-1}}(2N\epsilon)$ to $0 < \beta < 1$ and then inverting $\bar{F}_{\chi^2_{n-1}}$ would yield an ambiguity ball radius $\epsilon_\beta^N(\mathrm{KL})$.

2.6.6 χ^2 and modified χ^2

The χ^2 distance is given by $d_{\chi^2}(p,q) = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{p_i}$ [7, Table 2]. The convention is to interpret $0/x = 0$ for any real number x. It is then easy to see that $d(q,q) = 0$, for all $q \in \Delta^n$. From [22, page 429], we have, $\|p-q\|_1 = 2d_{\mathrm{TV}}(p,q) \leq \sqrt{d_{\chi^2}(p,q)}$. Thus, Assumption 2 holds with $\psi(z) = \sqrt{z}$, for $z \in \mathbb{R}_+$. We are not aware of a concentration inequality for this χ^2 distance. Thus, we are unable to ascertain whether or not Assumption 3 holds. Nevertheless, it is known that the statistic $Nd_{\chi^2}(q, \hat{q}^N)$ converges in distribution to a χ^2 distribution with $n-1$ degrees of freedom [47, Theorem 3.1]. Thus, for large N, we have,

$$\mathbb{P}\left[d_{\chi^2}(q, \hat{q}^N) \leq \epsilon\right] \approx 1 - \bar{F}_{\chi^2_{n-1}}(N\epsilon). \tag{2.37}$$

Thus, as explained before, the function $\bar{F}_{\chi^2_{n-1}}(N\epsilon)$ can be equated to $0 < \beta < 1$ to recover an ambiguity ball radius.

The modified χ^2 distance is given by $d_{\bar{\chi}^2}(p,q) = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i}$ [7, Table 2]. Thus, $d_{\bar{\chi}^2}(p,q) = d_{\chi^2}(q,p)$. In addition to the aforementioned interpretation that $0/x = 0$ for any real number x, we also interpret $x/0 = \infty$ for any nonzero real number x. The former interpretation again ensures that $d(q,q) = 0$, for all $q \in \Delta^n$. Again, Assumption 2 holds with $\psi(z) = \sqrt{z}$, for $z \in \mathbb{R}_+$. In this thesis, we will only work with pmfs $p \in \Delta^n$ such that $p_i > 0$, for all $i = 1, \ldots, n$, owing to the full support assumption. Thus, if $\hat{q}_i^N = 0$ for any $i = 1, \ldots, n$, then $d_{\bar{\chi}^2}(p, \hat{q}^N) = \infty$, for all $p \neq \hat{q}^N$. Thus, for any $\epsilon > 0$, $\hat{q}^N$ is the only pmf in the ambiguity ball in this scenario. We are not aware of any concentration inequality for this distance. However, [47, Theorem 3.1] again implies that $N d_{\bar{\chi}^2}(q, \hat{q}^N)$ converges in distribution to a χ^2 distribution with $n-1$ degrees of freedom. An identical counterpart of the approximate concentration formula (2.37) can thus be utilized.

2.7 Robust value iteration and solution of inner-optimization problems

We implemented robust value iteration [32, page 267] to solve problem (2.17) in our computational experiments in Section 2.8. The procedure is briefly outlined here for completeness. Starting with an initial guess vector $V_0 \in \mathbb{R}^n$, the decision-maker then calculates a sequence of vectors $V_1, V_2, \cdots \in \mathbb{R}^n$ via the recursion

$$V_{t+1}(i) = \max_{a \in A} \underbrace{\inf_{p \in \mathcal{P}^N(i,a;\epsilon)} \left[\sum_{j=1}^{n} p(j|i,a) r(j|i,a) + \alpha \sum_{j=1}^{n} p(j|i,a) V_t(j) \right]}_{\text{inner-optimization problem}}, \quad i = 1, \ldots, n. \tag{2.38}$$

It is shown in [32] that the sequence V_t converges to a $V^* \in \mathbb{R}^n$ such that $\tilde{J}^N(\epsilon) = \sum_{i=1}^{n} f(i) V^*(i)$ is the optimal value in problem (2.17). An important step in this process is the solution of the inner-optimization problem.

The inner-optimization problems in value iteration (2.38) can be compactly represented

as

$$\inf_{p\in\mathbb{R}^n} \left\{ \sum_{j=1}^{n} c_j p_j \middle| d(p,q) \le \epsilon,\ \sum_{j=1}^{n} p_j = 1,\ p_j \ge 0,\ j = 1,\dots,n \right\}. \tag{2.39}$$

Here, $c_1, \cdots, c_n$ are known constants; $p_1, \cdots, p_n$ are decision variables; $\epsilon > 0$ is a fixed radius of an ambiguity ball; and $q \in \Delta^n$ is a fixed pmf. A direct comparison with (2.38) reveals that $c_j \equiv r(j|i,a) + \alpha V_t(j)$; $p_j \equiv p(j|i,a)$; and $q_j \equiv \hat{p}^N(j|i,a)$. This problem has a linear objective, and is convex if $d(p,q)$ is convex in p. In that case, it can, in principle, be efficiently solved using standard convex optimization algorithms. However, in many cases, a tailored approach can be devised. Such tailored approaches are available for Kullback-Leibler [32, Lemma 4.1]; modified χ^2 [32, Lemma 4.2]; TV [32, Lemma 4.3]; and Burg [64, Proposition 2]. For Hellinger and χ^2, (2.39) can be formulated as a two-variable concave maximization problem [7, Corollary 3].

Definition (2.33) of Wasserstein distance as the minimum cost in a transportation problem reduces (2.39) to the linear program

$$\min_{x,\ p} \left\{ \sum_{j=1}^{n} c_j p_j \middle| \sum_{i=1}^{n}\sum_{j=1}^{n} |i-j| x_{ij} \le \epsilon;\ \sum_{i=1}^{n} x_{ij} = q_j,\ j = 1,\dots,n; \right.$$
$$\sum_{j=1}^{n} x_{ij} = p_i,\ i = 1,\dots,n; \sum_{j=1}^{n} p_j = 1,\ p_j \ge 0,\ j = 1,\dots,n;$$
$$\left. x_{ij} \ge 0,\ i = 1,\dots,n, j = 1,\dots,n \right\}.$$

Using duality, this linear program can be reformulated as the problem

$$\max_{y\ge 0} \left\{ -y\epsilon + \sum_{i=1}^{n} \left(\min_{j\in\{1,\dots,n\}} \left[c_j + |i-j|\, y \right] \right) q_i \right\} \tag{2.40}$$

of maximizing a concave function of one variable [41, Theorem 3.6]. Further, if $c_j \ge 0$ for all $j \in \{1,\dots,n\}$, then the objective function in (2.40) is strictly decreasing over $y \ge$

$\max_{j\in\{1,\cdots,n\}} c_j$. This can be seen as follows. For any $j \neq i$, $c_j + |i-j|\,y \geq |i-j|\,y \geq y \geq \max_{j\in\{1,\cdots,n\}} c_j$. However, if $j = i$, then $c_j + |i-j|\,y = c_i \leq \max_{j\in\{1,\cdots,n\}} c_j$. Thus, the objective function in problem (2.40) reduces to $-y\epsilon + \sum_{i=1}^{n} c_i q_i$, for $y \geq \max_{j\in\{1,\cdots,n\}} c_j$. Hence, we can solve (2.40) simply by discretizing y over the interval $\left[0, \max_{j\in\{1,\ldots,n\}} c_j\right]$ or by any one-dimensional search procedure for maximizing concave functions.

2.8 Computational experiments

In this section, we report computational experiments on two infinite-horizon problems: a randomly generated MDP instance and a machine replacement MDP from the literature [12, 67, 26]. For both problems, we conducted experiments with three distance functions: TV, Burg, and Wasserstein. For TV and Burg, we used the methods described in [32, Lemma 4.3] and [64, Proposition 2], respectively, to solve the inner-optimization problems. For Wasserstein, we solved formulation (2.40) by discretizing the single variable therein, because the cost coefficients in our inner-optimization problem were nonnegative. An initial guess of 0 was employed for value iteration. The initial state pmf was uniform over $S = \{1, \cdots, n\}$. For both problems, we studied two issues: impact of sample-size on robust and out-of-sample values; and out-of-sample performance and reliability for small sample-sizes no more than n. Since our experimental procedure for studying these issues was identical across the two problems, it is described first below.

2.8.1 Impact of sample-size on robust and out-of-sample values

We studied convergence behavior of the robust optimal value $\tilde{J}^N(\epsilon^N)$ and the out-of-sample value $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ as a function of $N \in \{10, 50, 100, 1000, 10000, 100000, 500000\}$, for $\gamma = 0.1$ corresponding to a confidence level of 90%. Based on Theorem 6, this calls for $\beta = 1 - (0.9)^{1/nm}$. This value of β was used for calculating ambiguity ball radii for each state-action pair. The ambiguity ball radius for the TV distance was calculated by substituting this β in formula (2.28). For Burg, we slightly modified the radius in formula (2.30) by using a lower bound of

$(N+1)^n$ on $\binom{N+n-1}{n-1}$ to obtain $\epsilon_\beta^N(\text{Burg}) = \frac{1}{N}\ln(1/\beta) + n\frac{\ln(N+1)}{N}$. For Wasserstein, motivated by formula (2.35), we set the radius to $1/\sqrt{N}$ regardless of β, because the values of constants c_1, c_2 in formula (2.35) are not available. For each value of N, we generated an empirical estimate of the true transition pmf by sampling N realizations of the next state reached from each state-action pair. This empirical estimate and the above radii characterize the ambiguity balls for each state-action pair, for each distance function. These were utilized within the robust value iteration procedure to obtain a robust optimal policy $\hat{\pi}^N$ and the robust value function $\tilde{J}^N(\epsilon^N)$. Using $\hat{\pi}^N$ and the true transition probabilities, we computed $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ using Bellman's linear equations of policy evaluation [50, Equation (6.1.9)]. We independently repeated this procedure 100 times, for each distance. We plotted the averages and the 10^{th} and 90^{th} percentiles of the resulting robust optimal values $\tilde{J}^N(\epsilon^N)$ and out-of-sample values $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ versus N.

2.8.2 Out-of-sample performance and reliability for small sample-sizes

We fixed sample-sizes to small values ($\leq n$) and varied the radius ϵ of the ambiguity set to understand its effect on the out-of-sample value of the robust optimal policy $\hat{\pi}^N$. We measured this effect in two ways. The first is through the relative closeness of $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ to J^*, and the second is via the probability of $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ exceeding the robust optimal value $\tilde{J}^N(\epsilon)$. We conducted experiments using various sample-sizes N. For each fixed sample-size N and each fixed radius ϵ, we computed 100 independent realizations of $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ and $\tilde{J}^N(\epsilon)$. We then calculated the average of $((J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}})/J^*) \times 100$ and the 10^{th} and 90^{th} percentiles over these independent trials. This was our first performance metric. For the second performance metric, we counted how many times $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ exceeded $\tilde{J}^N(\epsilon)$ and divided this count with 100. We denote this ratio by $\#\left\{J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \geq \tilde{J}^N(\epsilon)\right\}/100$, and call it "reliability". We used this as an estimate of the probability of the corresponding event. We plotted these two performance metrics versus different radii. The radii were picked as follows. For TV and Burg, we used $\beta = 1 - (0.9)^{1/nm}$ corresponding to a 90% confidence level to compute the largest possible radius on the x-axis via (2.28) and (2.30), respectively (for Burg, (2.30) was approximated

as $\epsilon_\beta^N(\text{Burg}) = \frac{1}{N}\ln(1/\beta) + n\frac{\ln(N+1)}{N}$). These radii were then successively halved to obtain the tick-marks on the x-axis. For TV, the radius was capped at 1 since that is the maximum possible TV distance. For the Wasserstein distance, since the constants c_1, c_2 needed to compute the upper bound given by (2.35) are not available, we picked the Wasserstein radii from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$.

2.8.3 Description of two examples and simulation results

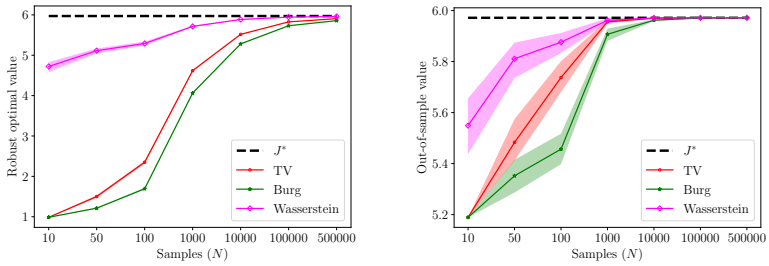
Problem-specific details and simluation results for our two computational examples are provided here.

Randomly generated MDP instance

We generated a single true infinite-horizon MDP with $n = 30$ states, $m = 10$ actions, and discount factor $\alpha = 0.9$. The rewards for each state-action pair were generated randomly from the Uniform[0,1] distribution. The transition probabilities for every state-action pair were also generated randomly from the Uniform[0,1] distribution and then normalized to obtain a pmf. The resulting true MDP was fixed throughout.

Figure 2.2 plots the the robust optimal values $\tilde{J}^N(\epsilon^N)$ and out-of-sample values $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ versus N, using $\beta = 1 - (0.9)^{1/300} = 0.00035114$. This figure is consistent with the results of Theorem 4 and Theorem 5 in that the two values are seen to converge to the true optimal value J^* as N grows. However, note that the three lines are not directly comparable since they use different radii to characterize ambiguity sets, and the "best" values of these radii are not known.

Figure 2.3 plots the two out-of-sample performance metrics described in Section 2.8.2 versus different radii, in 9 panels arranged in a 3×3 format. The rows correspond to $N = 10, 20, 30$ from top to bottom, and the columns correspond to the TV, Burg, and Wasserstein distances. Since the numerical ranges for the two metrics are quite different, we label them separately on the left and right sides of the panel's y-axis. The figure shows that both performance metrics improve, that is, the percentage relative difference with J^*



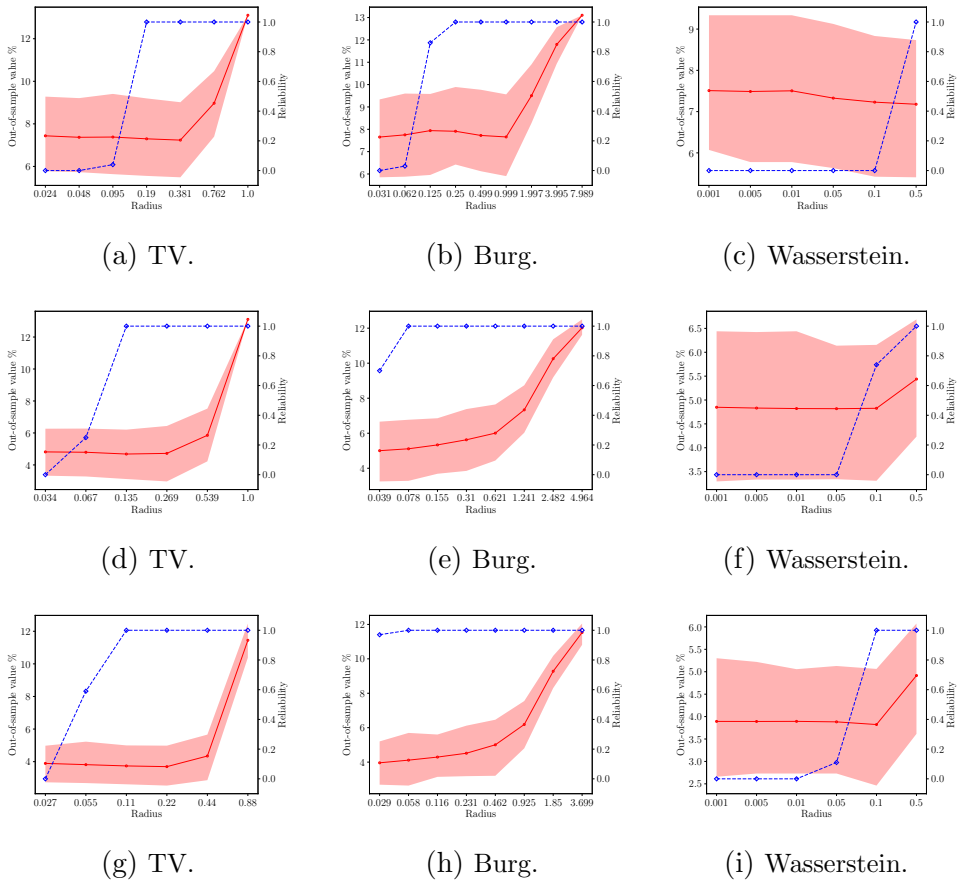
(a) $\tilde{J}^N(\epsilon^N)$ vs N. (b) $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ vs N.

Figure 2.2: Convergence of robust optimal values $\tilde{J}^N(\epsilon^N)$ and out-of-sample values $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ to the optimal value J^* of the true MDP, for the MDP instance from Section 2.8.3.

decreases and reliability increases, as the sample-size increases. The percentage relative difference is less than 5% for reliability values close to 1, despite the small sample-size of $N = 30$ (which equals the number of states in the problem). The figure also shows that the percentage relative difference with J^* starts deteriorating (that is, increasing) after the reliability reaches 1. There also seems to be an interval of radii where the percentage relative difference is low and the reliability is high. This phenomenon was observed for the Wasserstein distance in the case of single-stage optimization in [19] and the stochastic control problem in [70].

Machine repair

This example was introduced by Delage and Mannor [12] and also studied by [26, 67]. The goal is to design a policy for operating a machine that can be in one of $n = 10$ states representing its condition. The machine operator can choose from $m = 2$ actions — "do nothing" and "repair". The discount factor was $\alpha = 0.8$. The details regarding the rewards and transition probabilities are available in [67]. In this example, the set of feasible actions depends on the current state of the machine. Similarly, the set of states that the machine can occupy in the next time-stage depends on its current state and the action chosen. We accordingly restricted the set of feasible pmfs for the inner-optimization problems in our



(a) TV. (b) Burg. (c) Wasserstein.

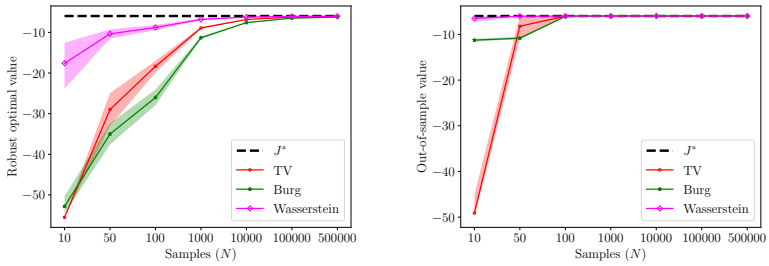
(d) TV. (e) Burg. (f) Wasserstein.

(g) TV. (h) Burg. (i) Wasserstein.

Figure 2.3: Out-of-sample value $((J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}})/J^*) \times 100$ (left axis, solid line) and reliability $\#\left\{J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \geq \tilde{J}^N(\epsilon)\right\}/100$ (right axis, dashed line) as a function of radius ϵ, for the MDP instance from Section 2.8.3. Panels (a) - (c) $N = 10$; (d) - (f) $N = 20$; (g) - (i) $N = 30$.

calculations.

Figure 2.4 plots the robust optimal values $\tilde{J}^N(\epsilon^N)$ and out-of-sample values $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ versus N, with $\beta = 1-(0.9)^{1/20} = 0.005254$. Figure 2.5 plots the two performance metrics described in Section 2.8.2 versus different radii. The rows correspond to $N = 5, 8, 10$ from top to bottom. The qualitative trends in these two figures can be interpreted similar to Figures 2.2 and 2.3, respectively.



(a) $\tilde{J}^N(\epsilon^N)$ vs N. (b) $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ vs N.

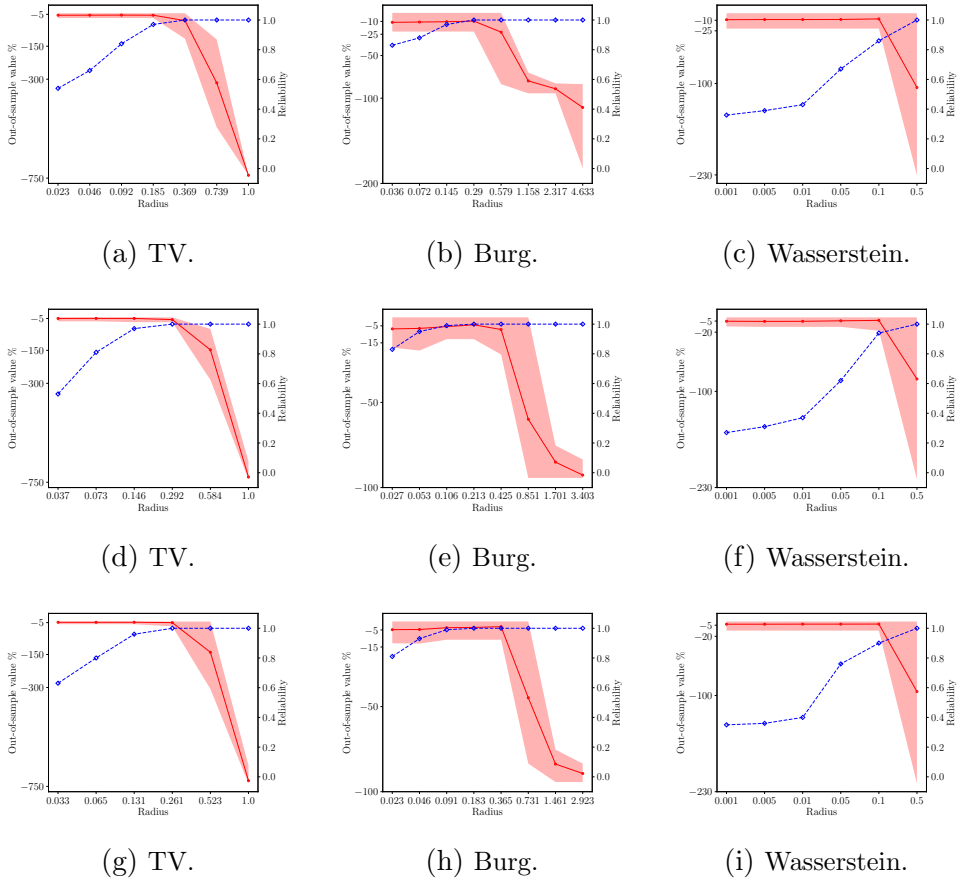
Figure 2.4: Convergence of robust optimal values $\tilde{J}^N(\epsilon^N)$ and out-of-sample values $J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}}$ to the optimal value J^* of the true MDP, for the machine repair example in Section 2.8.3.

2.9 Conclusions

We presented an axiomatic framework for rectangular RMDPs with data-driven, distance-based, ambiguity sets. The two axioms (Assumption 1 and Assumption 3) comprise an essentially minimal set of conditions that impart these RMDPs four strong properties. These are: convergence of robust optimal values to the true stochastic optimum; eventual optimality of robust optimal policies to the true stochastic problem; a probabilistic rate of value convergence to the true optimum; and a guarantee that the out-of-sample value will be larger than the robust optimal value with a high probability. Several well-studied distance functions satisfy these two axioms and also produce tractable inner optimization problems. Computational results suggest that this provides broad choices when attempting to trade-off conservativeness of robust optimal policies against their reliability by tuning ambiguity ball radii.

A potential problem for future research is to try to extend our axiomatic analysis to multi-stage stochastic programs. Another interesting future direction is to apply our framework for partially observable MDPs.

Recall that a more technical name for the RMDPs studied in this chapter is (s, a)-*rectangular* RMDP, which emphasizes the multiplicative separability of ambiguity sets across



(a) TV. (b) Burg. (c) Wasserstein.

(d) TV. (e) Burg. (f) Wasserstein.

(g) TV. (h) Burg. (i) Wasserstein.

Figure 2.5: Out-of-sample value $((J^* - J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}})/J^*) \times 100$ (left axis, solid line) and reliability $\#\left\{J^{\hat{\pi}^N}_{P^{\hat{\pi}^N}} \geq \tilde{J}^N(\epsilon)\right\}/100$ (right axis, dashed line) as a function of radius ϵ, for the machine repair example from Section 2.8.3. Panels (a) - (c) $N = 5$; (d) - (f) $N = 8$; (g) - (i) $N = 10$. The percentage values on the left axis are negative because J^* is negative.

state-action pairs. These were the earliest RMDPs studied in the literature. Other more general notions such as s-rectangularity [67] and k-rectangularity [42] have appeared in the recent literature. Both theory and computation are more challenging in these other rectangular RMDPs. For instance, in s-rectangular RMDPs, optimal policies may be randomized, which introduces hurdles in our proof technique for asymptotic convergence. An investigation of such issues in the context of s-RMDPs is pursued in Chapter 3. Another more challenging

problem is to generalize our analysis to RMDPs with continuous state- and action-spaces. We study that problem in Chapter 4.

Chapter 3

ROBUST MARKOV DECISION PROCESSES WITH DATA-DRIVEN, DISTANCE-BASED S-RECTANGULAR AMBIGUITY SETS

3.1 Introduction

In this chapter, we study infinite-horizon s-rectangular RMDPs. Before we formally define the problem, recall from Section 2.5, that an infinite-horizon stationary MDP is defined by the tuple $\langle S, A, \mathbf{P}, \mathbf{r}, \alpha, f\rangle$. Here, $S \stackrel{\text{def}}{=} \{1, \ldots, n\}$ is the finite state-space and $A \stackrel{\text{def}}{=} \{1, \ldots, m\}$ is the finite action-space. Matrix $\mathbf{P} \stackrel{\text{def}}{=} \{p(j|i,a)\}_{i\in S, a\in A}^{j\in S}$ stores the transition probabilities and matrix $\mathbf{r} \stackrel{\text{def}}{=} \{r(j|i,a)\}_{i\in S, a\in A}^{j\in S}$ stores the rewards earned, for entering state j at the end of a time-stage upon choosing action $a \in A$ in state i at the beginning of that time-stage. Finally, $\alpha \in (0,1)$ is the discount factor and f is the probability mass function (pmf) of the initial state over S. A stationary Markovian randomized policy π assigns a pmf over A to each state $i \in S$. That is, $\pi(a|i)$ is the probability that the decision-maker chooses action $a \in A$ when the system is in state i, regardless of the time-stage. Upon executing such a policy, the decision-maker earns an expected total discounted reward of

$$J_{\mathbf{P}}(\pi) \stackrel{\text{def}}{=} \sum_{i=1}^{n} f(i) J_{\mathbf{P}}(\pi; i), \tag{3.1}$$

where

$$J_{\mathbf{P}}(\pi; i) \stackrel{\text{def}}{=} \lim_{T\to\infty} \left\{ \mathop{\mathbb{E}}_{\mathbf{P}^\pi} \left[\sum_{t=1}^{T} \alpha^{t-1} r(\mathbf{s}_{t+1}|\mathbf{s}_t, \pi(\mathbf{s}_t)) \Big| \mathbf{s}_1 = i \right] \right\}, \text{ for } i = 1:n. \tag{3.2}$$

Here, the subscript $\mathbf{P}^\pi$ on the expectation operator $\mathbb{E}$ is matrix of size $n \times n$ with (i,j)th entry $\sum_{a \in A} \pi(a|i)p(j|i,a)$, for $i,j \in S$. That is, it is the expectation of the transition probability vector $p(j|i,\cdot) \in \mathbb{R}^m$ with respect to the pmf $\pi(\cdot|i) \in \Delta^m$. Here, Δ^m denotes the probability simplex in $\mathbb{R}^m$. The subscript $\mathbf{P}^\pi$ emphasizes that the expectation $\underset{\mathbf{P}^\pi}{\mathbb{E}}$ in (3.2) is with respect to the stochastic trajectory $(\mathbf{s}_2, \mathbf{s}_3, \cdots)$ of states induced by $\mathbf{P}^\pi$. The conditioning $\mathbf{s}_1 = i$ signifies that the initial state is i. The uncountable set of all randomized, stationary policies is denoted as Π_R. The decision-maker can maximize the expected total discounted reward by solving the problem

$$J^* \stackrel{\text{def}}{=} \max_{\pi \in \Pi_R} \; J_{\mathbf{P}}(\pi). \tag{3.3}$$

The infinite-horizon stationary MDP admits an optimal policy that is deterministic and stationary [50]. That is, a policy where, for each state i, the pmf $\pi(\cdot|i)$ puts positive probability mass on a single action in A.

Like in the previous chapter, we assume that the "true" transition probabilities $\mathbf{P}$ are unknown. In this chapter, we use a s-rectangular RMDP (henceforth, s-RMDP) to address the ambiguity present in the transition probabilities. Recall from Chapter 1, the ambiguity set in s-RMDP is multiplicatively separable only across states rather than state-action pairs and an ambiguity set satisfying such a property is called s-rectangular. Note that the s-rectangular ambiguity set can be expressed as $\mathcal{P} = \bigtimes_{s \in S} \mathcal{P}_s$. Here, $\mathcal{P}_s \subseteq \bigtimes_{a \in A} \Delta^n$ is the ambiguity set of m-tuples $(q(\cdot|s,1), \ldots, q(\cdot|s,m))$ of transition pmfs for state $s \in S$. As such, $\mathcal{P}_s$ denotes the ambiguity set of transition pmfs across all actions in state s. We note here for emphasis that the ambiguity sets for (s,a)-RMDPs and s-RMDPs are both subsets of $\bigtimes_{s \in S, a \in A} \Delta^n$.

The concept of s-RMDPs was studied formally in [67] and various properties of such RMDPs were established in that paper. There has been an increased interest in developing efficient algorithms for solving s-RMDPs with ambiguity sets constructed using specific distance functions [29, 30]. These algorithms exploit concrete algebraic properties of those distance functions to develop efficient algorithms. A generic convex programming formulation was developed in [27] to solve arbitrary (not necessarily distance-based) (s,a)- and

s-RMDPs by using the theory of regularized operators. However, none of these works analyze the data-driven properties of distance-based s-RMDPs nor study the relative conservativeness between (s,a)- versus s-RMDPs, as done in this chapter.

3.2 Contributions of this chapter

We begin with a detailed overview of the main contributions of our work.

3.2.1 A family of distance-based s-RMDPs

We study a family of s-RMDPs characterized by ambiguity sets indexed with $\rho \in [1,\infty]$. Let $\|x\|_\rho$ denote the ℓ_ρ-norm of any $x \in \mathbb{R}^m$. That is,

$$\|x\|_\rho \stackrel{\text{def}}{=} \begin{cases} \left(\sum_{i=1}^m |x_i|^\rho\right)^{1/\rho}, \text{ for } \rho \in [1,\infty) \text{ and} \\ \max_{1\le i\le m} |x_i|, \text{ for } \rho = \infty. \end{cases}$$

Recall that the maximum norm $\|x\|_\infty$ as defined above also equals $\lim_{\rho\to\infty}\left(\sum_{i=1}^m |x_i|^\rho\right)^{1/\rho}$. Let $d : \Delta^n \times \Delta^n \to \{\mathbb{R}_+ \cup \infty\}$ be a nonnegative extended real-valued distance function (need not be a metric). Let $\hat{p}(\cdot|s,a) \in \Delta^n$ be given reference transition pmfs corresponding to states $s \in S$ and actions $a \in A$. Use the shorthand $\vec{d}_s(q,\hat{p})$ to denote the m-tuple of distances $(d(q(\cdot|s,1),\hat{p}(\cdot|s,1)),\ldots,d(q(\cdot|s,m),\hat{p}(\cdot|s,m)))$, corresponding to any $s \in S$. That is, in this tuple, the component associated with action $a \in A$ equals the distance to some transition pmf $q(\cdot|s,a) \in \Delta^n$ from the reference pmf $\hat{p}(\cdot|s,a) \in \Delta^n$. Now, for any $\epsilon \in [0,\infty)$, we introduce the ambiguity set

$$\begin{aligned} \mathcal{P}(\rho,\epsilon) &\stackrel{\text{def}}{=} \bigtimes_{s\in S} \mathcal{P}_s(\rho,\epsilon), \text{ where} \\ \mathcal{P}_s(\rho,\epsilon) &\stackrel{\text{def}}{=} \left\{(q(\cdot|s,1),\ldots,q(\cdot|s,m)) \in \bigtimes_{a\in A}\Delta^n \,\middle|\, \left\|\vec{d}_s(q,\hat{p})\right\|_\rho \le \epsilon\right\}. \end{aligned} \quad (3.4)$$

That is, ambiguity sets $\mathcal{P}_s(\rho, \epsilon)$ are sublevel sets of the ℓ_ρ-norm composed with the m-tuple of distances. With a slight abuse of terminology, we refer to ϵ as the ambiguity set radius. The value of the corresponding s-RMDP is then given by

$$\tilde{J}_\rho(\epsilon) \stackrel{\text{def}}{=} \sup_{\pi \in \Pi_R} \tilde{J}_\rho(\pi, \epsilon), \text{ where } \tilde{J}_\rho(\pi, \epsilon) \stackrel{\text{def}}{=} \inf_{P \in \mathcal{P}(\rho, \epsilon)} J_P(\pi). \tag{3.5}$$

We refer to this as the distance-based s-RMDP problem (or simply the s-RMDP problem, if the context is clear). Moreover, $\tilde{J}_\rho(\epsilon)$ is called the robust optimal value. Throughout this chapter, we restrict the policy space to Π_R. In an s-RMDP, this is not without loss of optimality, because history-dependent randomized policies might earn better values [67, Table 1]. We will rely on a sufficient condition that guarantees the existence of an optimal policy in Π_R, in Section 3.3 and beyond. A policy achieving the maximum in (3.5) will then be termed a robust optimal policy.

Setting $\rho = \infty$ in (3.4) recovers precisely the distance-based (s, a)-rectangular ambiguity set studied, for example, in Chapter 2. To see this, note that

$$\begin{aligned}
\mathcal{P}(\infty, \epsilon) &= \bigtimes_{s \in S} \mathcal{P}_s(\infty, \epsilon) \\
&= \bigtimes_{s \in S} \Big\{ (q(\cdot|s,1), \ldots, q(\cdot|s,m)) \in \bigtimes_{a \in A} \Delta^n \Big| \max_{a \in A} |d(q(\cdot|s,a), \hat{p}(\cdot|s,a))| \leq \epsilon \Big\} \\
&= \bigtimes_{s \in S} \bigtimes_{a \in A} \Big\{ q(\cdot|s,a) \in \Delta^n \Big| d(q(\cdot|s,a), \hat{p}(\cdot|s,a)) \leq \epsilon \Big\} \stackrel{\text{def}}{=} \bigtimes_{s \in S, a \in A} \mathcal{P}_{sa}(\epsilon).
\end{aligned} \tag{3.6}$$

We refer to (3.6) as the distance-based (s, a)-rectangular ambiguity set (or simply (s, a)-rectangular ambiguity set if the context is unambiguous). The corresponding optimization problem (3.5) for calculating $\tilde{J}_\infty(\epsilon)$ is thereby called the distance-based (s, a)-RMDP (or simply (s, a)-RMDP if the context is clear).

The modeling framework in [29] is a special case of our s-RMDPs with the total variation (TV) or a weighted TV distance and $\rho = 1$; similarly, in [30], it is ϕ-divergence with $\rho = 1$.

Recall that $\|\cdot\|_{\rho_1} \geq \|\cdot\|_{\rho_2}$, for all $1 \leq \rho_1 \leq \rho_2 \leq \infty$. For any $\epsilon \in [0, \infty)$, the family

$\{\mathcal{P}(\rho,\epsilon)\}_{\rho\in[1,\infty]}$ is thus a nondecreasing collection, with respect to inclusion, of distance-based s-rectangular ambiguity sets. Hence, the distance-based (s,a)-rectangular ambiguity set $\mathcal{P}(\infty,\epsilon)$ is the largest member of this family. Further, this (s,a)-rectangular ambiguity set can be recovered as the Kuratowski limit (see Definition 1) of the family $\{\mathcal{P}(\rho,\epsilon)\}_{\rho\in[1,\infty)}$, if the distance function is lower semicontinuous (lsc), convex, and $d(x,x)=0$, for all $x\in\Delta^n$ (Assumption 4). This result is proven in Lemma 13 by showing that the (s,a)-rectangular ambiguity set $\mathcal{P}(\infty,\epsilon)$ equals the closure of $\bigcup_{\rho\in[1,\infty)}\mathcal{P}(\rho,\epsilon)$. Assumption 4 holds for many well-known distance functions such as TV, Burg entropy, Kullback-Leibler (KL), Hellinger, χ^2, modified χ^2, and Wasserstein. We also demonstrate via an example in Example 8 that the Kuratowski limit property fails to hold if we drop Assumption 4.

Our framework for constructing a family of s-rectangular ambiguity sets via the composition of ℓ_ρ norms and distance functions makes an explicit connection between distance-based (s,a)- and s-rectangular ambiguity sets. This perspective allows us to rigorously analyze relative conservativeness properties of (s,a)- versus s-RMDPs from this family, as described in the following subsections.

3.2.2 Relative conservativeness between (s,a)- versus s-RMDPs

In Section 3.3, we constructively establish in Lemma 8 that, for any instance of an s-RMDP, there is an instance of an (s,a)-RMDP with a robust optimal value that is at least as good; and vice versa. This suggests that appropriate caution should be exercised before interpreting too broadly any anecdotal claims that (s,a)-rectangular RMDPs are more conservative than s-rectangular ones. Such statements are rooted in the fact that the class of (s,a)-rectangular ambiguity sets is larger than that of s-rectangular ones, thereby forcing the decision-maker to hedge against more transition pmfs in the (s,a)-rectangular framework per the worst-case viewpoint. In our context, although it is true that $\mathcal{P}(\rho,\epsilon)\subseteq\mathcal{P}(\infty,\epsilon)$, for any $\rho\in[1,\infty)$ and $\epsilon\in[0,\infty)$, Lemma 8 concretely overcomes potential conservatism by choosing an appropriate smaller radius for the (s,a)-rectangular ambiguity set. The lemma

thus motivates the following question: given an instance of an s-RMDP, does there exist an instance of an (s,a)-RMDP with an identical robust optimal value? We answer this in the affirmative in Theorem 8. There, we establish under Assumption 4 that, for any s-RMDP with a fixed $\rho \in [1,\infty)$ and a fixed $\epsilon \in [0,\infty)$, there exists an $\epsilon^* \in [\epsilon/m^{1/\rho}, \epsilon]$ such that the (s,a)-RMDP value $\tilde{J}_\infty(\epsilon^*)$ equals $\tilde{J}_\rho(\epsilon)$. Theorem 8 is established using the intermediate value theorem, after first proving via Berge's maximum theorem that the (s,a)-robust optimal value $\tilde{J}_\infty(\epsilon)$ is continuous in ϵ. We also construct a counterexample in Example 9 to demonstrate that Assumption 4 cannot be dropped from the hypothesis of Theorem 8. Recalling that $\{\mathcal{P}(\rho,\epsilon)\}_{\rho\in[1,\infty]}$ is a nondecreasing collection for any fixed $\epsilon \in [0,\infty)$, we know that $\tilde{J}_{\rho_2}(\epsilon) \leq \tilde{J}_{\rho_1}(\epsilon)$, for any $1 \leq \rho_1 \leq \rho_2 \leq \infty$. Thus, a similar more general result about matching robust optimal values can be derived for any pair $1 \leq \rho_1 \leq \rho_2 \leq \infty$. We omit that for brevity.

3.2.3 Data-driven s-RMDPs: asymptotic and finite sample properties

Section 3.4 focuses on a data-driven version of our s-RMDP problem. That is, the reference transition pmf $\hat{p}(\cdot|s,a)$ for each $s \in S$ and each $a \in A$ equals the empirical estimate $\hat{p}^N(\cdot|s,a)$ of the true transition pmf $p(\cdot|s,a)$. This empirical estimate is constructed based on N independent observations of the next state reached upon choosing action $a \in A$ in state $s \in S$. We study the asymptotic (as sample-size $N \to \infty$) and finite sample properties of the resulting data-driven s-RMDP. We prove in Theorem 9 that robust optimal values of the data-driven s-RMDPs converge almost surely to the optimal value of the true MDP, as $N \to \infty$. The out-of-sample value is defined as the expected total discounted reward earned over an infinite horizon by a robust optimal policy under the true transition pmf. We prove in Theorem 9 that the out-of-sample values also converge almost surely to the true optimal value as $N \to \infty$. Both these properties are proved under Assumption 1. For finite sample-sizes, we show in Theorem 10 that the robust optimal value serves as a lower bound on the out-of-sample value, with a high probability. This probabilistic performance guarantee on out-of-sample values is proven under the concentration inequality for the distance function as

stated in Assumption 3. Under a mild strengthening of the generalized Pinsker's inequality (Assumption 2), we derive in Theorem 11 a probabilistic rate of convergence for the robust and out-of-sample values to the true optimal value. As such, Section 3.4 establishes that our data-driven s-RMDPs satisfy all properties that are deemed essential for any formal data-driven robust optimization framework (see [19, 64, 60] for detailed descriptions of these essential properties).

Similar asymptotic and finite sample properties were established, under the generalized Pinsker's inequality and the concentration inequality, for data-driven distance-based (s, a)-RMDPs in the previous chapter. In this chapter, we extend those results to data-driven distance-based s-RMDPs. However, the proof techniques here are different and more involved. These more sophisticated proofs are needed because of some fundamental differences in the structure of (s, a)- versus s-RMDPs. For instance, while there exists a deterministic policy that is optimal for an (s, a)-RMDP, all optimal policies for an s-RMDP may be randomized. Consequently, while proofs of asymptotic convergence relied on a simple interchange of limit and maximization operations in the (s, a)-RMDPs (Theorem 4 and Theorem 5), here we rely on a variation of Dini's theorem that connects continuous convergence of functions with their uniform convergence (see Proposition 1).

We also describe why the probabilistic performance guarantee on out-of-sample values in Theorem 10 leaves the door open for viewing (s, a)-RMDPs as actually less conservative than all s-RMDPs within our family. Specifically, among all s-RMDPs that satisfy a constant probability out-of-sample performance guarantee, the (s, a)-RMDP possesses the best robust optimal value. A similar criterion was employed to find an "optimal" data-driven robust optimization framework for single-stage stochastic programs in [60, 64]. This provides additional support for our aforementioned suggestion to exercise caution while interpreting too broadly the anecdotal claim that (s, a)-RMDPs are more conservative than s-RMDPs.

Finally, in Section 3.5, we conduct computational experiments using the TV distance and $\rho = 1$ on a machine repair example from the literature. Note that $\rho = 1$ corresponds to the smallest s-rectangular ambiguity set. The results of those experiments align with the theory

developed in this chapter.

3.3 Matching robust optimal values of (s,a)- and s-RMDPs

We first establish the following lemma.

Lemma 8. Fix any $\rho \in [1,\infty)$ and $\epsilon \in [0,\infty)$. Then, for any $k \in \{0,1,2,\ldots\}$, we have, $\tilde{J}_\infty\left(\frac{\epsilon}{m^{k/\rho}}\right) \leq \tilde{J}_\rho\left(\frac{\epsilon}{m^{k/\rho}}\right) \leq \tilde{J}_\infty\left(\frac{\epsilon}{m^{(k+1)/\rho}}\right)$. In other words, $\tilde{J}_\infty(\epsilon) \leq \tilde{J}_\rho(\epsilon) \leq \tilde{J}_\infty\left(\frac{\epsilon}{m^{1/\rho}}\right) \leq \tilde{J}_\rho\left(\frac{\epsilon}{m^{1/\rho}}\right) \leq \tilde{J}_\infty\left(\frac{\epsilon}{m^{2/\rho}}\right) \leq \tilde{J}_\rho\left(\frac{\epsilon}{m^{2/\rho}}\right) \leq \ldots$.

Proof. We establish the result for $k=0$. The result for general k then follows by iteratively repeating the same argument. Since $\frac{\|\cdot\|_\rho}{m^{1/\rho}} \leq \|\cdot\|_\infty \leq \|\cdot\|_\rho$, we have $\mathcal{P}\left(\infty, \frac{\epsilon}{m^{1/\rho}}\right) \subseteq \mathcal{P}(\rho,\epsilon) \subseteq \mathcal{P}(\infty,\epsilon)$. Therefore, $\inf_{P\in\mathcal{P}(\infty,\epsilon/m^{1/\rho})} J_P(\pi) \geq \inf_{P\in\mathcal{P}(\rho,\epsilon)} J_P(\pi) \geq \inf_{P\in\mathcal{P}(\infty,\epsilon)} J_P(\pi)$, for all $\pi \in \Pi_R$. Taking a supremum over Π_R in this chain of inequalities establishes $\tilde{J}_\infty\left(\frac{\epsilon}{m^{1/\rho}}\right) \geq \tilde{J}_\rho(\epsilon) \geq \tilde{J}_\infty(\epsilon)$. □

In this lemma, a sequence of (s,a)- and s-RMDPs was constructed by repeatedly scaling down the radius ϵ by a factor $m^{1/\rho}$. The lemma shows that robust optimal values cannot be employed to characterize relative conservativeness between s- and (s,a)-RMDPs within our family, since the decision-maker has the flexibility to choose the ambiguity set radii. We next establish the main result of this section — given an instance of an s-RMDP, there exists an (s,a)-RMDP with an appropriate ambiguity radius such that the robust optimal values of the two RMPDs are in fact identical. This more refined result relies on the following assumption on the distance function.

Assumption 4. The distance function $d : \Delta^n \times \Delta^n \to \{\mathbb{R} \cup \infty\}$ satisfies: (a) $d(x,y)$ is lower semicontinuous (lsc) in $x \in \Delta^n$ for every $y \in \Delta^n$; (b) $d(x,y)$ is convex in $x \in \Delta^n$ for every $y \in \Delta^n$; and (c) $d(x,x) = 0$ for every $x \in \Delta^n$.

Assumption 4 holds for many distance functions in the literature, including TV, Burg entropy, Kullback-Leibler (KL), Hellinger, χ^2, modified χ^2, and Wasserstein. One implication

of Assumption 4(c) is that the ambiguity sets are nonempty for every $\epsilon \in [0, \infty)$ since they always contain the reference pmf $\hat{p}$.

The role of Assumption 4 is multifold. In addition to being critical for studying relative conservativeness between (s, a)- versus s-RMDPs, it also plays a more fundamental role. This assumption guarantees that there is no loss of optimality in restricting attention to Π_R as in problem (3.5), instead of considering the larger and more complex class of history dependent randomized policies. In particular, Assumption 4 guarantees via Lemma 9 that ambiguity sets $\mathcal{P}(\rho, \epsilon)$ defined in (3.4) are compact and convex. This is sufficient for the existence of a stationary Markovian optimal policy for an s-RMDP [67, Table 1]. We emphasize, however, that such a policy may need to be a randomized one [67, Table 1]. This is in contrast to (s, a)-RMDPs, where the existence of a deterministic stationary Markovian optimal policy is guaranteed without any additional assumptions on the ambiguity sets [32]. Specifically, lower semicontinuity and convexity of the distance function were not assumed when studying (s, a)-RMDPs in Chapter 2.

Lemma 9. Suppose Assumption 4(a) holds. Then, $\mathcal{P}(\rho, \epsilon)$ is a compact subset of $\bigtimes_{s\in S, a\in A} \Delta^n$, for every $\rho \in [1, \infty]$ and $\epsilon \in [0, \infty)$. Similarly, if Assumption 4(b) holds, then $\mathcal{P}(\rho, \epsilon)$ is a convex subset of $\bigtimes_{s\in S, a\in A} \Delta^n$, for every $\rho \in [1, \infty]$ and $\epsilon \in [0, \infty)$.

Proof. Fix $\rho \in [1, \infty]$ and $\epsilon \in [0, \infty)$.

Observe that $\mathcal{P}_s(\rho, \epsilon)$ is closed for every $s \in S$ because it equals the sublevel set of an ℓ_ρ-norm composed with an lsc distance function. We omit the details of that argument here. Moreover, $\mathcal{P}_s(\rho, \epsilon)$ is bounded for every $s \in S$ because it is a subset of $\bigtimes_{a\in A} \Delta^n$. It is, therefore compact. This implies that $\mathcal{P}_s(\rho, \epsilon) = \bigtimes_{s\in S} \mathcal{P}_s(\rho, \epsilon)$ is also compact as claimed.

We prove that $\mathcal{P}_s(\rho, \epsilon)$ is a convex subset of $\bigtimes_{a\in A} \Delta^n$, for each $s \in S$. This would imply that $\mathcal{P}(\rho, \epsilon) = \bigtimes_{s\in S} \mathcal{P}_s(\rho, \epsilon)$ is also convex, as claimed. Thus, fix any $\lambda \in [0, 1]$; and any two m-tuples $(q^1(\cdot|s, 1), \ldots, q^1(\cdot|s, m))$ and $(q^2(\cdot|s, 1), \ldots, q^2(\cdot|s, m))$ of transition pmfs in $\mathcal{P}_s(\rho, \epsilon)$, for a particular state $s \in S$. Then, convexity of the distance function implies that $\vec{d}_s(\lambda q^1 + (1-\lambda)q^2, \hat{p}) \leq \lambda \vec{d}_s(q^1, \hat{p}) + (1-\lambda)\vec{d}_s(q^2, \hat{p})$, where the ordering is with respect to each

component of the vector. Since d is nonnegative, $\left\|\vec{d}_s(\lambda q^1 + (1-\lambda)q^2, \hat{p})\right\|_\rho \leq \lambda \left\|d_s(q^1, \hat{p})\right\|_\rho +$ $(1-\lambda)\left\|\vec{d}_s(q^2, \hat{p})\right\|_\rho \leq \lambda\epsilon + (1-\lambda)\epsilon = \epsilon$. This proves that $\lambda q^1 + (1-\lambda)q^2 \in \mathcal{P}_s(\rho, \epsilon)$ as required for convexity. □

We now consider any fixed $\rho \in [1, \infty)$ and $\epsilon \in [0, \infty)$. Under Assumption 4, we demonstrate the existence of a distance-based (s, a)-RMDP whose robust optimal value equals $\tilde{J}_\rho(\epsilon)$ — the robust optimal value of the s-RMDP (3.5). To assist with our quest for a distance-based (s, a)-RMDP with this property, recall from Lemma 8, $\tilde{J}_\infty\left(\frac{\epsilon}{m^{1/\rho}}\right) \geq \tilde{J}_\rho(\epsilon) \geq \tilde{J}_\infty(\epsilon)$. Observe that the (s, a)-robust optimal values $\tilde{J}_\infty(\cdot)$ are nonincreasing in radii since $\mathcal{P}(\infty, \epsilon_1) \subseteq \mathcal{P}(\infty, \epsilon_2)$ for all $\epsilon_1 \leq \epsilon_2$. As a result, if there is an (s, a)-RMDP whose robust optimal value equals $\tilde{J}_\rho(\epsilon)$, its ambiguity set radius must belong to the interval $[\epsilon/m^{1/\rho}, \epsilon]$. This allows us to reduce the search for the existence of an (s, a)-rectangular ambiguity set to the search for an $\epsilon^* \in [\epsilon/m^{1/\rho}, \epsilon]$. We thus wish to derive conditions that guarantee the existence of an $\epsilon^* \in [\epsilon/m^{1/\rho}, \epsilon]$ such that $\tilde{J}_\infty(\epsilon^*) = \tilde{J}_\rho(\epsilon)$. A sufficient condition is the continuity of $\tilde{J}_\infty(\epsilon)$ in ϵ over $[0, \infty)$, since we can then apply the intermediate value theorem. Therefore, a bulk of the remainder of this section focuses on establishing this continuity.

To establish continuity of $\tilde{J}_\infty(\epsilon)$ in ϵ over $[0, \infty)$, we will apply Berge's maximum theorem. Toward this end, we view the (s, a)-RMDP as an optimization problem parameterized by ϵ. In particular, the feasible regions of this problem is characterized by radius ϵ. We establish that these feasible regions, that is, the (s, a)-rectangular ambiguity sets, converge in the Kuratowski sense (Definition 1 below). This is then utilized to prove continuity of $\tilde{J}_\infty(\epsilon)$ in ϵ over $[0, \infty)$ by building upon the notion of continuity of set-valued mappings called correspondences (Definition 2 below).

For any $x \in \mathbb{R}^n$ and $\delta > 0$ let $B_\delta(x) \stackrel{\text{def}}{=} \{y \in \mathbb{R}^n | \left\|y - x\right\|_\infty < \delta\}$ denote an open ball of radius δ with respect to the maximum norm in $\mathbb{R}^n$, centered at x. Let $\mathbb{N}$ denote the set of positive integers.

Definition 1 (Kuratowski limits)**.** [39, Chapter 29]. Let $\{A_k\}$ be a sequence of sets in $\mathbb{R}^n$.

1. Lower limit. A point $x \in \mathbb{R}^d$ belongs to the lower limit of the sequence $\{A_k\}$ if, for

every $\delta > 0$, there exists some $j \in \mathbb{N}$ such that $B_\delta(x) \cap A_k \neq \emptyset$ for all $k \geq j$. We then write $x \in \text{Li } A_k$.

2. Upper limit. A point $x \in \mathbb{R}^d$ belongs to the upper limit of the sequence $\{A_k\}$ if, for every $\delta > 0$, we have that $B_\delta(x) \cap A_k \neq \emptyset$ for infinitely many k. We then write $x \in \text{Ls } A_k$.

3. Limit. From the above two definitions, Li $A_k \subseteq$ Ls A_k. If Li $A_k =$ Ls A_k, we say that Lim A_k exists and define it as Lim $A_k =$ Li $A_k =$ Ls A_k.

Lemma 10 (Kuratowski limits for products)**.** [39, Chapter 29, VI]. Suppose $\{A_k\}$ and $\{B_k\}$ are two sequence of nonempty compact sets in $\mathbb{R}^j$ and $\mathbb{R}^l$, respectively. If Lim $A_k = A$ and Lim $B_k = B$, then Lim $(A_k \times B_k) = A \times B$.

Lemma 11. Suppose $\{\epsilon^k\}$ is a sequence of nonnegative reals such that $\epsilon^k \to \epsilon$. Suppose Assumption 4 holds. Then, Lim $\mathcal{P}(\infty, \epsilon^k) = \mathcal{P}(\infty, \epsilon)$.

Proof. We show that Ls $\mathcal{P}_{sa}(\epsilon^k) \subseteq \mathcal{P}_{sa}(\epsilon)$ and $\mathcal{P}_{sa}(\epsilon) \subseteq$ Li $\mathcal{P}_{sa}(\epsilon^k)$, for all $s \in S$ and $a \in A$. This means that Lim $\mathcal{P}_{sa}(\epsilon^k) = \mathcal{P}_{sa}(\epsilon)$. The result then follows from Lemma 10 since $\mathcal{P}(\infty, \epsilon) = \bigtimes_{s \in S, a \in A} \mathcal{P}_{sa}(\epsilon)$ and $\mathcal{P}(\infty, \epsilon^k) = \bigtimes_{s \in S, a \in A} \mathcal{P}_{sa}(\epsilon^k)$, for all $k \in \mathbb{N}$.

We first establish that Ls $\mathcal{P}_{sa}(\epsilon^k) \subseteq \mathcal{P}_{sa}(\epsilon)$. Let $q \in$ Ls $\mathcal{P}_{sa}(\epsilon^k)$. Then, for every $j \in \mathbb{N}$, $B_{1/j}(q) \cap \mathcal{P}_{sa}(\epsilon^k) \neq \emptyset$ for infinitely many $k \in \mathbb{N}$. Hence, there exists a sequence $\{q^j\}$, such that $q^j \in B_{1/j}(q) \cap \mathcal{P}_{sa}(\epsilon^{k_j})$ for all $j \in \mathbb{N}$. Since $q^j \in B_{1/j}(q)$, we know that $\|q^j - q\|_\infty < 1/j$ for all $j \in \mathbb{N}$. Thus, $q^j \to q$. We also know that $d(q^j, \hat{p}) \leq \epsilon^{k_j}$ for all $j \in \mathbb{N}$ because $q^j \in \mathcal{P}_{sa}(\epsilon^{k_j})$. Since $\epsilon^k \to \epsilon$ and the distance function is lsc by Assumption 4(a), we have $d(q, \hat{p}) \leq \liminf_{j \to \infty} d(q^j, \hat{p}) \leq \liminf_{j \to \infty} \epsilon^{k_j} = \epsilon$. Hence, $q \in \mathcal{P}_{sa}(\epsilon)$, and thus Ls $\mathcal{P}_{sa}(\epsilon^k) \subseteq \mathcal{P}_{sa}(\epsilon)$.

We now prove that $\mathcal{P}_{sa}(\epsilon) \subseteq$ Li $\mathcal{P}_{sa}(\epsilon^k)$. Consider a $q \in \mathcal{P}_{sa}(\epsilon)$ and an arbitrary $\delta > 0$. If $q = \hat{p}$, then, by Assumption 4(c), we have $d(q, \hat{p}) = 0 \leq \epsilon^k$ for all $k \in \mathbb{N}$. Hence, $\mathcal{P}_{sa}(\epsilon^k) \cap B_\delta(q) \neq \emptyset$ for all $k \in \mathbb{N}$. Therefore, assume that $q \neq \hat{p}$ in the rest of this proof. We know that $q \in \mathcal{P}_{sa}(\epsilon) \cap B_\delta(q)$ by construction. There exists at least one point, distinct from

q, in $\mathcal{P}_{sa}(\epsilon) \cap B_\delta(q)$. To see this, consider the open line segment between q and $\hat{p}$ written as $\Lambda_{q,\hat{p}} \stackrel{\text{def}}{=} \{\lambda q + (1-\lambda)\hat{p} | \lambda \in (0,1)\}$. Since $\delta > 0$, there is at least one point $q' \neq q$ on $\Lambda_{q,\hat{p}}$ such that $q' \in B_\delta(q)$. But, in fact, since $\mathcal{P}_{sa}(\epsilon)$ is a convex set and $q, \hat{p} \in \mathcal{P}_{sa}(\epsilon)$, we know that $\Lambda_{q,\hat{p}} \subseteq \mathcal{P}_{sa}(\epsilon)$. Thus, $q' \neq q$ must also belong to $\mathcal{P}_{sa}(\epsilon) \cap B_\delta(q)$. Since q' is on $\Lambda_{q,\hat{p}}$, we express it as $q' = \lambda q + (1-\lambda)\hat{p}$ for some specific $\lambda \in (0,1)$. Recall that $d(q,\hat{p})$ is convex in q by Assumption 4(b), and $d(\hat{p},\hat{p}) = 0$ by Assumption 4(c). Moreover, $d(q,\hat{p}) \leq \epsilon$ from (3.6) because $q \in \mathcal{P}_{sa}(\epsilon)$. Hence,

$$d(q',\hat{p}) = d(\lambda q + (1-\lambda)\hat{p}, \hat{p}) \leq \lambda d(q,\hat{p}) + (1-\lambda) d(\hat{p},\hat{p}) \leq \lambda\epsilon + (1-\lambda)0 = \lambda\epsilon.$$

We consider two cases depending on whether $\epsilon = 0$ or $\epsilon > 0$. If $\epsilon = 0$, then $d(q',\hat{p}) \leq \lambda\epsilon = 0 \leq \epsilon^k$ for all $k \in \mathbb{N}$. This proves that $q' \in \mathcal{P}_{sa}(\epsilon^k)$ for all $k \in \mathbb{N}$, and hence $\mathcal{P}_{sa}(\epsilon^k) \cap B_\delta(q) \neq \emptyset$ for all $k \in \mathbb{N}$. Suppose $\epsilon > 0$. Since $\lambda \in (0,1)$, we have $d(q',\hat{p}) \leq \lambda\epsilon < \epsilon$. Hence, there exists $\eta > 0$, such that, $d(q',\hat{p}) = \epsilon - \eta$. Since $\epsilon^k \to \epsilon$, there exists a $j \in \mathbb{N}$, such that, $\epsilon - \eta < \epsilon^k < \epsilon + \eta$ for all $k \geq j$. Therefore, $d(q',\hat{p}) = \epsilon - \eta < \epsilon^k$ for all $k \geq j$. This establishes that $q' \in \mathcal{P}_{sa}(\epsilon^k)$ for all $k \geq j$, and hence $\mathcal{P}_{sa}(\epsilon^k) \cap B_\delta(q) \neq \emptyset$ for all $k \geq j$. That is, in both cases $\epsilon = 0$ and $\epsilon > 0$, there exists a $j \in \mathbb{N}$, such that, $\mathcal{P}_{sa}(\epsilon^k) \cap B_\delta(q) \neq \emptyset$ for all $k \geq j$. □

Our application of Berge's maximum theorem will rely upon the following notion of continuity of set-valued mappings.

Definition 2 (Continuity of correspondences)**.** [2, Theorem 17.20, 17.21]. Let $\Gamma : X \rightrightarrows Y$ be a correspondence (also known as set-valued mapping) between two metric spaces.

1. Γ is upper hemicontinuous at $x \in X$ if $\Gamma(x)$ is compact and for every $\{x_n\}$ in X with $x_n \to x$ and every $\{y_n\}$, such that, $y_n \in \Gamma(x_n)$ for each $n \geq 1$, there exists a subsequence $\{y_{n_m}\}$ with $y_{n_m} \to y \in \Gamma(x)$. Γ is called upper hemicontinuous if it is upper hemicontinuous for all $x \in X$.

2. Γ is lower hemicontinuous at $x \in X$ if for every sequence $\{x_n\}$ in X, such that, $x_n \to x$ and all $y \in \Gamma(x)$, there exists a subsequence $\{x_{n_m}\}$ and a sequence $\{y_m\}$ such that

$y_m \in \Gamma(x_{n_m})$ for each $m \geq 1$ and $y_m \to y$. Γ is called lower hemicontinuous if it is lower hemicontinuous for all $x \in X$.

3. Γ is continuous at x if it is upper hemicontinuous and lower hemicontinuous at x. Γ is called continuous if it is upper hemicontinuous and lower hemicontinuous.

Lemma 12. Suppose Assumption 4 holds. Then, $\tilde{J}_\infty(\epsilon)$ is continuous in ϵ over $[0, \infty)$.

Proof. We prove that $\tilde{J}_\infty(\pi, \epsilon)$ is continuous in (π, ϵ) over $\Pi_R \times [0, \infty)$. The result then follows from Berge's maximum theorem since Π_R is a compact set and $\tilde{J}_\infty(\epsilon) = \max_{\pi \in \Pi_R} \tilde{J}_\infty(\pi, \epsilon)$ from the outer problem in (3.5).

To show that $\tilde{J}_\infty(\pi, \epsilon)$ is continuous, we again apply Berge's maximum theorem. To achieve this, we view $\tilde{J}_\infty(\pi, \epsilon)$ as the optimal value of a parametric problem, namely, the inner optimization problem in (3.5). Specifically, (π, ϵ) are seen as the parameters and $P \in \mathcal{P}(\infty, \epsilon)$ as of course the variable in this problem. Further, $\mathcal{P}(\infty, \epsilon)$ is viewed as a correspondence, that is, a set-valued mapping of ϵ from $[0, \infty)$ into $\bigtimes_{i \in S, a \in A} \Delta^n$. Perhaps with some abuse of notation, we represent this as $\mathcal{P}_\infty : [0, \infty) \rightrightarrows \bigtimes_{i \in S, a \in A} \Delta^n$. Now, to apply Berge's theorem, we need to verify three properties: (i) the correspondence $\mathcal{P}_\infty$ is compact-valued; (ii) the correspondence $\mathcal{P}_\infty$ is continuous; and (iii) the objective function $J_P(\pi)$ of the inner optimization problem is continuous in (P, π). The first property is established in Lemma 9. The third property is established in [67, Proposition 3]. Thus, the rest of this proof focuses on establishing the second property. Specifically, we need to show that $\mathcal{P}_\infty$ is both upper and lower hemicontinuous.

For upper hemicontinuity, let ϵ^k be a sequence of nonnegative reals such that $\epsilon^k \to \epsilon$. Let q^k be a sequence of pmfs such that $q^k \in \mathcal{P}_\infty(\epsilon^k)$ for all $k \in \mathbb{N}$. Since $\mathcal{P}_\infty(\epsilon^k)$ is a subset of the compact set $\bigtimes_{s \in S, a \in A} \Delta^n$, there exists a subsequence $\{q^{k_j}\}$ of $\{q^k\}$ such that $q^{k_j} \to q \in \bigtimes_{s \in S, a \in A} \Delta^n$. Therefore, by Assumption 4(a), $d(q(\cdot|s,a), \hat{p}(\cdot|s,a)) \leq \liminf_{j \to \infty} d(q^{k_j}(\cdot|s,a), \hat{p}(\cdot|s,a))$, for all $s \in S,\ a \in A$. Since $q^{k_j} \in \mathcal{P}_\infty(\epsilon^{k_j})$ for all $j \in \mathbb{N}$, we have $d(q^{k_j}(\cdot|s,a), \hat{p}(\cdot|s,a)) \leq \epsilon^{k_j}$ for all $j \in \mathbb{N}$ and for all $s \in S$ and $a \in A$. Hence, for all $s \in S$ and $a \in A$, we

have $d(q(\cdot|s,a),\hat{p}(\cdot|s,a)) \leq \liminf_{j\to\infty} d(q^{k_j}(\cdot|s,a),\hat{p}(\cdot|s,a)) \leq \liminf_{j\to\infty} \epsilon^{k_j} = \epsilon$. This shows that $q \in \mathcal{P}_\infty(\epsilon)$, establishing upper hemicontinuity.

For lower hemicontinuity, again let $\epsilon^k \to \epsilon$ be a sequence of nonnegative reals. Let $q \in \mathcal{P}_\infty(\epsilon)$. Since $\epsilon^k \to \epsilon$ and Assumption 4 holds, $\text{Lim}\, \mathcal{P}_\infty(\epsilon^k) = \mathcal{P}_\infty(\epsilon)$ from Lemma 11. Then, for each $j \in \mathbb{N}$, we have $B_{1/j}(q) \cap \mathcal{P}_\infty(\epsilon^k) \neq \emptyset$ for infinitely many k. Hence, there exists a sequence $\{q^j\}$ such that $q^j \in B_{1/j}(q) \cap \mathcal{P}_\infty(\epsilon^{k_j})$ for each $j \in \mathbb{N}$. Since $\|q^j - q\|_\infty < 1/j$ for all $j \in \mathbb{N}$, we have $q^j \to q$. Hence $\mathcal{P}$ is lower hemicontinuous. □

We are now ready to state and prove the main result of this section — the existence of an (s,a)-RMDP whose robust optimal value equals the robust optimal value of an s-RMDP.

Theorem 8. Fix any $\rho \in [1,\infty)$ and $\epsilon \in [0,\infty)$. Suppose Assumption 4 holds. Then, there exists an $\epsilon^* \in [\epsilon/m^{1/\rho}, \epsilon]$ such that $\tilde{J}_\infty(\epsilon^*) = \tilde{J}_\rho(\epsilon)$.

Proof. Since Assumption 4 holds, $\tilde{J}_\infty(\epsilon)$ is continuous over $[0,\infty)$ from Lemma 12. Further, from Lemma 8, $\tilde{J}_\infty(\epsilon/m^{1/\rho}) \geq \tilde{J}_\rho(\epsilon) \geq \tilde{J}_\infty(\epsilon)$. Hence, by the intermediate value theorem, there exists an $\epsilon^* \in [\epsilon/m^{1/\rho}, \epsilon]$ such that $\tilde{J}_\infty(\epsilon^*) = \tilde{J}_\rho(\epsilon)$. □

Since the above theorem is an existence result, it leaves open the question of whether or not one can identify, at least algorithmically and approximately, an ϵ^* such that $\tilde{J}_\infty(\epsilon^*) = \tilde{J}_\rho(\epsilon)$. We provide insight into this matter. Since $\tilde{J}_\infty(\cdot)$ is nonincreasing, we can perform a binary search over $[\epsilon/m^{1/\rho}, \epsilon]$ in an attempt to find an appropriate ϵ^*. Specifically, for any $\delta > 0$, the complexity of finding a radius ϵ^δ such that $\left|\tilde{J}_\infty(\epsilon^\delta) - \tilde{J}_\rho(\epsilon)\right| \leq \delta$ is $O\left(L \log_2 \frac{\tilde{J}_\infty(\epsilon/m^{1/\rho}) - \tilde{J}_\infty(\epsilon)}{\delta}\right)$. Here, L is a uniform (over $\epsilon' \in [\epsilon/m^{1/\rho}, \epsilon]$) upper bound on the complexity of computing $\tilde{J}_\infty(\epsilon')$. This requirement of a uniform upper bound L on the complexity of solving an (s,a)-rectangular RMDP is mild and does hold for some well-known distance functions. For instance, the complexity is independent of ϵ' for the TV distance [32, Lemma 4.3] and modified χ^2 distance [32, Lemma 4.2]. For the 1-Wasserstein distance, the complexity is nonincreasing in ϵ' but only has a minimal dependence on ϵ' (see Section 2.6.4). For KL divergence, again the complexity is nonincreasing in ϵ'.

The discussion in this section did not rely on how the reference pmfs $\hat{p}(\cdot|s,a)$, for $s \in S$ and $a \in A$, were constructed. In Section 3.4, we provide more detailed analyses for the case where these reference pmfs equal data-driven empirical estimates of the true transition pmfs as in the previous chapter. Before that, we present counterexample which demonstrates that Assumption 4 cannot be dropped from the hypothesis of Theorem 8

3.3.1 Counterexamples when distance function violates Assumption 4

We first prove that the (s,a)-rectangular ambiguity set is the Kuratowski limit of the family of s-rectangular ambiguity sets.

Lemma 13. Suppose the distance function satisfies Assumption 4. Then, for any $\epsilon \in [0,\infty)$, the Kuratowski limit of the family $\{\mathcal{P}(\rho,\epsilon)\}_{\rho\in[1,\infty)}$ equals $\mathcal{P}(\infty,\epsilon)$.

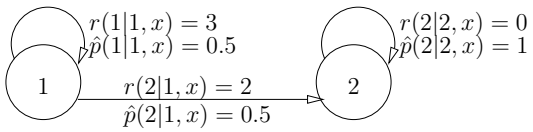
Proof. Since $\{\mathcal{P}(\rho,\epsilon)\}_{\rho\in[1,\infty)}$ is an increasing family, from [39, Chapter 29, VI], the Kuratowski limit of $\{\mathcal{P}(\rho,\epsilon)\}_{\rho\in[1,\infty)}$ equals $\mathbf{cl}\left(\bigcup_{\rho\in[1,\infty)} \mathcal{P}(\rho,\epsilon)\right)$. Hence, in the rest of the proof we establish that $\mathbf{cl}\left(\bigcup_{\rho\in[1,\infty)} \mathcal{P}(\rho,\epsilon)\right) = \mathcal{P}(\infty,\epsilon)$.

Since $\mathcal{P}(\rho,\epsilon) \subseteq \mathcal{P}(\infty,\epsilon)$ for all $\rho \in [1,\infty)$ and $\mathcal{P}(\infty,\epsilon)$ is compact by Lemma 9, we have, $\mathbf{cl}\left(\bigcup_{\rho\in[1,\infty)} \mathcal{P}(\rho,\epsilon)\right) \subseteq \mathcal{P}(\infty,\epsilon)$. For the other direction, let $q \in \mathcal{P}(\infty,\epsilon)$ and consider arbitrary $\delta > 0$. Note that q can be decomposed as $q = (q_1, q_2, \ldots, q_n) \in \times_{s\in A}(\times_{a\in A}\Delta^n)$. Similarly, decompose $\hat{p}$ as $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n) \in \times_{s\in A}(\times_{a\in A}\Delta^n)$. Suppose there exists $s \in S$, such that, $q_s \neq \hat{p}_s$. Using a similar argument as in the proof of Lemma 11, we can prove the existence of a $q'_s \in \mathcal{P}_s(\infty,\epsilon) \cap B_\delta(q_s)$, such that, $q'_s \neq q_s$, and $d(q'_s(\cdot|s,a), \hat{p}(\cdot|s,a)) \leq \epsilon - \eta$ for all $a \in A$, and some $\eta > 0$. Since $\|\cdot\|_\infty = \lim_{\rho\to\infty}\|\cdot\|_\rho$, there exists $\rho_1 \in [1,\infty)$, such that, for all $\rho \geq \rho_1$, we have $\left\|\vec{d}_s(q',\hat{p})\right\|_\rho < \left\|\vec{d}_s(q',\hat{p})\right\|_\infty + \eta \leq \epsilon - \eta + \eta = \epsilon$. This proves that $q'_s \in \bigcap_{\rho\in[\rho_1,\infty)} \mathcal{P}_s(\rho,\epsilon)$. Construct a new vector $q'' = (q''_1, q''_2, \ldots, q''_n) \in \times_{s\in A}(\times_{a\in A}\Delta^n)$, where $q''_s = q_s$, if $q_s = \hat{p}_s$, and $q''_s = q'_s$, otherwise. Since the number of states are finite, there exists large enough $\bar{\rho} \in [1,\infty)$, such that, $q'' \in \bigcap_{\rho\in[\bar{\rho},\infty)} \mathcal{P}(\rho,\epsilon) \subseteq \bigcup_{\rho\in[1,\infty)} \mathcal{P}(\rho,\epsilon)$. Also, by construction

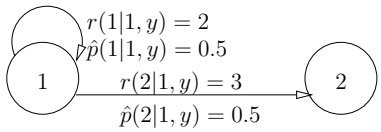
$q'' \in B_\delta(q)$. Therefore, $B_\delta(q) \cap \bigcup_{\rho\in[1,\infty)} \mathcal{P}(\rho,\epsilon) \neq \emptyset$, and hence $q \in \mathbf{cl}\left(\bigcup_{\rho\in[1,\infty)} \mathcal{P}(\rho,\epsilon)\right)$. $\square$

Now, we demonstrate via an example that Assumption 4 cannot be dropped from the hypothesis of Lemma 13 and Theorem 8 (equality of robust optimal values). We first describe the construction of the robust MDP instance which will serve as a counterexample for both these results when Assumption 4 is violated.

Consider a MDP instance where $S = \{1, 2\},\ A = \{x, y\}$, and the discount factor $\alpha = 0.8$. We assume that both actions are allowed in state 1 while only action x is allowed in state 2. We further assume that state 2 is an absorbing state. The reference transition pmf for state $s = 1$ and action $a \in \{x, y\}$ is given as $\hat{p}(\cdot|1, a) = (0.5, 0.5)$. Since state 2 is absorbing, its corresponding reference transition pmf is $\hat{p}(\cdot|2, x) = (0, 1)$. The rewards are given as $r(1|1, x) = 3;\ r(2|1, x) = 2;\ r(1|1, y) = 2;\ r(2|1, y) = 3;\ r(2|2, x) = 0$. Finally, the initial state pmf is $f = (1, 0)$. This MDP is illustrated in Figure 3.1.



(a) Rewards and reference transition probabilities under action x.



(b) Rewards and reference transition probabilities under action y.

Figure 3.1: A schematic of the MDP constructed for counterexample.

Next, consider the distance function

$$d(p, q) = \begin{cases} 0, \text{ if } p = q \\ 1, \text{ otherwise} \end{cases}, \quad \forall p, q \in \Delta^2. \tag{3.7}$$

The above distance function is lsc but nonconvex, thereby violating Assumption 4. Since state 2 is absorbing, its corresponding ambiguity set is given as $\mathcal{P}_2(\rho, \epsilon) = \mathcal{P}_{(2,x)}(\epsilon) = \{(0, 1)\}\ \forall \rho \in [1, \infty)$ and $\epsilon \in [0, \infty)$. For state 1, the ambiguity set is constructed using

the above distance function with $\epsilon = 1$. Therefore, for any $\rho \in [1, \infty)$, the s-recatngular ambiguity set is

$$\begin{aligned}
\mathcal{P}(\rho, \epsilon) &= \mathcal{P}_1(\rho, \epsilon) \times \mathcal{P}_2(\rho, \epsilon) = \mathcal{P}_1(\rho, \epsilon) \times \{(0,1)\}, \text{ where} \qquad (3.8)\\
\mathcal{P}_1(\rho, \epsilon) &= \Big\{(q(\cdot|1,x), q(\cdot|1,y)) \in \Delta^2 \times \Delta^2 \Big| \Big(d^\rho(q(\cdot|1,x), \hat{p}(\cdot|1,x)) + \\
&\qquad\qquad d^\rho(q(\cdot|1,y), \hat{p}(\cdot|1,y)) \Big)^{1/\rho} \leq \epsilon \Big\} \\
&= \big(\{(0.5, 0.5)\} \times \Delta^2\big) \cup \big(\Delta^2 \times \{(0.5, 0.5)\}\big).
\end{aligned}$$

Similarly, the (s,a)-rectangular ambiguity set is

$$\begin{aligned}
\mathcal{P}(\infty, \epsilon) &= \mathcal{P}_{(1,x)}(\epsilon) \times \mathcal{P}_{(1,y)}(\epsilon) \times \mathcal{P}_{(2,x)}(\epsilon) \qquad (3.9)\\
&= \Big\{q(\cdot|1,x) \Big| d(q(\cdot|1,x), \hat{p}(\cdot|1,x)) \leq \epsilon \Big\} \times \\
&\qquad\qquad \Big\{q(\cdot|1,y) \Big| d(q(\cdot|1,y), \hat{p}(\cdot|1,y)) \leq \epsilon \Big\} \times \{(0,1)\} \\
&= \Delta^2 \times \Delta^2 \times \{(0,1)\}.
\end{aligned}$$

Example 8 (Counterexample to Lemma 13)**.** Let $\mathcal{P}(\infty, \epsilon)$ and $\{\mathcal{P}(\rho, \epsilon)\}_{\rho \in [1,\infty)}$ be the (s,a)- and s-rectangular ambiguity sets as defined in (3.9) and (3.8), repectively. We demonstrate that $\mathcal{P}(\infty, \epsilon)$ cannot be recovered as the Kuratowski limit of the family $\{\mathcal{P}(\rho, \epsilon)\}_{\rho \in [1,\infty)}$. From the proof of Lemma 13, this is equivalent to $\mathbf{cl}\left(\bigcup_{\rho \in [1,\infty)} \mathcal{P}(\rho, \epsilon)\right) \neq \mathcal{P}(\infty, \epsilon)$. Towards that end, let $q = (1,0,1,0,0,1) = \{(1,0)\} \times \{(1,0)\} \times \{(0,1)\}$ from $\mathcal{P}_{(1,x)}(\epsilon) \times \mathcal{P}_{(1,y)}(\epsilon) \times \mathcal{P}_{(2,x)}(\epsilon)$. Hence, $\|q - q'\|_\infty \geq 0.5$ for any $\rho \in [1, \infty)$, $q' \in \mathcal{P}(\rho, \epsilon) = \mathcal{P}_1(\rho, \epsilon) \times \mathcal{P}_2(\rho, \epsilon)$. Therefore, $B_\delta(q) \cap \left(\bigcup_{\rho \in [1,\infty)} \mathcal{P}(\rho, \epsilon)\right) = \emptyset$ for any $\delta \leq 0.5$. Hence, $q \notin \mathbf{cl}\left(\bigcup_{\rho \in [1,\infty)} \mathcal{P}(\rho, \epsilon)\right)$.

Example 9 (Counterexample to Theorem 8)**.** We prove that it is not possible to match the robust optimal values of the (s,a)- and s-RMDP when Assumption 4 is violated. Towards that end, consider the RMDPs described in the begining of this section. We first prove that for any $\rho \in [1, \infty)$, the robust optimal value of the s-RMDP, $\tilde{J}_\rho(\epsilon) = \frac{2 + 1.25\sqrt{4.64}}{3.6 - \sqrt{4.64}} \approx 3.24536$. In

the next step, we calculate the robust optimal value of the (s,a)-RMDP. Since the number of actions $m = 2$, the corresponding range of radii for the (s,a)-rectangular ambiguity set is $[\epsilon/2^{1/\rho}, \epsilon]$ for any $\rho \in [1,\infty)$. Finally, we demonstrate that $\tilde{J}_\infty(\tilde{\epsilon}) \neq \tilde{J}_\rho(\epsilon)$ for any $\rho \in [1,\infty),\ \tilde{\epsilon} \in [\epsilon/2^{1/\rho}, \epsilon]$.

Step 1: $\tilde{J}_\rho(\epsilon) = \frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}} \approx 3.24536$ for all $\rho \in [1,\infty)$. Fix any $\rho \in [1,\infty)$. Since only action x is allowed in state 2, we have $\pi(2) = x$ for any $\pi \in \Pi_R$. Since state 1 has 2 actions, any policy $\pi(\cdot|1)$ corresponding to state 1 can be represented as $\pi(\cdot|1) = (\lambda, 1-\lambda)$ where $\lambda \in [0,1]$. We first provide an upper bound on $\tilde{J}_\rho(\epsilon)$. Consider the following transition pmfs

$$\mathbf{P_1} : p(\cdot|x,1) = (0.5, 0.5);\ p(\cdot|y,1) = (0,1);\ p(\cdot|2,x) = (0,1)$$
$$\mathbf{P_2} : p(\cdot|x,1) = (0,1);\ p(\cdot|y,1) = (0.5,0.5);\ p(\cdot|2,x) = (0,1).$$

From the definition of $\mathcal{P}(\rho,\epsilon)$ in (3.8), $\mathbf{P_1}$ and $\mathbf{P_2}$ belong to $\mathcal{P}(\rho,\epsilon)$. Hence,

$$\tilde{J}_\rho(\epsilon) = \max_{\pi\in\Pi_R} \inf_{P\in\mathcal{P}(\rho,\epsilon)} J_P(\pi) \leq \max_{\pi\in\Pi_R} \inf_{P\in\{\mathbf{P_1},\mathbf{P_2}\}} J_P(\pi).$$

Using Bellman's policy evaluation equations [50, Equation 6.1.9], for any policy $\pi \in \Pi_R$, we have, $J^\pi_{\mathbf{P_1}} = (3-0.5\lambda)/(1-0.4\lambda);\ J^\pi_{\mathbf{P_2}} = (2.5-0.5\lambda)/(0.6+0.4\lambda)$. Therefore,

$$\tilde{J}_\rho(\epsilon) \leq \max_{\lambda\in[0,1]} \min\left\{\frac{3-0.5\lambda}{1-0.4\lambda}, \frac{2.5-0.5\lambda}{0.6+0.4\lambda}\right\} = \frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}} \approx 3.24536. \tag{3.10}$$

Now, for $\lambda^* = \frac{2.4-\sqrt{4.64}}{0.8} \approx 0.307417$, consider the policy $\pi^* \in \Pi_R$ where $\pi^*(\cdot|1) = (\lambda^*, 1-\lambda^*)$. We prove that the value of this policy on the s-RMDP, namely, $\tilde{J}_\rho(\pi^*,\epsilon) = \frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}}$. This along with the upper bound in (3.10) establishes that $\tilde{J}_\rho(\epsilon) = \frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}}$. Suppose the minimum in the inner optimization problem occurs at $(p^*(\cdot|1,x), p^*(\cdot|1,y)) \in \Delta^2 \times \{(0.5,0.5)\}$. Parametrize $p^*(\cdot|1,x)$ as $p^*(\cdot|1,x) = (p_1^*, 1-p_1^*)$ where $p_1^* \in [0,1]$. Using Bellman's policy evaluation equations [50, Equation 6.1.9], the value of π^* as a function of p_1^* equals $\frac{2+1.25\sqrt{4.64}+(6+2.5\sqrt{4.64})p_1^*}{3.6-\sqrt{4.64}-(2.4-\sqrt{4.64})p_1^*}$. This is an increasing function in p_1^*, and hence minimized

at $p_1^* = 0$. Therefore, the corresponding value at $p_1^* = 0$ is $\frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}} \approx 3.24536$. Now, suppose the minimum in the inner optimization problem occurs at $(p^*(\cdot|1,x), p^*(\cdot|1,y)) \in \{(0.5,0.5)\} \times \Delta^2$. Using a similar approach, the value of π^* equals $\frac{3+1.25\sqrt{4.64}+(4-2.5\sqrt{4.64})p_1^*}{-0.4+\sqrt{4.64}+(3.2-2\sqrt{4.64})p_1^*}$ where $p_1^* \in [0,1]$. Again, this is an increasing function in p_1^*, and hence minimized at $p_1^* = 0$. Therefore, the corresponding value at $p_1^* = 0$ is $\frac{3+1.25\sqrt{4.64}}{-0.4+\sqrt{4.64}} \approx 3.24536$. Hence,

$$\tilde{J}_\rho(\pi^*, \epsilon) = \min\left\{\frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}}, \frac{3+1.25\sqrt{4.64}}{-0.4+\sqrt{4.64}}\right\} = \frac{2+1.25\sqrt{4.64}}{3.6-\sqrt{4.64}} \approx 3.24536.$$

Step 2: Computing $\tilde{J}_\infty(\tilde{\epsilon})$ for any $\rho \in [1,\infty)$, $\tilde{\epsilon} \in [\epsilon/2^{1/\rho}, \epsilon]$. Fix any $\rho \in [1,\infty)$. Since $\epsilon = 1$, we have, $\epsilon/2^{1/\rho} < 1$. From the definition of the distance function in (3.7), we have $\mathcal{P}_{(1,x)}(\tilde{\epsilon}) = \mathcal{P}_{(1,y)}(\tilde{\epsilon}) = \{(0.5,0.5)\}$ for all $\tilde{\epsilon} \in [\epsilon/2^{1/\rho}, \epsilon)$. Hence, $\tilde{J}_\infty(\tilde{\epsilon}) = 4.16666$ for all $\tilde{\epsilon} \in [\epsilon/2^{1/\rho}, \epsilon)$. When $\tilde{\epsilon} = \epsilon = 1$, the ambiguity set takes the form as in (3.9). Using robust value iteration or robust policy iteration [32], we get $\tilde{J}_\infty(\epsilon) = 3$. Therefore,

$$\tilde{J}_\infty(\tilde{\epsilon}) = \begin{cases} 4.16666 & \text{if } \tilde{\epsilon} \in [\epsilon/2^{1/\rho}, \epsilon),\ \rho \in [1,\infty) \\ 3 & \text{if } \tilde{\epsilon} = \epsilon. \end{cases}$$

Step 3: Conclusion. From steps 1 and 2, we conclude that $\tilde{J}_\infty(\tilde{\epsilon}) \neq \tilde{J}_\rho(\epsilon)$ for any $\rho \in [1,\infty)$, $\tilde{\epsilon} \in [\epsilon/2^{1/\rho}, \epsilon]$.

Note that the s-rectangular ambiguity set $\mathcal{P}(\rho,\epsilon)$ in (3.8) is nonconvex. As a result, an optimal policy in the s-RMDP may need to be history dependent and randomized [67, Table 1]. The above counterexample only shows that, within our family, there is no (s,a)-RMDP instance with a robust optimal value equal to that of the s-RMDP, if we restrict the policy space of the s-RMDP to Π_R. However, we do not know if it is possible to match the robust optimal values when history dependent randomized policies are allowed for the s-RMDP. An investigation of this issue is beyond the scope of this work.

3.4 Asymptotic and finite-sample properties of data-driven s-RMDPs

Suppose the decision-maker has access to N independent observations of the next state reached upon choosing action a in state s, for all $s \in S$ and $a \in A$. We call this the training data. Let $\hat{p}^N(\cdot|s,a)$ denote an empirical estimate, computed based on this training data, of the true transition pmf $p(\cdot|s,a)$. Throughout this section, we use these empirical pmfs in place of the generic reference pmfs $\hat{p}$ utilized earlier. We therefore indicate all corresponding entities such as ambiguity sets and robust optimal values with a superscript N, and call the corresponding problem as the data-driven s-RMDP. For example, we write $\mathcal{P}_\rho^N(\epsilon)$ or $\tilde{J}_\rho^N(\epsilon)$.

We assume throughout this section that, for every $(s,a) \in S \times A$, the true transition pmf $p(\cdot|s,a)$ has full support S. This is done merely for notational convenience and all our proofs go through even when this assumption is not satisfied. Finally, recall that the strong law of large numbers (SLLN) guarantees that the empirical estimates of the true transition probabilities converge wp 1 to the true transition probabilities, as $N \to \infty$.

In Section 3.4.1, we study asymptotic properties of the data-driven s-RMDP as sample-size $N \to \infty$. In Sections 3.4.2 and 3.4.3, we focus on finite-sample properties.

3.4.1 Value convergence to true optimal value

We will rely on Assumption 1 throughout this section. Recall that the generalized Pinsker's inequality stated in Assumption 2 is a sufficient condition for Assumption 1.

We begin our asymptotic analyses with a simple lemma. It establishes that any sequence of transition probabilities from s-rectangular ambiguity sets converges to the true transition pmf, as long as radii of those ambiguity sets converge to 0.

Lemma 14. Fix any $\rho \in [1,\infty)$. Suppose Assumption 1 holds. Consider the s-rectangular ambiguity sets $\mathcal{P}_\rho^N(\epsilon^N)$ such that the radii $\lim_{N\to\infty} \epsilon^N = 0$. Let $P^N \in \mathcal{P}_\rho^N(\epsilon^N)$ be any sequence. Then, $P^N \xrightarrow{N,\text{wp1}} \mathbf{P}$, where $\mathbf{P}$ is the true transition probability matrix.

Proof. Similar to Lemma 2; hence omitted. □

We introduce additional notation before presenting the next result. For any policy $\sigma \in \Pi_R$ and any transition probabilities $\{q(j|i,a)\}_{i\in S, a\in A}^{j\in A}$, let

$$Q_{ij}^{\sigma} \stackrel{\text{def}}{=} \sum_{a\in A} \sigma(a|i) q(j|i,a) \qquad r_i^{\sigma} \stackrel{\text{def}}{=} \sum_{a\in A} \sigma(a|i) \sum_{j\in S} q(j|i,a) r(j|i,a). \tag{3.11}$$

That is, Q_{ij}^{σ} is the probability that the next state is j if policy σ is implemented in the current state i. Similarly, r_i^{σ} is the expected reward earned upon implementing policy σ in state i. We use r^{σ} to denote a vector with components r_i^{σ}, for $i = 1, \ldots, n$.

Proposition 1. Fix any $\rho \in [1, \infty)$. Suppose the radii of ambiguity sets $\mathcal{P}^N(\rho, \epsilon^N)$ are such that $\lim_{N\to\infty} \epsilon^N = 0$. If Assumption 1 holds, then $\tilde{J}_{\rho}^N(\pi, \epsilon^N) \xrightarrow{N,\text{wp1}} J_{\mathbf{P}}(\pi)$ uniformly over Π_R.

Proof. The sequence of radii ϵ^N depends only on the sample-size N. Thus, view $\tilde{J}_{\rho}^N(\cdot, \epsilon^N)$ as a sequence of functions from Π_R to $\mathbb{R}$, indexed by N. We will first show that this sequence of functions converges continuously to the function $J_{\mathbf{P}}(\cdot)$ from Π_R to $\mathbb{R}$. That is, we will show that, for any sequence of policies $\pi^N \in \Pi$ that converges to some $\pi \in \Pi_R$, the sequence $\tilde{J}_{\rho}^N(\pi^N, \epsilon^N)$ converges to $J_{\mathbf{P}}(\pi)$, wp1. This would imply that the convergence $\tilde{J}_{\rho}^N(\cdot, \epsilon^N) \xrightarrow{N,\text{wp1}} J_{\mathbf{P}}(\cdot)$ is uniform over Π_R because Π_R is compact. This conclusion is based on a classic result from real analysis, which states that a sequence of functions converging continuously over a compact set in fact converges uniformly [39, Chapter 21, X, Theorem 5].

Fix $\delta > 0$. By the definition of infimum, there exists a sequence $P^N \in \mathcal{P}^N(\rho, \epsilon^N)$ such that $\left|\tilde{J}_{\rho}^N(\pi^N, \epsilon^N) - J_{P^N}(\pi^N)\right| \leq \delta/2$ for all $N \geq 1$ (P^N is induced by π^N, however, we do not emphasize this in the notation to avoid clutter). By the triangle inequality, for all $N \geq 1$ we have,

$$\begin{aligned}\left|\tilde{J}_{\rho}^N(\pi^N, \epsilon^N) - J_P(\pi)\right| &\leq \left|\tilde{J}_{\rho}^N(\pi^N, \epsilon^N) - J_{P^N}(\pi^N)\right| + \left|J_{P^N}(\pi^N) - J_P(\pi)\right| \\ &\leq \delta/2 + \left|J_{P^N}(\pi^N) - J_P(\pi)\right|.\end{aligned}$$

We will establish that $\left|J_{P^N}(\pi^N) - J_P(\pi)\right| \xrightarrow{N,\text{wp1}} 0$. From the above inequality, this would imply that $\tilde{J}_{\rho}^N(\pi^N, \epsilon^N) \xrightarrow{N,\text{wp1}} J_P(\pi)$, as required.

From [50, Chapter 6], we know that $J_{\mathbf{P}}(\pi) = \sum_{i \in S} f(i) \left[(I - \alpha \mathbf{P}^{\pi})^{-1} r^{\pi} \right]_i$ and

$$J_{P^N}(\pi^N) = \sum_{i \in S} f(i) \left[\left(I - \alpha \left(P^N \right)^{\pi^N} \right)^{-1} r^{\pi^N} \right]_i .$$

Therefore,

$$\left| J_{P^N}(\pi^N) - J_{\mathbf{P}}(\pi) \right| \leq \sum_{i \in S} f(i) \left| \left[\left(I - \alpha \left(P^N \right)^{\pi^N} \right)^{-1} r^{\pi^N} \right]_i - \left[(I - \alpha \mathbf{P}^{\pi})^{-1} r^{\pi} \right]_i \right| . \quad (3.12)$$

From Lemma 14, we know that $P^N \xrightarrow{N,\text{wp1}} \mathbf{P}$ because $P^N \in \mathcal{P}^N(\rho, \epsilon^N)$ and $\epsilon^N \to 0$. Further, $\pi^N \to \pi$. Hence, from the definition of $\left(P^N \right)^{\pi^N}$ and r^{π^N} as in (3.11), we get

$$\left(I - \alpha \left(P^N \right)^{\pi^N} \right)^{-1} \xrightarrow{N,\text{wp1}} (I - \alpha \mathbf{P}^{\pi})^{-1} \quad \text{and} \quad r^{\pi^N} \xrightarrow{N,\text{wp1}} r^{\pi}.$$

Utilizing this in (3.12) proves that $\left| J_{P^N}(\pi^N) - J_{\mathbf{P}}(\pi) \right| \xrightarrow{N,\text{wp1}} 0$. □

We assume in the rest of the chapter that the distance function satisfies Assumption 4. As previously explained, this renders the ambiguity set compact and convex by Lemma 9, thus, guaranteeing the existence of a stationary Markovian optimal policy to the data-driven s-RMDP problem. We therefore introduce $\hat{\pi}_{\rho}^N \in \Pi_R$ to denote a policy optimal to the data-driven s-RMDP. This is called a robust optimal policy. Let $J_{\mathbf{P}}(\hat{\pi}_{\rho}^N)$ denote the value of this policy in the the MDP where the transition probability matrix is the true one, that is, $\mathbf{P}$. This is called the out-of-sample value of the robust optimal policy. The above proposition helps establish the main value convergence result of this section next.

Theorem 9. Fix any $\rho \in [1, \infty)$. Suppose the radii of the ambiguity balls are dependent on N such that $\lim_{N \to \infty} \epsilon^N = 0$. Suppose Assumption 4 and Assumption 1 hold. Then $\tilde{J}_{\rho}^N(\epsilon^N) \xrightarrow{N,\text{wp1}} J^*$. That is, we have robust optimal value convergence. Moreover, $J_{\mathbf{P}}(\hat{\pi}^N) \xrightarrow{N,\text{wp1}} J^*$. That is, we have out-of-sample value convergence.

Proof. Since $\tilde{J}^N_\rho(\epsilon^N) = \max_{\pi\in\Pi_R} \tilde{J}^N_\rho(\pi,\epsilon^N)$, we have

$$\lim_{N\to\infty} \tilde{J}^N_\rho(\epsilon^N) = \lim_{N\to\infty} \max_{\pi\in\Pi_R} \tilde{J}^N_\rho(\pi,\epsilon^N) = \max_{\pi\in\Pi_R} \lim_{N\to\infty} \tilde{J}^N_\rho(\pi,\epsilon^N) = \max_{\pi\in\Pi_R} J_{\mathbf{P}}(\pi) = J^*.$$

Here, the interchange of the limit and the maximum that yields the second and the third equalities is allowed by the uniform convergence result in Proposition 1. For the second claim, note that

$$\begin{aligned}
\left|J^* - J_{\mathbf{P}}(\hat{\pi}^N)\right| &\leq \left|J^* - \tilde{J}^N_\rho(\epsilon^N)\right| + \left|\tilde{J}^N_\rho(\epsilon^N) - J_{\mathbf{P}}(\hat{\pi}^N)\right| \\
&\overset{(a)}{=} \left|\max_{\pi\in\Pi_R} J_{\mathbf{P}}(\pi) - \max_{\pi\in\Pi_R} \tilde{J}^N_\rho(\pi,\epsilon^N)\right| + \left|\tilde{J}^N_\rho(\hat{\pi}^N_\rho,\epsilon^N) - J_{\mathbf{P}}(\hat{\pi}^N)\right| \\
&\leq \max_{\pi\in\Pi_R} \left|J_{\mathbf{P}}(\pi) - \tilde{J}^N_\rho(\pi,\epsilon^N)\right| + \max_{\pi\in\Pi_R} \left|\tilde{J}^N_\rho(\pi,\epsilon^N) - J_{\mathbf{P}}(\pi)\right| \\
&= 2\max_{\pi\in\Pi_R} \left|\tilde{J}^N_\rho(\pi,\epsilon^N) - J_{\mathbf{P}}(\pi)\right|.
\end{aligned}$$

The second term in (a) follows because $\hat{\pi}^N_\rho$ is optimal to the data-driven s-RMDP, and hence $\tilde{J}^N_\rho(\epsilon^N) = \tilde{J}^N_\rho(\hat{\pi}^N_\rho,\epsilon^N)$. The result then follows because the above upper bound converges to 0 wp1 by Proposition 1. □

This proof sheds light on the importance of our uniform convergence result from Proposition 1. The value convergence result for the (s,a)-RMDP case in the previous chapter did not need this type of uniform convergence. There, it was sufficient to limit attention to the finite set of deterministic stationary policies. Thus, an interchange of a limit and a maximum was trivially valid.

3.4.2 Probabilistic guarantee on the performance of robust optimal policy

Though Theorem 9 shows that $J_{\mathbf{P}}(\hat{\pi}^N_\rho)$ converges to the true optimal value wp1, we would also want $J_{\mathbf{P}}(\hat{\pi}^N_\rho)$ to be sufficiently high with a large enough probability (with respect to the sampling uncertainty in training data) for finite sample-sizes. In this section, we will prescribe a concrete way for the decision-maker to proceed so that the robust optimal value

$\tilde{J}_\rho^N(\epsilon^N)$ provides a high probability lower bound on $J_{\mathbf{P}}(\hat{\pi}_\rho^N)$. This result builds upon an intermediate lemma that we now establish.

Lemma 15. Fix any $\rho \in [1, \infty)$. Suppose Assumption 4 holds. Then, for any sample-size $N \geq 1$ and $\epsilon \in [0, \infty)$, we have,

$$\mathbb{P}\left[J_{\mathbf{P}}(\hat{\pi}_\rho^N) \geq \tilde{J}_\rho^N(\epsilon)\right] \geq \prod_{i=1}^{n}\prod_{a=1}^{m} \mathbb{P}\left[d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)) \leq \epsilon/m^{1/\rho}\right].$$

Proof. The proof is similar to Theorem 6 established for (s,a)-RMDPs in the previous chapter, that is, for $\rho = \infty$. We nevertheless include a proof here for completeness and to bring forth the minor algebraic differences between the two proofs. We introduce the Bellman evaluation operator $\Phi_{\hat{\pi}^N} : \mathbb{R}^n \to \mathbb{R}^n$ as

$$[\Phi_{\hat{\pi}^N} V](i) \stackrel{\text{def}}{=} \sum_{a \in A} \hat{\pi}^N(a|i) \underset{p(\cdot|i,a)}{\mathbb{E}} \left[r(\mathbf{s}|i,a) + \alpha V(\mathbf{s})\right], \text{ for } i = 1, 2, \ldots, n. \tag{3.13}$$

A standard monotonicity and successive approximation approach from stationary infinite-horizon MDPs [50, Chapter 6] guarantees that

$$\left\{[\Phi_{\hat{\pi}^N} \tilde{J}_\rho^N(\epsilon)](i) \geq \tilde{J}_\rho^N(i;\epsilon),\ i = 1, 2, \ldots, n\right\} \subseteq \left\{J_{\mathbf{P}}(\hat{\pi}_\rho^N; i) \geq \tilde{J}_\rho^N(\epsilon; i),\ i = 1, 2, \ldots, n\right\}. \tag{3.14}$$

This yields,

$$\begin{aligned}
&\mathbb{P}\left[J_{\mathbf{P}}(\hat{\pi}_\rho^N) \geq \tilde{J}_\rho^N(\epsilon)\right] = \mathbb{P}\left[\sum_{i=1}^{n} f(i) J_{\mathbf{P}}(\hat{\pi}_\rho^N; i) \geq \sum_{i=1}^{n} f(i) \tilde{J}_\rho^N(\epsilon; i)\right] \\
&\geq \mathbb{P}\left[J_{\mathbf{P}}(\hat{\pi}_\rho^N; i) \geq \tilde{J}_\rho^N(\epsilon; i),\ i = 1, 2, \ldots, n\right] \\
&\stackrel{(a)}{\geq} \mathbb{P}\left[[\Phi_{\hat{\pi}^N} \tilde{J}_\rho^N(\epsilon)](i) \geq \tilde{J}_\rho^N(i;\epsilon),\ i = 1, 2, \ldots, n\right] \\
&\stackrel{(b)}{=} \mathbb{P}\left[\sum_{a \in A} \hat{\pi}^N(a|i) \underset{p(\cdot|i,a)}{\mathbb{E}} \left[r(\mathbf{s}|i,a) + \alpha \tilde{J}_\rho^N(\epsilon; \mathbf{s})\right]\right.
\end{aligned}$$

$$\geq \inf_{\substack{(q(\cdot|i,1),\ldots,q(\cdot|i,m)) \\ \in \\ \mathcal{P}_i^N(\rho,\epsilon)}} \left(\sum_{a\in A} \hat{\pi}^N(a|i) \mathop{\mathbb{E}}_{q(\cdot|i,a)} \left[r(\mathbf{s}|i,a) + \alpha \tilde{J}_\rho^N(\epsilon;\mathbf{s}) \right] \right),\ i = 1,2,\ldots,n \Bigg]$$

$$\geq \mathbb{P}\left[(p(\cdot|i,1),\ldots,p(\cdot|i,m)) \in \mathcal{P}_i^N(\rho,\epsilon),\ i = 1,2,\ldots,n \right] \text{ (defn. of infimum)}$$

$$= \mathbb{P}\left[\left\| \vec{d}_i(p,\hat{p}^N) \right\|_\rho \leq \epsilon,\ i = 1,2,\ldots,n \right] \text{ (defn. of } \mathcal{P}_i^N(\rho,\epsilon))$$

$$\overset{(c)}{=} \prod_{i=1}^n \mathbb{P}\left[\left\| \vec{d}_i(p,\hat{p}^N) \right\|_\rho \leq \epsilon \right]$$

$$\geq \prod_{i=1}^n \mathbb{P}\left[\left\| \vec{d}_i(p,\hat{p}^N) \right\|_\infty \leq \epsilon/m^{1/\rho} \right] \text{ (since } \|\cdot\|_\rho \leq m^{1/\rho} \|\cdot\|_\infty \text{)}$$

$$\overset{(d)}{=} \prod_{i=1}^n \mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)) \leq \epsilon/m^{1/\rho},\ a = 1,2,\ldots,m \right]$$

$$\overset{(e)}{=} \prod_{i=1}^n \prod_{a=1}^m \mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)) \leq (\epsilon)/m^{1/\rho} \right],$$

as claimed. Here, inequality "(a)" follows from (3.14). Equality "(b)" follows from (3.13) and the fact that $\hat{\pi}^N(a|i)$ attain the maxima in robust Bellman's equations of optimality (Theorem 4 in [67]). In particular, the latter yields that, for all $i = 1,2,\ldots,n$,

$$\tilde{J}_\rho^N(i;\epsilon^N) = \inf_{\substack{(q(\cdot|i,1),\ldots,q(\cdot|i,m)) \\ \in \\ \mathcal{P}_i^N(\rho,\epsilon^N)}} \left(\sum_{a\in A} \hat{\pi}^N(a|i) \mathop{\mathbb{E}}_{q(\cdot|i,a)} \left[r(\mathbf{s}|i,a) + \alpha \tilde{J}^N(\mathbf{s};\epsilon^N) \right] \right).$$

Equality "(c)" holds because the sampled observations of the next state reached are independent across current states i. Equality "(d)" follows by noting that d is nonnegative. Finally, the equality in "(e)" follows because the samples are independent across actions. □

Our probabilistic guarantee will rely on the concentration inequality stated in Assumption 3.

Theorem 10. Fix any $\rho \in [1,\infty)$ and sample-size $N \geq 1$. Suppose Assumption 4 and Assumption 3 hold. Consider any fixed $\gamma \in (0,1)$, and let $0 < \beta(\gamma) = 1 - (1-\gamma)^{1/nm}$. Let $0 \leq \epsilon_{\beta(\gamma)}^N < \sup_{x,y\in\Delta^n} d(x,y)$ be as in Assumption 3. If the ambiguity set radius is $m^{1/\rho}\epsilon_{\beta(\gamma)}^N$,

then

$$\mathbb{P}\left[J_{\mathbf{P}}(\hat{\pi}_\rho^N) \geq \tilde{J}_\rho^N(m^{1/\rho}\epsilon_{\beta(\gamma)}^N)\right] \geq 1-\gamma. \tag{3.15}$$

Proof. From Lemma 15, we have

$$\begin{aligned}\mathbb{P}\left[J_{\mathbf{P}}(\hat{\pi}_\rho^N) \geq \tilde{J}_\rho^N(m^{1/\rho}\epsilon_{\beta(\gamma)}^N)\right] &\geq \prod_{i=1}^{n}\prod_{a=1}^{m}\mathbb{P}\left[d(p(\cdot|i,a),\hat{p}^N(\cdot|i,a)) \leq \epsilon_{\beta(\gamma)}^N)\right] \\ &\overset{(a)}{\geq} \prod_{i=1}^{n}\prod_{a=1}^{m}(1-\beta(\gamma)) \overset{(b)}{=} 1-\gamma.\end{aligned}$$

The inequality in "(a)" follows from (2.9) and the equality in "(b)" follows by the choice of $\beta(\gamma)$. □

The above theorem has the following practical utility. Consider any $\gamma \in (0,1)$ for a desired $1-\gamma$ probabilistic guarantee. Let $\beta(\gamma) = 1-(1-\gamma)^{1/nm}$. Suppose the decision-maker solves a data-driven distance-based s-RMDP using ambiguity sets $\mathcal{P}_s(\rho, m^{1/\rho}\epsilon_{\beta(\gamma)}^N)$ for states $s \in S$. Then, with at least $1-\gamma$ probability, the out-of-sample value $J_{\mathbf{P}}(\hat{\pi}_\rho^N)$ of a resulting robust optimal policy $\hat{\pi}_\rho^N$ would be at least $\tilde{J}_\rho^N(m^{1/\rho}\epsilon_{\beta(\gamma)}^N)$.

The above result was established for $\rho = \infty$ (corresponding to the (s,a)-RMDP in the previous chapter) without requiring Assumption 4 in Theorem 6. That is,

$$\mathbb{P}\left[J_{\mathbf{P}}(\hat{\pi}_\infty^N) \geq \tilde{J}_\infty^N(\epsilon_{\beta(\gamma)}^N)\right] \geq 1-\gamma. \tag{3.16}$$

Recall that $\tilde{J}_\infty^N(\epsilon_{\beta(\gamma)}^N) \geq \tilde{J}_\rho^N(m^{1/\rho}\epsilon_{\beta(\gamma)}^N)$ because $\mathcal{P}(\infty, \epsilon_{\beta(\gamma)}^N) \subseteq \mathcal{P}(\rho, m^{1/\rho}\epsilon_{\beta(\gamma)}^N)$, for all $\rho \in [1,\infty)$, as shown in Lemma 8. We use this ordering of robust optimal values to compare the lower bounds in (3.15) and (3.16). This comparison reveals that the (s,a)-RMDP framework, wherein $\rho = \infty$, provides the best lower bound on the out-of-sample performance of the robust optimal policy, since the probabilistic guarantee $1-\gamma$ does not depend on ρ. In this sense, the (s,a)-RMDP with an ambiguity radius dictated by the concentration inequality is the least conservative among all s-RMDPs that use ambiguity radii prescribed by the same concentration inequality. This idea of using the quality of out-of-sample probabilistic

guarantees to characterize relative conservativeness of robust optimization frameworks was proposed in the context of single-stage stochastic optimization problems in [60, 64].

3.4.3 Rate of convergence

In this section, we derive rates of convergence of the robust optimal and out-of-sample values to the true optimal value as a function of the sample-size N. The rate of convergence provides a quantitative version of the asymptotic convergence result established in Theorem 9. To begin with, note that there exists a constant $C > 0$ such that $\max_{\substack{i,j\in S\\ a\in A}} |r(j|i,a)| < C$. This follows since the state- and action-spaces are finite.

Theorem 11. Fix any $\rho \in [1,\infty)$ and sample-size $N \geq 1$. Suppose Assumption 4 holds. Let $\beta(N,\epsilon) \in [0,1]$ be such that

$$\mathbb{P}[d(p(\cdot|i,a), \hat{p}^N(|i,a)) \leq \epsilon] \geq 1 - \beta(N,\epsilon),\ \forall i \in S, a \in A,\ \epsilon \in [0,\infty). \tag{3.17}$$

Suppose that the distance function d satisfies the generalized Pinsker inequality stated in Assumption 2, and that the function ψ introduced there is nondecreasing. Suppose the decision-maker employs ambiguity radius ϵ. Then, with probability at least $(1-\beta(N,\epsilon/m^{1/\rho}))^{nm}$, we have,

$$\left|\tilde{J}^N_\rho(\epsilon) - J^*\right| \leq \frac{\left(\psi(\epsilon) + \psi\left(\frac{\epsilon}{m^{1/\rho}}\right)\right) C}{(1-\alpha)^2} \text{ and } J^* - J_{\mathbf{P}}(\hat{\pi}^N_\rho) \leq \frac{2\left(\psi(\epsilon) + \psi\left(\frac{\epsilon}{m^{1/\rho}}\right)\right) C}{(1-\alpha)^2}. \tag{3.18}$$

Proof. We first prove that

$$\left|\tilde{J}^N_\rho(\epsilon) - J^*\right| \leq \frac{\left(\psi(\epsilon) + \max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)))\right) C}{(1-\alpha)^2}, \text{ and} \tag{3.19}$$

$$J^* - J_{\mathbf{P}}(\hat{\pi}^N_\rho) \leq \frac{2\left(\psi(\epsilon) + \max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)))\right) C}{(1-\alpha)^2}. \tag{3.20}$$

Fix $\delta > 0$. For any $\pi \in \Pi_R$, let $P^{N,\pi,\delta} \in \mathcal{P}_\rho^N(\epsilon)$ be such that $\left|\tilde{J}_\rho^N(\pi,\epsilon) - J_{P^{N,\pi,\delta}}(\pi)\right| \leq \delta$. Such a $P^{N,\pi,\delta}$ exists by defintion of the infimum. Following a similar proof argument as done in the proof of Lemma 5 for the (s,a)-RMDP in the previous chapter, we have

$$\left|\tilde{J}_\rho^N(\epsilon^N) - J^*\right| \leq \frac{C}{(1-\alpha)^2}\Bigg\{ \max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a))) + \max_{\pi\in\Pi_R, i\in S} \sum_{a\in A} \pi(a|i)\psi(d(\hat{p}^N(\cdot|i,a), p^{N,\pi,\delta}(\cdot|i,a)))\Bigg\} + \delta. \tag{3.21}$$

Since $P^{N,\pi,\delta} \in \mathcal{P}_\rho^N(\epsilon)$ for all $\pi \in \Pi_R$, we get, from the definition of $\mathcal{P}_\rho^N(\epsilon)$, that

$$\left(\sum_{a\in A} d^\rho(\hat{p}^N(\cdot|i,a), p^{N,\pi,\delta}(\cdot|i,a))\right)^{1/\rho} \leq \epsilon \ \forall \pi \in \Pi_R, i \in S. \tag{3.22}$$

Since d is nonnegative, (3.22) implies that $d(\hat{p}^N(\cdot|i,a), p^{N,\pi,\delta}(\cdot|i,a)) \leq \epsilon$, for all $\pi \in \Pi_R$, $i \in S$, and $a \in A$. Since $\psi(\cdot)$ is nondecreasing and $\pi(\cdot|i)$ is a pmf over A for all $i \in S$, we have $\max_{\pi\in\Pi_R, i\in S} \sum_{a\in A} \pi(a|i)\psi(d(\hat{p}^N(\cdot|i,a), p^{N,\pi,\delta}(\cdot|i,a))) \leq \psi(\epsilon)$. Utilizing this in (3.21) and noting that $\delta > 0$ was arbitrary establishes (3.19). Regarding the bound in (3.20), recall from the proof of Theorem 9 that $\left|J^* - J_{\mathbf{P}}(\hat{\pi}^N)\right| \leq 2 \max_{\pi\in\Pi_R} \left|\tilde{J}_\rho^N(\pi,\epsilon^N) - J_{\mathbf{P}}(\pi)\right|$. The bound in (3.20) then follows by repeating the argument in the proof of (3.19).

The claim of the theorem now follows by using a proof argument similar to the proof of Theorem 7 for the (s,a)-RMDP from the previous chapter. We present it here for completeness. Observe, from bounds (3.19) and (3.20) that the two inequalities in (3.18) hold if the event $\left[\max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a))) \leq \psi(\epsilon/m^{1/\rho})\right]$ occurs. The conclusion of the theorem then follows because

$$\begin{aligned}
&\mathbb{P}\left[\max_{i\in S, a\in A} \psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a))) \leq \psi(\epsilon/m^{1/\rho})\right] \\
&= \mathbb{P}\left[\psi(d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a))) \leq \psi(\epsilon/m^{1/\rho}), \quad \forall i \in S, a \in A\right] \\
&\geq \mathbb{P}\left[d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)) \leq \epsilon/m^{1/\rho}, \quad \forall i \in S, a \in A\right]
\end{aligned}$$

$$= \prod_{i \in S} \prod_{a \in A} \mathbb{P}\left[d(p(\cdot|i,a), \hat{p}^N(\cdot|i,a)) \leq \epsilon/m^{1/\rho} \right] \geq (1 - \beta(N, \epsilon/m^{1/\rho}))^{nm}.$$

Here, the first inequality holds because ψ is nondecreasing. The second equality holds because samples are independent across S and A. The final inequality follows from (3.17) because $|S| = n$ and $|A| = m$. □

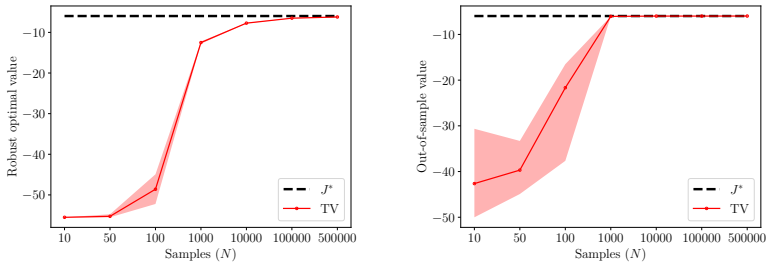
Existence of a $\beta(N, \epsilon) \in [0, 1]$ as in the statement of the above result is guaranteed because we allow for the edge-values of 0 and 1. More meaningful values of $\beta(N, \epsilon)$ would typically be reverse-engineered from a concentration inequality when one holds. We have included N, ϵ as arguments to emphasize that β would typically depend on these two quantities (among others that we suppress for brevity). We expect $\beta(N, \epsilon)$ to be nonincreasing in ϵ, because the larger the radius of the ambiguity ball centered at the empirical pmf, the larger the probability that the true pmf belongs to it. Per this intuition, we expect $\beta(N, \epsilon/m^{1/\rho})$ to be nonincreasing over $\rho \in [1, \infty)$. As such, we expect the probabilistic lower bound $(1 - \beta(N, \epsilon/m^{1/\rho}))^{nm}$ claimed in the above result to be nondecreasing in ρ. That is, we expect this lower bounding probability to be the smallest when $\rho = 1$. Moreover, the upper bounds in the inequalities in (3.18) are nondecreasing in ρ. That is, these upper bounds are the tightest when $\rho = 1$. In summary, roughly speaking, while the upper bounds in (3.18) deteriorate with increasing ρ, the corresponding probabilistic guarantee improves, and in this sense, it is not clear which ρ (if any) provides the best probabilistic rate of convergence.

3.5 Computational experiments

We present computational results on the machine repair MDP from Section 2.8.3. This MDP was introduced in [12] and studied by [26, 66, 67]. Recalling from Section 2.8.3, this MDP instance has $n = 10$ states indicating the operating condition of a machine. Every state has $m = 2$ available actions — do nothing or repair. The discount factor is $\alpha = 0.8$ and the initial state is chosen uniformly from $\{1, 2, \ldots, 10\}$. The details regarding the rewards and transition probabilities can be found in [67, Section 6]. The objective is to determine

a policy for running the machine so as to maximize the expected total discounted reward over an infinite horizon. We computationally demonstrate the asymptotic convergence of the robust and out-of-sample values of the data-driven s-RMDP as a function of the sample-size N. This serves as an empirical illustration of the asymptotic convergence result of Theorem 9. For our experiments, we used the TV distance and set $\rho = 1$. Note that the ambiguity set corresponding to $\rho = 1$ is the smallest set in the family $\{\mathcal{P}(\rho, \epsilon)\}_{\rho \in [1,\infty)}$, for any given distance and $\epsilon \in [0, \infty)$.

We solved the data-driven s-RMDP with sample-sizes $N \in \{10, 50, 100, 1000, 10000, 100000, 500000\}$. For each value of N, we computed the ambiguity set radius ϵ^N as described in Theorem 10 by setting $\gamma = 0.1$ corresponding to a confidence level of 90%. Recall that ambiguity set radii computed this way satisfy $\epsilon^N \to 0$ as $N \to \infty$ (see the discussion following Assumption 3). Next, for each value of N, we generated 100 independent instances of the RMDP from the true instance by constructing the empirical transition pmf based on N independent samples of the next state reached from each state-action pair. Finally, the RMDP was solved using robust value iteration [67] to obtain a robust optimal policy $\hat{\pi}^N$ and robust optimal value $\tilde{J}^N_\rho(\epsilon^N)$. Each iteration in the robust value iteration procedure requires the implementation of the robust Bellman operation, which was carried out using the algorithm described in [29]. Furthermore, in the machine repair MDP example, the state space of the system in the next time-stage depends on the its current state and action chosen. As a result, we restricted the set of feasible pmfs in the computation of the robust Bellman operator. Finally, the robust optimal policy $\hat{\pi}^N$ and the true transition probabilities $\mathbf{P}$ were substituted into Bellman's policy evaluation equations [50, Equation 6.1.9] to calculate the out-of-sample value, $J_{\mathbf{P}}(\hat{\pi}^N)$. The averages along with the 10^{th} and 90^{th} percentiles of the resulting robust optimal values $\tilde{J}^N_\rho(\epsilon^N)$ and out-of-sample values $J_{\mathbf{P}}(\hat{\pi}^N)$ versus N are plotted in Figure 3.2. This figure illustrates the results of Theorem 9 as both values can be observed to converge to the true optimal value J^* as $N \to \infty$.



(a) $\tilde{J}_\rho^N(\epsilon^N)$ vs N. (b) $J_{\mathbf{P}}(\hat{\pi}^N)$ vs N.

Figure 3.2: Asymtotic behavior of the robust and out-of-sample values as a function of sample-size N. The constant dashed line delineates the optimal value J^* of the true MDP.

3.6 Conclusions

We constructed a family of s-RMDPs, where the ambiguity sets were constructed by composing ℓ_ρ norms with distance functions, for $\rho \in [1, \infty]$. Distance-based (s, a)-RMDPs were recovered as a special case within this family, when $\rho = \infty$. This family enabled us to provide the first rigorous nuanced characterization of relative conservativeness between s- and (s, a)-RMDPs. Specifically, the following property holds despite the fact that an s-rectangular ambiguity set is contained in an (s, a)-rectangular one with the same radius: for any s-RMDP, there is an (s, a)-RMDP with robust optimal value that is at least as good, and vice versa. This occurs because the decision-maker can choose a smaller ambiguity radius for the (s, a)-RMDP than the s-RMDP. Consequently, relative conservativeness between s- and (s, a)-RMDPs cannot be characterized simply by looking at their robust optimal values. Further, we proved that the robust optimal value of any s-RMDP from our family equals the robust optimal value of some (s, a)-RMDP from the family. We also studied data-driven properties of s-RMDPs from this family. We demonstrated that these s-RMDPs enjoy three desirable properties: (i) robust and out-of-sample value convergence; (ii) a probabilistic rate of value convergence; and (iii) a guarantee that the out-of-sample value will be larger than the robust optimal value with a high probability. Finally, we discovered that (s, a)-RMDPs from our family exhibit a better theoretical out-of-sample probabilistic performance guar-

antee than s-RMDPs. In this sense, our (s, a)-RMDPs are actually less conservative than s-RMDPs, contrary to the general anecdotal belief in the literature.

Lemma 8 can be extended to an ordering of robust optimal value vectors (across different states) of s- and (s, a)-rectangular RMDPs, without taking an expectation with respect to the initial state pmf f. However, it might not be possible to extend Theorem 8 so as to match the components of such robust optimal value vectors. One possibility is to employ different ambiguity set radii for distinct state-action pairs in the (s, a)-RMDP. We defer this investigation to the future. Another potential direction for future research is to study whether or not better rates of convergence and out-of-sample performance guarantees can be derived by utilizing algebraic properties of specific distance functions. It would also be interesting to construct s-rectangular ambiguity sets using a broader class of functions $g : \mathbb{R}^m \to \mathbb{R}$ than the norms employed in this chapter for composition with distances. One can then attempt to extend our analyses here to the resulting more general family of s-RMDPs.

Chapter 4

ROBUST MARKOV DECISION PROCESSES ON GENERAL SPACES

4.1 Introduction

In this chapter, we study robust MDPs on uncountable state- and action-spaces. MDPs on uncountable state- and action-spaces also go by the name of stochastic control in the literature [9], and we follow that terminology throughout this chapter. An infinite horizon discrete-time stochastic control problem is described as follows [9]. A system is in state $x_t \in X$ at the beginning of stage $t \in \{1, 2, \ldots\}$. Upon observing the state of the system, an action $a_t \in A(x_t) \subseteq A$ is chosen. After choosing action $a_t \in A(x_t)$, two things happen: (i) the system stochastically transitions into a state $x_{t+1} \in X$ given by $x_{t+1} = F(x_t, a_t, \xi_t)$ where ξ_t is a Ξ-valued random variable, (ii) a cost of $c(x_t, a_t, \xi_t) \in \mathbb{R}$ is incurred. This process continues indefinitely and the future costs are discounted by $\alpha \in (0, 1)$. Here, $\{\xi_t\}$ is an i.i.d. sequence of Ξ-valued random variables with common distribution μ. Then, the goal in the discrete-time stochastic control problem is to minimize the expected total discounted cost over an infinite horizon.

A stationary, deterministic policy π is a measurable function from X to A such that $\pi(x) \in A(x)$ is the action prescribed in state $x \in X$ in every stage. The value of π when starting from state $x \in X$ equals

$$J(\pi, x) = \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} c(x_t, \pi(x_t), \xi_t) \Big| x_1 = x\right]. \tag{4.1}$$

Here, the expectation is with respect to the distribution induced by the policy π. Denote the set of all stationary, deterministic policies as Π. Then, the objective is to select a policy

$\pi \in \Pi$ to minimize $J(\pi, x)$ for all $x \in X$. That is, to choose a policy $\pi^* \in \Pi$ such that

$$J(\pi^*, x) \stackrel{\text{def}}{=} J^*(x) \stackrel{\text{def}}{=} \inf_{\pi \in \Pi} J(\pi, x) \ \forall x \in X. \tag{4.2}$$

In practice, the distribution μ of ξ may be unknown. As a result, the optimal value function in (4.2) cannot be determined. One way to circumvent this issue is via a robust approach. In this framework, the decision-maker assumes that the unknown distribution belongs to a certain ambiguity set and wishes to minimize the largest expected total discounted cost over all possible distributions from this ambiguity set.

Let $\{\widehat{\xi_1}, \ldots, \widehat{\xi_N}\} \subseteq \Xi$ be N i.i.d. samples of ξ.[1] Using these samples, we construct the empirical distribution of ξ, which is given as

$$\hat{\mu}^N \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \delta_{\widehat{\xi_i}}.$$

Here, δ_z is the Dirac measure concentrated at $z \in \Xi$. Using a nonnegative distance function (need not be a metric) $d : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to [0, \infty]$ and an $0 \leq \epsilon < \infty$, we construct the ambiguity set as

$$\mathcal{P}^N(\epsilon) \stackrel{\text{def}}{=} \{\nu \in \mathcal{M}(\Xi) : d(\nu, \hat{\mu}^N) \leq \epsilon\}. \tag{4.3}$$

Here, $\mathcal{M}(\Xi)$ is the set of all probability distributions over Ξ. Then, the robust stochastic control problem can be modeled as a two-player Markov (or stochastic) game with complete information between the decision-maker and a hypothetical adversary [25]. In order to describe the game, we first define a few concepts. Let $h_t = (x_1, a_1, \nu_1, \ldots, x_{t-1}, a_{t-1}, \nu_{t-1}, x_t)$ denote a history sequence up to stage t, where $x_t \in X$ and for all $1 \leq k \leq t-1$, $x_k \in X, a_k \in A(x_k), \nu_k \in \mathcal{P}^N(\epsilon)$. The set of all history sequences upto stage t is denoted as H_t. A strategy for the decision-maker $\pi = (\pi_1, \pi_2, \ldots)$ is a sequence of stochastic kernels where π_t is

[1] The implicit asumption here is that the random noise is observable. In many applications of interest, we can either directly observe the random noise or obtain it from the system equation $x_{t+1} = F(x_t, a_t, \xi_t)$ given the values of x_t, a_t, x_{t+1}.

a stochastic kernel from H_t to A, i.e., $\pi_t(A(x_t)|h_t) = 1$ for all $h_t \in H_t, t \in \{1, 2, \ldots, \}$. Denote by $h_t^{\#} = (h_t, a_t)$ an extended history sequence and let $H_t^{\#}$ be the set of extended history sequences up to stage t. A strategy $\gamma = (\gamma_1, \gamma_2, \ldots)$ for the adversary is a tuple of stochastic kernels where γ_t is a stochastic kernel from $H_t^{\#}$ to $\mathcal{P}^N(\epsilon)$, i.e., $\gamma_t(\mathcal{P}^N(\epsilon)|h_t^{\#}) = 1$ for all $h_t^{\#} \in H_t^{\#}, t \in \{1, 2, \ldots, \}$. The set of all strategies for the decision maker and the adversary are denoted by $\bar{\Pi}$ and Γ^N, respectively. The description of the game is then as follows. Let $\pi = (\pi_1, \pi_2, \ldots)$ be any strategy selected by the decision-maker while $\gamma = (\gamma_1, \gamma_2, \ldots)$ be any strategy employed by the adversary. Upon observing the history sequence h_t in stage t, the decision-maker selects an action a_t which is consistent with the strategy π. The adversary observes $h_t^{\#}$ and selects a distribution $\nu_t^N \in \mathcal{P}^N(\epsilon)$ consistent with the strategy γ. The decision-maker then incurs a cost of $c(x_t, a_t, \xi_t)$ where $\xi_t \in \Xi$, and the system transitions to state $x_{t+1} \in X$ with probability $Q_t^N(\cdot|x_t, a_t, \nu_t^N) \stackrel{\text{def}}{=} \nu_t^N \circ F^{-1}(x_t, a_t, \cdot)$. That is, for any measurable $D \subseteq X$, $Q_t^N(D|x_t, a_t, \nu_t^N) = \nu_t^N(\{z \in \Xi : F(x_t, a_t, z) \in D\})$. This process continues indefinitely. Assuming that the system starts in state $x \in X$, the expected total discounted cost incurred by the decision-maker is then

$$J(\pi, \gamma, x; \epsilon) \stackrel{\text{def}}{=} \mathbb{E}^{\pi,\gamma}\left[\sum_{t=1}^{\infty} \alpha^{t-1} c(x_t, \pi_t(s_t), \xi_t) \middle| x_1 = x\right],$$

where the expectation is taken with respect to the distribution induced by π and γ. The decision-maker's goal is to select a strategy $\pi \in \bar{\Pi}$ to minimize $\sup_{\gamma \in \Gamma^N} J(\pi, \gamma, x; \epsilon)$ for all $x \in X$. This can be formulated as the minmax optimization problem

$$\tilde{J}^{N,*}(x; \epsilon) \stackrel{\text{def}}{=} \inf_{\pi \in \bar{\Pi}} \tilde{J}^N(\pi, x; \epsilon) \ \forall x \in X, \text{ where } \tilde{J}^N(\pi, x; \epsilon) \stackrel{\text{def}}{=} \sup_{\gamma \in \Gamma^N} J(\pi, \gamma, x; \epsilon) \ \forall x \in X. \quad (4.4)$$

We refer to the function $\tilde{J}^{N,*}(\cdot; \epsilon)$ as the robust optimal value function and the strategy $\hat{\pi}^N \in \bar{\Pi}$ satisfying $\tilde{J}^{N,*}(x; \epsilon) = \min_{\pi \in \bar{\Pi}} \tilde{J}^N(\pi, x; \epsilon) \ \forall x \in X$ (assuming the minimum exists) as the robust optimal policy. The stochastic game framework described here is different from the modeling approach used for robust MDPs with finite state- and action-spaces [32, 67],

where the feasible decision for the adversary corresponds to picking a distribution from the ambiguity set. The stochastic game framework is needed to address the measure-theoretic complications which arise due to the uncountable nature of the state- and action-spaces.

To avoid trivialities, throughout this chapter, we assume that the ambiguity set, $\mathcal{P}^N(\epsilon)$, is nonempty with probability 1 (wp 1), for all sample-sizes N, and $0 \leq \epsilon < \infty$. A sufficient condition for the nonemptiness assumption is that $d(q,q) = 0$, for every $q \in \mathcal{M}(X)$. This guarantees that the center of the ambiguity ball belongs to the ambiguity ball. This assumption is satsified by many well known distances studied in the literature (see Section 4.4).

4.1.1 Notation and Preliminaries

Given a metric space X with metric ρ, let $C_b(X)$ be the set of real-valued continuous bounded functions over X. For any function $f \in C_b(X)$, define the supremum norm and the Lipschitz seminorm as $\|f\|_\infty \stackrel{\text{def}}{=} \sup_{x\in X} |f(x)|$ and $\|f\|_L \stackrel{\text{def}}{=} \sup_{x\neq y} |f(x) - f(y)| / \rho(x,y)$, respectively. Let $\|f\|_{BL} \stackrel{\text{def}}{=} \|f\|_L + \|f\|_\infty$. A sequence μ^k in $\mathcal{M}(X)$ is said to weakly converge to $\mu \in \mathcal{M}(X)$ if $\int_X f(z)\mu^k(dz) \to \int_X f(z)\mu(dz)$ for all $f \in C_b(X)$. Define $\beta : \mathcal{M}(X) \times \mathcal{M}(X) \to \mathbb{R}_+$ as

$$\beta(\mu,\nu) \stackrel{\text{def}}{=} \sup_{\substack{f:X\to\mathbb{R} \\ \|f\|_{BL}\leq 1}} \left| \int_X f(x)\mu(dx) - \int_X f(x)\nu(dx) \right|. \tag{4.5}$$

If X is a separable metric space, then β defined above is a metric on $\mathcal{M}(X)$, which metrizes the topology of weak convergence [16, Proposition 11.3.2, Theorem 11.3.3].

4.2 Related literature

The authors in [25] developed a generic framework to model and analyse two player stochastic games. The authors also illustrate their framework by applying it to model and analyse the robust stochastic control problem. Though we will be relying upon this framework, the work in this chapter differs from [25] in the construction of the ambiguity set. Also, [25] doesn't contain any data-driven analyses since their ambiguity set was not data-driven. The approach developed by [25] was used by [70] to study the data-driven robust stochastic control problem

where the ambiguity set is constructed using Wasserstein distance. The paper in [70] provides a probabilistic guarantee on the out-of-sample performance but no results on the asymptotic properties. Robust stochastic control with ambiguity sets contructed using Kullback-Leibler divergence was studied in [48]. Robust stochastic control problems using ambiguity sets based on TV distance was studied in [62]. This was extended to average cost problems in [63]. But, none of these papers provide any data-driven analyses. Robust stochastic control with application to inventory systems was studied in [57, 58]. However, no analysis on data-driven results was presented. Recently, the authors in [5] consider a finite horizon robust stochastic control problem, where the model assumptions slightly differ from the postulates formulated in [25]. They derive a robust Bellman equation and characterize the structural properties of the robust stochastic control problem under their set of assumptions. Again, no data-driven analysis was done in their work.

Convergence properties of stochastic control problems in the nonrobust setup is well studied. In those settings, typically, the true distribution of the random noise is replaced with an approximate distribution. The convergence behavior of the approximate stochastic control problem is then analysed under various model assumptions. Refer to the literature mentioned in [35] for an overview of the work in this direction. The paper in [35] undertakes a comprehensive study of the convergence behavior of the approximate stochastic control problem under various model assumptions. They also illustrate their convergence results in the data-driven context, wherein the approximate distribution of the random noise equals the empirical distribution constructed from i.i.d. samples. We remark that our work is different from [35] even though we also have similar convergence results. The main difference lies in the modeling framework, wherein [35] does not have a minmax formulation since they do not conisder a robust model setup. In this work, the robustness which manifests itself through the minmax formulation requires us to perform a more careful analysis in order to derive the convergence results. Another significant difference between our work and [35] is that we consider an explicit data-driven approach, thereby allowing us to obtain nonasymptotic results, like, the probabilistic performance guarantees under finite sample-sizes. Such finite-

sample results were not present in the work of [35].

4.3 Value function convergence and probabilistic performance guarantee

We first impose the following assumptions which ensure the existence of measurable selections and Bellman equations for the robust (and nonrobust) stochastic control problem. These assumptions were introduced in [25, Assumption 3.1], and we slightly modify them here in order to suit our data-driven context. In the rest of this chapter, we define $K \subseteq X \times A$ as $K \stackrel{\text{def}}{=} \{(x, a) | x \in X, a \in A(x)\}$.

Assumption 5. The following assumptions hold for the stochastic control problem.

(a) X, A, Ξ are measurable subsets of $\mathbb{R}^{n_1}, \mathbb{R}^{n_2}, \mathbb{R}^{n_3}$, respectively, where n_1, n_2, n_3 are positive integers.

(b) The set-valued mapping $A(x)$ from X to A is continuous (see Definition 2 for continuity of set-valued mapping).

(c) The disturbance space Ξ is compact.

(d) The system evolution function $F : K \times \Xi \to X$ is continuous over $K \times \Xi$.

(e) The single-stage cost function $c : K \times \Xi \to \mathbb{R}$ is continuous over $K \times \Xi$. Also, $B \stackrel{\text{def}}{=} \sup_{(x,a)\in K, \xi \in \Xi} |c(x, a, \xi)| < \infty$.

Since c is bounded, without loss of generality, we assume that it is nonnegative. Hence, throughout this chapter, c is nonnegative, continuous, and bounded over $K \times \Xi$. The continuity of $x \mapsto A(x)$ is a standard assumption and satisfied in many applications. The compactness of Ξ helps in establishing the compactness of the ambiguity set $\mathcal{P}^N(\epsilon)$, as long as the distance function d obeys certain continuity properties (Assumption 6). This compactness of the ambiguity set guarantees the existence of robust Bellman equations and robust optimal policies that are measurable [25, Assumption 3.1]. However, the ambiguity set can

be compact even without requiring Ξ to be compact, like, the Wasserstein ball [71, Theorem 1]. Apart from this technical consideration, it is not uncommon to find applications where the disturbance space is compact (see Section 4.5). The continuity of the function F will be used to establish the weak continuiuty of the transition kernel (see Lemma 18). Again, this weak continuity of transition kernel is a standard assumption [25, Assunption 3.1 c] which ensures the existence of measurable selections. Regarding the assumption on the single-stage cost function, refer to Remark 1 for a discussion.

Assumption 6. The distance function $d : \mathcal{M}(\Xi) \times \mathcal{M}(\Xi) \to \{\mathbb{R}_+ \cup \infty\}$ is such that $d(p, q)$ is weakly lower semicontinuous over p for every $q \in \mathcal{M}(\Xi)$, i.e., if p^N is a sequence in $\mathcal{M}(\Xi)$ which converges weakly to $p \in \mathcal{M}(\Xi)$, then $d(p, q) \leq \liminf_{N\to\infty} d(p^N, q)$ for all $q \in \mathcal{M}(X)$.

Lemma 16. Fix an integer $N \geq 1$ and $\epsilon \in [0, \infty)$. If Assumption 5 and Assumption 6 are satisfied, then $\mathcal{P}^N(\epsilon)$ is weakly compact.

Proof. Since Ξ is compact by Assumption 5, $\mathcal{M}(\Xi)$ is compact [9, Proposition 7.22]. The weak lower semicontinuity of d (by Assumption 6) ensures that $\mathcal{P}^N(\epsilon) = \{\nu \in \mathcal{M}(\Xi) : d(\nu, \hat{\mu}^N) \leq \epsilon\}$ is weakly closed. Hence, $\mathcal{P}^N(\epsilon)$ is weakly compact. □

The requirement that the distance function d is lower semicontinuous is mild and satisfied by many well known distance functions as listed in Section 4.4.

Before presenting the robust Bellman equation, we need to establish the weak continuity of the transition kernel, which is a sufficient condition for the existence of robust Bellman equation. For any fixed integer $N \geq 1$ and $\epsilon \in [0, \infty)$, the transition kernel is defined as $Q^N(\cdot|x, a, \nu^N) \stackrel{\text{def}}{=} \nu^N \circ F^{-1}(x, a, \cdot)$ for any $x \in A, a \in A(x)$, and $\nu^N \in \mathcal{P}^N(\epsilon)$. The transition kernel Q^N is said to be weakly continuous if for any continuous bounded function $u : X \to \mathbb{R}$, the function $(x, a, \nu) \mapsto \int_X u(z)Q^N(dz|x, a, \nu)$ is continuous over $K \times \mathcal{P}^N(\epsilon)$. In order to establish the weak continuity of the transition kernel, we need the following technical result, which will be heavily used in this work.

Lemma 17. [40, Theorem 3.5] [56, Theorem 3.5] Let S be a separable complete metric space. Let μ^k be a sequence in $\mathcal{M}(S)$ such that $\mu^k \to \mu \in \mathcal{M}(S)$ weakly. Let u^k be a sequence of real-valued functions defined on S and u be another real-valued function defined on S. If $\sup_{k\in\mathbb{N}} \left\|u^k\right\|_\infty < \infty$ and $u^k(x^k) \to u(x)$ for all $x^k \to x$, then $\int_S u^k(x)\mu^k(dx) \to \int_S u(x)\mu(dx)$.

Lemma 18. Fix an integer $N \geq 1$ and $\epsilon \in [0, \infty)$. If Assumption 5 holds, then, the transition kernel Q^N is weakly continuous.

Proof. A proof of this argument under a set of assumptions which is slightly different from Assumption 5 is established in [25, Proposition 6.2]. Our proof also follows a similar approach. We present it here for the sake of completeness.

Fix any sample-size $N \geq 1$. Let $u : X \to \mathbb{R}$ be any continuous bounded function. From the definition of Q^N we have

$$\int_X u(z)Q^N(dz|x,a,\nu) = \int_\Xi u(F(x,a,z))\nu(dz) \ \forall x \in X, a \in A(x), \nu \in \mathcal{P}^N(\epsilon).$$

Consider any sequence $(x^k, a^k, \nu^k) \to (x, a, \nu)$. From Assumption 5, F is continuous over $K \times \Xi$. Since u is continuous, $\lim_{k\to\infty} u(F(x^k, a^k, z^k)) = u(F(x, a, z))$ for all $z^k \to z$. The boundedness of u implies that $\sup_{k\in\mathbb{N}, z\in\Xi} \left|u(F(x^k, a^k, z))\right| < \infty$. Hence,

$$\begin{aligned}
\lim_{k\to\infty} \int_X u(z)Q^N(dz|x^k,a^k,\nu^k) &= \lim_{k\to\infty} \int_\Xi u(F(x^k,a^k,z))\nu^k(dz) \\
&= \int_\Xi u(F(x,a,z))\nu(dz) \\
&= \int_X u(z)Q^N(dz|x,a,\nu),
\end{aligned}$$

where the equality in the second line follows from Lemma 17. □

Theorem 12 (Robust Bellman equation and robust optimal policies — Theorem 4.1 and 4.2 in [25])**.** Fix any integer $N \geq 1$ and $\epsilon \in [0, \infty)$. For $v \in C_b(X)$, define the robust Bellman

operator $\tilde{\Phi}^N : C_b(X) \to C_b(X)$ as

$$(\tilde{\Phi}^N v)(x) \stackrel{\text{def}}{=} \min_{a \in A(x)} \sup_{\nu^N \in \mathcal{P}^N(\epsilon)} \left[\int_{\Xi} \left(c(x,a,z) + \alpha v(F(x,a,z)) \right) \nu^N(dz) \right] \ \forall x \in X. \tag{4.6}$$

For any $t \in \{1, 2, 3, \ldots\}$, define $\tilde{\Phi}_t^N : C_b(X) \to C_b(X)$ as $\tilde{\Phi}_t^N v \stackrel{\text{def}}{=} \tilde{\Phi}^N \tilde{\Phi}_{t-1}^N v$ where $\tilde{\Phi}_1^N v \stackrel{\text{def}}{=} \tilde{\Phi}^N v$.

If Assumption 5 and Assumption 6 hold, then

(a) $\tilde{\Phi}^N$ is a contraction mapping with respect to $\|\cdot\|_\infty$-norm and the modulus of contraction is α.

(b) There exists a $v^N \in C_b(X)$ and a deterministic stationary policy π^N such that

  (i) v^N is the unique solution of the fixed point equation $\tilde{\Phi}^N v = v$.

  (ii) v^N and π^N are the robust optimal value function and the robust optimal policy, respectively, i.e.,

$$v^N(x) = \tilde{J}^{N,*}(x;\epsilon) = \inf_{\pi \in \bar{\Pi}} \sup_{\gamma \in \Gamma^N} J(\pi, \gamma, x; \epsilon) = \sup_{\gamma \in \Gamma^N} J(\pi^N, \gamma, x; \epsilon) \ \forall x \in X.$$

(c) If $v \equiv \mathbf{0}$ (the identically zero function), then $\sup_{x \in X} \left| (\tilde{\Phi}_t^N v)(x) - \tilde{J}^{N,*}(x) \right| \leq \frac{B\alpha^t}{1-\alpha}$ $\forall t \in \mathbb{N}$. In other words, $\tilde{\Phi}_t^N \mathbf{0}$ converges to $\tilde{J}^{N,*}$ uniformly over X, and the convergence rate depends only on B and α.

Proof. The proof follows from [25, Theorem 4.1, Theorem 4.2] as long as the assumptions listed in [25, Assunmption 3.1] are satisfied. In our context, those assumptions are equivalent to Assumption 5 along with Lemma 16 and Lemma 18. Since Lemma 18 follows from Assumption 5, and Lemma 16 follows from Assumption 5 and Assumption 6, the claim of the theorem follows as long as Assumption 5 and Assumption 6 hold. □

Remark 1. The standard approach to deal with unbounded cost functions is to assume the existence of a continuous bounding function $w : X \to [1, \infty)$ such that $|c(x, a, \xi)| \leq Mw(x)$ for all $(x, a) \in K, \xi \in \Xi$. Apart from being continuous, w also needs to satisfy other regularity properties to ensure the existence of measurable selections and the Bellman equations [25, Assumption 3.1 (d), (e)]. Translated in our context, those assumptions state that for any sample-size $N \geq 1$ and $\epsilon \in [0, \infty)$, (i) $(x, a, \nu) \mapsto \int_\Xi w(F(x, a, z))\nu(dz)$ is continuous over $K \times \mathcal{P}^N(\epsilon)$, (ii) there exists a constant $\beta > 0$ such that $\int_\Xi w(F(x, a, z))\nu(dz) \leq \beta w(x)$ for all $(x, a) \in K, \nu \in \mathcal{P}^N(\epsilon)$. These conditions are hard to verify and it is unknown if they even hold. One approach to prove the above two claims is to let $\mathcal{P}^N(\epsilon)$ be a family of density functions (w.r.t. Lebesgue measure) that are continuous and then exploit the specific properties of $\mathcal{P}^N(\epsilon)$ to construct a bounding function satisfying the regularity conditions. See [25, Example 7.2] for more details.

Before proceeding further, we present a recursive characterization of the optimal value function, J^*, defined in (4.2). Theorem 12 can be viewed as a robust counterpart of the following result.

Theorem 13. [28, Chapter 4] For $v \in C_b(X)$, define the Bellman operator $\Phi : C_b(X) \to C_b(X)$ as

$$(\Phi v)(x) \stackrel{\text{def}}{=} \min_{a \in A(x)} \left[\int_\Xi \left(c(x, a, z) + \alpha v(F(x, a, z)) \right) \mu(dz) \right] \quad \forall x \in X. \tag{4.7}$$

For any $t \in \{1, 2, 3, \ldots\}$, define $\Phi_t : C_b(X) \to C_b(X)$ as $\Phi_t v \stackrel{\text{def}}{=} \Phi\Phi_{t-1}v$ where $\Phi_1 v \stackrel{\text{def}}{=} \Phi v$.

If Assumption 5 holds, then

(a) Φ is a contraction mapping with respect to $\|\cdot\|_\infty$-norm and the modulus of contraction is α.

(b) There exists a $v \in C_b(X)$ and a deterministic stationary policy π such that

   (i) v is the unique solution of the fixed point equation $\Phi v = v$.

(ii) v and π are the optimal value function and the optimal policy, respectively, i.e.,

$$v(x) = J^*(x) = \inf_{\pi \in \Pi} J(\pi, x) \forall x \in X.$$

(c) If $v \equiv \mathbf{0}$ (the identically zero function), then $\sup_{x \in X} |(\Phi_t v)(x) - J^*(x)| \leq \frac{B\alpha^t}{1-\alpha}$ $\forall t \in \mathbb{N}$. In other words, $\Phi_t \mathbf{0}$ converges to J^* uniformly over X, and the convergence rate depends only on B and α.

We need the following technical results which will be used later in order to establish the asymptotic convergence.

Lemma 19. If Assumption 5 holds, then for every $v \in C_b(X)$, the function $\Phi_t v$ is continuous and bounded over X for all $t \in \mathbb{N}$, i.e., $(\Phi_t v)(x^k) \to (\Phi_t v)(x)$ for all $x^k \to x$ and $\|\Phi_t v\|_\infty < \infty$.

Proof. We prove this using induction on $t \in \mathbb{N}$. Consider the base case which corresponds to $t = 1$. From Assumption 5(d),(e), the functions c and F are continuous over $K \times \Xi$. Also, c is bounded over $K \times \Xi$. Hence, $(x, a) \mapsto \int_\Xi (c(x, a, z) + \alpha v(F(x, a, z))) \, \mu(dz)$ is continuous over K by the dominated convergence theorem. Since $x \mapsto A(x)$ is continuous, the continuity of $x \mapsto (\Phi v)(x)$ follows by Berge's maximum theorem. Finally, Φv is bounded since c and v are bounded. Now consider $t \geq 1$ and assume that the result is true by the induction hypothesis for all $t' \leq t-1$. Since $\Phi_t v = \Phi \Phi_{t-1} v$ and $\Phi_{t-1} v$ is continuous and bounded by the indcution hypothesis, $(x, a) \mapsto \int_\Xi (c(x, a, z) + \alpha \Phi_{t-1} v((F(x, a, z)))) \, \mu(dz)$ is continuous over K by the dominated convergence theorem. The continuity follows from Berge's maximum theorem by noting that $x \mapsto A(x)$ is continuous. The boundedness follows by observing that c and $\Phi_{t-1} v$ are bounded. □

Lemma 20. If Assumption 5 holds, then J^* is continuous over X and $\|J^*\|_\infty \leq \frac{B}{1-\alpha}$.

Proof. From Theorem 13(c), the sequence of functions $\{\Phi_t \mathbf{0}\}$ converge to J^* uniformly over X. Since $\mathbf{0} \in C_b(X)$, from Lemma 19, $\Phi_t \mathbf{0}$ is continuous over X for each $t \in \mathbb{N}$. Hence, J^* is continuous over X. The boundedness on J^* follows from Assumption 5(e). □

The next lemma is a well known folk result. We present the proof for completeness.

Lemma 21. For any stationary, deterministic policy $\pi \in \Pi$, define the Bellman evaluation operator $\Phi_\pi : C_b(X) \times C_b(X)$ as

$$(\Phi_\pi v)(x) \stackrel{\text{def}}{=} \int_\Xi \left(c(x, \pi(x), z) + \alpha v(F(x, \pi(x), z))\right) \mu(dz) \ \forall x \in X. \tag{4.8}$$

For any $t \in \{1, 2, 3, \ldots\}$, define $\Phi_{t,\pi} : C_b(X) \to C_b(X)$ as $\Phi_{t,\pi} v \stackrel{\text{def}}{=} \Phi_\pi \Phi_{t-1,\pi} v$ where $\Phi_{1,\pi} v \stackrel{\text{def}}{=} \Phi_\pi v$. If Assumption 5 holds, then, for any $v \in C_b(X)$, we have, $\sup_{x \in X, \pi \in \Pi} |(\Phi_{t,\pi} v)(x) - J(\pi, x)| \leq \alpha^t \left(\frac{B}{1-\alpha} + \|v\|_\infty\right) \ \forall t \in \mathbb{N}$.

Proof. Fix any $\pi \in \Pi$. Let $x_1 \stackrel{\text{def}}{=} x$. Recursively expanding $\Phi_{t,\pi} v$, we have for any $t \in \mathbb{N}$,

$$\begin{aligned}
\Phi_{t,\pi} v(x) = \underbrace{\int_\Xi \int_\Xi \ldots \int_\Xi}_{t \text{ times}} \Bigg[\sum_{k=1}^{t} \alpha^{k-1} c(x_k, \pi(x_k), z_k) + \\
\alpha^t v(F(x_t, \pi(x_t)), z_t) \Bigg] \mu(dz_1) \mu(dz_2) \ldots \mu(dz_t) \\
\stackrel{(a)}{=} \mathbb{E}\left[\sum_{k=1}^{t} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) + \alpha^t v(F(x_t, \pi(x_t)), \xi_t) \Big| x_1 = x \right] \\
= \mathbb{E}\left[\sum_{k=1}^{\infty} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) + \alpha^t v(F(x_t, \pi(x_t)), \xi_t) \Big| x_1 = x \right] - \\
\mathbb{E}\left[\sum_{k=t+1}^{\infty} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) \Big| x_1 = x \right] \\
= \mathbb{E}\left[\sum_{k=1}^{\infty} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) \Big| x_1 = x \right] + \mathbb{E}\left[\alpha^t v(F(x_t, \pi(x_t)), \xi_t) \Big| x_1 = x \right] - \\
\mathbb{E}\left[\sum_{k=t+1}^{\infty} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) \Big| x_1 = x \right].
\end{aligned}$$

Here, the expectation in (a) and all the subsequent steps is with respect to the distribution induced by the policy π. The random states $x_2, x_3, \ldots$, are given as $x_k = F(x_{k-1}, \pi(x_{k-1}), \xi_{k-1})$

for all $k = 2, 3, \ldots$. Hence, from the definition of $J(\pi, x)$ in (4.1), we have for any $t \in \mathbb{N}$,

$$\begin{aligned}
|\Phi_{t,\pi} v(x) - J(\pi, x)| &= \Bigg| \mathbb{E}\left[\alpha^t v(F(x_t, \pi(x_t)), \xi_t) \middle| x_1 = x\right] - \\
&\qquad \mathbb{E}\left[\sum_{k=t+1}^{\infty} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) \middle| x_1 = x\right] \Bigg| \\
&\leq \left| \mathbb{E}\left[\alpha^t v(F(x_t, \pi(x_t)), \xi_t) \middle| x_1 = x\right] \right| + \\
&\qquad \left| \mathbb{E}\left[\sum_{k=t+1}^{\infty} \alpha^{k-1} c(x_k, \pi(x_k), \xi_k) \middle| x_1 = x\right] \right| \\
&\leq \alpha^t \left\| v \right\|_\infty + \frac{\alpha^t B}{1-\alpha}.
\end{aligned}$$

Since the above inequality holds for all $\pi \in \Pi$ and $x \in X$, the claim of the lemma follows. □

4.3.1 Value convergence to true optimal value

In this section, we study the convergence of the robust optimal value function defined in (4.4) as the sample-size N diverges to infinity. We are interested in checking if $\tilde{J}^{N,*}(x; \epsilon) \xrightarrow{N,\text{wp1}} J^*(x)\ \forall x \in X$. However, such a result would not be true if the radius of the ambiguity ball ϵ did not depend on N. Apart from the dependence of the radius on the sample-size, we also impose the following assumption on the distance function d used to construct the ambiguity set.

Assumption 7. There exists a continuous function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\psi(0) = 0$ such that $\beta(\mu, \nu) \leq \psi(d(\mu, \nu))$, for all $\mu, \nu \in \mathcal{M}(\Xi)$. Here, $\beta(\mu, \nu)$ is the metric on $\mathcal{M}(\Xi)$ as defined in (4.5).

Lemma 22. Suppose the radii of ambiguity sets are dependent on N such that $\lim_{N\to\infty} \epsilon^N = 0$. Let $\nu^N \in \mathcal{P}^N(\epsilon^N)$ be any sequence. If Assumption 7 holds, then, $\nu^N \xrightarrow{N,\text{wp1}} \mu$ weakly, where μ is the true distribution of ξ.

Proof. The β-metric on $\mathcal{M}(\Xi)$ defined in (4.5) metrizes the topology of weak convergence

([16, Theorem 11.3.3]). Hence, it is enough to show that $\beta(\nu^N, \mu) \xrightarrow{N,\text{wp1}} 0$. Since β is a metric,

$$\beta(\nu^N, \mu) \leq \beta(\nu^N, \hat{\mu}^N) + \beta(\hat{\mu}^N, \mu).$$

We show that each term above converges to 0, wp1, as $N \to \infty$. From ([16, Theorem 11.4.1]), the second term converges to 0, wp1, as $N \to \infty$. Since $\nu^N \in \mathcal{P}^N(\epsilon^N)$, we have that $d(\nu^N, \hat{\mu}^N) \leq \epsilon^N$. Further, $\lim_{N\to\infty} \epsilon^N = 0$ implies that $\lim_{N\to\infty} d(\nu^N, \hat{\mu}^N) = 0$. Hence, $\lim_{N\to\infty} \beta(\nu^N, \hat{\mu}^N) \leq \lim_{N\to\infty} \psi(d(\nu^N, \hat{\mu}^N)) = \psi(\lim_{N\to\infty} d(\nu^N, \hat{\mu}^N)) = \psi(0) = 0$. Here, the inequality follows from Assumption 7, the first equality follows by the continuity of ψ defined in Assumption 7, and the last equality by Assumption 7. □

Assumption 7 links the convergence of probability distributions with respect to the distane function, to the convergence with respect to the β-metric. Assumption 7 is satisfied by many well known distance functions studied in the literature. We provide a list of such distance functions in Section 4.4.

Next, we state and prove the asymptotic convergence of the robust value function to the true value function.

Theorem 14. Suppose the radii of ambiguity sets are dependent on N such that $\lim_{N\to\infty} \epsilon^N = 0$. If Assumption 5, Assumption 6, and Assumption 7 hold, then, the event $\left\{\tilde{J}^{N,*}(x;\epsilon^N) \xrightarrow{N} J^*(x)\ \forall x \in X\right\}$ holds with probability 1.

Proof. We use successive approximation to establish this claim. In fact, we prove a stronger claim $\left\{\tilde{J}^{N,*}(x^N;\epsilon^N) \xrightarrow{N} J^*(x)\ \forall x^N \to x\right\}$ holds with probability 1. The technique of successive approximation was also used in [35, Theorem 4.2] to prove the convergence in the non-robust setup. Further, in their work, they assumed that the action spaces $A(x) = A\ \forall x \in X$, while in our case $A(x)$ is nonconstant. As a result, our proof requires delicate arguments.

Let x^N be any sequence in X such that $x^N \to x \in X$. For any fixed $t \in \mathbb{N}$, we have for all $N \geq 1$,

$$\left|\tilde{J}^{N,*}(x^N;\epsilon^N) - J^*(x)\right| \leq \left|\tilde{J}^{N,*}(x^N;\epsilon^N) - (\tilde{\Phi}_t^N \mathbf{0})(x^N)\right| + \left|(\tilde{\Phi}_t^N \mathbf{0})(x^N) - (\Phi_t \mathbf{0})(x)\right| +$$

$$
\begin{aligned}
&|(\Phi_t\mathbf{0})(x) - J^*(x)| \\
\leq &\sup_{x\in X}\left|\tilde{J}^{N,*}(x;\epsilon^N) - (\tilde{\Phi}_t^N\mathbf{0})(x)\right| + \left|(\tilde{\Phi}_t^N\mathbf{0})(x^N) - (\Phi_t\mathbf{0})(x)\right| + \\
&\sup_{x\in X}|(\Phi_t\mathbf{0})(x) - J^*(x)| \\
\leq &\frac{B\alpha^t}{1-\alpha} + \left|(\tilde{\Phi}_t^N\mathbf{0})(x^N) - (\Phi_t\mathbf{0})(x)\right| + \frac{B\alpha^t}{1-\alpha} \\
= &\frac{2B\alpha^t}{1-\alpha} + \left|(\tilde{\Phi}_t^N\mathbf{0})(x^N) - (\Phi_t\mathbf{0})(x)\right|,
\end{aligned}
$$

where the inequality in the third line follows from Theorem 12(c) and Theorem 13(c), respectively. Note that B is finite since the single-stage cost function c is bounded by Assumption 5(e). Since $\alpha \in (0,1)$, there exists a $t \in \mathbb{N}$, such that, $\frac{2B\alpha^t}{1-\alpha}$ can be made arbitrarily small. Hence, $\tilde{J}^{N,*}(x^N;\epsilon^N) \to J^*(x)$ if $(\tilde{\Phi}_t^N\mathbf{0})(x^N) \to (\Phi_t\mathbf{0})(x)$ for all $t \in \mathbb{N}$.

Using induction, we prove that for each $t \in \mathbb{N}$,

1. $(\tilde{\Phi}_t^N\mathbf{0})(x^N) \xrightarrow{N,\text{wp1}} (\Phi_t\mathbf{0})(x)$.
2. $\sup_{N\in\mathbb{N}}\left\|\tilde{\Phi}_t^N\mathbf{0}\right\|_\infty < \infty$.

Consider the base case of the induction which corresponds to $t = 1$. From the definition of $\tilde{\Phi}^N$ in (4.6) and the boundedness of c by Assumption 5(e), the uniform boundedness of $\{\tilde{\Phi}^N\mathbf{0}\}$ is immediate. Next, for any $N \geq 1$, we have,

$$
\begin{aligned}
\left|(\tilde{\Phi}^N\mathbf{0})(x^N) - (\Phi\mathbf{0})(x)\right| &\leq \left|(\tilde{\Phi}^N\mathbf{0})(x^N) - (\Phi\mathbf{0})(x^N)\right| + \left|(\Phi\mathbf{0})(x^N) - (\Phi\mathbf{0})(x)\right| \\
&= \left|\min_{a\in A(x^N)}\sup_{\nu^N\in\mathcal{P}^N(\epsilon^N)}\left[\int_\Xi c(x^N,a,z)\nu^N(dz)\right] - \right. \\
&\qquad \left.\min_{a\in A(x^N)}\left[\int_\Xi c(x^N,a,z)\mu(dz)\right]\right| + \\
&\qquad \left|(\Phi\mathbf{0})(x^N) - (\Phi\mathbf{0})(x)\right| \\
&\leq \underbrace{\sup_{a\in A(x^N)}\left|\sup_{\nu^N\in\mathcal{P}^N(\epsilon^N)}\left[\int_\Xi c(x^N,a,z)\nu^N(dz)\right] - \int_\Xi c(x^N,a,z)\mu(dz)\right|}_{\text{Term 1}} +
\end{aligned}
$$

$$\underbrace{\left|(\Phi\mathbf{0})(x^N) - (\Phi\mathbf{0})(x)\right|}_{\text{Term 2}}.$$

We prove that each of Term 1 and Term 2 converges to 0, wp1, as $N \to \infty$. The convergence of Term 2 to 0 as $N \to \infty$ follows from Lemma 19. Regarding Term 1, suppose for a contradiction, it does not converge to 0. Hence, there exists $\delta > 0$, a subsequence x^{N_k} of x^N, and $a^{N_k} \in A(x^{N_k})$, such that,

$$\left| \sup_{\nu^{N_k} \in \mathcal{P}^{N_k}(\epsilon^{N_k})} \left[\int_\Xi c(x^{N_k}, a^{N_k}, z) \nu^{N_k}(dz) \right] - \int_\Xi c(x^{N_k}, a^{N_k}, z) \mu(dz) \right| > \delta \ \forall k \in \mathbb{N}. \tag{4.9}$$

Since $x^{N_k} \to x$ and $x \mapsto A(x)$ is continuous by Assumption 5(b), using the upper hemicontinuity of $x \mapsto A(x)$, there is a subsequence $a^{N_{k_i}}$ of a^{N_k}, such that, $a^{N_{k_i}} \to a \in A(x)$. By the definition of supremum, for any $\eta > 0$, there exists $\bar{\nu}^{N_{k_i}} \in \mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})$ for all $i \in \mathbb{N}$, such that, for all $i \geq 1$,

$$\begin{aligned} &\left| \sup_{\nu^{N_{k_i}} \in \mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})} \left[\int_\Xi c(x^{N_{k_i}}, a^{N_{k_i}}, z) \nu^{N_{k_i}}(dz) \right] - \int_\Xi c(x, a, z) \mu(dz) \right| \\ &\qquad < \\ &\eta/2 + \left| \int_\Xi c(x^{N_{k_i}}, a^{N_{k_i}}, z) \bar{\nu}^{N_{k_i}}(dz) - \int_\Xi c(x, a, z) \mu(dz) \right|. \end{aligned}$$

Since c is continuous and bounded over $K \times \Xi$, we have $c(x^{N_{k_i}}, a^{N_{k_i}}, z^i) \to c(x, a, z)$ for all $z^i \to z$. Since Assumption 7 holds and $\epsilon^N \to 0$, by Lemma 22, $\bar{\nu}^{N_{k_i}} \xrightarrow{i,\text{wp1}} \mu$ weakly. Therefore, by Lemma 17, $\left| \int_\Xi c(x^{N_{k_i}}, a^{N_{k_i}}, z) \bar{\nu}^{N_{k_i}}(dz) - \int_\Xi c(x, a, z) \mu(dz) \right| \xrightarrow{i,\text{wp1}} 0$. Hence, for all large enough i,

$$\left| \sup_{\nu^{N_{k_i}} \in \mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})} \left[\int_\Xi c(x^{N_{k_i}}, a^{N_{k_i}}, z) \nu^{N_{k_i}}(dz) \right] - \int_\Xi c(x, a, z) \mu(dz) \right| < \eta. \tag{4.10}$$

Since c is continuous and bounded, from the dominated convergence theorem,

$$\left| \int_{\Xi} c(x^{N_{k_i}}, a^{N_{k_i}}, z)\mu(dz) - \int_{\Xi} c(x, a, z)\mu(dz) \right| \xrightarrow{i} 0. \tag{4.11}$$

From (4.10) and (4.11), we have for all large enough i,

$$\left| \sup_{\nu^{N_{k_i}} \in \mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})} \left[\int_{\Xi} c(x^{N_{k_i}}, a^{N_{k_i}}, z)\nu^{N_{k_i}}(dz) \right] - \int_{\Xi} c(x^{N_{k_i}}, a^{N_{k_i}}, z)\mu(dz) \right| < 2\eta,$$

which contradicts (4.9) since $\eta > 0$ is arbitrary. Therefore, Term 1 converges to 0, wp1, thereby establishing the base case.

Now consider $t \geq 1$ and assume that the result is true by the induction hypothesis for all $t' \leq t-1$. Since $\tilde{\Phi}_t^N \mathbf{0} = \tilde{\Phi}^N \tilde{\Phi}_{t-1}^N \mathbf{0}$, the uniform boundedness (w.r.t. supremum norm) of $\left\{ \tilde{\Phi}_t^N \mathbf{0} \right\}$ follows by the boundedness of c and the uniform boundedness of $\{\tilde{\Phi}_{t-1}^N \mathbf{0}\}$. Next, for any $N \geq 1$, we have

$$\begin{aligned}
&\left| (\tilde{\Phi}_t^N \mathbf{0})(x^N) - (\Phi_t \mathbf{0})(x) \right| \\
&\quad \leq \\
&\left| (\tilde{\Phi}_t^N \mathbf{0})(x^N) - (\Phi_t \mathbf{0})(x^N) \right| + \left| (\Phi_t \mathbf{0})(x^N) - (\Phi_t \mathbf{0})(x) \right| \\
&\quad \leq \\
&\sup_{a \in A(x^N)} \left| \sup_{\nu^N \in \mathcal{P}^N(\epsilon^N)} \left[\int_{\Xi} \left(c(x^N, a, z) + \alpha(\tilde{\Phi}_{t-1}^N \mathbf{0})(F(x^N, a, z)) \right) \nu^N(dz) \right] - \right. \\
&\qquad \left. \int_{\Xi} \left(c(x^N, a, z) + \alpha(\Phi_{t-1} \mathbf{0})(F(x^N, a, z)) \right) \mu(dz) \right| + \\
&\left| (\Phi_t \mathbf{0})(x^N) - (\Phi_t \mathbf{0})(x) \right|.
\end{aligned}$$

Let **I** and **II** denote the first and second term in the last expression, respectively, i.e.,

$$\mathbf{I} \stackrel{\text{def}}{=} \sup_{a\in A(x^N)} \left| \sup_{\nu^N\in\mathcal{P}^N(\epsilon^N)} \left[\int_\Xi \left(c(x^N,a,z)+\alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^N,a,z))\right)\nu^N(dz)\right] - \int_\Xi \left(c(x^N,a,z)+\alpha(\Phi_{t-1}\mathbf{0})(F(x^N,a,z))\right)\mu(dz)\right|$$

$$\mathbf{II} \stackrel{\text{def}}{=} \left|(\Phi_t\mathbf{0})(x^N)-(\Phi_t\mathbf{0})(x)\right|.$$

We prove that **I** and **II** converge to 0, wp1, as $N\to\infty$. The convergence of **II** to 0 as $N\to\infty$ follows from Lemma 19. Regarding **I**, suppose for a contradiction, it does not converge to 0. Hence, there exists $\delta>0$, a subsequence x^{N_k} of x^N, and $a^{N_k}\in A(x^{N_k})$, such that,

$$\left| \sup_{\nu^{N_k}\in\mathcal{P}^{N_k}(\epsilon^{N_k})} \left[\int_\Xi \left(c(x^{N_k},a^{N_k},z)+\alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^{N_k},a^{N_k},z))\right)\nu^{N_k}(dz)\right] - \int_\Xi \left(c(x^{N_k},a^{N_k},z)+\alpha(\Phi_{t-1}\mathbf{0})(F(x^{N_k},a^{N_k},z))\right)\mu(dz)\right| > \delta \ \forall k\in\mathbb{N}. \tag{4.12}$$

Since $x^{N_k}\to x$ and $x\mapsto A(x)$ is continuous by Assumption 5(b), using the upper hemicontinuity of $x\mapsto A(x)$, there is a subsequence $a^{N_{k_i}}$ of a^{N_k}, such that, $a^{N_{k_i}}\to a\in A(x)$. By the definition of supremum, for any $\eta>0$, there exists $\bar{\nu}^{N_{k_i}}\in\mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})$ for all $i\in\mathbb{N}$, such that, for all $i\geq 1$,

$$\left| \sup_{\nu^{N_{k_i}}\in\mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})} \left[\int_\Xi \left(c(x^{N_{k_i}},a^{N_{k_i}},z)+\alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^{N_{k_i}},a^{N_{k_i}},z))\right)\nu^{N_{k_i}}(dz)\right] - \int_\Xi \left(c(x,a,z)+\alpha(\Phi_{t-1}\mathbf{0})(F(x,a,z))\right)\mu(dz)\right|$$

$$<$$

$$\eta/2+\left|\int_\Xi \left(c(x^{N_{k_i}},a^{N_{k_i}},z)+\alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^{N_{k_i}},a^{N_{k_i}},z))\right)\bar{\nu}^{N_{k_i}}(dz) - \int_\Xi \left(c(x,a,z)+\alpha(\Phi_{t-1}\mathbf{0})(F(x,a,z))\right)\mu(dz)\right|.$$

Since Assumption 7 holds and $\epsilon^N \to 0$, by Lemma 22, $\bar{\nu}^{N_{k_i}} \xrightarrow{i,\text{wp1}} \mu$ weakly. By the induction hypothesis, $\tilde{\Phi}^N_{t-1}\mathbf{0}(F(x^{N_{k_i}}, a^{N_{k_i}}, z^i)) \xrightarrow{i,\text{wp1}} (\Phi_{t-1}\mathbf{0})(F(x,a,z))$ for all $z^i \to z$. Again, by the induction hypothesis, $\left\{\tilde{\Phi}^N_{t-1}\mathbf{0}\right\}$ is uniformly bounded. Also, c is also continuous and bounded over $K \times \Xi$. Hence, by Lemma 17,

$$\left| \int_\Xi \left(c(x^{N_{k_i}}, a^{N_{k_i}}, z) + \alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^{N_{k_i}}, a^{N_{k_i}}, z)) \right) \bar{\nu}^{N_{k_i}}(dz) - \right.$$
$$\left. \int_\Xi \left(c(x,a,z) + \alpha(\Phi_{t-1}\mathbf{0})(F(x,a,z)) \right) \mu(dz) \right| \xrightarrow{i,\text{wp1}} 0.$$

Hence, for all large enough i,

$$\left| \sup_{\nu^{N_{k_i}} \in \mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})} \left[\int_\Xi \left(c(x^{N_{k_i}}, a^{N_{k_i}}, z) + \alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^{N_{k_i}}, a^{N_{k_i}}, z)) \right) \nu^{N_{k_i}}(dz) \right] - \right.$$
$$\left. \int_\Xi \left(c(x,a,z) + \alpha(\Phi_{t-1}\mathbf{0})(F(x,a,z)) \right) \mu(dz) \right| < \eta. \tag{4.13}$$

From Lemma 19, $\Phi_{t-1}\mathbf{0}$ is continuous and bounded over X. Also, c is continuous and bounded over X. Hence, by the dominated convergence theorem,

$$\left| \int_\Xi \left(c(x^{N_{k_i}}, a^{N_{k_i}}, z) + \alpha(\Phi_{t-1}\mathbf{0})(F(x^{N_{k_i}}, a^{N_{k_i}}, z)) \right) \mu(dz) - \right.$$
$$\left. \int_\Xi \left(c(x,a,z) + \alpha(\Phi_{t-1}\mathbf{0})(F(x,a,z)) \right) \mu(dz) \right| \xrightarrow{i} 0. \tag{4.14}$$

From (4.13) and (4.14), we have for all large enough i,

$$\left| \sup_{\nu^{N_{k_i}} \in \mathcal{P}^{N_{k_i}}(\epsilon^{N_{k_i}})} \left[\int_\Xi \left(c(x^{N_{k_i}}, a^{N_{k_i}}, z) + \alpha(\tilde{\Phi}^N_{t-1}\mathbf{0})(F(x^{N_{k_i}}, a^{N_{k_i}}, z)) \right) \nu^{N_{k_i}}(dz) \right] - \right.$$
$$\left. \int_\Xi \left(c(x^{N_{k_i}}, a^{N_{k_i}}, z) + \alpha(\Phi_{t-1}\mathbf{0})(F(x^{N_{k_i}}, a^{N_{k_i}}, z)) \right) \mu(dz) \right| < 2\eta,$$

which contradicts (4.12) since $\eta > 0$ is arbitrary. Therefore, $\mathbf{I}$ converges to 0, wp1, as $N \to \infty$. Hence, the induction step is proved. □

Corollary 1. Suppose the radii of ambiguity sets are dependent on N such that $\lim_{N\to\infty} \epsilon^N = 0$. Let $\rho \in \mathcal{M}(X)$ be an intial state distribution over X. If Assumption 5, Assumption 6, and Assumption 7 hold, then, $\int_X \tilde{J}^{N,*}(x;\epsilon^N)\rho(dx) \xrightarrow{N,\text{wp1}} \int_X J^*(x)\rho(dx)$.

Proof. Since Assumption 5, Assumption 6, and Assumption 7 hold, from Theorem 14, the event $\left\{\tilde{J}^{N,*}(x;\epsilon^N) \xrightarrow{N} J^*(x)\ \forall x \in X\right\}$ holds with probability 1. Also, the sequence $\{\tilde{J}^{N,*}\}$ is uniformly bounded (w.r.t. supremum norm) since c is bounded by Assumption 5(e). The result then follows by the dominated convergence theorem. □

Let $\hat{\pi}^N \in \Pi$ denote the optimal policy to the data-driven problem, referred as the robust optimal policy. Recall from Section 4.1, the value of this policy on the true problem when starting from state $x \in X$ is denoted as $J(\hat{\pi}^N, x)$. In other words, $J(\hat{\pi}^N, \cdot)$ is the value of policy $\hat{\pi}^N$ when the distribution of the random noise is the true distribution, μ. We refer to $J(\hat{\pi}^N, \cdot)$ as the out-of-sample value.

Theorem 15. Suppose the radii of ambiguity sets are dependent on N such that $\lim_{N\to\infty} \epsilon^N = 0$. If Assumption 5, Assumption 6, and Assumption 7 hold, then, the event $\left\{J(\hat{\pi}^N, x) \xrightarrow{N} J^*(x)\ \forall x \in X\right\}$ holds with probability 1.

Proof. The proof is similar to the convergence analysis established for the nonrobust case [35, Theorem 4.4]. We present the proof here for completeness and to highlight the minor differences.

Consider any $x^N \to x$. Since $J^* \in C_b(X)$ by Lemma 20, $\Phi_{t,\pi}J^*$ and $\Phi_t J^*$ are well defined for all $t \geq 1$ and $\pi \in \Pi$. Hence, for any fixed $t \in \mathbb{N}$, we have for all $N \geq 1$,

$$\begin{aligned}\left|J(\hat{\pi}^N, x^N) - J^*(x)\right| &\leq \left|J(\hat{\pi}^N, x^N) - (\Phi_{t,\hat{\pi}^N}J^*)(x^N)\right| + \left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N) - (\Phi_t J^*)(x)\right| + \\ &\qquad |(\Phi_t J^*)(x) - J^*(x)| \\ &\overset{(a)}{=} \left|J(\hat{\pi}^N, x^N) - (\Phi_{t,\hat{\pi}^N}J^*)(x^N)\right| + \left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N) - (\Phi_t J^*)(x)\right| + 0\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{x\in X,\pi\in\Pi} |J(\pi,x) - (\Phi_{t,\pi}J^*)(x)| + \left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N) - (\Phi_t J^*)(x)\right| \\
&\overset{(b)}{\leq} \alpha^t\left(\frac{B}{1-\alpha} + \|J^*\|_\infty\right) + \left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N) - (\Phi_t J^*)(x)\right| \\
&\overset{(c)}{\leq} \frac{2B\alpha^t}{1-\alpha} + \left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N) - (\Phi_t J^*)(x)\right|.
\end{aligned}
$$

Here, (a) follows from Theorem 13 since J^* is the fixed point of the operator Φ. The inequalities in (b) and (c) follow from Lemma 21 and Lemma 20, respectively. Since $\alpha \in (0,1)$, there exists a $t \in \mathbb{N}$, such that, $\frac{2B\alpha^t}{1-\alpha}$ can be made arbitrarily small. Hence,
$\left|J(\hat{\pi}^N, x^N) - J^*(x)\right| \xrightarrow{N,\text{wp1}} 0$ if $\left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N) - (\Phi_t J^*)(x)\right| \xrightarrow{N,\text{wp1}} 0$ for all $t \in \mathbb{N}$.

The almost sure convergence of $(\Phi_{t,\hat{\pi}^N}J^*)(x^N)$ to $(\Phi_t J^*)(x)$ will be established using induction on $t \in \mathbb{N}$. Before that, we prove that every convergent subsequence of $\hat{\pi}^N(x^N)$ converges to an optimal solution for the true problem. First, note that $\hat{\pi}^N(x^N)$ has a convergent subsequence. This follows since $x^N \to x$ and $x \mapsto A(x)$ is upper hemicontinuous. Next, every convergent subsequence of $\hat{\pi}^N(x^N)$ converges to an element in $A(x)$. To see this, suppose a subsequence of $\hat{\pi}^N(x^N)$ converges to $a \notin A(x)$. Then, using the upper hemicontinuity of $x \mapsto A(x)$, we can extract a subsequence of this convergent subsequence which converges to an element of $A(x)$. Now, we prove that every convergent subsequence of $\hat{\pi}^N(x^N)$ converges to a true optimal solution. Let $\hat{\pi}^{N_k}(x) \to a \in A(x)$. Since $\tilde{J}^{N_k,*}$ is the fixed point of $\tilde{\Phi}^{N_k}$, we have $\tilde{J}^{N_k,*}(x^{N_k};\epsilon^{N_k}) = (\tilde{\Phi}^{N_k}\tilde{J}^{N_k,*})(x^{N_k})$, and hence

$$
\begin{aligned}
&\tilde{J}^{N_k,*}(x^{N_k};\epsilon^{N_k}) = \\
&\qquad \sup_{\nu^{N_k}\in\mathcal{P}^{N_k}(\epsilon^{N_k})} \int_\Xi \left(c(x^{N_k},\hat{\pi}^{N_k}(x^{N_k}),z) + \alpha\tilde{J}^{N_k,*}(F(x^{N_k},\hat{\pi}^{N_k}(x^{N_k}),z);\epsilon^{N_k})\right)\nu^{N_k}(dz).
\end{aligned} \tag{4.15}
$$

The equality follows since $\hat{\pi}^{N_k}$ is the robust optimal policy. Taking limits $k \to \infty$ in (4.15), from Theorem 12, $\tilde{J}^{N_k,*}(x^{N_k};\epsilon^{N_k}) \xrightarrow{k,\text{wp1}} J^*(x)$. Regarding the expression in the rhs of (4.15), from the proof of Theorem 12, $\tilde{J}^{N_k,*}(F(x^{N_k},\hat{\pi}^{N_k}(x^{N_k}),z^k);\epsilon^{N_k}) \xrightarrow{k,\text{wp1}} J^*(F(x,a,z))$ for all

$z^k \to z$. Also, $c(x^{N_k}, \pi^{N_k}(x), z^k) \to c(x, a, z)$ for all $z^k \to z$ since c is continuous. Next, from the definition of $\tilde{J}^{N,*}(\cdot; \epsilon^N)$ in (4.4), the sequence $\left\{ \tilde{J}^{N_k,*}(\cdot; \epsilon^{N_k}) \right\}$ is uniformly bounded since c is bounded. Hence, the sequence $\left\{ c + J^{N_k,*}(\cdot; \epsilon^{N_k}) \right\}$ is uniformly bounded. Finally, since $\epsilon^N \to 0$, from Lemma 22, $\nu^{N_k} \xrightarrow{k,\text{wp1}} \mu$ weakly. Therefore, using an argument as done in the proof of Theorem 14,

$$\sup_{\nu^{N_k} \in \mathcal{P}^{N_k}(\epsilon^{N_k})} \int_\Xi \left(c(x^{N_k}, \hat{\pi}^{N_k}(x^{N_k}), z) + \alpha \tilde{J}^{N_k,*}(F(x^{N_k}, \hat{\pi}^{N_k}(x^{N_k}), z); \epsilon^{N_k}) \right) \nu^{N_k}(dz)$$
$$\xrightarrow{k,\text{wp1}} \int_\Xi \left(c(x, a, z) + \alpha J^*(F(x, a, z)) \right) \mu(dz).$$

Hence, $J^*(x) = \int_\Xi \left(c(x, a, z) + \alpha J^*(F(x, a, z)) \right) \mu(dz)$ almost surely, thereby proving that every convergent subsequence of $\hat{\pi}^{N_k}(x)$ converges to a true optimal solution, wp1.

We now proceed towards establishing the almost sure convergence of $(\Phi_{t,\hat{\pi}^N} J^*)(x)$ to $(\Phi_t J^*)(x)$ by using induction on $t \in \mathbb{N}$. Consider the base case $t = 1$. For all $N \geq 1$,

$$\left| (\Phi_{\hat{\pi}^N} J^*)(x^N) - (\Phi J^*)(x) \right| = \left| \int_\Xi \left(c(x^N, \hat{\pi}^N(x^N), z) + \alpha J^*(F(x^N, \hat{\pi}^N(x^N), z)) \right) \mu(dz) - \right.$$
$$\left. \min_{a \in A(x)} \int_\Xi \left(c(x, a, z) + \alpha J^*(F(x, a, z)) \right) \mu(dz) \right|$$

Supose for a contradiction $|(\Phi_{\hat{\pi}^N} J^*)(x) - (\Phi J^*)(x)|$ does not converge to 0. Hence, there exists $\delta > 0$, subsequences x^{N_k} and $\hat{\pi}^{N_k}(x^{N_k})$, such that,

$$\left| \int_\Xi \left(c(x^{N_k}, \hat{\pi}^{N_k}(x^{N_k}), z) + \alpha J^*(F(x^{N_k}, \hat{\pi}^{N_k}(x^{N_k}), z)) \right) \mu(dz) - \right.$$
$$\left. \min_{a \in A(x)} \int_\Xi \left(c(x, a, z) + \alpha J^*(F(x, a, z)) \right) \mu(dz) \right| > \delta \ \forall k \in \mathbb{N}. \tag{4.16}$$

Since $x^N \to x$, using the upper hemicontinuity of $x \mapsto A(x)$, the subsequence $\hat{\pi}^{N_k}(x^{N_k})$ has a convergent subsequence $\hat{\pi}^{N_{k_i}}(x^{N_{k_i}})$ which converges to a true optimal solution $\pi(x) \in A(x)$. Also, J^* is continuous and bounded over X by Lemma 20. Hence, by the dominated

convergence theorem,

$$\lim_{i\to\infty}\int_{\Xi}\left(c(x^{N_{k_i}},\hat{\pi}^{N_{k_i}}(x^{N_{k_i}}),z)+\alpha J^*(F(x^{N_{k_i}},\hat{\pi}^{N_{k_i}}(x^{N_{k_i}}),z))\right)\mu(dz)$$
$$=\min_{a\in A(x)}\int_{\Xi}\left(c(x,a,z)+\alpha J^*(F(x,a,z))\right)\mu(dz),$$

which contradicts (4.16). Hence, $|(\Phi_{\hat{\pi}^N}J^*)(x)-(\Phi J^*)(x)|\xrightarrow{N,\text{wp1}}0$, thereby establishing the base case. Now consider $t\geq 1$ and assume that the result is true by the induction hypothesis for all $t'\leq t-1$. Hence, for all $N\geq 1$,

$$\left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N)-(\Phi_t J^*)(x)\right|=$$
$$\left|\int_{\Xi}\left(c(x^N,\hat{\pi}^N(x^N),z)+\alpha(\Phi_{t-1,\hat{\pi}^N}J^*)(F(x^N,\hat{\pi}^N(x^N),z))\right)\mu(dz)-\right.$$
$$\left.\min_{a\in A(x)}\int_{\Xi}\left(c(x,a,z)+\alpha(\Phi_{t-1}J^*)(F(x,a,z))\right)\mu(dz)\right|$$

Supose for a contradiction $\left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N)-(\Phi_t J^*)(x)\right|$ does not converge to 0. Hence, there exists $\delta>0$, subsequences x^{N_k} and $\hat{\pi}^{N_k}(x^{N_k})$, such that,

$$\left|\int_{\Xi}\left(c(x^{N_k},\hat{\pi}^{N_k}(x^{N_k}),z)+\alpha(\Phi_{t-1,\hat{\pi}^{N_k}}J^*)(F(x^{N_k},\hat{\pi}^{N_k}(x^{N_k}),z))\right)\mu(dz)-\right.$$
$$\left.\min_{a\in A(x)}\int_{\Xi}\left(c(x,a,z)+\alpha(\Phi_{t-1}J^*)(F(x,a,z))\right)\mu(dz)\right|>\delta\ \forall k\in\mathbb{N}. \tag{4.17}$$

Since $x^N\to x$, using the upper hemicontinuity of $x\mapsto A(x)$, the subsequence $\hat{\pi}^{N_k}(x^{N_k})$ has a convergent subsequence $\hat{\pi}^{N_{k_i}}(x^{N_{k_i}})$ which converges to a true optimal solution $\pi(x)\in A(x)$. By the induction hypothesis,
$(\Phi_{t-1,\hat{\pi}^{N_{k_i}}}J^*)((F(x^{N_{k_i}},\hat{\pi}^{N_{k_i}}(x^{N_{k_i}}),z^i)))\xrightarrow{i,\text{wp1}}(\Phi_{t-1}J^*)(F(x,\pi(x),z))$ for all $z^i\to z$. Next, the sequence $\left\{\Phi_{t-1,\hat{\pi}^N}J^*\right\}$ is uniformly bounded (w.r.t. supremum norm). Since c is also

continuous and bounded, by Lemma 17,

$$\lim_{i\to\infty}\int_{\Xi}\Big(c(x^{N_{k_i}},\hat{\pi}^{N_{k_i}}(x^{N_{k_i}}),z)+\alpha(\Phi_{t-1,\hat{\pi}^{N_{k_i}}}J^*)(F(x^{N_{k_i}},\hat{\pi}^{N_{k_i}}(x^{N_{k_i}}),z))\Big)\,\mu(dz)$$
$$=\min_{a\in A(x)}\int_{\Xi}\left(c(x,a,z)+\alpha(\Phi_{t-1}J^*)(F(x,a,z))\right)\mu(dz),$$

which contradicts (4.17). Therefore, $\left|(\Phi_{t,\hat{\pi}^N}J^*)(x^N)-(\Phi_tJ^*)(x)\right| \xrightarrow{N,\text{wp1}} 0$. Hence, the induction step is proved. □

Corollary 2. Suppose the radii of ambiguity sets are dependent on N such that $\lim_{N\to\infty}\epsilon^N=0$. Let $\rho\in\mathcal{M}(X)$ be an intial state distribution over X. If Assumption 5, Assumption 6, and Assumption 7 hold, then, $\int_X J(\hat{\pi}^N,x)\rho(dx)\xrightarrow{N,\text{wp1}}\int_X J^*(x)\rho(dx)$.

Proof. Since Assumption 5, Assumption 6, and Assumption 7 hold, from Theorem 14, the event $\left\{J(\hat{\pi}^N,x)\xrightarrow{N}J^*(x)\ \forall x\in X\right\}$ holds with probability 1. Also, the sequence $\{J(\hat{\pi}^N,\cdot)\}$ is uniformly bounded (w.r.t. supremum norm) since c is bounded by Assumption 5(e). The result then follows by the dominated convergence theorem. □

4.3.2 Probabilistic guarantee on the performance of robust optimal policy

Although Theorem 15 shows that $J(\hat{\pi}^N,\cdot)$ converges to the true optimal value, we might still want $J(\hat{\pi}^N,\cdot)$ to be sufficiently small with a large enough probability (with respect to the uncertainty in the sampled training data) for finite sample-sizes. In order to derive such a claim, we first need the following concentration inequality assumption on the distance function.

Assumption 8. Fix an integer $N\geq 1$ and $q\in\mathcal{M}(\Xi)$. Let $\hat{q}^N$ be the empirical distribution of q. Then, for every $\beta\in(0,1)$, there exists an $0<\epsilon_\beta^N<\sup_{\nu_1,\nu_2\in\mathcal{M}(\Xi)}d(\nu_1,\nu_2)$, such that,

$$\mathbb{P}[d(q,\hat{q}^N)\leq\epsilon_\beta^N]\geq 1-\beta. \tag{4.18}$$

Remark 2. The concentration inequality stated above is a generic statement and it might need additional assumptions on Ξ and q for it to hold. Further, the set of additional assumptions might vary across distance functions. The reason for omitting the specifics from Assumption 8 is to demonstrate that the probabilistic performance guarantee can be obtained, as long as the distance function satisfies a generic concentration inequality. We also emphasize that the parameter ϵ_β^N might depend on some properties of Ξ and q which are typically independent of the sample-size N.

Our next result establishes that the robust optimal value $\tilde{J}^{N,*}(x, \epsilon^N)$ provides an upper bound on $J(\hat{\pi}^N, x)$ with arbitrarily high probability.

Theorem 16. Fix the sample-size $N \geq 1$. Consider any fixed $\gamma \in (0, 1)$. Suppose the true distribution, μ, of the random noise, ξ, satisfies Assumption 8. Let $0 < \epsilon_\gamma^N < \sup_{\nu_1, \nu_2 \in \mathcal{M}(\Xi)} d(\nu_1, \nu_2)$ be as in (4.18). Then,

$$\mathbb{P}\left[J(\hat{\pi}^N, x) \leq \tilde{J}^{N,*}(x; \epsilon_\gamma^N) \ \forall x \in X\right] \geq 1 - \gamma. \tag{4.19}$$

Proof. The proof is similar to [70, Theorem 3] who established a probabilistic performance guarantee for Wasserstein distance. Using a standard monotonicity and successive approximation argument, we have,

$$\left\{(\Phi_{\hat{\pi}^N} \tilde{J}^{N,*})(x) \leq \tilde{J}^{N,*}(x; \epsilon_\gamma^N) \ \forall x \in X\right\} \subseteq \left\{J(\hat{\pi}^N, x) \leq \tilde{J}^{N,*}(x; \epsilon_\gamma^N) \ \forall x \in X\right\}. \tag{4.20}$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left[J(\hat{\pi}^N, x) \leq \tilde{J}^{N,*}(x; \epsilon_\gamma^N) \ \forall x \in X\right] \\
&\qquad \geq \\
&\mathbb{P}\left[(\Phi_{\hat{\pi}^N} \tilde{J}^{N,*})(x) \leq \tilde{J}^{N,*}(x; \epsilon_\gamma^N) \ \forall x \in X\right] \\
&\qquad \overset{(a)}{=}
\end{aligned}$$

$$\mathbb{P}\left[\int_{\Xi}\left(c(x,\hat{\pi}^N(x),z)+\alpha\tilde{J}^{N,*}(F(x,\hat{\pi}^N(x),z))\right)\mu(dz)\leq \tilde{J}^{N,*}(x;\epsilon_\gamma^N)\ \forall x\in X\right]$$

$$\overset{(b)}{=}$$

$$\mathbb{P}\left[\int_{\Xi}\left(c(x,\hat{\pi}^N(x),z)+\alpha\tilde{J}^{N,*}(F(x,\hat{\pi}^N(x),z))\right)\mu(dz)\leq \right.$$

$$\left.\sup_{\nu^N\in\mathcal{P}^N(\epsilon_\gamma^N)}\left[\int_{\Xi}\left(c(x,\hat{\pi}^N(x),z)+\alpha\tilde{J}^{N,*}(F(x,\hat{\pi}^N(x),z))\right)\nu^N(dz)\right]\ \forall x\in X\right]$$

$$\geq \mathbb{P}\left[\mu\in\mathcal{P}^N(\epsilon_\gamma^N)\right]\overset{(c)}{=}\mathbb{P}\left[d(\mu,\hat{\mu}^N)\leq\epsilon_\gamma^N\right]\overset{(d)}{\geq}1-\gamma.$$

Equality (a) follows from defintion of $\Phi_{\hat{\pi}^N}$ in (4.8). The equality in (b) follows since $\tilde{J}^{N,*}$ is fixed point of $\tilde{\Phi}^N$ and $\hat{\pi}^N$ is the robust optimal policy. The equality in (c) follows by definition of $\mathcal{P}^N(\epsilon_\gamma^N)$ and inequality (d) follows by choice of ϵ_γ^N and Assumption 8. □

Remark 3. The modeling framework and the results in this section also hold for the case of finite-horizon problems as long as Assumption 5, Assumption 6, and Assumption 7 are satisfied.

4.4 Distances satisfying Assumption 6 and Assumption 7

We provide a list of distance well known distance functions that satisfy Assumption 6 and Assumption 7. Recall that β-metric defined in (4.5) is given as

$$\beta(\mu,\nu)\overset{\text{def}}{=}\sup_{\substack{f:\Xi\to\mathbb{R}\\ \|f\|_{BL}\leq 1}}\left|\int_X f(x)\mu(dx)-\int_X f(x)\nu(dx)\right|.$$

Throughout this section, we let ρ be a metric which metrizes the topology on Ξ.

4.4.1 Prokhorov distance

For any $T\subset\Xi$ and $\delta>0$, define $T^\delta\overset{\text{def}}{=}\{z\in\Xi:\exists y\in T \text{ such that } \rho(z,y)<\delta\}$. The Prokhorov distance between $\mu_1,\mu_2\in\mathcal{M}(\Xi)$ is given by

$d_{\mathrm{P}} \stackrel{\text{def}}{=} \inf\left\{\delta > 0 : \mu_1(T) \leq \mu_2(T^\delta) + \delta \ \forall \text{ Borel sets } T\right\}$ [16, Section 11.3]. Also, d_{P} is a metric which metrizes the topology of weak convergence [16, Theorem 11.3.1, Theorem 11.3.3]. Hence, d_{P} is weakly continuous over $\mathcal{M}(\Xi) \times \mathcal{M}(\Xi)$, thereby satisfying Assumption 6. Since $\beta(\mu_1, \mu_2) \leq 2d_{\mathrm{P}}(\mu_1, \mu_2)$ [15, Corollary 2], d_{P} satisfies Assumption 7 with $\psi(t) = 2t$.

4.4.2 Total Variation distance

The toal variation (TV) distance, $d_{\mathrm{TV}}(\mu_1, \mu_2)$, between $\mu_1, \mu_2 \in \mathcal{M}(\Xi)$ is given as $d_{\mathrm{TV}}(\mu_1, \mu_2) \stackrel{\text{def}}{=} \frac{1}{2}\sup\left\{\left|\int_\Xi f(z)\mu_1(dz) - \int_\Xi f(z)\mu_2(dz)\right| | f : \Xi \to \mathbb{R},\ f \text{ is measurable }, \|f\|_\infty \leq 1\right\}$. Since $\{f : \Xi \to \mathbb{R} |\, \|f\|_{BL} \leq 1\} \subseteq \{f : \Xi \to \mathbb{R} | f \text{ is measurable } \|f\|_\infty \leq 1\}$, we have, $\beta(\mu_1, \mu_2) \leq 2d_{\mathrm{TV}}(\mu_1, \mu_2)$ for all $\mu_1, \mu_2 \in \mathcal{M}(\Xi)$. Hence, Assumption 7 holds with $\psi(t) = 2t$. TV distance also admits an alternate representation given by $d_{\mathrm{TV}}(\mu_1, \mu_2) \stackrel{\text{def}}{=} \int_\Xi \left|\frac{d\mu_1}{d\mu_2}(z) - 1\right| \mu_2(dz)$, where $\frac{d\mu_1}{d\mu_2}$ is the Radon-Nikodyn derivative of μ_1 with respect to μ_2. Hence, by [3, Theorem 2.34], it is weakly lower semicontinuous over $\mathcal{M}(\Xi) \times \mathcal{M}(\Xi)$, thereby satisfying Assumption 6.

4.4.3 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence between $\mu_1, \mu_2 \in \mathcal{M}(\Xi)$ is given as $d_{\mathrm{KL}}(\mu_1, \mu_2) \stackrel{\text{def}}{=} \int_\Xi \log\left(\frac{d\mu_1}{d\mu_2}(z)\right) \mu_1(dz)$ if μ_1 is absolutely continuous with respect to μ_2, and ∞ otherwise [22, page 422]. Here, $\frac{d\mu_1}{d\mu_2}$ is the Radon-Nikodyn derivative of μ_1 with respect to μ_2. The Pinsker's inequality [22, page 429] states that $d_{\mathrm{TV}}(\mu_1, \mu_2) \leq \sqrt{d_{\mathrm{KL}}(\mu_1, \mu_2)/2}$. In the previous subsection, we established that $\beta(\mu_1, \mu_2) \leq 2d_{\mathrm{TV}}(\mu_1, \mu_2)$. Hence, $\beta(\mu_1, \mu_2) \leq \sqrt{2d_{\mathrm{KL}}(\mu_1, \mu_2)}$. Therefore, d_{KL} satisfies Assumption 7 with $\psi(t) = \sqrt{2t}$. Since d_{KL} is weakly lower semicontinuous over $\mathcal{M}(\Xi) \times \mathcal{M}(\Xi)$ [49, Theorem 1], it satisfies Assumption 6.

4.4.4 χ^2 distance

The χ^2 distance between $\mu_1, \mu_2 \in \mathcal{M}(\Xi)$ is given as $d_{\chi^2}(\mu_1, \mu_2) \stackrel{\text{def}}{=} \int_\Xi \left[\left(\frac{d\mu_1}{d\mu_2}(z) - 1\right)^2\right] \mu_2(dz)$ if μ_1 is absolutely continuous with respect to μ_2, and ∞ otherwise [22, page 425]. Here, $\frac{d\mu_1}{d\mu_2}$ is the Radon-Nikodyn derivative of μ_1 with respect to μ_2. From [22, page 429], we have,

$d_{\mathrm{TV}}(\mu_1,\mu_2) \leq \frac{1}{2}\sqrt{d_{\chi^2}(\mu_1,\mu_2)}$. Hence, $\beta(\mu_1,\mu_2) \leq \sqrt{d_{\chi^2}(\mu_1,\mu_2)}$. Therefore, d_{χ^2} satisfies Assumption 7 with $\psi(t) = \sqrt{t}$. Also, d_{χ^2} satisfies Assumption 6 since it is weakly lower semicontinuous over $\mathcal{M}(\Xi) \times \mathcal{M}(\Xi)$ [3, Theorem 2.34].

4.4.5 Wasserstein distance

For any $p \in [1,\infty)$, the p-Wasserstein distance between $\mu_1, \mu_2 \in \mathcal{M}_p(\Xi)$ is defined as $d_{\mathrm{W}(p)}(\mu_1,\mu_2) \stackrel{\text{def}}{=} \inf_{\theta\in\Theta(\mu_1,\mu_2)} \left(\int_{\Xi\times\Xi} \rho^p(z_1,z_2)d\theta(z_1,z_2)\right)^{1/\rho}$, where $\Theta(\mu_1,\mu_2)$ is the set of all probability distributions supported on $\Xi\times\Xi$ with marginals μ_1 and μ_2 [65, Chapter 6]. Using the Kantrovich-Rubinstein duality formula [65, Remark 6.5], $d_{\mathrm{W}(1)}$ can be equivalently expressed as $d_{\mathrm{W}(1)} = \sup\left\{\left|\int_\Xi f(z)\mu_1(dz) - \int_\Xi f(z)\mu_2(dz)\right| | f:\Xi\to\mathbb{R},\ \|f\|_L \leq 1\right\}$. Since $\{f:\Xi\to\mathbb{R}|\,\|f\|_{BL}\leq 1\} \subseteq \{f:\Xi\to\mathbb{R}|\,\|f\|_L \leq 1\}$, we have, $\beta(\mu_1,\mu_2) \leq d_{\mathrm{W}(1)}(\mu_1,\mu_2)$ for all $\mu_1,\mu_2 \in \mathcal{M}(\Xi)$. Since $d_{\mathrm{W}(p)}(\mu_1,\mu_2) \leq d_{\mathrm{W}(q)}(\mu_1,\mu_2)$ for all $1 \leq p < q$ and $\mu_1,\mu_2 \in \mathcal{M}(\Xi)$ [65, Remark 6.6], for any $p \in [1,\infty)$, we have, $\beta(\mu_1,\mu_2) \leq d_{\mathrm{W}(p)}(\mu_1,\mu_2)$ for all $\mu_1,\mu_2 \in \mathcal{M}(\Xi)$. Hence, for any $p \in [1,\infty)$, the p-Wasserstein distance satisfies Assumption 7 with $\psi(t) = t$. Regarding the weak lower semicontinuity, $d_{\mathrm{W}(p)}$ is weakly lower semicontinuous over $\mathcal{M}(\Xi)\times\mathcal{M}(\Xi)$ [65, Remark 6.12], thereby satisfying Assumption 6.

Finally, we note that Wasserstein distance admits a concentration inequality of the form stated in Assumption 8, as long as the disturbance space and the true probability distribution satisfy certain additional properties (see [20, Theorem 2]). However, if the disturbance space is compact, then these properties hold for all probability measures on the disturbance space.

4.4.6 β-metric

This is the β-metric defined in (4.5). Since β-metric metrizes the topology of weak convergence [16, Theorem 11.3.3], Assumption 6 and Assumption 7 trivially hold. Finally, the β-metric also admits a concentration inequality of the form stated in Assumption 8 for compact disturbance spaces. To see this, first note that for any $p \in [1,\infty)$, from Section 4.4.5, $\beta(\mu_1,\mu_2) \leq d_{\mathrm{W}(p)}(\mu_1,\mu_2)$ for all $\mu_1,\mu_2 \in \mathcal{M}(\Xi)$. Hence, for any finite sample-size $N \geq 1$, and $\nu \in \mathcal{M}(\Xi)$, we have, $\mathbb{P}\left[\beta(\nu,\hat{\nu}^N) \leq \delta\right] \geq \mathbb{P}\left[d_{\mathrm{W}(p)}(\nu,\hat{\nu}^N) \leq \delta\right]$ for all $\delta > 0$. Here, $\hat{\nu}^N$

is the empirical distribution of ν. Therefore, a concentration inequality for β-metric can be obtained from the concentration inequality for p-Wasserstein distance.

4.5 Applications

In this sections, we present applications which satisfy the model assumptions listed in Assumption 5. Therefore, the results mentioned in Section 4.3 hold true for such applications, provided the distance function employed satisfies Assumption 6 and Assumption 7.

4.5.1 Stochastic inventory control with bounded demand

We consider a stochastic inventory control problem with finte time horizon $T < \infty$. The random demand $D_1, D_2, \ldots, D_T$ in time periods $1, 2, \ldots, T$ are independent random variables with support in a compact set $[0, M] \subseteq \mathbb{R}_+$, where $M \in \mathbb{R}_+$. The unit production cost, holding cost, and backorder cost are denoted by c_p, c_h, and c_b, respectively. Without loss of generality, we take the terminal cost to be 0. Therefore, we can restrict the order quantity in any period to the interval $[0, TM]$. The inventory x_{t+1} in period $t+1$ is given as $x_{t+1} = x_t + a_t - D_t$, where x_t and a_t are the inventory and order quantity in period t, respectively. The single-stage cost is given by $c(x_t, a_t, D_t) = c_p a_t + c_h \max(0, x_t + a_t - D_t) + c_b \max(0, -(x_t + a_t - D_t))$. Finally, we assume that the discount factor is 1 for all time periods. This comes without any loss of generality. This problem can be modeled as a stochastic control problem which is as follows. For all $t \in \{1, 2, \ldots, T\}$, the state space X_t is the inventory level in time period t. The action space $A(x_t) = A = [0, TM]$ for all $x_t \in X_t$ and $t \in \{1, 2, \ldots, T\}$. The disturbance space $\Xi = [0, M]$. The system evolution function is $F(x_t, a_t, \xi_t) = x_t + a_t - \xi_t$ for all $t \in \{1, 2, \ldots, T\}$. The single-stage cost function is $c(x, a, \xi) = c_p a + c_h \max(0, x + a - \xi) + c_b \max(0, -(x + a - \xi))$. Clearly, this problem satisfies all the assumptions listed in Assumption 5. Therefore, if we choose a distance function satisfying Assumption 6 and Assumption 7, then all the results established in Section 4.3 hold for the stochastic inventory control problem.

4.5.2 Optimal asset selling problem

An optimal asset selling problem is described as follows [50, Section 3.4]. A system is in state $x_t \in X_t$ at the beginning of stage t. At each stage, there are two actions available at every state: to accept the given offer or to reject it. If the decision-maker accepts in state x_t in stage t, a reward of $g_t(x_t)$ is received and the system stops. On the other hand, if the decision-maker decides to reject when in state x_t, a cost of $f_t(x_t)$ is incurred and the system transitions to state $x_{t+1} \in X_{t+1}$ determined by a stochastic state evolution function. This continues utill the end of stage T, and if the system does not stop utill the end of stage T, then a terminal reward of $h(x_{T+1})$ is received where $x_{T+1} \in X_{T+1}$.

A stochastic control formulation of the optimal asset selling problem is as follows. The state space is given as $X_t \cup \{\Delta\}$ for all $t = 1, 2, \ldots, T+1$ where Δ denotes the termination state. We interchangeably use the term state space to denote X_t and $X_t \cup \{\Delta\}$. The exact meaning will be clear from the context. Let $A = \{0, 1\}$ denote the two actions where 0 indicates to accept and 1 indicates to reject. The action space is given as $A(x_t) = \{0, 1\}$ if $x_t \in X_t$ and $\{1\}$ if $x_t \in \Delta$, for all $t = 1, 2, \ldots, T$. The single-stage reward function is given as

$$r_t(x_t, a_t) = \begin{cases} -f_t(x_t) & x_t \in X_t,\ a_t = 1 \\ g_t(x_t) & x_t \in X_t,\ a_t = 0 \\ 0 & x_t = \Delta, \end{cases} \quad t = 1, 2, \ldots, T. \tag{4.21}$$

$$r_{T+1}(x_{T+1}) = h(x_{T+1}) \quad \text{if } x_{T+1} \in X_{T+1} \text{ and } 0 \text{ if } x_{T+1} = \Delta.$$

Let the random offer ξ_t take values in Ξ_t, which corresponds to the disturbance space. We assume that $\xi_1, \xi_2, \ldots, \xi_T$ are independent. For $t = 1, 2, \ldots, T$, and $\tilde{F}_t : X_t \times \Xi_t \to X_{t+1}$, the

system's stochastic state evolution function $F_t : X_t \cup \{\Delta\} \times A \times \Xi_t \to X_{t+1} \cup \{\Delta\}$ is given as

$$x_{t+1} \stackrel{\text{def}}{=} F_t(x_t, a_t, \xi_t) \stackrel{\text{def}}{=} \begin{cases} \tilde{F}_t(x_t, \xi_t) & x_t \in X_t, a_t = 1 \\ \Delta & x_t \in X_t, a_t = 0 \text{ or } x_t = \Delta, a_t = 1. \end{cases} \quad t = 1, 2, \ldots, T. \tag{4.22}$$

This generic formulation of the optimal asset selling problem captures a wide variety of real world scenarios and many of those models satisfy Assumption 5. To begin with, it is not uncommon in applications to have $X_t = \Xi_t = \Theta$ for all $t = 1, 2, \ldots, T$, where Θ is any compact subset of $\mathbb{R}$. Moving on, the continuation cost f_t in many instances is a fixed constant. Further, it is fairly common to have g_t and h to be continuous and bounded functions. Regarding the system evolution function, the two common selling mechanisms in practice are asset selling without recall and with recall, respectively. In asset selling without recall, rejected offers cannot be considered in the future. This can be modeled as $\tilde{F}_t(x_t, \xi_t) = \xi_t$. In asset selling with recall, past offers can also be considered in the future. This can be modeled as $\tilde{F}_t(x_t, \xi_t) = \max(x_t, \xi_t)$. Hence, the system evolution function can be shown to be continuous in both these cases. In summary, a large class of optimal asset selling problem satisfies Assumption 5. Hence, by using an appropriate distance function satisfying Assumption 6 and Assumption 7, the results established in Section 4.3 hold for a wide class of optimal asset selling problem.

4.5.3 Mold level control problem

This problem is from [25, Example 7.1]. Unlike the stochastic inventory control and optimal asset selling problem, this is an infinite horizon stochastic control problem. In this problem, the state variable corresponds to the height of an object, which we desire to keep as close as possible to a nominal height x^*. The state space is given as $X \stackrel{\text{def}}{=} [x^* - l, x^* + h] \subseteq \mathbb{R}$, where l, h are known positive numbers. The action space $A(x) \subseteq A \stackrel{\text{def}}{=} [-h, l]$ is given as $A(x) \stackrel{\text{def}}{=} [x^* - x, 0]$ if $x \geq x^*$ and $A(x) \stackrel{\text{def}}{=} [0, x^* - x]$ if $x \leq x^*$. The disturbance space

is $\Xi \stackrel{\text{def}}{=} [-\bar{\xi}, \bar{\xi}]$, where $0 < \bar{\xi} \leq \min(l, h)$. The system evolution is given as $F(x, a, \xi) \stackrel{\text{def}}{=} x + a - \xi$ and the single-stage cost is $c(x, a, \xi) \stackrel{\text{def}}{=} (x + a - x^*)^2 + \xi^2$. This problem satisfies all the assumptions listed in Assumption 5. Hence, all the results established in Section 4.3 hold for this problem if the distance function used to construct the ambiguity set satisfies Assumption 6 and Assumption 7.

4.5.4 Wind energy management

This is a finite-horizon stochastic control problem, which was introduced in [4] and also studied by [5]. In this problem, the owner of a wind power facility needs to announce the amount of energy to be produced before every period. The wind energy in every period is a random variable with values in $[0, M]$, where $0 < M < \infty$. The per unit energy price is $r > 0$ and the per unit penalty for failing to supply the committed amount of energy is $p > 0$. It is assumed that excess energy will be stored in a battery with finite capacity $E > 0$. Therefore, if sufficient wind is not produced in any period, the owner can potentially avoid paying the penalty by using the available energy stored in the battery to cover for the shortage. This can be modeled as a stochastic control problem, where $X = [0, E],\ A(x) = A = [0, M],\ \Xi = [0, M]$. The system evolution function is given as

$$F(x, a, \xi) \stackrel{\text{def}}{=} \begin{cases} \min(x + \xi - a, M) & \text{if } \xi \geq a \\ \max(x + \xi - a, 0) & \text{if } \xi \leq a. \end{cases}$$

Finally, the single-stage cost is given as

$$c(x, a, \xi) \stackrel{\text{def}}{=} -ar + \begin{cases} 0 & \text{if } \xi \geq a \\ (r + p)\max(a - x - \xi, 0) & \text{if } \xi \leq a. \end{cases}$$

Clearly, this problem satisfies all the assumptions listed in Assumption 5. Therefore, if we choose a distance function satisfying Assumption 6 and Assumption 7, then all the results

established in Section 4.3 hold for the robust version of this problem.

4.6 Conclusions

We considered stochastic control problem with unknown noise distribution and derived a robust counterpart of the problem by using the framework developed for modeling and analysing two player stochastic games. Under suitable model assumptions, we studied robust stochastic control problem with data-driven distance-based ambiguity sets. We presented an axiomatic characterization of the distance function (Assumption 6, Assumption 7, Assumption 8) which enabled us to establish the convergence of robust and out-of-sample values to the true optimal value, and a guarantee that the out-of-sample value will be smaller than the robust optimal value with a high probability. We also identified well known distance functions that satisfy the aximoatic characterization. Finally, we presented applications which fit into our modeling framework.

Unlike the previous two chapters, we do not have a probabilistic rate of convergence result in this chapter. Deriving such a result is challenging in this context for many reasons. First of all, we need to establish a counterpart of the simulation lemma (Lemma 4) for uncountable spaces. Apart from this, we also need to underatand the structure of the optimal solution in the inner optimization problem, so that it could be meaningfully combined along with the simulation lemma. An investigation of these issues is a potential problem for future research. Another interesting future direction is to try to extend the robust framework for partially observable MDPs.

BIBLIOGRAPHY

[1] A Agarwal, S Kakade, and L F Yang. Model-based reinforcement learning with a generative model is minimax optimal. *Proceedings of Machine Learning Research*, 125:1–17, 2020.

[2] C D Aliprantis and K C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Berlin and Heidelberg, third edition, 2006.

[3] L Ambrosio, N Fusco, and D Pallara. *Functions of bounded variation and free discontinuity problems*. Courier Corporation, 2000.

[4] C L Anderson, N Burke, and M Davison. Optimal management of wind energy with storage: Structural implications for policy and market design. *Journal of Energy Engineering*, 141(1):B4014002, 2015.

[5] N Bäuerle and A Glauner. Distributionally robust markov decision processes and their connection to risk measures. *Mathematics of Operations Research*, 47(3):1757–1780, 2022.

[6] G Bayraksan and D K Love. Data-Driven Stochastic Programming Using Phi-Divergences. *INFORMS Tutorials in Operations Research*, 2015.

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

[8] D Berend and A Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013.

[9] D P Bertsekas and S E Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Academic Press, New York, 1978.

[10] D Bertsimas, V Gupta, and N Kallus. Robust sample average approximation. *Mathematical Programming*, 171:217–282, 2018.

[11] Z Chen, P Yu, and W B Haskell. Distributionally robust optimization for sequential decision-making. *Optimization*, 68(12):2397–2426, 2019.

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

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

[14] E Derman and S Mannor. Distributional robustness and regularization in reinforcement learning. *arXiv preprint arXiv:2003.02894*, 2020.

[15] R M Dudley. Distances of probability measures and random variables. *The Annals of Mathematical Statistics*, 39(5):1563–1572, 1968.

[16] R M Dudley. *Real analysis and probability*. CRC Press, 2018.

[17] D Duque and D Morton. Distributionally robust stochasatic dynamic programming. *SIAM Journal on Optimization*, 30(4), 2020.

[18] E Erdogan and G N Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107:37–61, 2006.

[19] P M Esfahani and D Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171:115–166, 2017.

[20] N Fournier and A Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[21] R Gao and A J Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance. *Mathematics of Operations Reearch*, 48(2):603–655, 2022.

[22] A L Gibbs and F E Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[23] P Glasserman and X Xu. Robust Portfolio Control with Stochastic Factor Dynamics. *Operations Research*, 61(4):874–893, 2013.

[24] J Goh and M Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58:902–917, 2010.

[25] J I Gonzalez-Trejo, O Hernandez-Lerma, and L F Hoyos-Reyes. Minimax Control of Discrete-Time Stochastic Systems. *SIAM Journal on Control and Optimization*, 41(5):1626–1659, 2003.

[26] V Goyal and J Grand-Clement. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

[27] J Grand-Clément and M Petrik. On the convex formulations of robust Markov decision processes. *arXiv preprint arXiv:2209.10187*, 2022.

[28] O Hernández-Lerma and J B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.

[29] C P Ho, M Petrik, and W Wiesemann. Partial policy iteration for L1-robust Markov decision processes. *Journal of Machine Learning Research*, 22:1–46, 2021.

[30] C P Ho, M Petrik, and W Wiesemann. Robust phi-divergence MDPs. *arXiv preprint arXiv:2205.14202*, 2022.

[31] Q Huang, Q-S Jia, and X Guan. Robust Scheduling of EV Charging Load With Uncertain Wind Power Integration. *IEEE Transactions on Smart Grid*, 9(2):1043–1054, 2018.

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

[33] R Jiang and Y Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158:291–327, 2016.

[34] R Jiang and Y Guan. Risk-Averse Two-Stage Stochastic Program with Distributional Ambiguity. *Operations Research*, 66(5):1390–1405, 2018.

[35] A D Kara and S Yuksel. Robustness to Incorrect System Models in Stochastic Control. *SIAM Journal on Control and Optimization*, 58(2):1144–1182, 2020.

[36] D L Kaufman and A J Schaefer. Robust modified policy iteration. *INFORMS Journal on Computing*, 25(3):396–410, 2013.

[37] D Klabjan, D Simchi-Levi, and M Song. Robust Stochastic Lot-Sizing by Means of Histograms. *Production and Operations Management*, 22(3):691–710, 2013.

[38] M J Kochenderfer and J P Chryssanthacopoulos. Robust airborne collision avoidance through dyanmic programming. Project Report ATC-371, Lincoln Laboratory, MIT, Lexington, MA, USA, January 2011.

[39] K Kuratowski. *Topology*, volume 1. Elsevier, 2014.

[40] H-J Langen. Convergence of dynamic programming models. *Mathematics of Operations Reearch*, 6(4):493–512, 1981.

[41] F Luo and S Mehrotra. Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Letters*, 14:2565–2594, 2020.

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

[43] J Mardia, J Jiao, E Tanczos, R A Nowak, and T Weissman. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9:813–850, 2020.

[44] C McDiarmid. On the method of bounded differences. In J Siemons, editor, *LMS Lecture Notes Series*, volume 141, pages 148–188, San Mateo, CA, USA, 1989. Morgan Kaufmann.

[45] H Nakao, R Jiang, and S Shen. Distributionally Robust Partially Observable Markov Decision Process with Moment-Based Ambiguity. *SIAM Journal on Optimization*, 31(1):461–488, 2019.

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

[47] L Pardo. *Statistical Inference Based on Divergence Measures*. Chapman & Hall/CRC, Boca Raton, FL, USA, 2006.

[48] I R Petersen, M R James, and P Dupuis. Minimax optimal control of stochastic uncertain systems with relative entropy constraints. *IEEE Transactions on Automatic Control*, 45(3):398–412, 2000.

[49] E Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

[50] M L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, USA, 2014.

[51] H Rahimian and S Mehrotra. Distributionally Robust Optimization: A Review. *https://arxiv.org/abs/1908.05659*, 2019.

[52] A Rajeswaran, I Mordatch, and V Kumar. A game theoretic framework for model based Reinforcement Learning. In *International Conference on Machine Learning*, volume 119, pages 7953–7963, Virtual, 2020. PMLR.

[53] P Rusmevichientong and H Topaloglu. Assortment Optimization in Revenue Management Under the Multinomial Logit Choice Model. *Operations Research*, 60(4):865–882, 2012.

[54] J K Satia and R E Lave, Jr. Markov decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, 1973.

[55] H E Scarf. The min-max solution of an inventory problem. Technical Report P-910, The RAND Corporation, June 1957.

[56] R Serfozo. Convergence of Lebesgue integrals with varying measures. *Sankhya: The Indian Journal of Statistics, Series A*, 44(3):380–402, 1982.

[57] A Shapiro. Distributionally robust optimal control and mdp modeling. *Operations Research Letters*, 49(5):809–814, 2021.

[58] A Shapiro. Distributionally robust modeling of optimal control. *Operations Research Letters*, 50(5):561–567, 2022.

[59] A Shapiro and S Ahmed. On a class of minimax stochastic programs. *SIAM Journal on Optimization*, 14(4):1237–1249, 2004.

[60] T Sutter, B P G Van Parys, and D Kuhn. A general framework for optimal data-driven optimization. https://arxiv.org/pdf/2010.06606.pdf, 2021.

[61] A Szulga. On minimal metrics in the space of random variables. *Theory of Probability and its Applications*, 27(2):424–430, 1983.

[62] I Tzortis, C D Charalambous, and T Charalambous. Dynamic programming subject to total variation distance ambiguity. *SIAM Journal on Control and Optimization*, 53(4):2040–2075, 2015.

[63] I Tzortis, C D Charalambous, and T Charalambous. Infinite horizon average cost dynamic programming subject to total variation distance ambiguity. *SIAM Journal on Control and Optimization*, 57(4):2843–2872, 2019.

[64] B P G Van Parys, P M Esfahani, and D Kuhn. From data to decisions: distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

[65] C Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[66] J Wang, R Gao, and H Zha. Reliable off-policy evaluation for reinforcement learning. *Operations Research*, 2022.

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

[68] W Wiesemann, D Kuhn, and M Sim. Distributionally robust convex optimization. *Operations Research*, 62:1358–1376, 2014.

[69] I Yang. A Convex Optimization Approach to Distributionally robust Markov decision processes with Wasserstein distance. *IEEE Control Systems Letters*, 1(1):164–169, 2017.

[70] I Yang. Wasserstein Distributionally Robust Stochastic Control: A Data-Driven Approach. *IEEE Transactions on Automatic Control*, 66(8):3863–3870, 2021.

[71] M-C Yue, D Kuhn, and W Wiesemann. On linear optimization over Wasserstein balls. *Mathematical Programming*, pages 1–16, 2021.

[72] Y Zhang. *Robust optimal control for medical treatment decisions*. PhD thesis, North Carolina State University, Raleigh, NC, USA, 2014.

[73] C Zhao and Y Guan. Data-driven risk-averse stochastic optimization with Wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018.