arXiv:2308.05520v1 [math.OC] 10 Aug 2023

BOUNDING THE DIFFERENCE BETWEEN THE VALUES OF ROBUST AND NON-ROBUST MARKOV DECISION PROBLEMS

ARIEL NEUFELD[1], JULIAN SESTER[2]

August 11, 2023

[1]*NTU Singapore, Division of Mathematical Sciences,*
21 Nanyang Link, Singapore 637371.
[2]*National University of Singapore, Department of Mathematics,*
21 Lower Kent Ridge Road, 119077.

Abstract. In this note we provide an upper bound for the difference between the value function of a distributionally robust Markov decision problem and the value function of a non-robust Markov decision problem, where the ambiguity set of probability kernels of the distributionally robust Markov decision process is described by a Wasserstein-ball around some reference kernel whereas the non-robust Markov decision process behaves according to a fixed probability kernel contained in the ambiguity set. Our derived upper bound for the difference between the value functions is dimension-free and depends linearly on the radius of the Wasserstein-ball.

Keywords: Markov Decision Process, Wasserstein Uncertainty, Distributionally Robust Optimization, Reinforcement Learning

1. Introduction

Markov decision processes enable to model non-deterministic interactions between an agent and its environment within a tractable stochastic framework. At each time t an agent observes the current state and takes an action which leads to an immediate reward. The goal of the agent then is to optimize its expected cumulative reward. Mathematically, Markov decision problems are solved based on a dynamic programming principle, whose framework builds the fundament of many reinforcement learning algorithms such as, e.g., the Q-learning algorithm. We refer to [4], [9], [20], [21] for the theory of Markov decision processes and to [1], [5], [6], [11], [10], [12], [16], [24], [27] for their applications, especially in the field of reinforcement learning.

In the classical setup for Markov decision problems, the transition kernel describing the transition probabilities of the underlying Markov decision processes is given. Economically, this means that the agent possesses the knowledge of the true distribution of the underlying process, an assumption which typically cannot be justified in practice. To address this issue, academics have recently introduced a robust version of the Markov decision problem accounting for a possible misspecification of the assumed underlying probability kernel that describes the dynamics of the state process. Typically, one assumes that the agent possesses a good guess of the true but to the agent unknown probability kernel, but due to her uncertainty decides to consider the worst case among all laws which lie within a ball of certain radius around the estimated probability kernel with respect to some distance, e.g. the Wasserstein distance or the KL-distance. We refer to [2], [3], [7], [8], [13], [14], [15], [17], [18], [19], [22], [23], [25], [26], [28], [29], [30], [31], and [32] for robust Markov decision problems and corresponding reinforcement learning based algorithms to solve them. In this note, the goal is to analyze the difference between the value function of the corresponding Markov decision problem with respect to the true (but to the agent unknown) probability kernel and the one of the robust Markov decision problem defined with respect to some Wasserstein-ball around the by the agent estimated transition kernel. Note that the estimated transition kernel does not necessarily need to coincide with the true probability kernel, however we assume that the agent's guess is good enough that the true probability kernel lies within the Wasserstein-ball around the estimated probability kernel.

Under some mild assumptions, we obtain in Theorem 3.1 an *explicit* upper bound for the difference between the value function of the robust and the non-robust Markov decision problem which only depends on the radius ε of the Wasserstein-ball, the discount factor α, and the Lipschitz constants of the reward function and the true transition kernel. In particular, we obtain that the difference of the two value functions only grows *at most linearly in the radius* ε and *does not depend on the dimensions of the underlying state and action space.*

The remainder of this note is as follows. In Section 2 we introduce the underlying setting which is used to derive our main result reported in Section 3. The proof of the main result and auxiliary results necessary for the proof are reported in Section 4.

2. Setting

We first present the underlying setting to define both robust and non-robust Markov decision processes which we then use to compare their respective value functions.

2.1. **Setting.** As state space we consider a closed subset $\mathcal{X} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, equipped with its Borel σ-field $\mathcal{F}_\mathcal{X}$, which we use to define the infinite Cartesian product

$$\Omega := \mathcal{X}^{\mathbb{N}_0} = \mathcal{X} \times \mathcal{X} \times \cdots$$

and the σ-field $\mathcal{F} := \mathcal{F}_\mathcal{X} \otimes \mathcal{F}_\mathcal{X} \otimes \cdots$. For any $q \in \mathbb{N}$, we denote by $\mathcal{M}_1^q(\mathcal{X})$ the set of probability measures on $\mathcal{X}$ with finite q-moments and write $\mathcal{M}_1(\mathcal{X}) := \mathcal{M}_1^1(\mathcal{X})$ for brevity. We define on Ω the infinite horizon stochastic process $(X_t)_{t\in\mathbb{N}_0}$ via the canonical process $X_t((\omega_0, \omega_1, \ldots, \omega_t, \ldots)) := \omega_t$ for $(\omega_0, \omega_1, \ldots, \omega_t, \ldots) \in \Omega$, $t \in \mathbb{N}_0$.

To define the set of controls (also called actions) we fix a compact set $A \subseteq \mathbb{R}^m$ for some $m \in \mathbb{N}$, and set

$$\begin{aligned}\mathcal{A} :&= \{\mathbf{a} = (a_t)_{t\in\mathbb{N}_0} \mid (a_t)_{t\in\mathbb{N}_0} : \Omega \to A;\ a_t \text{ is } \sigma(X_t)\text{-measurable for all } t \in \mathbb{N}_0\} \\ &= \left\{(a_t(X_t))_{t\in\mathbb{N}_0} \;\middle|\; a_t : \mathcal{X} \to A \text{ Borel measurable for all } t \in \mathbb{N}_0\right\}.\end{aligned}$$

Next, we define the q-Wasserstein-distance $d_{W_q}(\cdot,\cdot)$ for some $q \in \mathbb{N}$. For any $\mathbb{P}_1, \mathbb{P}_2 \in \mathcal{M}_1^q(\mathcal{X})$ let $d_{W_q}(\mathbb{P}_1, \mathbb{P}_2)$ be defined as

$$d_{W_q}(\mathbb{P}_1, \mathbb{P}_2) := \left(\inf_{\pi\in\Pi(\mathbb{P}_1,\mathbb{P}_2)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^q \mathrm{d}\pi(x,y)\right)^{1/q},$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$, and where $\Pi(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of joint distributions of $\mathbb{P}_1$ and $\mathbb{P}_2$. Moreover, we denote by τ_q the Wasserstein q - topology induced by the convergence w.r.t. d_{W_q}.

To define an ambiguity set of probability kernels, we impose the following standing assumption on a reference probability kernel.

Standing Assumption 2.1 (Assumption on the center of the ambiguity set)**.** *Let* $q \in \mathbb{N}$. *Let*

$$\mathcal{X} \times A \ni (x,a) \mapsto \widehat{\mathbb{P}}(x,a) \in (\mathcal{M}_1(\mathcal{X}), \tau_q)$$

be continuous with finite q-th moments.

Then, we define as ambiguity set of probability kernels

$$\mathcal{X} \times A \ni (x,a) \twoheadrightarrow \mathcal{P}(x,a) := \mathcal{B}_\varepsilon^{(q)}\left(\widehat{\mathbb{P}}(x,a)\right) := \left\{\mathbb{P} \in \mathcal{M}_1(\mathcal{X}) \;\middle|\; d_{W_q}(\mathbb{P}, \widehat{\mathbb{P}}(x,a)) \le \varepsilon\right\}, \tag{2.1}$$

where $\mathcal{B}_\varepsilon^{(q)}\left(\widehat{\mathbb{P}}(x,a)\right)$ denotes the q-Wasserstein-ball (also called Wasserstein-ball of order q) with ε-radius and center $\widehat{\mathbb{P}}(x,a)$.

Under these assumptions we define for every $x \in \mathcal{X}, \mathbf{a} \in \mathcal{A}$ the set of admissible measures on $(\Omega, \mathcal{F})$ by

$$\begin{aligned}\mathfrak{P}_{x,\mathbf{a}} := \Big\{\delta_x \otimes \mathbb{P}_0 \otimes \mathbb{P}_1 \otimes \cdots \;\Big|\; & \text{ for all } t \in \mathbb{N}_0:\ \mathbb{P}_t : \mathcal{X} \to \mathcal{M}_1(\mathcal{X}) \text{ Borel-measurable,} \\ & \text{ and } \mathbb{P}_t(\omega_t) \in \mathcal{P}(\omega_t, a_t(\omega_t)) \text{ for all } \omega_t \in \mathcal{X}\Big\},\end{aligned}$$

where the notation $\mathbb{P} = \delta_x \otimes \mathbb{P}_0 \otimes \mathbb{P}_1 \otimes \cdots \in \mathfrak{P}_{x,\mathbf{a}}$ abbreviates

$$\mathbb{P}(B) := \int_\mathcal{X} \cdots \int_\mathcal{X} \cdots \mathbb{1}_B\left((\omega_t)_{t\in\mathbb{N}_0}\right) \cdots \mathbb{P}_{t-1}(\omega_{t-1}; \mathrm{d}\omega_t) \cdots \mathbb{P}_0(\omega_0; \mathrm{d}\omega_1)\delta_x(\mathrm{d}\omega_0), \qquad B \in \mathcal{F}.$$

2.2. **Problem Formulation.** Let $r : \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ be some *reward function.* We assume from now on that it fulfils the following assumptions.

Standard Assumption 2.2 (Assumptions on the reward function and the discount factor)**.**

(i) *The map*

$$\mathcal{X} \times A \times \mathcal{X} \ni (x_0, a, x_1) \mapsto r(x_0, a, x_1) \in \mathbb{R}$$

is Lipschitz continuous with constant $L_r > 0$.

(ii) *There exists some* $C_r \geq 1$ *such that for all* $x_0, x_1 \in \mathcal{X}$, $a \in A$

$$|r(x_0, a, x_1)| \leq C_r. \tag{2.2}$$

(iii) *We fix an associated* discount factor $\alpha < 1$ *which satisfies*

$$0 < \alpha < 1.$$

Our goal is to compare the *value* of the robust Markov decision problem with the *value* of the non-robust Markov decision problem. To define the robust value function, for every initial value $x \in \mathcal{X}$, one maximizes the expected value of $\sum_{t=0}^{\infty} \alpha^t r(X_t, a_t, X_{t+1})$ under the worst case measure from $\mathfrak{P}_{x,\mathbf{a}}$ over all possible actions $\mathbf{a} \in \mathcal{A}$. More precisely, we introduce the robust value function by

$$\mathcal{X} \ni x \mapsto V(x) := \sup_{\mathbf{a} \in \mathcal{A}} \inf_{\mathbb{P} \in \mathfrak{P}_{x,\mathbf{a}}} \left(\mathbb{E}_{\mathbb{P}} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, a_t, X_{t+1}) \right] \right). \tag{2.3}$$

To define the non-robust value function under the true but to the agent unknown probability kernel $\mathbb{P}^{\text{true}}$ contained in the ambiguity set $\mathcal{P}$, we impose the following assumptions on $\mathbb{P}^{\text{true}}$.

Standing Assumption 2.3. *Fix* $q \in \mathbb{N}$. *The probability kernel* $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{true}}(x, a) \in \mathcal{P}(x, a)$ *is* L_P*-Lipschitz with constant*

$$0 \leq \alpha \cdot L_P < 1, \tag{2.4}$$

where $0 < \alpha < 1$ *is defined in Assumption 2.2 (iii), i.e., we have*

$$d_{W_q}\left(\mathbb{P}^{\text{true}}(x, a), \mathbb{P}^{\text{true}}(x', a')\right) \leq L_P \left(\|x - x'\| + \|a - a'\|\right) \text{ for all } x, x' \in \mathcal{X},\ a, a' \in A. \tag{2.5}$$

Then, we introduce the non-robust value function under the true (but to the agent unknown) transition kernel by

$$\mathcal{X} \ni x \mapsto V^{\text{true}}(x) := \sup_{\mathbf{a} \in \mathcal{A}} \left(\mathbb{E}_{\mathbb{P}^{\text{true}}_{x,\mathbf{a}}} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, a_t, X_{t+1}) \right] \right), \tag{2.6}$$

where we denote for any $x \in \mathcal{X}$, $\mathbf{a} \in \mathcal{A}$,

$$\mathbb{P}^{\text{true}}_{x,\mathbf{a}} := \delta_x \otimes \mathbb{P}^{\text{true}} \otimes \mathbb{P}^{\text{true}} \otimes \mathbb{P}^{\text{true}} \otimes \mathbb{P}^{\text{true}} \cdots \in \mathcal{M}_1(\Omega).$$

Note that Assumption 2.1–2.3 ensures that the dynamic programming principle holds for both the robust and non-robust Markov decision problem, see [18, Theorem 2.7].

3. Main Result

As a main result we establish a bound on the difference between the value function of the Markov decision process with fixed reference measure defined in (2.6), and the value function of the robust Markov decision process defined in (2.3).

Theorem 3.1. *Let Assumption 2.1, Assumption 2.2, and Assumption 2.3 hold true. Then, for any* $x_0 \in \mathcal{X}$ *we have*

$$0 \leq V^{\text{true}}(x_0) - V(x_0) \leq 2L_r\varepsilon\,(1 + \alpha) \sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} (L_P)^j < \infty. \tag{3.1}$$

We highlight that the upper bound from (3.1) depends only on ε, α, and the Lipschitz-constants L_r and L_P. In particular, the upper bound depends linearly on the radius ε of the Wasserstein-ball and is independent of the current state x_0 and the dimensions d and m of the state and action space, respectively.

Remark 3.2. *The assertion from Theorem 3.1 also carries over to the case of autocorrelated time series where one assumes that the past $h \in \mathbb{N} \cap [2, \infty)$ values of a time series $(Y_t)_{t \in \{-h, \cdots -1, 0, 1 \dots\}}$ taking values in some closed subset $\mathcal{Y}$ of $\mathbb{R}^D$ for some $D \in \mathbb{N}$ may have an influence on the next value. This can be modeled by defining the state process $X_t := (Y_{t-h+1}, \dots, Y_t) \in \mathcal{Y}^h =: \mathcal{X}$, $t \in \mathbb{N}_0$. In this setting, the subsequent state $X_{t+1} = (Y_{t-h+2}, \dots, Y_{t+1})$ shares $h-1$ components with the preceding state $X_t = (Y_{t-h+1}, \dots, Y_t)$ and uncertainty is only inherent in the last component Y_{t+1}. Thus, we consider a reference kernel of the form $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{true}}(x, a) = \delta_{\pi(x)} \otimes \widetilde{\mathbb{P}}^{\text{true}}(x, a) \in \mathcal{M}_1(\mathcal{X})$, where $\widetilde{\mathbb{P}}^{\text{true}}(x, a) \in \mathcal{M}_1(\mathcal{Y})$ and $\mathcal{X} \ni (x_1, \dots, x_h) \mapsto \pi(x) := (x_2, \dots, x_h)$ denotes the projection on the last $h-1$ components. In this setting, for $q \in \mathbb{N}$ and $\varepsilon > 0$, the ambiguity set is given by*

$$\mathcal{X} \times A \ni (x, a) \twoheadrightarrow \mathcal{P}(x, a) := \Big\{ \mathbb{P} \in \mathcal{M}_1(\mathcal{X}) \ \ s.t. \\ \mathbb{P} = \delta_{\pi(x)} \otimes \widetilde{\mathbb{P}} \text{ for some } \widetilde{\mathbb{P}} \in \mathcal{M}_1(\mathcal{Y}) \text{ with } W_q(\widetilde{\mathbb{P}}, \widetilde{\mathbb{P}}^{\text{true}}(x, a)) \leq \varepsilon \Big\}.$$

The described setting is discussed in more detail in [18, Section 3.3] *or* [17, Section 2.2]. *Typical applications can be found in finance and include portfolio optimization, compare* [18, Section 4].

4. Proof of the main result

In Section 4.1 we provide several auxiliary lemmas which are necessary to establish the proof of Theorem 3.1 reported in Section 4.2.

4.1. Auxiliary Results.

Lemma 4.1. *Let $q \in \mathbb{N}$. Let $r : \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ satisfy Assumption 2.2. Let $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{true}}(x, a)(\mathrm{d}X_1^{x,a}) \in (\mathcal{M}_1^q(\mathcal{X}), \tau_q)$ satisfy Assumption 2.3. For any*[1] *$v \in C_b(\mathcal{X}, \mathbb{R})$ define*

$$\mathcal{T}^{\text{true}}(x_0) := \sup_{a \in A} \mathbb{E}_{\mathbb{P}^{\text{true}}(x_0, a)} \left[r(x_0, a, X_1^{x_0, a}) + \alpha v(X_1^{x_0, a}) \right], \qquad x_0 \in \mathcal{X}. \tag{4.1}$$

Then, for any $v \in C_b(\mathcal{X}, \mathbb{R})$ being L_r-Lipschitz, $n \in \mathbb{N}$, $x_0, x_0' \in \mathcal{X}$, we have

$$\left| \left(\mathcal{T}^{\text{true}}\right)^n v(x_0) - \left(\mathcal{T}^{\text{true}}\right)^n v(x_0') \right| \leq L_r \left(1 + L_P (1 + \alpha) \sum_{i=0}^{n-1} \alpha^i L_P^i \right) \|x_0 - x_0'\|. \tag{4.2}$$

Proof. For any $x_0, x_0' \in \mathcal{X}$, $a \in A$ let $\Pi^{\text{true}}_{x_0, a, x_0'}(\mathrm{d}X_1^{x_0, a}, \mathrm{d}X_1^{x_0', a}) \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ denote an optimal coupling between $\mathbb{P}^{\text{true}}(x_0, a)$ and $\mathbb{P}^{\text{true}}(x_0', a)$ w.r.t. d_{W_1}, i.e.,

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{X}} \left\| X_1^{x_0, a} - X_1^{x_0', a} \right\| \Pi^{\text{true}}_{x_0, a, x_0'}(\mathrm{d}X_1^{x_0, a}, \mathrm{d}X_1^{x_0', a}) &= d_{W_1}(\mathbb{P}^{\text{true}}(x_0, a), \mathbb{P}^{\text{true}}(x_0', a)) \\ &\leq d_{W_q}(\mathbb{P}^{\text{true}}(x_0, a), \mathbb{P}^{\text{true}}(x_0', a)). \end{aligned} \tag{4.3}$$

[1] We denote here and in the following by $C_b(\mathcal{X}, \mathbb{R})$ the set of continuous and bounded functions from $\mathcal{X}$ to $\mathbb{R}$.

We prove the claim by induction. We start with the base case $n = 1$, and compute by using the Lipschitz continuity of the functions r and v and of $\mathbb{P}^{\text{true}}$ that

$$\begin{aligned}
&\left| \left(\mathcal{T}^{\text{true}}\right) v(x_0) - \left(\mathcal{T}^{\text{true}}\right) v(x_0') \right| \\
&= \left| \sup_{a\in A} \mathbb{E}_{\mathbb{P}^{\text{true}}(x_0,a)} \left[r(x_0, a, X_1^{x_0,a}) + \alpha v(X_1^{x_0,a})\right] - \sup_{a\in A} \mathbb{E}_{\mathbb{P}^{\text{true}}(x_0',a)} \left[r(x_0', a, X_1^{x_0',a}) + \alpha v(X_1^{x_0',a})\right] \right| \\
&\leq \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left| r(x_0, a, X_1^{x_0,a}) + \alpha v(X_1^{x_0,a}) - r(x_0', a, X_1^{x_0',a}) - \alpha v(X_1^{x_0',a}) \right| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}) \\
&\leq L_r\|x_0 - x_0'\| + L_r(1+\alpha) \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left\| X_1^{x_0,a} - X_1^{x_0',a} \right\| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}) \\
&\leq L_r\|x_0 - x_0'\| + L_r(1+\alpha) \sup_{a\in A} d_{W_q}\left(\mathbb{P}^{\text{true}}(x_0,a), \mathbb{P}^{\text{true}}(x_0',a)\right) \\
&\leq L_r\|x_0 - x_0'\| + L_r(1+\alpha) L_P \|x_0 - x_0'\| \\
&= L_r\big(1 + (1+\alpha)L_P\big)\|x_0 - x_0'\|.
\end{aligned}$$

We continue with the induction step. Hence, let $n \in \mathbb{N} \cap [2, \infty)$ be arbitrary and assume that (4.2) holds for $n - 1$. Then, we compute

(4.4)

$$\begin{aligned}
&\left| \left(\mathcal{T}^{\text{true}}\right)^n v(x_0) - \left(\mathcal{T}^{\text{true}}\right)^n v(x_0') \right| \\
&\leq \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \Big| r(x_0, a, X_1^{x_0,a}) + \alpha(\mathcal{T}^{\text{true}})^{n-1} v(X_1^{x_0,a}) \\
&\qquad\qquad - r(x_0', a, X_1^{x_0',a}) - \alpha(\mathcal{T}^{\text{true}})^{n-1} v(X_1^{x_0',a}) \Big| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}) \\
&\leq L_r\|x_0 - x_0'\| + L_r \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left\| X_1^{x_0,a} - X_1^{x_0',a} \right\| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}) \\
&\qquad + \alpha \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left| (\mathcal{T}^{\text{true}})^{n-1} v(X_1^{x_0,a}) - (\mathcal{T}^{\text{true}})^{n-1} v(X_1^{x_0',a}) \right| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}).
\end{aligned}$$

Applying the induction hypothesis to (4.4) therefore yields

$$\begin{aligned}
&\left| \left(\mathcal{T}^{\text{true}}\right)^n v(x_0) - \left(\mathcal{T}^{\text{true}}\right)^n v(x_0') \right| \\
&\leq L_r\|x_0 - x_0'\| + L_r \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left\| X_1^{x_0,a} - X_1^{x_0',a} \right\| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}) \\
&\qquad + \alpha L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^i L_P^i \right) \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left\| X_1^{x_0,a} - X_1^{x_0',a} \right\| \Pi^{\text{true}}_{x_0,a,x_0'}(\mathrm{d}X_1^{x_0,a}, \mathrm{d}X_1^{x_0',a}) \\
&\leq L_r\|x_0 - x_0'\| + L_r \cdot L_P \|x_0 - x_0'\| + \alpha L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^i L_P^i \right) L_P \|x_0 - x_0'\| \\
&= L_r \left(1 + (1+\alpha)L_P + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^{i+1} L_P^{i+1} \right) \|x_0 - x_0'\| \\
&= L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-1} \alpha^i L_P^i \right) \|x_0 - x_0'\|.
\end{aligned}$$

□

Lemma 4.2. *Let Assumption 2.1, Assumption 2.2, and Assumption 2.3 hold true. Moreover, let*

$$\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{wc}}(x, a) \in \mathcal{P}(x, a)$$

denote another probability kernel contained in $\mathcal{P}(x,a)$ *for each* $x, a, \in \mathcal{X} \times A$*. Furthermore, for any* $v \in C_b(\mathcal{X}, \mathbb{R})$ *define*

$$\mathcal{T}^{\mathrm{wc}} v(x_0) := \mathbb{E}_{\mathbb{P}^{\mathrm{wc}}(x_0,a)}\left[r(x_0, a, X_1^{\mathrm{wc}}) + \alpha v(X_1^{\mathrm{wc}})\right], \qquad x_0 \in \mathcal{X}. \tag{4.5}$$

Then, for any $v \in C_b(\mathcal{X}, \mathbb{R})$ *being* L_r*-Lipschitz,* $n \in \mathbb{N}$*,* $x_0 \in \mathcal{X}$*, we have*

$$\left|\left(\mathcal{T}^{\mathrm{wc}}\right)^n v(x_0) - \left(\mathcal{T}^{\mathrm{true}}\right)^n v(x_0)\right| \leq 2L_r\varepsilon\,(1+\alpha)\sum_{i=0}^{n-1}\alpha^i\sum_{j=0}^{i}(L_P)^j, \tag{4.6}$$

where $\mathcal{T}^{\mathrm{true}}$ *is defined in* (4.1)*.*

Proof. For any $x_0 \in \mathcal{X}$, $a \in A$, let $\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}}) \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ denote an optimal coupling between and $\mathbb{P}^{\mathrm{wc}}(x_0,a)$ and $\mathbb{P}^{\mathrm{true}}(x_0,a)$ w.r.t. d_{W_1}. Then, since both $\mathbb{P}^{\mathrm{wc}}(x_0,a), \mathbb{P}^{\mathrm{true}}(x_0,a) \in \mathcal{B}_\varepsilon^{(q)}\left(\widehat{\mathbb{P}}(x,a)\right)$ we have

$$\begin{aligned}\int_{\mathcal{X}\times\mathcal{X}}\left\|X_1^{\mathrm{wc}} - X_1^{\mathrm{true}}\right\|\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}}) &= d_{W_1}(\mathbb{P}^{\mathrm{wc}}(x_0,a), \mathbb{P}^{\mathrm{true}}(x_0,a))\\ &\leq d_{W_q}(\mathbb{P}^{\mathrm{wc}}(x_0,a), \mathbb{P}^{\mathrm{true}}(x_0,a)) \leq 2\varepsilon.\end{aligned} \tag{4.7}$$

We prove the claim by induction. To this end, we start with the base case $n = 1$, and compute by using (4.7) and the Lipschitz continuity of r, v, and of $\mathbb{P}^{\mathrm{true}}$ that

$$\begin{aligned}&\left|\left(\mathcal{T}^{\mathrm{wc}}\right) v(x_0) - \left(\mathcal{T}^{\mathrm{true}}\right) v(x_0)\right|\\ &= \left|\sup_{a\in A}\mathbb{E}_{\mathbb{P}^{\mathrm{wc}}(x_0,a)}\left[r(x_0,a,X_1^{\mathrm{wc}}) + \alpha v(X_1^{\mathrm{wc}})\right] - \sup_{a\in A}\mathbb{E}_{\mathbb{P}^{\mathrm{true}}(x_0,a)}\left[r(x_0,a,X_1^{\mathrm{true}}) + \alpha v(X_1^{\mathrm{true}})\right]\right|\\ &\leq \sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}}\left|r(x_0,a,X_1^{\mathrm{wc}}) + \alpha v(X_1^{\mathrm{wc}}) - r(x_0,a,X_1^{\mathrm{true}}) - \alpha v(X_1^{\mathrm{true}})\right|\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}})\\ &\leq L_r(1+\alpha)\sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}}\left\|X_1^{\mathrm{wc}} - X_1^{\mathrm{true}}\right\|\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}})\\ &\leq L_r(1+\alpha)\sup_{a\in A} d_{W_q}(\mathbb{P}^{\mathrm{wc}}(x_0,a), \mathbb{P}^{\mathrm{true}}(x_0,a)) \leq L_r(1+\alpha)\cdot 2\varepsilon.\end{aligned}$$

We continue with the induction step. Therefore, let $n \in \mathbb{N} \cap [2, \infty)$ be arbitrary and assume that (4.6) holds for $n-1$. Then, we compute

$$\begin{aligned}&\left|\left(\mathcal{T}^{\mathrm{wc}}\right)^n v(x_0) - \left(\mathcal{T}^{\mathrm{true}}\right)^n v(x_0)\right|\\ &\leq \sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}}\Big|r(x_0,a,X_1^{\mathrm{wc}}) + \alpha\left(\mathcal{T}^{\mathrm{wc}}\right)^{n-1} v(X_1^{\mathrm{wc}})\\ &\qquad\qquad - r(x_0,a,X_1^{\mathrm{true}}) - \alpha\left(\mathcal{T}^{\mathrm{true}}\right)^{n-1} v(X_1^{\mathrm{true}})\Big|\,\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}})\end{aligned} \tag{4.8}$$

$$\begin{aligned}&\leq \sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}}\left|r(x_0,a,X_1^{\mathrm{wc}}) - r(x_0,a,X_1^{\mathrm{true}})\right|\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}})\\ &\qquad + \alpha\sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}}\left|\left(\mathcal{T}^{\mathrm{true}}\right)^{n-1} v(X_1^{\mathrm{wc}}) - \left(\mathcal{T}^{\mathrm{true}}\right)^{n-1} v(X_1^{\mathrm{true}})\right|\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}}) \qquad (4.9)\\ &\qquad + \alpha\sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}}\left|\left(\mathcal{T}^{\mathrm{wc}}\right)^{n-1} v(X_1^{\mathrm{wc}}) - \left(\mathcal{T}^{\mathrm{true}}\right)^{n-1} v(X_1^{\mathrm{wc}})\right|\Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}}, \mathrm{d}X_1^{\mathrm{true}}). \qquad (4.10)\end{aligned}$$

Applying Lemma 4.1 to (4.9) and the induction hypothesis to (4.10) together with (4.7) therefore yields

$$\begin{aligned}
&\left|\left(\mathcal{T}^{\mathrm{wc}}\right)^n v(x_0)-\left(\mathcal{T}^{\mathrm{true}}\right)^n v(x_0)\right| \\
&\leq L_r \sup_{a\in A}\int_{\mathcal{X}\times\mathcal{X}} \left\|X_1^{\mathrm{true}}-X_1^{\mathrm{wc}}\right\| \Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}},\mathrm{d}X_1^{\mathrm{true}}) \\
&\quad +\alpha L_r\left(1+L_P(1+\alpha)\sum_{i=0}^{n-2}\alpha^i L_P^i\right)\int_{\mathcal{X}\times\mathcal{X}} \left\|X_1^{\mathrm{true}}-X_1^{\mathrm{wc}}\right\| \Pi_{x_0,a}(\mathrm{d}X_1^{\mathrm{wc}},\mathrm{d}X_1^{\mathrm{true}}) \\
&\quad +\alpha\left(2L_r\varepsilon(1+\alpha)\sum_{i=0}^{n-2}\alpha^i\sum_{j=0}^{i}L_P^j\right) \\
&\leq L_r\cdot 2\varepsilon+\alpha L_r\left(1+L_P(1+\alpha)\sum_{i=0}^{n-2}\alpha^i L_P^i\right)2\varepsilon+\alpha\left(L_r\cdot 2\varepsilon(1+\alpha)\sum_{i=0}^{n-2}\alpha^i\sum_{j=0}^{i}L_P^j\right) \\
&=2L_r\varepsilon(1+\alpha)\left(1+\alpha L_P\sum_{i=0}^{n-2}\alpha^i L_P^i+\sum_{i=0}^{n-2}\alpha^{i+1}\sum_{j=0}^{i}L_P^j\right) \\
&=2L_r\varepsilon(1+\alpha)\left(\sum_{i=0}^{n-1}\alpha^i L_P^i+\sum_{i=1}^{n-1}\alpha^i\sum_{j=0}^{i-1}L_P^j\right)=2L_r\varepsilon(1+\alpha)\left(\sum_{i=0}^{n-1}\alpha^i\sum_{j=0}^{i}L_P^j\right).
\end{aligned}$$

□

Lemma 4.3. *Let $0<\alpha<1$ and $L_P\geq 0$ satisfy $\alpha\cdot L_P<1$. Then*

$$\sum_{i=0}^{\infty}\alpha^i\sum_{j=0}^{i}(L_P)^j<\infty. \tag{4.11}$$

Proof. Note that

$$0\leq\sum_{i=0}^{\infty}\alpha^i\sum_{j=0}^{i}(L_P)^j\leq\sum_{i=0}^{\infty}(i+1)\cdot\alpha^i\max\{1,L_P\}^i=:\sum_{i=0}^{\infty}a_i, \tag{4.12}$$

with $a_i=(i+1)\cdot\alpha^i\max\{1,L_P\}^i$. Moreover

$$\frac{a_{i+1}}{a_i}=\frac{(i+2)\cdot\alpha^{i+1}\max\{1,L_P\}^{i+1}}{(i+1)\cdot\alpha^i\max\{1,L_P\}^i}=\frac{i+2}{i+1}\cdot\alpha\cdot\max\{1,L_P\}\rightarrow\alpha\cdot\max\{1,L_P\}<1 \text{ as } i\rightarrow\infty.$$

Hence, d'Alembert's criterion implies that $\sum_{i=0}^{\cdot} a_i$ converges absolutely. Thus, by (4.12), we have $\sum_{i=0}^{\infty}\alpha^i\sum_{j=0}^{i}(L_P)^j<\infty$. □

4.2. **Proof of Theorem 3.1.** First note that as by assumption $\mathbb{P}^{\mathrm{true}}(x,a)\in\mathcal{P}(x,a)$ for all $(x,a)\in\mathcal{X}\times A$, we have

$$0\leq V^{\mathrm{true}}(x_0)-V(x_0) \text{ for all } x_0\in\mathcal{X}.$$

To compute the upper bound, we fix any $v\in C_b(\mathcal{X},\mathbb{R})$ which is L_r-Lipschitz and we define the operator $\mathcal{T}^{\mathrm{true}}$ by (4.1). Then, by [18, Theorem 2.7 (ii)], we have

$$V^{\mathrm{true}}(x_0)=\lim_{n\rightarrow\infty}\left(\mathcal{T}^{\mathrm{true}}\right)^n v(x_0),\qquad V(x_0)=\lim_{n\rightarrow\infty}(\mathcal{T})^n v(x_0) \tag{4.13}$$

for all $x_0\in\mathcal{X}$ and for $\mathcal{T}$ as defined in [18, Equation (8)]. Moreover, by [18, Theorem 2.7 (iii)], there exists a *worst case* transition kernel $\mathcal{X}\times A\ni(x,a)\mapsto\mathbb{P}^{\mathrm{wc}}(x,a)$ with $\mathbb{P}^{\mathrm{wc}}(x,a)\in\mathcal{P}(x,a)$ for all $(x,a)\in\mathcal{X}\times A$ such that, by denoting for any $\mathbf{a}=(a_t)_{t\in\mathbb{N}_0}\in\mathcal{A}$

$$\mathbb{P}^{\mathrm{wc}}_{x_0,\mathbf{a}}:=\delta_{x_0}\otimes\mathbb{P}^{\mathrm{wc}}\otimes\mathbb{P}^{\mathrm{wc}}\otimes\mathbb{P}^{\mathrm{wc}}\otimes\mathbb{P}^{\mathrm{wc}}\cdots\in\mathcal{M}_1(\Omega),$$

we have

$$V(x_0)=\sup_{\mathbf{a}\in\mathcal{A}}\mathbb{E}_{\mathbb{P}^{\mathrm{wc}}_{x_0,\mathbf{a}}}\left[\sum_{t=0}^{\infty}\alpha^t r(X_t,a_t(X_t),X_{t+1})\right]=\lim_{n\rightarrow\infty}(\mathcal{T}^{\mathrm{wc}})^n v(x_0),\qquad x_0\in\mathcal{X}, \tag{4.14}$$

where $\mathcal{T}^{\mathrm{wc}}$ is defined in (4.5). Therefore by (4.13), (4.14), Lemma 4.2, and Lemma 4.3, we have for all $x_0 \in \mathcal{X}$ that

$$\begin{aligned} V^{\mathrm{true}}(x_0) - V(x_0) &= \lim_{n\to\infty} \left(\mathcal{T}^{\mathrm{true}}\right)^n v(x_0) - \lim_{n\to\infty} \left(\mathcal{T}^{\mathrm{wc}}\right)^n v(x_0) \\ &\leq \lim_{n\to\infty} \left| \left(\mathcal{T}^{\mathrm{true}}\right)^n v(x_0) - \left(\mathcal{T}^{\mathrm{wc}}\right)^n v(x_0) \right| \\ &\leq 2L_r\varepsilon\,(1+\alpha) \lim_{n\to\infty} \sum_{i=0}^{n-1} \alpha^i \sum_{j=0}^{i} L_P^j = 2L_r\varepsilon\,(1+\alpha) \sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} L_P^j < \infty. \end{aligned}$$

Acknowledgments

We thank an anonymous referee of the paper [17] who raised a question that led to this note. Moreover, financial support by the MOE AcRF Tier 1 Grant *RG74/21* and by the Nanyang Assistant Professorship Grant (NAP Grant) *Machine Learning based Algorithms in Finance and Insurance* is gratefully acknowledged.

References

[1] Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement learning for mean field games, with applications to economics. *arXiv preprint arXiv:2106.13755*, 2021.
[2] Nicole Bäuerle and Alexander Glauner. Distributionally robust Markov decision processes and their connection to risk measures. *Mathematics of Operations Research*, 2021.
[3] Nicole Bäuerle and Alexander Glauner. Q-learning for distributionally robust Markov decision processes. In *Modern Trends in Controlled Stochastic Processes:*, pages 108–128. Springer, 2021.
[4] Nicole Bäuerle and Ulrich Rieder. *Markov decision processes with applications to finance*. Springer Science & Business Media, 2011.
[5] Jay Cao, Jacky Chen, John Hull, and Zissis Poulos. Deep hedging of derivatives using reinforcement learning. *The Journal of Financial Data Science*, 3(1):10–27, 2021.
[6] Arthur Charpentier, Romuald Elie, and Carl Remlinger. Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38, 2021.
[7] Zhi Chen, Pengqian Yu, and William B Haskell. Distributionally robust optimization for sequential decision-making. *Optimization*, 68(12):2397–2426, 2019.
[8] Laurent El Ghaoui and Arnab Nilim. Robust solutions to Markov decision problems with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.
[9] Eugene A Feinberg and Adam Shwartz. *Handbook of Markov decision processes: methods and applications*, volume 40. Springer Science & Business Media, 2012.
[10] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
[11] Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *The Journal of Machine Learning Research*, 21(1):6742–6804, 2020.
[12] Esther Levin, Roberto Pieraccini, and Wieland Eckert. Using markov decision process for learning dialogue strategies. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 1, pages 201–204. IEEE, 1998.
[13] Mengmeng Li, Tobias Sutter, and Daniel Kuhn. Policy gradient algorithms for robust mdps with non-rectangular uncertainty sets. *arXiv preprint arXiv:2305.19004*, 2023.
[14] Zijian Liu, Qinxun Bai, Jose Blanchet, Perry Dong, Wei Xu, Zhengqing Zhou, and Zhengyuan Zhou. Distributionally robust Q-learning. In *International Conference on Machine Learning*, pages 13623–13643. PMLR, 2022.
[15] Shie Mannor, Ofir Mebel, and Huan Xu. Robust MDPs with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.
[16] Mausam Natarajan and Andrey Kolobov. *Planning with Markov decision processes: An AI perspective*. Springer Nature, 2022.
[17] Ariel Neufeld and Julian Sester. Robust Q-learning algorithm for markov decision processes under wasserstein uncertainty. *arXiv preprint arXiv:2210.00898*, 2022.
[18] Ariel Neufeld, Julian Sester, and Mario Šikić. Markov decision processes under model uncertainty. *Mathematical Finance*, 33(3):618–665, 2023.
[19] Kishan Panaganti and Dileep Kalathil. Sample complexity of robust reinforcement learning with a generative model. In *International Conference on Artificial Intelligence and Statistics*, pages 9582–9602. PMLR, 2022.
[20] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.
[21] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[22] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributional robust batch contextual bandits. *arXiv preprint arXiv:2006.05630*, 2020.

[23] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*, pages 8884–8894. PMLR, 2020.

[24] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[25] Kerem Uğurlu. Robust optimal control using conditional risk mappings in infinite horizon. *Journal of Computational and Applied Mathematics*, 344:275–287, 2018.

[26] Yue Wang and Shaofeng Zou. Policy gradient method for robust reinforcement learning. *arXiv preprint arXiv:2205.07344*, 2022.

[27] Douglas J White. A survey of applications of markov decision processes. *Journal of the operational research society*, 44(11):1073–1096, 1993.

[28] Wolfram Wiesemann, Daniel Kuhn, and Berç Rustem. Robust Markov decision processes. *Mathematics of Operations Research*, 38(1):153–183, 2013.

[29] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations research*, 62(6):1358–1376, 2014.

[30] Huan Xu and Shie Mannor. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

[31] Wenhao Yang, Liangyu Zhang, and Zhihua Zhang. Towards theoretical understandings of robust markov decision processes: Sample complexity and asymptotics. *arXiv preprint arXiv:2105.03863*, 2021.

[32] Zhengqing Zhou, Zhengyuan Zhou, Qinxun Bai, Linhai Qiu, Jose Blanchet, and Peter Glynn. Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3331–3339. PMLR, 2021.