J. Appl. Probab. **62,** 558–571 (2025)
*doi:*10.1017/jpr.2024.88

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

ARIEL NEUFELD,* *Nanyang Technological University Singapore*
JULIAN SESTER,** *National University of Singapore*

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

2020 Mathematics Subject Classification: Primary 90C40
Secondary 60J05

1. Introduction

Markov decision processes enable the modeling of 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 is the foundation of many reinforcement learning algorithms such as, e.g., the Q-learning algorithm. See [5, 10, 25, 26] for the theory of Markov decision processes, and [1, 6, 7, 11, 12, 15, 20, 29, 33] 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

Received 20 August 2023; accepted 30 August 2024.
* Postal address: Division of Mathematical Sciences, 21 Nanyang Link, Singapore 637371. Email: ariel.neufeld@ntu.edu.sg
** Postal address: Department of Mathematics, 21 Lower Kent Ridge Road, Singapore 119077. Email: jul_ses@nus.edu.sg

describes the dynamics of the state process. Typically, we assume 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 Kullback–Leibler distance. See [3, 4, 8, 9, 16–18, 21, 23, 24, 27, 28, 30, 32, 34–37, 39] 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 that 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.

Similar, while not identical, research questions have been studied in [2, 13, 14, 19, 38], mainly focusing on establishing stability results for value functions with respect to the choice of the underlying transition probability. In [19, Theorem 4.2], the author presents a state-dependent bound on the difference between iterations of value functions (obtained via the so-called value iteration algorithm) of two Markov decision processes, implying that these iterations depend continuously on the transition kernels. As a refinement of [19, Theorem 4.2] and also of the related result obtained in [13, Theorem 2.2.8], the result from [38, Theorem 6.2] shows that in a finite time horizon setting the difference between the value functions of two Markov decision processes with different transition probabilities can be bounded by an expression depending on a certain tailored distance between the transition probabilities. In [14], the author proposes a semi-metric for Markov processes which allows us to determine bounds for certain types of linear stochastic optimization problems, cf. [14, Theorem 3]. The authors of [2] study the sensitvity of multi-period stochastic optimization problems over a finite time horizon with respect to the underlying probability distribution in the so-called adapted Wasserstein distance. They show in [2, Theorem 2.4] that the value function of their robust optimization problem, with the corresponding ambiguity set being a Wasserstein ball around a reference measure, can be approximated by the corresponding value function of the non-robust optimization problem defined with respect to the reference measure plus an explicit correction term. Vaguely speaking (as the optimization problem in [2] is technically speaking not comparable to our setting), this is similar to our analysis in the special case where our reference measure coincides with the true measure.

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 used to derive our main result, which is reported in Section 3. The proof of the main result and auxiliary results necessary for the proof are reported in Section 4.

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

2. Setting

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

2.1. Setting

As the 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 $X_t((\omega_0, \omega_1, \dots, \omega_t, \dots)) := \omega_t$ for $(\omega_0, \omega_1, \dots, \omega_t, \dots) \in \Omega$, $t \in \mathbb{N}_0$.

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

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

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 with respect to d_{W_q}.

To define an ambiguity set of probability kernels, we first fix throughout this paper some $q \in \mathbb{N}$ and $\varepsilon > 0$. Then, we define, as an ambiguity set of probability kernels,

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

with respect to some center $\mathcal{X} \times A \ni (x, a) \mapsto \widehat{\mathbb{P}}(x, a) \in \left(\mathcal{M}_1^q(\mathcal{X}), \tau_q\right)$, meaning that $\mathcal{B}_\varepsilon^{(q)}(\widehat{\mathbb{P}}(x, a))$ denotes the q-Wasserstein ball (also called the Wasserstein ball of order q) with radius ε and center $\widehat{\mathbb{P}}(x, a)$.

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

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

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

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

2.1. Problem formulation and standing assumptions

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

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

Assumption 2.1. (Assumptions on the reward function) *The reward function* $r\colon \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ *satisfies the following:*

(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) *If* $\mathcal{X}$ *is unbounded and* $q \in \mathbb{N}$ *defined in* (2.1) *satisfies* $q = 1$, *then we additionally assume that* $\sup_{x_0, x_1 \in \mathcal{X}, a \in A} |r(x_0, a, x_1)| < \infty$.

Note that Assumption 2.1(i) implies that the reward r is bounded whenever $\mathcal{X}$ is bounded.

Next, we impose the following standing assumption on our reference probability kernel modeled by the center of the q-Wasserstein ball.

Assumption 2.2. (Assumption on the center of the ambiguity set.) *Let* $q \in \mathbb{N}$ *be defined in* (2.1). *Then the center* $\mathcal{X} \times A \ni (x, a) \mapsto \widehat{\mathbb{P}}(x, a) \in (\mathcal{M}_1^q(\mathcal{X}), \tau_q)$ *satisfies the following:*

(i) *The map* $\mathcal{X} \times A \ni (x, a) \mapsto \widehat{\mathbb{P}}(x, a) \in (\mathcal{M}_1^q(\mathcal{X}), \tau_q)$ *is continuous.*

(i′) *If the reward function* r *is unbounded, then we assume instead of (i) the stronger assumption that* $\widehat{\mathbb{P}}$ *is Lipschitz continuous, i.e. that there exists* $L_{\widehat{\mathbb{P}}} > 0$ *such that*

$$d_{W_q}(\widehat{\mathbb{P}}(x, a), \widehat{\mathbb{P}}(x', a')) \leq L_{\widehat{\mathbb{P}}}\big(\|x - x'\| + \|a - a'\|\big) \quad \text{for all } x, x' \in \mathcal{X},\ a, a' \in A.$$

Finally, we assume the following on the discount factor $\alpha \in (0, 1)$.

Assumption 2.3. (Assumption on the discount factor) *Let* $q \in \mathbb{N}$, $\varepsilon > 0$ *be defined as in* (2.1), *and* $\mathcal{X} \times A \ni (x, a) \mapsto \widehat{\mathbb{P}}(x, a) \in (\mathcal{M}_1^q(\mathcal{X}), \tau_q)$ *be as defined in Assumption* 2.2. *Then the discount factor* α *satisfies* $0 < \alpha < 1/C_P$, *where* $1 \leq C_P < \infty$ *is defined by*

$$C_P = \begin{cases} \max\left\{1 + \varepsilon + \sup\limits_{a \in A} \inf\limits_{x \in \mathcal{X}} \left\{ \displaystyle\int_{\mathcal{X}} \|z\| \, \widehat{\mathbb{P}}(x, a)(\mathrm{d}z) + L_{\widehat{\mathbb{P}}} \|x\| \right\},\ L_{\widehat{\mathbb{P}}} \right\} & \text{if } r \text{ is unbounded,} \\ 1 & \text{otherwise.} \end{cases}$$

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 $x \in \mathcal{X}$, we maximize 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

$$\mathcal{X} \ni x \mapsto V(x) := \sup_{\mathbf{a} \in \mathcal{A}} \inf_{\mathbb{P} \in \mathfrak{P}_{x,\mathbf{a}}} \left(\mathbb{E}_{\mathbb{P}} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, a_t, X_{t+1}) \right] \right). \tag{2.2}$$

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 $\mathbb{P}^{\text{true}}$.

Assumption 2.4. (Assumptions on the true probability kernel) *Let* $q \in \mathbb{N}$ *be as defined in* (2.1). *Then the true (but unknown) probability kernel* $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{true}}(x, a) \in (\mathcal{M}_1^q(\mathcal{X}), \tau_q)$ *satisfies the following:*

(i) $\mathbb{P}^{\text{true}}(x, a) \in \mathcal{P}(x, a)$ *for all* $(x, a) \in \mathcal{X} \times A$.

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

(ii) $\mathbb{P}^{\text{true}}$ *is* L_P*-Lipschitz with constant* $0 \le L_P < 1/\alpha$*, where* $0 < \alpha < 1$ *is defined in Assumption 2.3, i.e. we have*

$$d_{W_q}(\mathbb{P}^{\text{true}}(x, a), \mathbb{P}^{\text{true}}(x', a')) \le L_P(\|x - x'\| + \|a - a'\|)$$

for all $x, x' \in \mathcal{X}$*,* $a, a' \in A$*.*

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

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

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

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

Note that Assumptions 2.1–2.4 ensure that the dynamic programming principle holds for both the robust and non-robust Markov decision problem; see [23, 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 (2.3), and the value function of the robust Markov decision process defined in (2.2).

Theorem 3.1. *Let all Assumptions 2.1–2.4 hold.*

(i) Then, for any $x_0 \in \mathcal{X}$*,*

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

(ii) Moreover, in the special case that $\mathbb{P}^{\text{true}} = \widehat{\mathbb{P}}$*, for any* $x_0 \in \mathcal{X}$*,*

$$0 \le V^{\text{true}}(x_0) - V(x_0) \le L_r\varepsilon(1 + \alpha) \sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} (L_P)^j < \infty. \tag{3.2}$$

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 3.1. 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 $(Y_t)_{t \in \{-h, \dots, -1, 0, 1, \dots\}}$ 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 $X_t = (Y_{t-h+1}, \dots, Y_t)$

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

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 $\mathcal{X} \ni (x_1, \ldots, x_h) \mapsto \pi(x) := (x_2, \ldots, x_h)$ 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

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

The described setting is discussed in more detail in [23, Section 3.3] or [21, Section 2.2]. Typical applications can be found in finance and include portfolio optimization; cf. [23, Section 4].

Example 3.1. (*Coin toss.*) To illustrate the applicability of Theorem 3.1, we study an example similar to the one provided in [21, Example 4.1]. To this end, we consider an agent who at each time tosses 10 coins and observes the number of heads. Thus, we model the environment by a state space $\mathcal{X} := \{0, \ldots, 10\}$. Prior to the toss, the agent can bet whether in the next toss of 10 coins the sum of heads will be smaller ($a = -1$) or larger ($a = 1$) than the previous toss. She gains \$! if the bet is correct, and in turn has to pay \$1 if it is not (without being rewarded/punished if the sum of heads remains the same). Moreover, the agent can also decide not to bet for the toss (by choosing $a = 0$). We model this via the reward function

$$\mathcal{X} \times A \times \mathcal{X} \ni (x, a, x') \mapsto r(x, a, x') := a\mathbf{1}_{\{x < x'\}} - a\mathbf{1}_{\{x > x'\}},$$

where the possible actions are given by $A := \{-1, 0, 1\}$. The reference measure in this setting assumes a fair coin, and therefore (independent of the state action pair) is a binomial distribution with $n = 10$, $p = 0.5$, i.e.

$$\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{true}}(x, a) = \widehat{\mathbb{P}}(x, a) := \text{Bin}\,(10, 0.5).$$

In the described setting it is easy to see that r is Lipschitz continuous with Lipschitz constant

$$L_r = \left(\max_{\substack{y_0, y_0', x_1, x_1' \in \mathcal{X},\, b, b' \in A \\ (y_0, b, x_1) \neq (y_0', b', x_1')}} \frac{|r(y_0, b, x_1) - r(y_0', b', x_1')|}{\|y_0 - y_0'\| + \|b - b'\| + \|x_1 - x_1'\|} \right) = 1.$$

Moreover, we have $L_P = 0$. In Figure 1 we plot the corresponding upper bound from (3.2) against the difference $V^{\text{true}}(x_0) - V(x_0)$ for different initial values x_0 and different levels of ε used for the computation of V with $\alpha = 0.45$. The value functions are computed using the robust Q-learning algorithm proposed in [21]. The code used can be found at https://github.com/juliansester/MDP_Bound.

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, which is reported in Section 4.2.

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

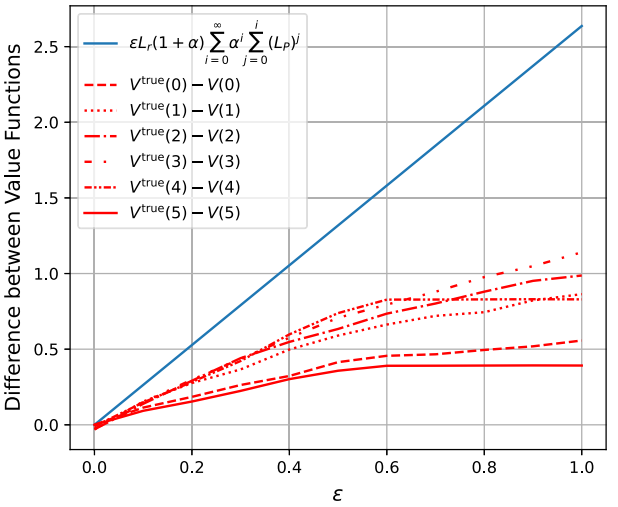

FIGURE 1. The difference between the non-robust and the robust value function compared with the upper bound from (3.2) in the setting described in Example 3.1 with $\varepsilon > 0$ and for different initial values of the Markov decision process. Initial values larger than 5 are omitted due to the setting-specific symmetry $V(x_0) - V^{\text{true}}(x_0) = V(10 - x_0) - V^{\text{true}}(10 - x_0)$ for $x_0 \in \{0, 1, \dots, 10\}$.

4.1. Auxiliary results

Lemma 4.1. *Let* $r\colon \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ *satisfy Assumption* 2.1. *Let* $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\text{true}}(x, a)(\mathrm{d}x_1) \in (\mathcal{M}_1^q(\mathcal{X}), \tau_q)$ *satisfy Assumption* 2.4. *For any* $v \in C_b(\mathcal{X}, \mathbb{R})$, *where here and in the following we denote by* $C_b(\mathcal{X}, \mathbb{R})$ *the set of continuous and bounded functions from* $\mathcal{X}$ *to* $\mathbb{R}$, *define*

$$\mathcal{T}^{\text{true}} v(x_0) := \sup_{a \in A} \int_{\mathcal{X}} (r(x_0, a, x_1) + \alpha v(x_1)) \mathbb{P}^{\text{true}}(x_0, a)(\mathrm{d}x_1), \qquad x_0 \in \mathcal{X}. \tag{4.1}$$

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*

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

Proof. For any $x_0, x_0' \in \mathcal{X}$ and $a \in A$, let $\Pi^{\text{true}}(\mathrm{d}x_1, \mathrm{d}x_1') \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ denote an optimal coupling between $\mathbb{P}^{\text{true}}(x_0, a)$ and $\mathbb{P}^{\text{true}}(x_0', a)$ with respect to d_{W_1}, i.e.

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{X}} \|x_1 - x_1'\| \Pi^{\text{true}}(\mathrm{d}x_1, \mathrm{d}x_1') &= d_{W_1}(\mathbb{P}^{\text{true}}(x_0, a), \mathbb{P}^{\text{true}}(x_0', a)) \\ &\le d_{W_q}(\mathbb{P}^{\text{true}}(x_0, a), \mathbb{P}^{\text{true}}(x_0', a)), \end{aligned}$$

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

where the inequality follows from Hölder's inequality (see, e.g., [31, Remark 6.6]). 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

$$\begin{aligned}
\left|(\mathcal{T}^{\text{true}})v(x_0) - (\mathcal{T}^{\text{true}})v\left(x_0'\right)\right| &= \left| \sup_{a\in A} \int_{\mathcal{X}} (r(x_0, a, x_1) + \alpha v(x_1))\mathbb{P}^{\text{true}}(x_0, a)(\mathrm{d}x_1) \right. \\
&\qquad \left. - \sup_{a\in A} \int_{\mathcal{X}} \left(r\left(x_0', a, x_1'\right) + \alpha v\left(x_1'\right)\right) \mathbb{P}^{\text{true}}\left(x_0', a\right)\left(\mathrm{d}x_1'\right) \right| \\
&\leq \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} |r(x_0, a, x_1) + \alpha v(x_1) - r\left(x_0', a, x_1'\right) - \alpha v\left(x_1'\right)| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right) \\
&\leq L_r\|x_0 - x_0'\| + L_r(1+\alpha) \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_1'\| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right) \\
&\leq L_r\|x_0 - x_0'\| + L_r(1+\alpha) \sup_{a\in A} d_{W_q}\left(\mathbb{P}^{\text{true}}(x_0, a), \mathbb{P}^{\text{true}}\left(x_0', a\right)\right) \\
&\leq L_r\|x_0 - x_0'\| + L_r(1+\alpha)L_P\|x_0 - x_0'\| \\
&= L_r(1 + (1+\alpha)L_P)\|x_0 - x_0'\|.
\end{aligned}$$

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

$$\begin{aligned}
\left|(\mathcal{T}^{\text{true}})^n v(x_0) - (\mathcal{T}^{\text{true}})^n v\left(x_0'\right)\right| &\leq \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left| r(x_0, a, x_1) + \alpha(\mathcal{T}^{\text{true}})^{n-1} v(x_1) \right. \\
&\qquad \left. - r\left(x_0', a, x_1'\right) - \alpha(\mathcal{T}^{\text{true}})^{n-1} v\left(x_1'\right) \right| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right) \\
&\leq L_r\|x_0 - x_0'\| + L_r \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_1'\| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right) \\
&\quad + \alpha \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \left|(\mathcal{T}^{\text{true}})^{n-1} v(x_1) - (\mathcal{T}^{\text{true}})^{n-1} v\left(x_1'\right)\right| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right).
\end{aligned}$$

Applying the induction hypothesis to this therefore yields

$$\begin{aligned}
&\left|(\mathcal{T}^{\text{true}})^n v(x_0) - (\mathcal{T}^{mathrmtrue})^n v\left(x_0'\right)\right| \\
&\leq L_r\|x_0 - x_0'\| + L_r \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_1'\| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right) \\
&\quad + \alpha L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^i L_P^i\right) \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_1'\| \Pi^{\text{true}}\left(\mathrm{d}x_1, \mathrm{d}x_1'\right) \\
&\leq L_r\|x_0 - x_0'\| + L_r \cdot L_P\|x_0 - x_0'\| + \alpha L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^i L_P^i\right) L_P\|x_0 - x_0'\| \\
&= L_r \left(1 + (1+\alpha)L_P + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^{i+1} L_P^{i+1}\right) \|x_0 - x_0'\| \\
&= L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-1} \alpha^i L_P^i\right) \|x_0 - x_0'\|.
\end{aligned}$$

□

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

Lemma 4.2. *Let Assumptions 2.1 and 2.4 hold. Moreover, let* $\mathcal{X} \times A \ni (x, a) \mapsto \mathbb{P}^{\mathrm{wc}}(x, a) \in \mathcal{P}(x, a)$ *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*

$$\mathcal{T}^{\mathrm{wc}} v(x_0) := \sup_{a \in A} \int_{\mathcal{X}} (r(x_0, a, x_1) + \alpha v(x_1)) \mathbb{P}^{\mathrm{wc}}(x_0, a)(\mathrm{d}x_1), \qquad x_0 \in \mathcal{X}. \tag{4.3}$$

(i) Then, for any $v \in C_b(\mathcal{X}, \mathbb{R})$ *that is* L_r*-Lipschitz,* $n \in \mathbb{N}$, *and* $x_0 \in \mathcal{X}$,

$$\left|(\mathcal{T}^{\mathrm{wc}})^n v(x_0) - (\mathcal{T}^{\mathrm{true}})^n v(x_0)\right| \leq 2L_r \varepsilon (1 + \alpha) \sum_{i=0}^{n-1} \alpha^i \sum_{j=0}^{i} (L_P)^j, \tag{4.4}$$

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

(ii) Moreover, in the special case that $\mathbb{P}^{\mathrm{true}} = \widehat{\mathbb{P}}$, *we obtain, for any* $x_0 \in \mathcal{X}$,

$$\left|(\mathcal{T}^{\mathrm{wc}})^n v(x_0) - (\mathcal{T}^{\mathrm{true}})^n v(x_0)\right| \leq L_r \varepsilon (1 + \alpha) \sum_{i=0}^{n-1} \alpha^i \sum_{j=0}^{i} (L_P)^j. \tag{4.5}$$

Proof. (i) For any $x_0 \in \mathcal{X}$ and $a \in A$, let $\Pi(\mathrm{d}x_1, \mathrm{d}x_1') \in \mathcal{M}_1(\mathcal{X} \times \mathcal{X})$ denote an optimal coupling between $\mathbb{P}^{\mathrm{wc}}(x_0, a)$ and $\mathbb{P}^{\mathrm{true}}(x_0, a)$ with respect to d_{W_1}. Then, since both $\mathbb{P}^{\mathrm{wc}}(x_0, a), \mathbb{P}^{\mathrm{true}}(x_0, a) \in \mathcal{B}_\varepsilon^{(q)}(\widehat{\mathbb{P}}(x_0, a))$ we have

$$\begin{aligned} \int_{\mathcal{X} \times \mathcal{X}} \|x_1 - x_1'\| \Pi(\mathrm{d}x_1, \mathrm{d}x_1') &= d_{W_1}(\mathbb{P}^{\mathrm{wc}}(x_0, a), \mathbb{P}^{\mathrm{true}}(x_0, a)) \\ &\leq d_{W_q}(\mathbb{P}^{\mathrm{wc}}(x_0, a), \mathbb{P}^{\mathrm{true}}(x_0, a)) \leq 2\varepsilon, \end{aligned} \tag{4.6}$$

where the first inequality follows from Hölder's inequality (see, e.g., [31, Remark 6.6]). We prove the claim by induction. To this end, we start with the base case $n = 1$, and compute by using (4.6) and the Lipschitz continuity of r, v, and $\mathbb{P}^{\mathrm{true}}$ that

$$\begin{aligned} \left|(\mathcal{T}^{\mathrm{wc}})v(x_0) - (\mathcal{T}^{\mathrm{true}})v(x_0)\right| &= \Big| \sup_{a \in A} \int_{\mathcal{X}} (r(x_0, a, x_1) + \alpha v(x_1)) \mathbb{P}^{\mathrm{wc}}(x_0, a)(\mathrm{d}x_1) \\ &\qquad - \sup_{a \in A} \int_{\mathcal{X}} \big(r\big(x_0, a, x_1'\big) + \alpha v\big(x_1'\big)\big) \mathbb{P}^{\mathrm{true}}(x_0, a)\big(\mathrm{d}x_1'\big) \Big| \\ &\leq \sup_{a \in A} \int_{\mathcal{X} \times \mathcal{X}} |r(x_0, a, x_1) + \alpha v(x_1) - r\big(x_0, a, x_1'\big) - \alpha v\big(x_1'\big)| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \\ &\leq L_r(1 + \alpha) \sup_{a \in A} \int_{\mathcal{X} \times \mathcal{X}} \|x_1 - x_1'\| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \\ &\leq L_r(1 + \alpha) \sup_{a \in A} d_{W_q}(\mathbb{P}^{\mathrm{wc}}(x_0, a), \mathbb{P}^{\mathrm{true}}(x_0, a)) \leq L_r(1 + \alpha) \cdot 2\varepsilon. \end{aligned}$$

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

We continue with the induction step. Therefore, let $n \in \mathbb{N} \cap [2, \infty)$ be arbitrary and assume that (4.4) holds for $n-1$. Then, we compute

$$\begin{aligned}
&\left|(\mathcal{T}^{\mathrm{wc}})^n v(x_0) - (\mathcal{T}^{\mathrm{true}})^n v(x_0)\right| \\
&\quad \le \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \big| r(x_0, a, x_1) + \alpha (\mathcal{T}^{\mathrm{wc}})^{n-1} v(x_1) \\
&\qquad\qquad - r\big(x_0, a, x_1'\big) - \alpha (\mathcal{T}^{\mathrm{true}})^{n-1} v\big(x_1'\big)\big| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \\
&\quad \le \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} |r(x_0, a, x_1) - r\big(x_0, a, x_1'\big)| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \\
&\qquad + \alpha \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \big|(\mathcal{T}^{\mathrm{true}})^{n-1} v(x_1) - (\mathcal{T}^{\mathrm{true}})^{n-1} v\big(x_1'\big)\big| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \qquad (4.7) \\
&\qquad + \alpha \sup_{a\in A} \int_{\mathcal{X}\times\mathcal{X}} \big|(\mathcal{T}^{\mathrm{wc}})^{n-1} v(x_1) - (\mathcal{T}^{\mathrm{true}})^{n-1} v(x_1)\big| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big). \qquad (4.8)
\end{aligned}$$

Applying Lemma 4.1 to (4.7) and the induction hypothesis to (4.8) together with (4.6) therefore yields

$$\begin{aligned}
&\left|(\mathcal{T}^{\mathrm{wc}})^n v(x_0) - (\mathcal{T}^{\mathrm{true}})^n v(x_0)\right| \\
&\quad \le L_r \sup_{\in A} \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_1'\| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \\
&\qquad + \alpha L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^i L_P^i\right) \int_{\mathcal{X}\times\mathcal{X}} \|x_1 - x_1'\| \Pi\big(\mathrm{d}x_1, \mathrm{d}x_1'\big) \\
&\qquad + \alpha \left(2L_r \varepsilon (1+\alpha) \sum_{i=0}^{n-2} \alpha^i \sum_{j=0}^{i} L_P^j\right) \\
&\quad \le L_r \cdot 2\varepsilon + \alpha L_r \left(1 + L_P(1+\alpha) \sum_{i=0}^{n-2} \alpha^i L_P^i\right) 2\varepsilon + \alpha \left(L_r \cdot 2\varepsilon (1+\alpha) \sum_{i=0}^{n-2} \alpha^i \sum_{j=0}^{i} L_P^j\right) \\
&\quad = 2L_r\varepsilon(1+\alpha) \left(1 + \alpha L_P \sum_{i=0}^{n-2} \alpha^i L_P^i + \sum_{i=0}^{n-2} \alpha^{i+1} \sum_{j=0}^{i} L_P^j\right) \\
&\quad = 2L_r\varepsilon(1+\alpha) \left(\sum_{i=0}^{n-1} \alpha^i L_P^i + \sum_{i=1}^{n-1} \alpha^i \sum_{j=0}^{i-1} L_P^j\right) = 2L_r\varepsilon(1+\alpha) \left(\sum_{i=0}^{n-1} \alpha^i \sum_{j=0}^{i} L_P^j\right).
\end{aligned}$$

(ii) In the case $\mathbb{P}^{\mathrm{true}} = \widehat{\mathbb{P}}$ we have, for any $x_0 \in \mathcal{X}$ and $a \in A$ that

$$d_{W_q}(\mathbb{P}^{\mathrm{wc}}(x_0, a), \mathbb{P}^{\mathrm{true}}(x_0, a)) \le \varepsilon, \qquad (4.9)$$

since the ambiguity set $\mathcal{P}(x_0, a)$ is centered around $\mathbb{P}^{\mathrm{true}}(x_0, a) = \widehat{\mathbb{P}}(x_0, a)$. Hence, replacing (4.6) by (4.9) and then following the proof of (i) shows the assertion. □

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

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

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

Proof. Note that

$$0 \leq \sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} (L_P)^j \leq \sum_{i=0}^{\infty} (i+1) \cdot \alpha^i \max\{1, L_P\}^i =: \sum_{i=0}^{\infty} a_i, \tag{4.10}$$

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

$$\frac{a_{i+1}}{a_i} = \frac{(i+2) \cdot \alpha^{i+1} \max\{1, L_P\}^{i+1}}{(i+1) \cdot \alpha^i \max\{1, L_P\}^i} = \frac{i+2}{i+1} \cdot \alpha \cdot \max\{1, L_P\} \to \alpha \cdot \max\{1, L_P\} < 1$$

as $i \to \infty$. Hence, d'Alembert's criterion implies that $\sum_{i=0}^{\cdot} a_i$ converges absolutely. Thus, by (4.10), we have $\sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} (L_P)^j < \infty$. □

Lemma 4.4. *Let Assumptions 2.1–2.3 hold. Then* $\mathcal{P}(x, a) := \mathcal{B}_\varepsilon^{(q)}(\widehat{\mathbb{P}}(x, a))$ *as defined in* (2.1) *satisfies* [23, *Standing Assumption* 2.2] *and the reward function* $r\colon \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ *together with the discount factor* $0 < \alpha < 1$ *satisfy* [23, *Standing Assumption* 2.4]. *As a consequence,* [23, *Theorem* 2.7] *then directly implies that the dynamic programming principle holds for the robust Markov decision problem defined in* (2.2).

Proof. First, if $r\colon \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ is bounded, then Assumptions 2.1–2.3 allow us to use [23, Proposition 3.1], which immediately ensures that the result holds true with respect to $p = 0$ and $C_P = 1$ in the notation of [23, Standing Assumptions 2.2 2.4].

Now, assume for the rest of this proof that $r\colon \mathcal{X} \times A \times \mathcal{X} \to \mathbb{R}$ is unbounded. Then, by Assumption 2.1(ii) we have that $q \in [2, \infty) \cap \mathbb{N}$. In this case, let $p = 1$ in the notation of [23, Standing Assumptions 2.2 and 2.4]. Then our Assumptions 2.1 and 2.3 immediately ensure that [23, Standing Assumption 2.4] holds. Moreover, by our Assumption 2.2, we directly obtain from [22, Proposition 4.1] that [23, Standing Assumption 2.2(i)] holds. Therefore, it remains to verify [23, Standing Assumptions 2.2(ii)]. To that end, let

$$C_P := \max\left\{1 + \varepsilon + \sup_{a \in A} \inf_{x' \in \mathcal{X}} \left\{ \int_{\mathcal{X}} \|z\| \, \widehat{\mathbb{P}}(x', a)(\mathrm{d}z) + L_{\widehat{\mathbb{P}}} \|x'\| \right\}, L_{\widehat{\mathbb{P}}} \right\} < \infty. \tag{4.11}$$

Indeed, note that $C_P < \infty$, as Assumption 2.2 ensures that the map

$$\mathcal{X} \times A \ni (x', a) \mapsto \int_{\mathcal{X}} \|z\| \, \widehat{\mathbb{P}}(x', a)(\mathrm{d}z) + L_{\widehat{\mathbb{P}}} \|x'\| \in [0, \infty)$$

is continuous. This implies that the map

$$A \ni a \mapsto \inf_{x' \in \mathcal{X}} \left\{ \int_{\mathcal{X}} \|z\| \, \widehat{\mathbb{P}}(x', a)(\mathrm{d}z) + L_{\widehat{\mathbb{P}}} \|x'\| \right\} \in [0, \infty)$$

is upper semicontinuous, which in turns ensures that C_P is finite as A is compact. Now, let $(x, a) \in \mathcal{X} \times A$ and $\mathbb{P} \in \mathcal{P}(x, a) = \mathcal{B}_\varepsilon^{(q)}(\widehat{\mathbb{P}}(x, a))$ be arbitrarily chosen. Then by following the calculations in [22, Proof of Proposition 4.1, (6.34)] (with $p = 1$ in the notation of [22]), using the Lipschitz continuity of $\widehat{\mathbb{P}}$ we obtain, for any arbitrary $x' \in \mathcal{X}$, that

$$\int_{\mathcal{X}} 1 + \|y\| \, \mathbb{P}(\mathrm{d}y) \leq 1 + \varepsilon + \int_{\mathcal{X}} \|z\| \, \widehat{\mathbb{P}}(x', a)(\mathrm{d}z) + L_{\widehat{\mathbb{P}}}(\|x'\| + \|x\|).$$

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

Since $x' \in \mathcal{X}$ was arbitrarily chosen, we see from (4.11) that

$$\int_{\mathcal{X}} 1 + \|y\| \, \mathbb{P}(\mathrm{d}y) \le 1 + \varepsilon + \sup_{a \in A} \inf_{x' \in \mathcal{X}} \left\{ \int_{\mathcal{X}} \|z\| \, \widehat{\mathbb{P}}(x', a)(\mathrm{d}z) + L_{\widehat{\mathbb{P}}} \|x'\| \right\} + L_{\widehat{\mathbb{P}}} \|x\| \le C_P(1 + \|x\|),$$

which shows that [23, Standing Assumption 2.2(ii)] indeed holds. □

4.2. Proof of Theorem 3.1

Proof. (i) 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 $0 \le V^{\mathrm{true}}(x_0) - V(x_0)$ 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 Lemma 4.4 and [23, Theorem 2.7(ii)], we have

$$V^{\mathrm{true}}(x_0) = \lim_{n \to \infty} (\mathcal{T}^{\mathrm{true}})^n v(x_0), \qquad V(x_0) = \lim_{n \to \infty} (\mathcal{T})^n v(x_0) \tag{4.12}$$

for all $x_0 \in \mathcal{X}$ and for $\mathcal{T}$ as defined in [23, (8)]. Moreover, by [23, 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

$$V(x_0) = \sup_{\mathbf{a} \in \mathcal{A}} \mathbb{E}_{\mathbb{P}^{\mathrm{wc}}_{x_0, \mathbf{a}}} \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, a_t(X_t), X_{t+1}) \right] = \lim_{n \to \infty} (\mathcal{T}^{\mathrm{wc}})^n v(x_0), \qquad x_0 \in \mathcal{X}, \tag{4.13}$$

where $\mathcal{T}^{\mathrm{wc}}$ is as defined in (4.3). Therefore, by (4.12), (4.13), Lemma 4.2, and Lemma 4.3, we have, for all $x_0 \in \mathcal{X}$,

$$\begin{aligned} V^{\mathrm{true}}(x_0) - V(x_0) &= \lim_{n \to \infty} (\mathcal{T}^{\mathrm{true}})^n v(x_0) - \lim_{n \to \infty} (\mathcal{T}^{\mathrm{wc}})^n v(x_0) \\ &\le \lim_{n \to \infty} \left| (\mathcal{T}^{\mathrm{true}})^n v(x_0) - (\mathcal{T}^{\mathrm{wc}})^n v(x_0) \right| \\ &\le 2L_r \varepsilon (1 + \alpha) \lim_{n \to \infty} \sum_{i=0}^{n-1} \alpha^i \sum_{j=0}^{i} L_P^j \\ &= 2L_r \varepsilon (1 + \alpha) \sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} L_P^j < \infty. \end{aligned} \tag{4.14}$$

(ii) In the case $\mathbb{P}^{\mathrm{true}} = \widehat{\mathbb{P}}$, due to Lemma 4.2(ii), we may use (4.5) and replace the inequality (4.14) by

$$V^{\mathrm{true}}(x_0) - V(x_0) \le L_r \varepsilon (1 + \alpha) \lim_{n \to \infty} \sum_{i=0}^{n-1} \alpha^i \sum_{j=0}^{i} L_P^j = L_r \varepsilon (1 + \alpha) \sum_{i=0}^{\infty} \alpha^i \sum_{j=0}^{i} L_P^j < \infty. \qquad \square$$

Acknowledgements

We thank an anonymous referee of [21] who raised a question that led to this note.

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

Funding information

The first author gratefully acknowledges 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*. The second author gratefully acknowledges financial support by the NUS Start-Up Grant *Tackling model uncertainty in Finance with machine learning*.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

Supplementary material

The supplementary material for this article (Python code for Example 3.1 can be found at https://github.com/juliansester/MDP_Bound.

References

[1] Angiuli, A., Fouque, J.-P. and Lauriere, M. (2021). Reinforcement learning for mean field games, with applications to economics. Preprint, arXiv:2106.13755.
[2] Bartl, D. and Wiesel, J. (2022). Sensitivity of multiperiod optimization problems in adapted Wasserstein distance. Preprint, arXiv:2208.05656.
[3] Bäuerle, N. and Glauner, A. (2021), Q-learning for distributionally robust Markov decision processes. In *Modern Trends in Controlled Stochastic Processes:*, eds A. Piunovsky and Y. Zhang. Springer, New York, pp. 108–128.
[4] Bäuerle, N. and Glauner, A. (2022). Distributionally robust Markov decision processes and their connection to risk measures. *Math. Operat. Res.* **47**, 1757–1780.
[5] Bäuerle, N. and Rieder, U. (2011). *Markov Decision Processes with Applications to Finance*. Springer, New York.
[6] Cao, J., Chen, J., Hull, J. and Poulos, Z. (2021). Deep hedging of derivatives using reinforcement learning. *J Financial Data Sci.* **3**, 10–27.
[7] Charpentier, A., Elie, R. and Remlinger, C. (2021). Reinforcement learning in economics and finance. *Comput. Economics* **62**, 425–462.
[8] Chen, Z., Yu, P. and Haskell, W. B. (2019). Distributionally robust optimization for sequential decision-making. *Optimization* **68**, 2397–2426.
[9] El Ghaoui, L. and Nilim, A. (2005). Robust solutions to Markov decision problems with uncertain transition matrices. *Operat. Res.* **53**, 780–798.
[10] Feinberg, E. A. and Shwartz, A. (2012). *Handbook of Markov Decision Processes: Methods and Applications* (Int. Ser. Operat. Res. Manag. Sci. 40). Springer, New York.
[11] Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996). Reinforcement learning: A survey. *J. Artificial Intell. Res.* **4**, 237–285.
[12] Kallus, N. and Uehara, M. (2020). Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Mach. Learn. Res.* **21**, 6742–6804.
[13] Kern, P. (2020). Sensitivity and statistical inference in Markov decision models and collective risk models. PhD dissertation, Saarland University.
[14] Kiszka, A. and Wozabal, D. (2022). A stability result for linear Markovian stochastic optimization problems. *Math. Program.* **191**, 871–906.
[15] Levin, E., Pieraccini, R. and Eckert, W. (1998). Using Markov decision process for learning dialogue strategies. In *Proc. 1998 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vol. 1. IEEE, Piscataway, NJ, pp. 201–204.
[16] Li, M., Sutter, T. and Kuhn, D. (2023). Policy gradient algorithms for robust MDPs with non-rectangular uncertainty sets. Preprint, arXiv:2305.19004.
[17] Liu, Z., Bai, Q., Blanchet, J., Dong, P., Xu, W., Zhou, Z. and Zhou, Z. (2022). Distributionally robust Q-learning. *Proc. Mach. Learn. Res.* 162, 13623–13643.

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press

[18] MANNOR, S., MEBEL, O. AND XU, H. (2016). Robust MDPs with k-rectangular uncertainty. *Math. Operat. Res.* **41**, 1484–1509.
[19] MÜLLER, A. (1997). How does the value function of a Markov decision process depend on the transition probabilities? *Math. Operat. Res.* **22**, 872–885.
[20] NATARAJAN, M. AND KOLOBOV, A. (2022). *Planning with Markov decision processes: An AI perspective*. Springer, New York.
[21] NEUFELD, A. AND SESTER, J. (2024). Robust Q-learning algorithm for Markov decision processes under Wasserstein uncertainty. *Automatica* **168**, 111825.
[22] NEUFELD, A. AND SESTER, J. (2024). Non-concave distributionally robust stochastic control in a discrete time finite horizon setting. Preprint, arXiv:2404.05230.
[23] NEUFELD, A., SESTER, J. AND ŠIKIĆ, M. (2023). Markov decision processes under model uncertainty. *Math. Finance* **33**, 618–665.
[24] PANAGANTI, K. AND KALATHIL, D. (2022). Sample complexity of robust reinforcement learning with a generative model. *Proc. Mach. Learn. Res*. 151, 9582–9602.
[25] PUTERMAN, M. L. (1990). Markov decision processes. In *Handbooks in Operations Research and Management Science*, Vol. **2**. Elsevier, Amsterdam, pp. 331–434.
[26] PUTERMAN, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley, Chichester.
[27] SI, N., ZHANG, F., ZHOU, Z. AND BLANCHET, J. (2020). Distributionally robust policy evaluation and learning in offline contextual bandits. *Proc. Mach. Learn. Res*. 119, 8884–8894.
[28] SI, N., ZHANG, F., ZHOU, Z. AND BLANCHET, J. (2023). Distributional robust batch contextual bandits. *Manag. Sci.* **69**, 5772–5793.
[29] SUTTON, R. S. AND BARTO, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press, Cambridge, MA.
[30] UĞURLU K. (2018). Robust optimal control using conditional risk mappings in infinite horizon. *J. Comput. Appl. Math.* **344**, 275–287.
[31] VILLANI, C. (2008). *Optimal Transport: Old and New* (Grundlehren der mathematischen Wissenschaften **338**). Springer, Berlin.
[32] WANG, Y. AND ZOU, S. (2022). Policy gradient method for robust reinforcement learning. Preprint, arXiv:2205.07344.
[33] WHITE, D. J. (1993). A survey of applications of Markov decision processes. *J. Operat. Res. Soc.* **44**, 1073–1096.
[34] WIESEMANN, W., KUHN, D. AND RUSTEM, B. (2013). Robust Markov decision processes. *Math. Operat. Res.* **38**, 153–183.
[35] WIESEMANN, W., KUHN, D. AND SIM, M. (2014). Distributionally robust convex optimization. *Operat. Res.* **62**, 1358–1376.
[36] XU, X. AND MANNOR, S. (2012). Distributionally robust Markov decision processes. *Math. Operat. Res.* **37**, 288–300.
[37] YANG, W., ZHANG, L. AND ZHANG, Z. (2022). Towards theoretical understandings of robust Markov decision processes: Sample complexity and asymptotics. *Ann. Statist.* **50**, 3223–3248.
[38] ZÄHLE, H. (2022). A concept of copula robustness and its applications in quantitative risk management. *Finance Stoch.* **26**, 825–875.
[39] ZHOU, Z., ZHOU, Z., BAI, Q., QIU, L., BLANCHET, L. AND GLYNN, P. (2021). Finite-sample regret bound for distributionally robust offline tabular reinforcement learning. *Proc. Mach. Learn. Res*. 130, 3331–3339.

https://doi.org/10.1017/jpr.2024.88 Published online by Cambridge University Press