COMISEF Computational Optimization Methods in Statistics, Econometrics and Finance

- Marie Curie Research and Training Network funded by the EU Commission through MRTN-CT-2006-034270 -

COMISEF WORKING PAPERS SERIES

WPS-034 05/05/2010

Robust Markov Decision Processes

W. Wiesemann
D. Kuhn
B. Rustem

Robust Markov Decision Processes

Wolfram Wiesemann, Daniel Kuhn and Berç Rustem

May 5, 2010

Abstract

Markov decision processes (MDPs) are powerful tools for decision making in uncertain dynamic environments. However, the solutions of MDPs are of limited practical use due to their sensitivity to distributional model parameters, which are typically unknown and have to be estimated by the decision maker. To counter the detrimental effects of estimation errors, we consider robust MDPs that offer probabilistic guarantees in view of the unknown parameters. To this end, we assume that an observation history of the MDP is available. Based on this history, we derive a confidence region that contains the unknown parameters with a pre-specified probability $1 - \beta$. Afterwards, we determine a policy that attains the highest worst-case performance over this confidence region. By construction, this policy achieves or exceeds its worst-case performance with a confidence of at least $1 - \beta$. Our method involves the solution of tractable conic programs of moderate size.

Notation For a finite set $\mathcal{X} = \{1, \ldots, X\}$, $\mathcal{M}(\mathcal{X})$ denotes the probability simplex in $\mathbb{R}^X$. An $\mathcal{X}$-valued random variable χ has distribution $m \in \mathcal{M}(\mathcal{X})$, denoted by $\chi \sim m$, if $\mathbb{P}(\chi = x) = m_x$ for all $x \in \mathcal{X}$. By default, all vectors are column vectors. We denote by e_k the kth canonical basis vector, while e denotes the vector whose components are all ones. In both cases, the dimension will usually be clear from the context. For square matrices A and B, the relation $A \succeq B$ indicates that the matrix $A - B$ is positive semidefinite. We denote the space of symmetric $n \times n$ matrices by $\mathbb{S}^n$. The declaration $f : X \overset{\mathrm{c}}{\mapsto} Y$ ($f : X \overset{\mathrm{a}}{\mapsto} Y$) implies that f is a continuous (affine) function from X to Y. For a matrix A, we denote its ith row by $A_{i\cdot}^\top$ (a row vector) and its jth column by $A_{\cdot j}$.

1 Introduction

Markov decision processes (MDPs) provide a versatile model for sequential decision making under uncertainty, which accounts for both the immediate effects and future ramifications of decisions. In the past sixty years, MDPs have been successfully applied to numerous areas, ranging from inventory control and investment planning to studies in economics and behavioural ecology [4, 19].

In this paper, we study MDPs with a finite state space $\mathcal{S} = \{1, \ldots, S\}$, a finite action space $\mathcal{A} = \{1, \ldots, A\}$, and a discrete but infinite planning horizon $\mathcal{T} = \{0, 1, 2, \ldots\}$. Without loss of generality (w.l.o.g.), we assume that every action is admissible in every state. The initial state is random and follows the probability distribution $p_0 \in \mathcal{M}(\mathcal{S})$. If action $a \in \mathcal{A}$ is chosen in state $s \in \mathcal{S}$, the subsequent state is determined by the conditional probability distribution $p(\cdot|s, a) \in \mathcal{M}(\mathcal{S})$. We condense these conditional distributions to the transition kernel $P \in [\mathcal{M}(\mathcal{S})]^{S \times A}$, where $P_{sa} := p(\cdot|s, a)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$. The decision maker receives an expected reward of $r(s, a, s') \in \mathbb{R}_+$ if action $a \in \mathcal{A}$ is chosen in state $s \in \mathcal{S}$ and the subsequent state is $s' \in \mathcal{S}$. W.l.o.g., we assume that all rewards are non-negative. The MDP is controlled through a policy $\pi = (\pi_t)_{t \in \mathcal{T}}$, where $\pi_t : (\mathcal{S} \times \mathcal{A})^{t-1} \times \mathcal{S} \mapsto \mathcal{M}(\mathcal{A})$. $\pi_t(\cdot|s_0, a_0, \ldots, s_{t-1}, a_{t-1}; s_t)$ represents the probability distribution over $\mathcal{A}$ according to which the next action is chosen if the current state is s_t and the state-action history is given by $(s_0, a_0, \ldots, s_{t-1}, a_{t-1})$. Together with the transition kernel P, π induces a stochastic process $(s_t, a_t)_{t \in \mathcal{T}}$ on the space $(\mathcal{S} \times \mathcal{A})^\infty$ of sample paths. We use the notation $\mathbb{E}^{P,\pi}$ to denote expectations with respect to this process. Throughout this paper, we evaluate policies in view of their expected total reward under the discount factor $\lambda \in (0, 1)$:

$$\mathbb{E}^{P,\pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 \sim p_0 \right] \tag{1}$$

For a fixed policy π, the *policy evaluation problem* asks for the value of expression (1). The *policy improvement problem*, on the other hand, asks for a policy π that maximises (1).

Most of the literature on MDPs assumes that the expected rewards r and the transition kernel P are known, with a tacit understanding that they have to be estimated in practice. However, it is well-known that the expected total reward (1) can be very sensitive to small changes in r and P [15]. Thus, decision makers are confronted with two different sources of uncertainty. On one hand, they face *internal variation* due to the stochastic nature of MDPs. On the other hand, they need to cope with *external variation* because the estimates for r and P deviate from their true values. In this paper, we assume that the decision maker is risk-neutral to internal variation but risk-averse to external variation. This is justified if the MDP runs for a long time, or if many instances of the same MDP run in parallel [15]. We focus on external variation in P and assume r to be known. Indeed, the expected total reward (1) is typically more sensitive to P, and the inclusion of reward variation is straightforward [7, 15].

Let P^0 be the unknown true transition kernel of the MDP. Since the expected total reward of a policy depends on P^0, we cannot evaluate expression (1) under external variation. Iyengar [11] and Nilim and El Ghaoui [17] therefore suggest to find a policy that guarantees the highest expected total reward at a given confidence level. To this end, they determine a policy π that maximises the worst-case objective

$$z^* = \inf_{P \in \mathcal{P}} \mathbb{E}^{P,\pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \,\middle|\, s_0 \sim p_0 \right], \tag{2}$$

where the uncertainty set $\mathcal{P}$ is the Cartesian product of independent marginal sets $\mathcal{P}_{sa} \subseteq \mathcal{M}(\mathcal{S})$ for each $(s, a) \in \mathcal{S} \times \mathcal{A}$. In the following, we call such uncertainty sets *rectangular*. Problem (2) determines the worst-case expected total reward of π if the transition kernel can vary freely within $\mathcal{P}$. In analogy to our earlier definitions, the *robust policy evaluation problem* evaluates expression (2) for a fixed policy π, while the *robust policy improvement problem* asks for a policy that maximises (2). The optimal value z^* in (2) provides a lower bound on the expected total reward of π if the true transition kernel P^0 is contained in the uncertainty set $\mathcal{P}$. Hence, if $\mathcal{P}$ is a confidence region that contains P^0 with probability $1 - \beta$, then the policy π guarantees an expected total reward of at least z^* at a confidence level $1 - \beta$. To construct an uncertainty set $\mathcal{P}$ with this property, [11] and [17] assume that independent transition samples are available for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. Under this assumption, the samples for each state-action pair follow independent multinomial distributions whose (unknown) parameters coincide with the entries of P^0. One can then employ standard statistical techniques to derive a confidence region for P^0. If we project this confidence region onto the marginal sets $\mathcal{P}_{sa}$, then z^* provides the desired probabilistic lower bound on the expected total reward of π.

In this paper, we alter two key assumptions of the outlined procedure. Firstly, we assume that the decision maker cannot obtain independent transition samples for the state-action pairs. Instead, she has merely access to an observation history $(s_1, a_1, \ldots, s_n, a_n) \in (\mathcal{S} \times \mathcal{A})^n$ generated by the MDP under some known policy. Secondly, we relax the assumption of rectangular uncertainty sets. In the following, we briefly motivate these changes and give an outlook on their consequences.

Although transition sampling has theoretical appeal, it is often prohibitively costly or even infeasible in practice. To obtain independent samples for each state-action pair, one needs to repeatedly direct the MDP into any of its states and record the transitions resulting from different actions. In particular, one cannot use the transition frequencies of an observation history because those frequencies violate the independence assumption stated above. The availability of an observation history, on the other hand, seems much more realistic in practice. Observation histories introduce a number of theoretical challenges, such as the lack of observations for some transitions and stochastic dependencies between the transition frequencies. We will apply results from statistical inference on Markov chains to address these issues.

The restriction to rectangular uncertainty sets has been introduced in [11] and [17] to facilitate computational tractability. Under the assumption of rectangularity, the robust policy evaluation and improvement problems can be solved efficiently with a modified value or policy iteration. This implies, however, that non-rectangular uncertainty sets have to be projected onto the marginal sets $\mathcal{P}_{sa}$. Not only does this 'rectangularisation' unduly increase the level of conservatism, but it also creates a number of undesirable side-effects that we discuss in Section 2. In this paper, we show that the robust policy evaluation and improvement problems remain tractable for uncertainty sets that exhibit a milder form of rectangularity, and we develop a polynomial time solution method. On the other hand, we prove that the robust policy evaluation and improvement problems are intractable for non-rectangular uncertainty sets. For this setting, we formulate conservative approximations of the policy evaluation and improvement problems. We bound the optimality gap incurred from solving those approximations, and we outline how our approach can be generalised to a hierarchy of increasingly accurate approximations.

The contributions of this paper can be summarised as follows.

1. We analyse a new class of uncertainty sets, which contains the above defined rectangular uncertainty sets as a special case. We show that the optimal policies for this class are randomised but memoryless. We develop algorithms that solve the robust policy evaluation and improvement problems over these uncertainty sets in polynomial time.

2. It is stated in [17] that the robust policy evaluation and improvement problems "seem to be hard to solve" for non-rectangular uncertainty sets. We prove that both problems are indeed strongly $\mathcal{NP}$-hard. We develop a hierarchy of increasingly accurate conservative approximations, together with bounds on the incurred optimality gap.

3. We present a method to construct uncertainty sets from observation histories. In contrast, existing approaches rely on transition sampling, which is often too costly or infeasible in practice. Our approach allows to account for different types of a priori information about the transition kernel, which helps to reduce the size of the uncertainty set. We also investigate the convergence behaviour of our uncertainty set when the length of the observation history increases.

The study of robust MDPs with rectangular uncertainty sets dates back to the seventies, see [2, 9, 21, 25] and the surveys in [11, 17]. However, most of the early contributions do not address the construction of suitable uncertainty sets. In [15], Mannor *et al.* approximate the bias and variance of the expected total reward (1) if the unknown model parameters are replaced with estimates. Delage and Mannor [7] use these approximations to solve a chance-constrained policy improvement problem in a Bayesian setting. Recently, alternative performance criteria have been suggested to address external variation, such as the worst-case expected utility and regret measures. We refer to [18, 26] and the references cited therein. Note that we could address external variation by encoding the unknown model parameters into the states of a partially observable MDP (POMDP) [16]. However, the optimisation of

POMDPs becomes challenging even for small state spaces. In our case, the augmented state space would become very large, which renders optimisation of the resulting POMDPs prohibitively expensive.

The remainder of the paper is organised as follows. Section 2 defines and analyses the classes of robust MDPs that we consider. Sections 3 and 4 study the robust policy evaluation and improvement problems, respectively. Section 5 constructs uncertainty sets from observation histories. We illustrate our method in Section 6, where we apply it to the machine replacement problem. We conclude in Section 7.

Remark 1.1 (Finite Horizon MDPs) *Throughout the paper, we outline how our results extend to finite horizon MDPs. In this case, we assume that $\mathcal{T} = \{0, 1, 2, \ldots, T\}$ with $T < \infty$ and that $\mathcal{S}$ can be partitioned into nonempty disjoint sets $\{\mathcal{S}_t\}_{t \in \mathcal{T}}$ such that at period t the system is in one of the states in $\mathcal{S}_t$. We do not discount rewards in finite horizon MDPs. If the MDP reaches a terminal state $s \in \mathcal{S}_T$, an expected reward of $\mathrm{r}_s \in \mathbb{R}_+$ is received. We assume that $p_0(s) = 0$ for $s \notin \mathcal{S}_1$.*

2 Robust Markov Decision Processes

This section studies properties of the robust policy evaluation and improvement problems. Both problems are concerned with *robust MDPs*, for which the transition kernel is only known to be an element of an uncertainty set $\mathcal{P} \subseteq [\mathcal{M}(\mathcal{S})]^{S \times A}$. We assume that the initial state distribution p_0 is known.

We start with the robust policy evaluation problem. We define the structure of the uncertainty sets that we consider, as well as different types of rectangularity that can be imposed to facilitate computational tractability. Afterwards, we discuss the robust policy improvement problem. We define several policy classes that are commonly used in MDPs, and we investigate the structure of optimal policies for different types of rectangularity. We close with a complexity result for the robust policy evaluation problem. Since the remainder of this paper almost exclusively deals with the robust versions of the policy evaluation and improvement problems, we may suppress the attribute 'robust' in the following.

2.1 The Robust Policy Evaluation Problem

Consider the policy evaluation problem (2), where we replace the uncertainty set $\mathcal{P}$ with

$$\mathcal{P} := \Big\{ P \in [\mathcal{M}(\mathcal{S})]^{S \times A} \; : \; \exists \, \xi \in \Xi \;\text{ such that }\; P_{sa} = p^{\xi}(\cdot|s,a) \;\; \forall \, (s,a) \in \mathcal{S} \times \mathcal{A} \Big\}. \tag{3a}$$

Here, we assume that Ξ is a subset of $\mathbb{R}^q$ and that $p^{\xi}(\cdot|s,a)$, $(s,a) \in \mathcal{S} \times \mathcal{A}$, is an affine function from Ξ to $\mathcal{M}(\mathcal{S})$ that satisfies $p^{\xi}(\cdot|s,a) := k_{sa} + K_{sa}\xi$ for some $k_{sa} \in \mathbb{R}^S$ and $K_{sa} \in \mathbb{R}^{S \times q}$. We also stipulate that

$$\Xi := \left\{ \xi \in \mathbb{R}^q \; : \; \xi^\top O_l \, \xi + o_l^\top \xi + \omega \geq 0 \;\; \forall \, l = 1, \ldots, L \right\}, \tag{3b}$$

where $O_l \in \mathbb{S}^q$ satisfies $O_l \preceq 0$. We assume that Ξ is bounded and that it contains a Slater point $\overline{\xi} \in \mathbb{R}^q$ which satisfies $\overline{\xi}^\top O_l \, \overline{\xi} + o_l^\top \overline{\xi} + \omega > 0$ for all l. Our definition of Ξ encompasses all compact subsets of $\mathbb{R}^q$ that have a nonempty interior and that result from finite intersections of closed halfspaces and ellipsoids.

Example 2.1 *Consider a robust infinite horizon MDP with three states and one action. The transition probabilities are defined through*

$$p^\xi(1|s,1) = \frac{1}{3} + \frac{\xi_1}{3}, \quad p^\xi(2|s,1) = \frac{1}{3} + \frac{\xi_2}{3} \quad \textit{and} \quad p^\xi(3|s,1) = \frac{1}{3} - \frac{\xi_1}{3} - \frac{\xi_2}{3} \quad \textit{for } s \in \{1,2,3\},$$

where $\xi = (\xi_1, \xi_2)$ is only known to satisfy $\xi_1^2 + \xi_2^2 \leq 1$ and $\xi_1 \leq \xi_2$. We can model this MDP through

$$\Xi = \left\{ \xi \in \mathbb{R}^2 \, : \, \xi_1^2 + \xi_2^2 \leq 1, \, \xi_1 \leq \xi_2 \right\}, \quad k_{s1} = \frac{1}{3}\mathrm{e} \quad \textit{and} \quad K_{s1} = \frac{1}{3} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \quad \textit{for } s \in \{1,2,3\}.$$

Note that the mapping K cannot be absorbed in the definition of Ξ without violating the Slater condition.

We say that an uncertainty set $\mathcal{P}$ is (s,a)*-rectangular* if

$$\mathcal{P} = \mathop{\times}_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{P}_{sa}, \qquad \text{where} \qquad \mathcal{P}_{sa} := \{P_{sa} \, : \, P \in \mathcal{P}\} \text{ for } (s,a) \in \mathcal{S} \times \mathcal{A}.$$

Likewise, we say that an uncertainty set $\mathcal{P}$ is *s-rectangular* if

$$\mathcal{P} = \mathop{\times}_{s \in \mathcal{S}} \mathcal{P}_s, \qquad \text{where} \qquad \mathcal{P}_s := \{(P_{s1}, \ldots, P_{sA}) \, : \, P \in \mathcal{P}\} \text{ for } s \in \mathcal{S}.$$

For any uncertainty set $\mathcal{P}$, we call $\mathcal{P}_{sa}$ and $\mathcal{P}_s$ the *marginal uncertainty sets* (or simply marginals). For our definition (3) of $\mathcal{P}$, we have $\mathcal{P}_{sa} = \left\{ p^\xi(\cdot|s,a) \, : \, \xi \in \Xi \right\}$ and $\mathcal{P}_s = \left\{ \left(p^\xi(\cdot|s,1), \ldots, p^\xi(\cdot|s,A) \right) \, : \, \xi \in \Xi \right\}$, respectively. Note that all transition probabilities $p^\xi(\cdot|s,a)$ can vary freely within their marginals $\mathcal{P}_{sa}$ if the uncertainty set is (s,a)-rectangular. In contrast, the transition probabilities $\left\{ p^\xi(\cdot|s,a) \, : \, a \in \mathcal{A} \right\}$ for different actions in the same state may be dependent in an s-rectangular uncertainty set. By definition, (s,a)-rectangularity implies s-rectangularity. (s,a)-rectangular uncertainty sets have been introduced in [11, 17], whereas the notion of s-rectangularity seems to be new. Note that our definition (3) of $\mathcal{P}$ does not impose any kind of rectangularity. Indeed, the uncertainty set in Example 2.1 is not s-rectangular. The following example shows that rectangular uncertainty sets can result in crude approximations of the decision maker's knowledge about the true transition kernel P^0.

Example 2.2 (Rectangularity) *Consider the robust infinite horizon MDP that is shown in Figure 1. The uncertainty set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realisations $\xi \in$*

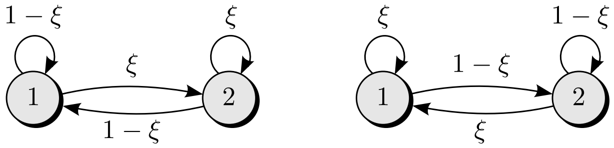

Figure 1: *MDP with two states and two actions. The left and right charts present the transition probabilities for actions 1 and 2, respectively. In both diagrams, nodes correspond to states and arcs to transitions. We label each arc with the probability of the associated transition. We suppress* p_0 *and the expected rewards.*

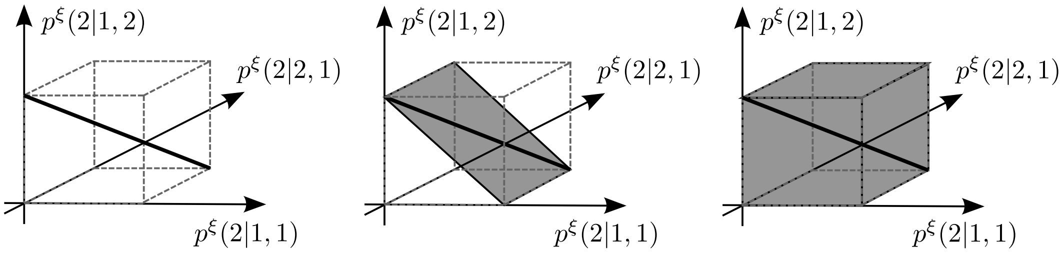

Figure 2: *Illustration of* $\mathcal{P}$ *(left chart) and the smallest* s*-rectangular (middle chart) and* (s,a)*-rectangular (right chart) uncertainty sets that contain* $\mathcal{P}$*. The charts show three-dimensional projections of* $\mathcal{P} \subset \mathbb{R}^8$*. The thick line represents* $\mathcal{P}$*, while the shaded areas visualise the corresponding rectangular uncertainty sets. Figure 1 implies that* $p^\xi(2|1,1) = \xi$*,* $p^\xi(2|1,2) = 1-\xi$ *and* $p^\xi(2|2,1) = \xi$*. The dashed lines correspond to the unit cube in* $\mathbb{R}^3$*.*

$[0,1]$. *This MDP can be assigned an uncertainty set of the form (3). Figure 2 visualises* $\mathcal{P}$ *and the smallest* s*-rectangular and* (s,a)*-rectangular uncertainty sets that contain* $\mathcal{P}$*.*

From now on, we always consider uncertainty sets of the form (3). We may sometimes call a generic uncertainty set *non-rectangular* to emphasise that it is neither s- nor (s,a)-rectangular.

2.2 The Robust Policy Improvement Problem

We now consider the policy improvement problem, which asks for a policy that maximises the worst-case expected total reward (2) over an uncertainty set of the form (3). Remember that a policy π represents a sequence of functions $(\pi_t)_{t\in\mathcal{T}}$ that map state-action histories to probability distributions over $\mathcal{A}$. In its most general form, such a policy is *history dependent*, that is, at any time period t the policy may assign a different probability distribution to each state-action history $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}; s_t)$.

Due to the storage requirements of history dependent policies, one typically prefers more 'economical' policy classes. A policy π is called *Markovian* if π_t is determined by s_t and t for all $t \in \mathcal{T}$. A Markovian policy π is called *stationary* if π_t is solely determined by s_t for all $t \in \mathcal{T}$. In finite horizon MDPs, Markovian and stationary policies are equally expressive since the sets $\mathcal{S}_t$ are disjoint. In infinite horizon MDPs, however, stationary policies form a strict subset of the class of Markovian policies. A policy π is called *deterministic* if π_t places all probability mass on one action for each $t \in \mathcal{T}$; otherwise, π is called *randomised*. In the following, we will focus on stationary policies due to their favourable storage requirements. We denote by Π the set of all randomised stationary policies for a given MDP instance.

It is well-known that non-robust finite and infinite horizon MDPs always allow for a deterministic stationary policy that maximises the expected total reward (1). Optimal policies can be determined via value or policy iteration, or via linear programming. Finding an optimal policy, as well as evaluating (1) for a given stationary policy, can be done in polynomial time. For a detailed discussion, see [4, 19, 22].

To date, the literature on robust MDPs has focused on (s, a)-rectangular uncertainty sets. For this class of uncertainty sets, it is shown in [11, 17] that the worst-case expected total reward (2) is maximised by a deterministic stationary policy π for finite and infinite horizon MDPs. Optimal policies can be determined via extensions of the value and policy iteration. For some uncertainty sets, finding an optimal policy, as well as evaluating (2) for a given stationary policy, can be achieved in polynomial time. Moreover, the policy improvement problem satisfies the following saddle point condition:

$$\sup_{\pi \in \Pi} \inf_{P \in \mathcal{P}} \mathbb{E}^{P,\pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 \sim p_0 \right] = \inf_{P \in \mathcal{P}} \sup_{\pi \in \Pi} \mathbb{E}^{P,\pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 \sim p_0 \right] \quad (4)$$

A similar result for robust finite horizon MDPs is discussed in [17].

We now show that the benign structure of optimal policies over (s, a)-rectangular uncertainty sets partially extends to the broader class of s-rectangular uncertainty sets.

Proposition 2.3 (s-Rectangular Uncertainty Sets) *Consider the policy improvement problem for a finite or infinite horizon MDP over an s-rectangular uncertainty set of the form (3).*

(a) There is always an optimal policy that is stationary.

(b) It is possible that all optimal stationary policies are randomised.

Proof As for claim (a), consider a finite horizon MDP with an s-rectangular uncertainty set. By construction, the probabilities associated with transitions emanating from state $s \in \mathcal{S}$ are independent from those emanating from any other state $s' \in \mathcal{S}$, $s' \neq s$. Moreover, each state s is visited at most once since the sets $\mathcal{S}_t$ are disjoint. Hence, any knowledge about past transition probabilities cannot contribute to better decisions in future time periods, which implies that stationary policies are optimal.

Consider now an infinite horizon MDP with an s-rectangular uncertainty set. Appendix A shows that the saddle point condition (4) extends to s-rectangular uncertainty sets. For any fixed transition kernel $P \in \mathcal{P}$, the supremum over all stationary policies on the right-hand side of (4) is equivalent to the supremum over all history dependent policies. By weak duality, the right-hand side of (4) thus represents an upper bound on the worst-case expected total reward of any history dependent policy. Since there is a stationary policy whose worst-case expected total reward on the left-hand side of (4) attains this upper bound, claim (a) follows.

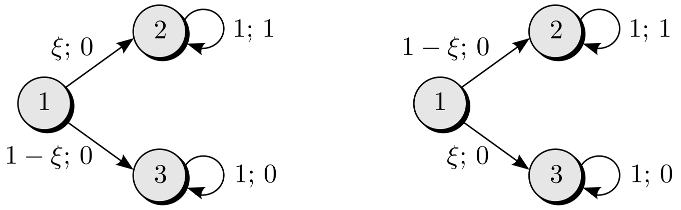

Figure 3: *MDP with three states and two actions. The left and right figures present the transition probabilities and expected rewards for actions* 1 *and* 2*, respectively. The first and second expressions in the arc labels correspond to the probabilities and expected rewards of the associated transitions, respectively. Apart from that, the same drawing conventions as in Figure 1 are used. The initial state distribution* p_0 *places unit mass on state* 1.

As for claim (b), consider the robust infinite horizon MDP that is visualised in Figure 3. The uncertainty set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realisations $\xi \in [0,1]$. This MDP can be assigned an s-rectangular uncertainty set of the form (3). Since the transitions are independent of the chosen actions from time 1 onwards, a policy is completely determined by the decision $\beta = \pi_0(1|1)$ at time 0. The worst-case expected total reward is

$$\min_{\xi \in [0,1]} \left[\beta\xi + (1-\beta)(1-\xi)\right] \frac{\lambda}{1-\lambda} = \min\left\{\beta, 1-\beta\right\} \frac{\lambda}{1-\lambda}.$$

Over $\beta \in [0,1]$, this expression has its unique maximum at $\beta^* = 1/2$, that is, the optimal policy is randomised. If we replace the self-loops with expected terminal rewards of $\mathbf{r}_2 := 1$ and $\mathbf{r}_3 := 0$, then we obtain an example of a robust *finite* horizon MDP whose optimal policy is randomised. ■

Figure 3 illustrates the counterintuitive result that randomisation is superfluous for (s,a)-rectangular uncertainty sets. If we project the uncertainty set $\mathcal{P}$ associated with Figure 3 onto its marginals $\mathcal{P}_{sa}$, then the transition probabilities in the left chart become independent of those in the right chart. In this case, any policy results in an expected total reward of zero, and randomisation becomes ineffective.

We now show that in addition to randomisation, the optimal policy may require history dependence if the uncertainty set lacks s-rectangularity.

Proposition 2.4 (General Uncertainty Sets) *For finite and infinite horizon MDPs, the policy improvement problem over non-rectangular uncertainty sets is in general solved by non-Markovian policies.*

Proof Consider the robust infinite horizon MDP with six states and two actions that is visualised in Figure 4. The uncertainty set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realisations $\xi \in [0,1]$. This MDP can be assigned an uncertainty set of the form (3). Since the transitions do not depend on the chosen actions except for π_2, a policy is completely determined by the decision $\beta = (\beta_1, \beta_2)$, where $\beta_1 = \pi_2(1|1, a_0, 2, a_1; 4)$ and $\beta_2 = \pi_2(1|1, a_0, 3, a_1; 4)$.

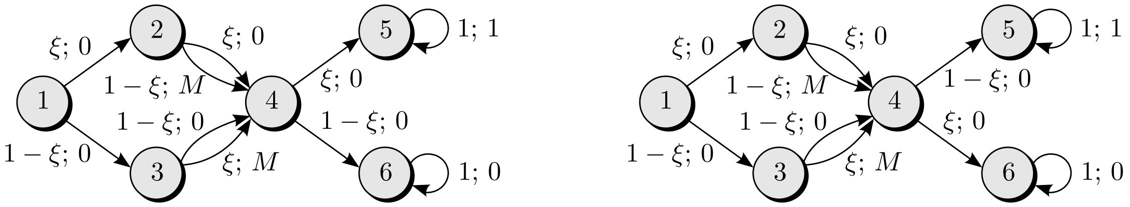

Figure 4: *MDP with six states and two actions. The initial state distribution* p_0 *places unit mass on state* 1*. The same drawing conventions as in Figure 3 are used.*

The conditional probability to reach state 5 is $\varphi_1(\xi) := \beta_1\xi + (1-\beta_1)(1-\xi)$ if state 2 is visited and $\varphi_2(\xi) := \beta_2\xi + (1-\beta_2)(1-\xi)$ if state 3 is visited, respectively. Thus, the expected total reward is

$$2\lambda\xi(1-\xi)M + \frac{\lambda^3}{1-\lambda}\left[\xi\,\varphi_1(\xi) + (1-\xi)\,\varphi_2(\xi)\right],$$

which is concave in ξ for all $\beta \in [0,1]^2$ if $M \geq \lambda^2/(1-\lambda)$. Thus, the worst (minimal) expected total reward is incurred for $\xi^* \in \{0,1\}$, independently of $\beta \in [0,1]^2$. Hence, the worst-case expected total reward is

$$\min_{\xi\in\{0,1\}} \frac{\lambda^3}{1-\lambda}\left[\xi\,\varphi_1(\xi) + (1-\xi)\,\varphi_2(\xi)\right] = \frac{\lambda^3}{1-\lambda}\min\left\{\beta_1, 1-\beta_2\right\},$$

and the unique maximiser of this expression is $\beta = (1,0)$. We conclude that in state 4, the optimal policy chooses action 1 if state 2 has been visited and action 2 otherwise. Hence, the optimal policy is history dependent. If we replace the self-loops with expected terminal rewards of $\mathfrak{r}_5 := \lambda^3/(1-\lambda)$ and $\mathfrak{r}_6 := 0$, then we can extend the result to robust finite horizon MDPs. ■

Although the policy improvement problem over non-rectangular uncertainty sets is in general solved by non-Markovian policies, we will restrict ourselves to stationary policies in the remainder. Thus, we will be interested in the best deterministic or randomised stationary policies for robust MDPs.

2.3 Complexity of the Robust Policy Evaluation Problem

We show that the policy evaluation problem over non-rectangular uncertainty sets is strongly $\mathcal{NP}$-hard. To this end, we will reduce the evaluation of (2) to the 0/1 Integer Programming (IP) problem [8]:

0/1 Integer Programming.

Instance. Given are $F \in \mathbb{Z}^{m\times n}$, $g \in \mathbb{Z}^m$, $c \in \mathbb{Z}^n$, $\zeta \in \mathbb{Z}$.

Question. Is there a vector $x \in \{0,1\}^n$ such that $Fx \leq g$ and $c^\top x \leq \zeta$?

Assume that $x \in [0,1]^n$ constitutes a fractional vector that satisfies $Fx \leq g$ and $c^\top x \leq \zeta$. The following lemma shows that we can obtain an integral vector $y \in \{0,1\}^n$ that satisfies $Fy \leq g$ and

$c^\top y \leq \zeta$ by rounding x if its components are 'close enough' to zero or one.

Lemma 2.5 *Let $0 < \epsilon \leq \min\{\epsilon_F, \epsilon_c\}$, where $0 < \epsilon_F < \min_i \left\{ \left(\sum_j |F_{ij}| \right)^{-1} \right\}$ and $0 < \epsilon_c < \left(\sum_j |c_j| \right)^{-1}$. Assume that $x \in ([0,\epsilon] \cup [1-\epsilon, 1])^n$ satisfies $Fx \leq g$ and $c^\top x \leq \zeta$. Then $Fy \leq g$ and $c^\top y \leq \zeta$ for $y \in \{0,1\}^n$, where $y_j := 1$ if $x_j \geq 1 - \epsilon$ and $y_j := 0$ otherwise.*

Proof By construction, $F_{i\cdot}^\top y \leq F_{i\cdot}^\top x + \sum_j |F_{ij}| \epsilon_F < F_{i\cdot}^\top x + 1 \leq g_i + 1$ for all $i \in \{1, \ldots, m\}$. Similarly, we have that $c^\top y \leq c^\top x + \sum_j |c_j| \epsilon_c < c^\top x + 1 \leq \zeta + 1$. Due to the integrality of F, g, c, ζ and y, we therefore conclude that $Fy \leq g$ and $c^\top y \leq \zeta$. ■

We can now prove strong $\mathcal{NP}$-hardness of the policy evaluation problem.

Theorem 2.6 *Deciding whether the worst-case expected total reward (2) over an uncertainty set of the form (3) exceeds a given value γ is strongly $\mathcal{NP}$-hard for deterministic as well as randomised stationary policies and for finite as well as infinite horizon MDPs.*

Proof Let us fix an IP instance specified through F, g, c and ζ. W.l.o.g., we can assume that $\zeta \leq \sum_j [c_j]^+$ because all feasible IP solutions are binary. We construct a reduction to a robust infinite horizon MDP as follows. The states are $\mathcal{S} = \left\{ b_j, b_j^1, b_j^0 \,:\, j = 1, \ldots, n \right\} \cup \{c_0, \tau\}$, there is only one action, and $\lambda \in (0,1)$ can be chosen freely. The state transitions and expected rewards are illustrated in Figure 5. The uncertainty set $\mathcal{P}$ contains all transition kernels associated with $\xi \in [0,1]^n$ that satisfy $F\xi \leq g$. We choose $M > \left(\lambda n \sum_j |c_j| \right) / \left(2\epsilon^2 \right)$, where ϵ is chosen as in Lemma 2.5, and set $\gamma := \lambda^2 \zeta$. Following our discussion in Section 2.1, the described MDP instance can be constructed in polynomial time with respect to the size of the IP instance (which we henceforth abbreviate as 'in polynomial time').[1]

We show that the answer to the IP instance is affirmative if and only if the worst-case expected total reward (2) does not exceed γ. Indeed, assume that the answer to the IP instance is affirmative, that is, there is a vector $x \in \{0,1\}^n$ that satisfies $Fx \leq g$ and $c^\top x \leq \zeta$. The transition kernel associated with $\xi = x$ is contained in $\mathcal{P}$ and leads to an expected total reward of $\lambda^2 c^\top \xi \leq \lambda^2 \zeta = \gamma$. This implies that the worst-case expected total reward (2) does not exceed γ either. Conversely, assume that (2) does not exceed γ. For the constructed MDP, the expected total reward (1) is continuous in ξ. Since $\mathcal{P}$ is compact, we can therefore assume that the value of (2) is attained by a transition kernel associated with some $\xi^* \in \Xi$. By construction of Ξ, ξ^* satisfies $\xi^* \in [0,1]^n$ and $F\xi^* \leq g$. Assume that $\xi_q^* \notin ([0,\epsilon] \cup [1-\epsilon, 1])$ for some $q \in \{1, \ldots, n\}$. In this case, the expected total reward under ξ^* is greater than or equal to $2\lambda \xi_q^* (1 - \xi_q^*) M/n - \lambda^2 \sum_j [-c_j]^+ > \lambda^2 \sum_j [c_j]^+ \geq \gamma$, which contradicts our assumption. We have thus established that $\xi^* \in ([0,\epsilon] \cup [1-\epsilon, 1])^n$. Under the transition kernel associated with ξ^*, the expected

[1] Note that the set Ξ associated with the MDP instance might not contain a Slater point. However, one can decide in polynomial time whether the system of linear equations $Fx \leq g$, $x \in [0,1]^n$ is strictly feasible. If this is not the case, one can furthermore reduce the system to a strictly feasible one in polynomial time.

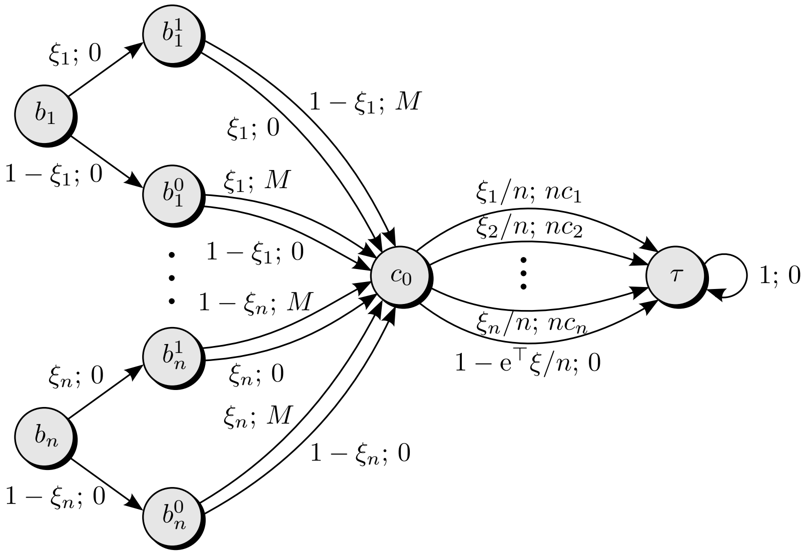

Figure 5: *MDP with $3n+2$ states and one action. The distribution p_0 places a probability mass of $1/n$ on each state b_j, $j=1,\ldots,n$. The drawing conventions from Figure 3 are used.*

reward in periods 0 and 1 is guaranteed to be non-negative, while the expected reward from period 2 onward amounts to $\lambda^2 c^\top \xi^*$. Since the expected total reward under ξ^* does not exceed γ, we therefore have that $\lambda^2 c^\top \xi^* \leq \gamma = \lambda^2 \zeta$, which implies that $c^\top \xi^* \leq \zeta$. Hence, we can apply Lemma 2.5 to obtain a vector $\xi' \in \{0,1\}^n$ that also satisfies $F\xi' \leq g$ and $c^\top \xi' \leq \zeta$. We have thus shown that the answer to the IP instance is affirmative if and only if the worst-case expected total reward (2) does not exceed γ.

If we could decide in polynomial time whether the worst-case expected total reward of the constructed MDP exceeds γ, we could also decide IP in polynomial time. Since IP is strongly $\mathcal{NP}$-hard [8], we conclude that the policy evaluation problem (2) is strongly $\mathcal{NP}$-hard for MDPs with a single action and uncertainty sets of the form (3). Since the policy space of the constructed MDP reduces to a singleton, our proof applies to robust MDPs with deterministic and randomised stationary policies. If we remove the self-loop emanating from state τ, introduce a terminal reward $\mathbf{r}_\tau := 0$ and multiply the rewards in period t with λ^{-t}, our proof furthermore applies to robust finite horizon MDPs. ■

Remark 2.7 *Theorem 2.6 remains valid if definition (3) is altered to require that $O_l = 0$ and $o_l \in \{0,1\}^q$. This follows from the fact that IP remains strongly $\mathcal{NP}$-hard if F and g are binary, see [8].*

Remark 2.8 *Throughout this section we assumed that $\mathcal{P}$ is a convex set of the form (3). If we extend our analysis to nonconvex uncertainty sets, then we obtain the results in Table 1. Note that the complexity of the policy evaluation and improvement problems will be discussed in Sections 2.3, 3 and 4.*

uncertainty set $\mathcal{P}$	optimal policy	complexity
(s, a)-rectangular, convex	deterministic, stationary	polynomial
(s, a)-rectangular, nonconvex	deterministic, stationary	strongly $\mathcal{NP}$-hard
s-rectangular, convex	randomised, stationary	polynomial
s-rectangular, nonconvex	randomised, history dependent	strongly $\mathcal{NP}$-hard
non-rectangular, convex	randomised, history dependent	strongly $\mathcal{NP}$-hard

Table 1: *Properties of infinite horizon MDPs with different uncertainty sets. From left to right, the columns describe the structure of the uncertainty set, the structure of the optimal policy, and the complexity of the policy evaluation and improvement problems over randomised stationary policies. Each uncertainty set is of the form (3). For nonconvex uncertainty sets, we do not require the matrices O_l in (3b) to be negative semidefinite. The properties of finite horizon MDPs are similar, the only difference being that MDPs with s-rectangular nonconvex uncertainty sets are optimised by randomised stationary policies.*

3 Robust Policy Evaluation

It is shown in [11, 17] that the worst-case expected total reward (2) can be calculated in polynomial time for certain types of (s, a)-rectangular uncertainty sets. We extend this result to the broader class of s-rectangular uncertainty sets in Section 3.1. On the other hand, Theorem 2.6 shows that the evaluation of (2) is strongly $\mathcal{NP}$-hard for non-rectangular uncertainty sets. We therefore develop conservative approximations for the policy evaluation problem over general uncertainty sets in Section 3.2. We bound the optimality gap that is incurred by solving these approximations, and we outline how these approximations can be refined. Although this section primarily sets the stage for the policy improvement problem, we stress that policy evaluation is an important problem in its own right. For example, it finds frequent use in labour economics, industrial organisation and marketing [15].

Our solution approaches for s-rectangular and non-rectangular uncertainty sets rely on the reward to-go function. For a stationary policy π, we define the *reward to-go* function $v : \Pi \times \Xi \mapsto \mathbb{R}^S$ through

$$v_s(\pi; \xi) = \mathbb{E}^{p^\xi, \pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \,\middle|\, s_0 = s \right] \qquad \text{for } s \in \mathcal{S}. \tag{5}$$

$v_s(\pi; \xi)$ represents the expected total reward under the transition kernel p^ξ and the policy π if the initial state is $s \in \mathcal{S}$. The reward to-go function allows us to express the worst-case expected total reward as

$$\inf_{\xi \in \Xi} \mathbb{E}^{p^\xi, \pi} \left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \,\middle|\, s_0 \sim p_0 \right] = \inf_{\xi \in \Xi} \left\{ p_0^\top v(\pi; \xi) \right\}. \tag{6}$$

We simplify our notation by defining the Markov reward process (MRP) induced by p^ξ and π. MRPs are Markov chains which pay a state-dependent reward at each time period. In our case, the MRP is given by the transition kernel $\widehat{P} : \Pi \times \Xi \overset{\text{a}}{\mapsto} \mathbb{R}^{S \times S}$ and the expected state rewards $\widehat{r} : \Pi \times \Xi \overset{\dashv}{\mapsto} \mathbb{R}^S$ defined through

$$\widehat{P}_{ss'}(\pi;\xi) := \sum_{a\in\mathcal{A}} \pi(a|s)\, p^{\xi}(s'|s,a) \tag{7a}$$

$$\text{and} \qquad \widehat{r}_s(\pi;\xi) := \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p^{\xi}(s'|s,a)\, r(s,a,s'). \tag{7b}$$

Note that $\widehat{r}(\pi;\xi) \geq 0$ for each $\pi \in \Pi$ and $\xi \in \Xi$ since all expected rewards $r(s,a,s')$ were assumed to be non-negative. For $s, s' \in \mathcal{S}$, $\widehat{P}_{ss'}(\pi;\xi)$ denotes the probability that the next state of the MRP is s', given that the MRP is currently in state s. Likewise, $\widehat{r}_s(\pi;\xi)$ denotes the expected reward that is received in state s. By taking the expectation with respect to the sample paths of the MRP and reordering terms, we can reformulate the reward to-go function (5) as

$$v(\pi;\xi) = \sum_{t=0}^{\infty} \left[\lambda\, \widehat{P}(\pi;\xi)\right]^t \widehat{r}(\pi;\xi), \tag{8}$$

see [19]. The following proposition brings together several results about v that we will use later on.

Proposition 3.1 *The reward to-go function v has the following properties.*

(a) v is Lipschitz continuous on $\Pi \times \Xi$.

(b) For given $\pi \in \Pi$ and $\xi \in \Xi$, $w \in \mathbb{R}^S$ satisfies $w = \widehat{r}(\pi;\xi) + \lambda\, \widehat{P}(\pi;\xi)\, w$ if and only if $w = v(\pi;\xi)$.

(c) For given $\pi \in \Pi$ and $\xi \in \Xi$, if $w \in \mathbb{R}^S$ satisfies $w \leq \widehat{r}(\pi;\xi) + \lambda\, \widehat{P}(\pi;\xi)\, w$, then $w \leq v(\pi;\xi)$.

Proof For a square matrix $A \in \mathbb{R}^{n\times n}$, let $\mathrm{Adj}(A)$ and $\det(A)$ denote the adjugate matrix and the determinant of A, respectively. From equation (8), we see that

$$v(\pi;\xi) \;=\; \left[I - \lambda\, \widehat{P}(\pi;\xi)\right]^{-1} \widehat{r}(\pi;\xi) \;=\; \frac{\mathrm{Adj}\left(I - \lambda\, \widehat{P}(\pi;\xi)\right) \widehat{r}(\pi;\xi)}{\det\left(I - \lambda\, \widehat{P}(\pi;\xi)\right)} \qquad \forall\, \xi \in \Xi. \tag{9}$$

Here, the first identity follows from the matrix inversion lemma, see e.g. Theorem C.2 in [19], while the second equality is due to Cramer's rule. The adjugate matrix and the determinant in (9) constitute polynomials in π and ξ, and the matrix inversion lemma guarantees that the determinant is nonzero throughout Ξ. Hence, the fraction on the right hand-side of (9) has bounded first derivative on $\Pi \times \Xi$, which implies that it is Lipschitz continuous on $\Pi \times \Xi$. We have thus proven assertion (A).

Assertions (b) and (c) follow directly from Theorems 6.1.1 and 6.2.2 in [19], respectively. ■

Proposition 3.1 allows us to reformulate the worst-case expected total reward (6) as follows.

$$\begin{aligned}
\inf_{\xi \in \Xi} \left\{ p_0^\top v(\pi;\xi) \right\} &= \inf_{\xi \in \Xi} \sup_{w \in \mathbb{R}^S} \left\{ p_0^\top w \; : \; w \leq \widehat{r}(\pi;\xi) + \lambda \, \widehat{P}(\pi;\xi) \, w \right\} \\
&= \sup_{\vartheta : \Xi \mapsto \mathbb{R}^S} \left\{ \inf_{\xi \in \Xi} \left\{ p_0^\top \vartheta(\xi) \right\} \; : \; \vartheta(\xi) \leq \widehat{r}(\pi;\xi) + \lambda \, \widehat{P}(\pi;\xi) \, \vartheta(\xi) \;\; \forall \, \xi \in \Xi \right\} \\
&= \sup_{\vartheta : \Xi \overset{c}{\mapsto} \mathbb{R}^S} \left\{ \inf_{\xi \in \Xi} \left\{ p_0^\top \vartheta(\xi) \right\} \; : \; \vartheta(\xi) \leq \widehat{r}(\pi;\xi) + \lambda \, \widehat{P}(\pi;\xi) \, \vartheta(\xi) \;\; \forall \, \xi \in \Xi \right\} \qquad (10)
\end{aligned}$$

Here, the first equality follows from Proposition 3.1 (b)–(c) and non-negativity of p_0, while the last equality follows from Proposition 3.1 (a). Theorem 2.6 implies that (10) is intractable for general uncertainty sets. In the following, we approximate (10) by replacing the space of continuous functions in the outer supremum with the subspaces of constant, affine and piecewise affine functions. Since the policy π is fixed in this section, we may omit the dependence of v, $\widehat{P}$ and $\widehat{r}$ on π in the following.

3.1 Robust Policy Evaluation over s-Rectangular Uncertainty Sets

We show that the policy evaluation problem (10) is optimised by a constant reward to-go function if the uncertainty set $\mathcal{P}$ is s-rectangular. The result also points out an efficient method to solve problem (10).

Theorem 3.2 *For an s-rectangular uncertainty set $\mathcal{P}$, the policy evaluation problem (10) is optimised by the constant reward to-go function $\vartheta^*(\xi) := w^*$, $\xi \in \Xi$, where $w^* \in \mathbb{R}^S$ is the unique fixed point of the contraction mapping $\phi(\pi;\cdot) : \mathbb{R}^S \mapsto \mathbb{R}^S$ defined through*

$$\phi_s(\pi; w) := \min_{\xi^s \in \Xi} \left\{ \widehat{r}_s(\pi;\xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\pi;\xi^s) \, w \right\} \qquad \forall \, s \in \mathcal{S}. \qquad (11)$$

Remark 3.3 *A function $\varphi : \mathbb{R}^S \mapsto \mathbb{R}^S$ is called* contraction mapping *if there is some $\gamma \in [0,1)$ such that $\|\varphi(w) - \varphi(w')\| \leq \gamma \, \|w - w'\|$ for all $w, w' \in \mathbb{R}^S$. The iterated application of φ to any $w \in \mathbb{R}^S$ converges to the unique fixed point w^* that satisfies $w^* = \varphi(w^*)$, see [19].*

Proof of Theorem 3.2 We prove the assertion in two steps. We first show that w^* solves the restriction of the policy evaluation problem (10) to constant reward to-go functions:

$$\sup_{w \in \mathbb{R}^S} \left\{ p_0^\top w \; : \; w \leq \widehat{r}(\xi) + \lambda \widehat{P}(\xi) \, w \;\; \forall \, \xi \in \Xi \right\} \qquad (12)$$

Afterwards, we prove that the optimal values of (10) and (12) coincide for s-rectangular uncertainty sets.

In view of the first step, we note that the objective function of (12) is linear in w. Moreover, the feasible region of (12) is closed because it results from the intersection of closed halfspaces parametrised by $\xi \in \Xi$. Since $w = 0$ is feasible in (12), we can append the constraint $w \geq 0$ without changing the optimal value of (12). Hence, the feasible region is also bounded, and we can apply Weierstrass' extreme value theorem to replace the supremum in (12) with a maximum. Since each of the S one-dimensional inequality constraints in (12) has to be satisfied for all $\xi \in \Xi$, (12) is equivalent to

$$\max_{w \in \mathbb{R}^S} \left\{ p_0^\top w \, : \, w_s \leq \widehat{r}_s(\xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\xi^s) \, w \ \ \forall \, s \in \mathcal{S}, \ \xi^1, \ldots, \xi^S \in \Xi \right\}.$$

We can reformulate the semi-infinite constraints in this problem to obtain

$$\max_{w \in \mathbb{R}^S} \left\{ p_0^\top w \, : \, w_s \leq \min_{\xi^s \in \Xi} \left\{ \widehat{r}_s(\xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\xi^s) \, w \right\} \ \ \forall \, s \in \mathcal{S} \right\}. \tag{13}$$

Note that the constraints in (13) are equivalent to $w \leq \phi(\pi; w)$, where ϕ is defined in (11). One can adapt the results in [11, 17] to show that $\phi(\pi; \cdot)$ is a contraction mapping. Hence, the Banach fixed point theorem guarantees existence and uniqueness of $w^* \in \mathbb{R}^S$. This vector w^* is feasible in (13), and any feasible solution $w \in \mathbb{R}^S$ to (13) satisfies $w \leq \phi(\pi; w)$. According to Theorem 6.2.2 in [19], this implies that $w^* \geq w$ for every feasible solution w to (13). By non-negativity of p_0, w^* must therefore maximise (13). Since (12) and (13) are equivalent, we have thus shown that w^* maximises (12).

We now prove that the optimal values of (10) and (13) coincide if $\mathcal{P}$ is s-rectangular. Since (13) is maximised by the unique fixed point w^* of $\phi(\pi; \cdot)$, we can reexpress (13) as

$$\min_{w \in \mathbb{R}^S} \left\{ p_0^\top w \, : \, w_s = \min_{\xi^s \in \Xi} \left\{ \widehat{r}_s(\xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\xi^s) \, w \right\} \ \ \forall \, s \in \mathcal{S} \right\}.$$

This problem is equivalent to

$$\min_{w \in \mathbb{R}^S} \ \min_{\substack{\xi^s \in \Xi: \\ s \in \mathcal{S}}} \left\{ p_0^\top w \, : \, w_s = \widehat{r}_s(\xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\xi^s) \, w \ \ \forall \, s \in \mathcal{S} \right\}. \tag{14}$$

The s-rectangularity of the uncertainty set $\mathcal{P}$ implies that (14) can be reformulated as

$$\min_{w \in \mathbb{R}^S} \ \min_{\xi \in \Xi} \left\{ p_0^\top w \, : \, w_s = \widehat{r}_s(\xi) + \lambda \widehat{P}_{s\cdot}^\top(\xi) \, w \ \ \forall \, s \in \mathcal{S} \right\}. \tag{15}$$

For a fixed $\xi \in \Xi$, $w = v(\xi)$ is the unique feasible solution to (15), see Proposition 3.1 (b). By Weierstrass' extreme value theorem, (15) is therefore equivalent to the policy evaluation problem (10). ■

The fixed point w^* of the contraction mapping $\phi(\pi; \cdot)$ defined in (11) can be found by applying the following *robust value iteration*. We start with an initial estimate $w^1 := 0$. In the ith iteration, $i = 1, 2, \ldots$, we determine the updated estimate w^{i+1} via $w^{i+1} := \phi(\pi; w^i)$. Since $\phi(\pi; \cdot)$ is a contraction mapping, the Banach fixed point theorem guarantees that the sequence w^i converges to w^* at a geometric rate. The following corollary investigates the computational complexity of this approach.

Corollary 3.4 *If the uncertainty set $\mathcal{P}$ is s-rectangular, then problem (10) can be solved to any accuracy ϵ in polynomial time $\mathcal{O}\left(q^3 L^{3/2} S \log^2 \epsilon^{-1} + qAS^2 \log \epsilon^{-1}\right)$.*

Proof Assume that at each iteration i of the robust value iteration, we evaluate $\phi(\pi; w^i)$ to the accuracy $\delta := \epsilon(1-\lambda)^2/(4+4\lambda)$. We stop the algorithm as soon as $\left\|w^{N+1} - w^N\right\|_\infty \leq \epsilon(1-\lambda)/(1+\lambda)$ at some iteration N. This is guaranteed to happen within $\mathcal{O}\left(\log \epsilon^{-1}\right)$ iterations [19]. By construction, w^{N+1} is feasible for the policy evaluation problem (10), see [19]. We can adapt Theorem 5 from [17] to show that w^{N+1} satisfies $\left\|w^{N+1} - w^*\right\|_\infty \leq \epsilon$. Hence, w^{N+1} is also an ϵ-optimal solution to (10).

We now investigate the complexity of evaluating ϕ to the accuracy δ. Under mild assumptions, interior point methods can solve second-order cone programs of the form

$$\min_{x \in \mathbb{R}^n} \left\{ f^\top x \; : \; \|A_j x + b_j\|_2 \leq c_j^\top x + d_j \;\; \forall j = 1, \ldots, m \right\},$$

where $A_j \in \mathbb{R}^{n_j \times n}$, $b_j \in \mathbb{R}^{n_j}$, $c_j \in \mathbb{R}^n$ and $d_j \in \mathbb{R}$, $j = 1, \ldots, m$, to any accuracy δ in polynomial time $\mathcal{O}\left(\sqrt{m}\left[n^3 + n^2 \sum_j n_j\right] \log \delta^{-1}\right)$, see [14]. For $w \in \mathbb{R}^S$, we can evaluate $\phi(\pi; w)$ by solving the following second-order cone program:

$$\begin{aligned}
&\underset{\xi}{\text{minimise}} && \sum_{a \in \mathcal{A}} \pi(a|s) \left(k_{sa} + K_{sa}\xi\right)^\top (r_{sa} + \lambda w) && \text{(16a)} \\
&\text{subject to} && \xi \in \mathbb{R}^q && \\
& && \left\| \begin{bmatrix} \Omega_l \\ -o_l^\top \end{bmatrix} \xi + \begin{bmatrix} 0 \\ \frac{1-\omega_l}{2} \end{bmatrix} \right\|_2 \leq o_l^\top \xi + \frac{\omega_l + 1}{2} \qquad \forall l = 1, \ldots, L, && \text{(16b)}
\end{aligned}$$

where $(r_{sa})_{s'} := r(s, a, s')$ for $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and Ω_l satisfies $\Omega_l^\top \Omega_l = O_l$. We can determine each matrix Ω_l in time $\mathcal{O}\left(q^3\right)$ by a Cholesky decomposition, we can construct (16) in time $\mathcal{O}\left(qAS + q^2 L\right)$, and we can solve (16) to accuracy δ in time $\mathcal{O}\left(q^3 L^{3/2} \log \delta^{-1}\right)$. Each step of the robust value iteration requires the construction and solution of S such problems. Since the constraints of (16) only need to be generated once, this results in an iteration complexity of $\mathcal{O}\left(q^3 L^{3/2} S \log \delta^{-1} + qAS^2\right)$. The assertion now follows from the fact that the robust value iteration terminates within $\mathcal{O}\left(\log \epsilon^{-1}\right)$ iterations. ∎

Depending on the properties of Ξ defined in (3b), we can evaluate the mapping ϕ more efficiently. We refer to [11, 17] for a discussion of different numerical schemes.

Remark 3.5 (Finite Horizon MDPs) *For a finite horizon MDP, we can solve the policy evaluation problem (10) over an s-rectangular uncertainty set $\mathcal{P}$ via* robust backward induction *as follows. We start* *with $w^T \in \mathbb{R}^S$ defined through $w_s^T := \mathbf{r}_s$ if $s \in \mathcal{S}_T$; $:= 0$ otherwise. At iteration $i = T-1, T-2, \ldots, 1$, we determine w^i through $w_s^i := \widehat{\phi}_s(\pi; w^{i+1})$ if $s \in \mathcal{S}_i$; $:= w_s^{i+1}$ otherwise. The operator $\widehat{\phi}$ is defined as*

$$\widehat{\phi}_s(\pi; w) := \min_{\xi^s \in \Xi} \left\{ \widehat{r}_s(\pi; \xi^s) + \widehat{P}_{s\cdot}^\top(\pi; \xi^s)\, w \right\} \qquad \forall\, s \in \mathcal{S}.$$

An adaptation of Corollary 3.4 shows that we obtain an ϵ-optimal solution to the policy evaluation problem (10) in time $\mathcal{O}\left(q^3 L^{3/2} S \log \epsilon^{-1} + qAS^2\right)$ if we evaluate $\widehat{\phi}$ to the accuracy $\epsilon/(T-1)$.

3.2 Robust Policy Evaluation over Non-Rectangular Uncertainty Sets

If the uncertainty set $\mathcal{P}$ is non-rectangular, then Theorem 2.6 implies that constant reward to-go functions are no longer guaranteed to optimise the policy evaluation problem (10). Nevertheless, we can still use the robust value iteration to obtain a lower bound on the optimal value of (10).

Proposition 3.6 *Let $\mathcal{P}$ be a non-rectangular uncertainty set, and define $\overline{\mathcal{P}} := \times_{s \in \mathcal{S}} \mathcal{P}_s$ as the smallest s-rectangular uncertainty set that contains $\mathcal{P}$. The function $\vartheta^*(\xi) = w^*$ defined in Theorem 3.2 has the following properties.*

1. *The vector w^* solves the restriction (12) of the policy evaluation problem (10) that approximates the reward to-go function by a constant.*
2. *The function ϑ^* solves the exact policy evaluation problem (10) over $\overline{\mathcal{P}}$.*

Proof The first property follows from the fact that the first part of the proof of Theorem 3.2 does not depend on the structure of the uncertainty set $\mathcal{P}$. As for the second property, the proof of Theorem 3.2 shows that w^* minimises (14), irrespective of the structure of $\mathcal{P}$. The proof also shows that (14) is equivalent to the policy evaluation problem (10) if we replace $\mathcal{P}$ with $\overline{\mathcal{P}}$. ∎

Proposition 3.6 provides a dual characterisation of the robust value iteration. On one hand, the robust value iteration determines the exact worst-case expected total reward over the rectangularised uncertainty set $\overline{\mathcal{P}}$. On the other hand, the robust value iteration calculates a lower bound on the worst-case expected total reward over the original uncertainty set $\mathcal{P}$. Hence, rectangularising the uncertainty set is equivalent to replacing the space of continuous reward to-go functions in the policy evaluation problem (10) with the subspace of constant functions.

We obtain a tighter lower bound on the worst-case expected total reward (10) if we replace the space of continuous reward to-go functions with the subspaces of affine or piecewise affine functions. We use the following result to formulate these approximations as tractable optimisation problems.

Proposition 3.7 *For Ξ defined in (3b) and any fixed $S \in \mathbb{S}^q$, $s \in \mathbb{R}^q$ and $\sigma \in \mathbb{R}$, we have*

$$\exists\, \gamma \in \mathbb{R}^L_+ \; : \; \begin{bmatrix} \sigma & \frac{1}{2}s^\top \\ \frac{1}{2}s & S \end{bmatrix} - \sum_{l=1}^{L} \gamma_l \begin{bmatrix} \omega_l & \frac{1}{2}o_l^\top \\ \frac{1}{2}o_l & O_l \end{bmatrix} \succeq 0 \qquad \Longrightarrow \qquad \xi^\top S\, \xi + s^\top \xi + \sigma \geq 0 \quad \forall \xi \in \Xi. \quad (17)$$

Furthermore, the reversed implication holds if (C1) $L = 1$ or (C2) $S \succeq 0$.

Proof Implication (17) and the reversed implication under condition (C1) follow from the approximate and exact versions of the $\mathcal{S}$-Lemma, respectively (see e.g. Proposition 3.4 in [13]).

Assume now that (C2) holds. We define $f(\xi) := \xi^\top S\, \xi + s^\top \xi + \sigma$ and $g_l(\xi) := -\xi^\top O_l\, \xi - o_l^\top \xi - \omega_l$, $l = 1, \ldots, L$. Since f and $g := (g_1, \ldots, g_L)$ are convex, Farkas' Theorem [20] ensures that the system

$$f(\xi) < 0, \quad g(\xi) < 0, \quad \xi \in \mathbb{R}^q \qquad (18a)$$

has no solution if and only if there is a nonzero vector $(\kappa, \gamma) \in \mathbb{R}_+ \times \mathbb{R}^L_+$ such that

$$\kappa f(\xi) + \gamma^\top g(\xi) \geq 0 \quad \forall \xi \in \mathbb{R}^q. \qquad (18b)$$

Since Ξ contains a Slater point $\overline{\xi}$ that satisfies $\overline{\xi}^\top O_l\, \overline{\xi} + o_l^\top \overline{\xi} + \omega = -g_l(\overline{\xi}) > 0$, $l = 1, \ldots, L$, convexity of g and continuity of f allows us to replace the second strict inequality in (18a) with a less or equal constraint. Hence, (18a) has no solution if and only if f is non-negative on $\Xi = \{\xi \in \mathbb{R}^q : g(\xi) \leq 0\}$, that is, if the right-hand side of (17) is satisfied. We now show that (18b) is equivalent to the left-hand side of (17). Assume that there is a nonzero vector $(\kappa, \gamma) \geq 0$ that satisfies (18b). Note that $\kappa \neq 0$ since otherwise, (18b) would not be satisfied by the Slater point $\overline{\xi}$. Hence, a suitable scaling of γ allows us to set $\kappa := 1$. For our choice of f and g, this implies that (18b) is equivalent to

$$\begin{bmatrix} 1 \\ \xi \end{bmatrix}^\top \left(\begin{bmatrix} \sigma & \frac{1}{2}s^\top \\ \frac{1}{2}s & S \end{bmatrix} - \sum_{l=1}^{L} \gamma_l \begin{bmatrix} \omega_l & \frac{1}{2}o_l^\top \\ \frac{1}{2}o_l & O_l \end{bmatrix} \right) \begin{bmatrix} 1 \\ \xi \end{bmatrix} \geq 0 \qquad \forall \xi \in \mathbb{R}^q. \qquad (18b')$$

Since the above inequality is homogeneous of degree 2 in $\left[1,\ \xi^\top\right]^\top$, it extends to the whole of $\mathbb{R}^{q+1}$. Hence, (18b') is equivalent to the left-hand side of (17). ■

Proposition 3.7 allows us to bound the worst-case expected total reward (10) from below as follows.

Theorem 3.8 *Consider the following variant of the policy evaluation problem (10), which approximates* *the reward to-go function by an affine function,*

$$\sup_{\vartheta:\Xi \overset{\text{a}}{\mapsto} \mathbb{R}^S} \left\{ \inf_{\xi\in\Xi} \left\{p_0^\top \vartheta(\xi)\right\} \; : \; \vartheta(\xi) \le \widehat{r}(\xi) + \lambda \widehat{P}(\xi)\,\vartheta(\xi) \;\; \forall \xi \in \Xi \right\}, \tag{19}$$

as well as the semidefinite program

$$\begin{aligned}
\underset{\tau,w,W,\gamma,\Gamma}{\text{maximise}} \quad & \tau && \text{(20a)}\\
\text{subject to} \quad & \tau \in \mathbb{R}, \quad w \in \mathbb{R}^S, \quad W \in \mathbb{R}^{S\times q}, \quad \gamma \in \mathbb{R}^L_+, \quad \Gamma \in \mathbb{R}^{S\times L}_+ \\
& \begin{bmatrix} p_0^\top w - \tau & \frac{1}{2} p_0^\top W \\ \frac{1}{2} W^\top p_0 & 0 \end{bmatrix} - \sum_{l=1}^L \gamma_l \begin{bmatrix} \omega_l & \frac{1}{2} o_l^\top \\ \frac{1}{2} o_l & O_l \end{bmatrix} \succeq 0, && \text{(20b)}\\
& \sum_{a\in\mathcal{A}} \pi(a|s) \begin{bmatrix} k_{sa}^\top \left(r_{sa} + \lambda w\right) & \frac{1}{2}\left(r_{sa}^\top K_{sa} + \lambda\left[k_{sa}^\top W + w^\top K_{sa}\right]\right) \\ \frac{1}{2}\left(K_{sa}^\top r_{sa} + \lambda\left[W^\top k_{sa} + K_{sa}^\top w\right]\right) & \lambda K_{sa}^\top W \end{bmatrix} \\
& \qquad - \begin{bmatrix} w_s & \frac{1}{2} W_{s\cdot}^\top \\ \frac{1}{2}\left(W_{s\cdot}^\top\right)^\top & 0 \end{bmatrix} - \sum_{l=1}^L \Gamma_{sl} \begin{bmatrix} \omega_l & \frac{1}{2} o_l^\top \\ \frac{1}{2} o_l & O_l \end{bmatrix} \succeq 0 \qquad \forall\, s \in \mathcal{S}, && \text{(20c)}
\end{aligned}$$

where $(r_{sa})_{s'} := r(s,a,s')$ *for* $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$*. Let* $(\tau^*, w^*, W^*, \gamma^*, \Gamma^*)$ *denote an optimal solution to (20), and define* $\vartheta^* : \Xi \overset{\text{a}}{\mapsto} \mathbb{R}^S$ *through* $\vartheta^*(\xi) := w^* + W^*\xi$*. We have that:*

(a) If $L = 1$*, then (19) and (20) are* equivalent *in the following sense:* τ^* *coincides with the supremum of (19), and* ϑ^* *is feasible and optimal in (19).*

(b) If $L > 1$*, then (20) constitutes a* conservative approximation *for (19):* τ^* *provides a lower bound on the supremum of (19), and* ϑ^* *is feasible in (19) and satisfies* $\inf_{\xi\in\Xi}\left\{p_0^\top \vartheta^*(\xi)\right\} = \tau^*$.

Proof The approximate policy evaluation problem (19) can be written as

$$\sup_{\substack{w\in\mathbb{R}^S,\\ W\in\mathbb{R}^{S\times q}}} \left\{ \inf_{\xi\in\Xi} \left\{p_0^\top \left(w + W\xi\right)\right\} \; : \; w + W\xi \le \widehat{r}(\xi) + \lambda \widehat{P}(\xi)\left(w + W\xi\right) \;\; \forall \xi \in \Xi \right\}. \tag{21}$$

We first show that (21) is solvable. Since $p_0^\top\left(w + W\xi\right)$ is linear in (w, W) and continuous in ξ while Ξ is compact, $\inf_{\xi\in\Xi}\left\{p_0^\top\left(w + W\xi\right)\right\}$ is a concave and therefore continuous function of (w, W). Likewise, the feasible region of (21) is closed because it results from the intersection of closed halfspaces parametrised by $\xi \in \Xi$. However, the feasible region of (21) is *not* bounded because any non-positive constant reward to-go function, that is, any (w, W) with $w \in \mathbb{R}_-$ and $W = 0$, constitutes a feasible solution. However, since $(w, W) = (0, 0)$ is feasible, we can append the constraint $w + W\xi \ge 0$ for all $\xi \in \Xi$ without changing the optimal value of (21). Moreover, all expected rewards $r(s, a, s')$ are bounded from above by $\overline{r} := \max_{s,a,s'} \{r(s,a,s')\}$. Therefore, Proposition 3.1 (c) implies that any feasible solution (w, W) for (21) satisfies $w + W\xi \leq \overline{r}\mathrm{e}/(1-\lambda)$ for all $\xi \in \Xi$.

Our results so far imply that any feasible solution (w, W) for (21) satisfies $0 \leq w + W\xi \leq \overline{r}\mathrm{e}/(1-\lambda)$ for all $\xi \in \Xi$. We now show that this implies boundedness of the feasible region for (w, W). The existence of a Slater point $\overline{\xi}$ with $\overline{\xi}^\top O_l \overline{\xi} + o_l^\top \xi + \omega_l > 0$ for all $l = 1, \ldots, L$ guarantees that there is an ϵ-neighbourhood of $\overline{\xi}$ that is contained in Ξ. Hence, W must be bounded because all points ξ in this neighbourhood satisfy $0 \leq w + W\xi \leq \overline{r}\mathrm{e}/(1-\lambda)$. As a consequence, w is bounded as well since $0 \leq w + W\overline{\xi} \leq \overline{r}\mathrm{e}/(1-\lambda)$. Thus, the feasible region of (21) is bounded, and Weierstrass' extreme value theorem is applicable. Therefore, (21) is solvable. If we furthermore replace $\widehat{P}$ and $\widehat{r}$ with their definitions from (7) and go over to an epigraph formulation, we obtain

$$\begin{aligned}
&\underset{\tau, w, W}{\text{maximise}} \quad && \tau && \text{(22a)}\\
&\text{subject to} && \tau \in \mathbb{R}, \quad w \in \mathbb{R}^S, \quad W \in \mathbb{R}^{S \times q} && \\
& && \tau \leq p_0^\top (w + W\xi) \qquad \forall \xi \in \Xi && \text{(22b)}\\
& && w_s + W_{s\cdot}^\top \xi \leq \sum_{a \in \mathcal{A}} \pi(a|s) \left(k_{sa} + K_{sa}\xi\right)^\top \left(r_{sa} + \lambda \left[w + W\xi\right]\right) \qquad \forall \xi \in \Xi,\ s \in \mathcal{S}. && \text{(22c)}
\end{aligned}$$

Constraint (22b) is equivalent to constraint (20b) by Proposition 3.7 under condition (C2). Likewise, Proposition 3.7 guarantees that constraint (22c) is implied by constraint (20c). Moreover, if $L = 1$, condition (C1) of Proposition 3.7 is satisfied, and both constraints are equivalent. ■

We can employ conic duality [1, 14] to equivalently replace constraint (20b) with conic quadratic constraints. There does not seem to be a conic quadratic reformulation of constraint (20c), however.

Theorem 3.8 provides an exact (for $L = 1$) or conservative (for $L > 1$) reformulation for the approximate policy evaluation problem (19). Since (19) optimises only over affine approximations of the reward to-go function, Proposition 3.1 (c) implies that (19) provides a conservative approximation for the worst-case expected total reward (10). We will see below that both approximations are tight for s-rectangular uncertainty sets. First, however, we investigate the computational complexity of problem (20).

Corollary 3.9 *The semidefinite program (20) can be solved to any accuracy ϵ in polynomial time $\mathcal{O}\big((qS + LS)^{\frac{5}{2}}(q^2S + LS)\log \epsilon^{-1} + q^2AS^2\big)$.*

Proof The objective function and constraints of (20) can be constructed in time $\mathcal{O}\big(q^2AS^2 + q^2LS\big)$. Under mild assumptions, interior point methods can solve semidefinite programs of the type

$$\min_{x \in \mathbb{R}^n} \left\{ c^\top x \, : \, F_0 + \sum_{i=1}^{n} x_i F_i \succeq 0 \right\},$$

where $F_i \in \mathbb{S}^m$ for $i = 0, \ldots, n$, to accuracy ϵ in time $\mathcal{O}\big(n^2 m^{\frac{5}{2}} \log \epsilon^{-1}\big)$, see [23]. Moreover, if all matrices F_i possess a block-diagonal structure with blocks $G_{ij} \in \mathbb{S}^{m_j}$, $j = 1, \ldots, J$ with $\sum_j m_j = m$, then the computational effort can be reduced to $\mathcal{O}\big(n^2 m^{\frac{1}{2}} \sum_j m_j^2\big)$. Problem (20) involves $\mathcal{O}(qS + LS)$ variables. By exploiting the block-diagonal structure of (20), constraint (20b) gives rise to a single block of dimension $(q+1) \times (q+1)$, constraint set (20c) leads to S blocks of dimension $(q+1) \times (q+1)$ each, and non-negativity of γ and Γ results in L and SL one-dimensional blocks, respectively. ■

In Section 4 we discuss a method for constructing uncertainty sets from observation histories. Asymptotically, this method generates an uncertainty set Ξ that is described by a single quadratic inequality ($L = 1$), which means that problem (20) can be solved in time $\mathcal{O}\big(q^{\frac{9}{2}} S^{\frac{7}{2}} \log \epsilon^{-1} + q^2 A S^2\big)$. Note that q does not exceed $S(S-1)A$, the affine dimension of the space $[\mathcal{M}(\mathcal{S})]^{S \times A}$, unless some components of ξ are perfectly correlated. If information about the structure of the transition kernel is available, however, q can be much smaller. Section 6 provides an example in which q remains constant as the problem size (measured in terms of S, the number of states) increases.

The semidefinite program (20) is based on two approximations. It is a conservative approximation for problem (19), which itself is a restriction of the policy evaluation problem (10) to affine reward to-go functions. We now show that both approximations are tight for s-rectangular uncertainty sets.

Proposition 3.10 *Let $(\tau^*, w^*, W^*, \gamma^*, \Gamma^*)$ denote an optimal solution to the semidefinite program (20), and define $\vartheta^* : \Xi \mapsto \mathbb{R}^S$ through $\vartheta^*(\xi) := w^* + W^* \xi$. If the uncertainty set $\mathcal{P}$ is s-rectangular, then the optimal value of the policy evaluation problem (10) is τ^*, and ϑ^* is feasible and optimal in (10).*

Proof We show that any constant reward to-go function that is feasible for the policy evaluation problem (10) can be extended to a feasible solution of the semidefinite program (20) with the same objective value. The assertion then follows from the optimality of constant reward to-go functions for s-rectangular uncertainty sets, see Theorem 3.2, and the fact that (20) bounds (10) from below, see Theorem 3.8.

Assume that $\vartheta : \Xi \mapsto \mathbb{R}^S$ with $\vartheta(\xi) = c$ for all $\xi \in \Xi$ satisfies the constraints of the policy evaluation problem (10). We show that there is $\gamma \in \mathbb{R}_+^L$ and $\Gamma \in \mathbb{R}_+^{S \times L}$ such that $(\tau, w, W, \gamma, \Gamma)$ with $\tau := p_0^\top c$, $w := c$ and $W := 0$ satisfies the constraints of the semidefinite program (20). Since $\tau = \inf_{\xi \in \Xi} \big\{ p_0^\top \vartheta(\xi) \big\}$, ϑ in (10) and $(\tau, w, W, \gamma, \Gamma)$ in (20) clearly attain equal objective values.

By the proof of Theorem 3.8, there is $\gamma \in \mathbb{R}_+^L$ that satisfies constraint (20b) if and only if $\tau \leq p_0^\top (w + W\xi)$ for all $\xi \in \Xi$. Since $w + W\xi = c$ for all $\xi \in \Xi$ and $\tau = p_0^\top c$, such a γ indeed exists.

Let us now consider constraint set (20c). Since the constant reward to-go function $\vartheta(\xi) = c$ is feasible in the policy evaluation problem (10), we have for state $s \in \mathcal{S}$ that

$$c_s \leq \widehat{r}_s(\xi) + \lambda \widehat{P}_{s\cdot}^\top(\xi)\, c \qquad \forall\, \xi \in \Xi.$$

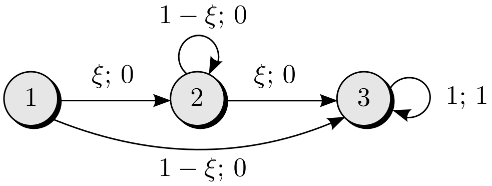

Figure 6: *MDP with three states and one action.* p_0 *places unit probability mass on state* 1. *The same drawing conventions as in Figure 3 are used.*

If we replace $\widehat{r}$ and $\widehat{P}$ with their definitions from (7), this is equivalent to

$$c_s \leq \sum_{a \in \mathcal{A}} \pi(a|s)(k_{sa} + K_{sa}\xi)^\top (r_{sa} + \lambda c) \qquad \forall \xi \in \Xi,$$

which is an instance of constraint (22c) where $w = c$ and $W = 0$. For this choice of (w, W), Proposition 3.7 under condition (C2) is applicable to constraint (22c). Hence, (22c) is satisfied if and only if there is $\Gamma_{s\cdot}^\top \in \mathbb{R}_+^{1 \times L}$ that satisfies constraint (20c). Since (22c) is satisfied, we conclude that we can indeed find γ and Γ such that $(\tau, w, W, \gamma, \Gamma)$ satisfies the constraints of the semidefinite program (20). ∎

Propositions 3.6 and 3.10 show that the lower bound provided by the robust value iteration is dominated by the bound obtained from the semidefinite program (20). The following example highlights that the quality of these bounds can differ substantially.

Example 3.11 *Consider the robust infinite horizon MDP that is visualised in Figure 6. The uncertainty set* $\mathcal{P}$ *encompasses all transition kernels that correspond to parameter realisations* $\xi \in [0, 1]$. *This MDP can be assigned an uncertainty set of the form (3). For* $\lambda := 0.9$, *the worst-case expected total reward is* $\lambda^2/(1 - \lambda) = 8.1$ *and is incurred under the transition kernel corresponding to* $\xi = 1$. *The solution of the semidefinite program (20) yields the (affine) approximate reward to-go function* $\vartheta^*(\xi) = (6.5,\ 9\xi,\ 10)^\top$ *and therefore provides a lower bound of* 6.5. *The unique solution to the fixed point equations* $w^* = \phi(w^*)$, *where* ϕ *is defined in (11), is* $w^* = (0,\ 0,\ 1/[1 - \lambda])$. *Hence, the best constant reward to-go approximation yields a lower bound of zero. Since all expected rewards are non-negative, this is a trivial bound. Intuitively, the poor performance of the constant reward to-go function is due to the fact that it considers separate worst-case parameter realisations for states 1 (*$\xi = 1$*) and 2 (*$\xi = 0$*).*

Example 3.11 shows that the semidefinite program (20) generically provides a strict lower bound on the worst-case expected total reward if the uncertainty set is non-rectangular. In such cases, we would like to estimate the incurred approximation error. Note that we obtain an *upper* (i.e., optimistic) bound on the worst-case expected total reward if we evaluate $p_0^\top v(\xi)$ for any single $\xi \in \Xi$. Let $\vartheta^*(\xi)$ denote an optimal affine approximation of the reward to-go function obtained from the semidefinite program (20). This ϑ^* can be used to obtain a suboptimal solution to $\arg\min \left\{ p_0^\top v(\xi) \,:\, \xi \in \Xi \right\}$ by solving

$\arg\min \left\{ p_0^\top \vartheta^*(\xi) \,:\, \xi \in \Xi \right\}$, which is a convex optimisation problem. Let ξ^* denote an optimal solution to this problem. We obtain an upper bound on the worst-case expected total reward by evaluating

$$p_0^\top v(\xi^*) \;=\; p_0^\top \sum_{t=0}^{\infty} \left[\lambda \widehat{P}(\xi^*) \right]^t \widehat{r}(\xi^*) \;=\; p_0^\top \left[I - \lambda \widehat{P}(\xi^*) \right]^{-1} \widehat{r}(\xi^*), \tag{23}$$

where the last equality follows from the matrix inversion lemma, see e.g. Theorem C.2 in [19]. We can thus estimate the approximation error of the semidefinite program (20) by evaluating the difference between (23) and the optimal value of (20). If this difference is large, the affine approximation of the reward to-go function may be too crude. In this case, one could use modern decision rule techniques [3, 10] to reduce the approximation error via piecewise affine approximations of the reward to-go function. Since the resulting generalisation requires no new ideas, we omit details for the sake of brevity.

Remark 3.12 (Finite Horizon MDPs) *Our results can be directly applied to finite horizon MDPs if we convert them to infinite horizon MDPs. To this end, we choose any discounting factor λ and multiply the rewards associated with transitions in period $t \in \mathcal{T}$ by λ^{-t}. Moreover, for every terminal state $s \in \mathcal{S}_T$, we introduce a deterministic transition to an auxiliary absorbing state and assign an action-independent expected reward of $\lambda^{-T} \mathbf{r}_s$. Note that in contrast to non-robust and rectangular MDPs, the approximate policy evaluation problem (20) does not decompose into separate subproblems for each time period $t \in \mathcal{T}$.*

4 Robust Policy Improvement

In view of (10), we can formulate the policy improvement problem as

$$\sup_{\pi \in \Pi} \; \sup_{\vartheta : \Xi \overset{c}{\mapsto} \mathbb{R}^S} \left\{ \inf_{\xi \in \Xi} \left\{ p_0^\top \vartheta(\xi) \right\} \,:\, \vartheta(\xi) \leq \widehat{r}(\pi;\xi) + \lambda \, \widehat{P}(\pi;\xi) \, \vartheta(\xi) \;\; \forall \, \xi \in \Xi \right\}. \tag{24}$$

Since π is no longer fixed in this section, we make the dependence of v, $\widehat{P}$ and $\widehat{r}$ on π explicit. Section 3 shows that the policy evaluation problem can be solved efficiently if the uncertainty set $\mathcal{P}$ is s-rectangular. We now extend this result to the policy improvement problem.

Theorem 4.1 *For an s-rectangular uncertainty set $\mathcal{P}$, the policy improvement problem (24) is optimised by the policy $\pi^* \in \Pi$ and the constant reward to-go function $\vartheta^*(\xi) := w^*$, $\xi \in \Xi$, that are defined as follows. The vector $w^* \in \mathbb{R}^S$ is the unique fixed point of the contraction mapping φ defined through*

$$\varphi_s(w) := \max_{\pi \in \Pi} \left\{ \phi_s(\pi; w) \right\} \qquad \forall \, s \in \mathcal{S}, \tag{25}$$

where ϕ is defined in (11). For each $s \in \mathcal{S}$, let $\pi^s \in \arg\max_{\pi \in \Pi} \left\{ \phi_s(\pi; w^*) \right\}$ denote a policy that attains *the maximum on the right-hand side of (25) for $w = w^*$. Then $\pi^*(a|s) := \pi^s(a|s)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Proof In analogy to the proof of Theorem 3.2, we can rewrite the policy improvement problem (24) as

$$\max_{\pi \in \Pi} \max_{w \in \mathbb{R}^S} \left\{ p_0^\top w \,:\, w_s \leq \widehat{r}_s(\pi; \xi^s) + \lambda \, \widehat{P}_{s\cdot}^\top(\pi; \xi^s) \, w \;\; \forall \, s \in \mathcal{S}, \; \xi^1, \ldots, \xi^S \in \Xi \right\}.$$

By definition of ϕ, the S semi-infinite constraints in this problem are equivalent to the constraint $w \leq \phi(\pi; w)$. If we interchange the order of the maximum operators, we can reexpress the problem as

$$\max_{w \in \mathbb{R}^S} \left\{ p_0^\top w \,:\, \exists \, \pi \in \Pi \;\text{ such that }\; w \leq \phi(\pi; w) \right\}. \tag{26}$$

Note that ϕ_s only depends on the components $\pi(\cdot|s)$ of π. Hence, we have $w^* = \phi(\pi^*; w^*)$, and π^* and w^* are feasible in (26). One can adapt the results in [11, 17] to show that φ is a contraction mapping. Since $w^* = \varphi(w^*)$ and every feasible solution w to (26) satisfies $w \leq \varphi(w)$, Theorem 6.2.2 in [19] therefore implies that $w^* \geq w$ for all feasible vectors w. By non-negativity of p_0, π^* and w^* must then be optimal in (26). The assertion now follows from the equivalence of (24) and (26). ∎

The fixed point w^* of the contraction mapping φ defined in (25) can be found via robust value iteration, see Section 3.1. The following result analyses the complexity of this method.

Corollary 4.2 *The fixed point w^* of the contraction mapping φ defined in (25) can be determined to any accuracy ϵ in polynomial time $\mathcal{O}\left((q + A + L)^{1/2}(qL + A)^3 S \log^2 \epsilon^{-1} + qAS^2 \log \epsilon^{-1}\right)$.*

Proof We apply the robust value iteration presented in Section 3.1 to the contraction mapping φ. To evaluate $\varphi_s(w)$, we solve the following semi-infinite optimisation problem:

$$\begin{aligned}
&\underset{\tau, \pi}{\text{maximise}} && \tau && \text{(27a)}\\
&\text{subject to} && \tau \in \mathbb{R}, \quad \pi \in \mathbb{R}^A \\
& && \tau \leq \sum_{a \in \mathcal{A}} \pi_a (k_{sa} + K_{sa}\xi)^\top (r_{sa} + \lambda w) \qquad \forall \, \xi \in \Xi, && \text{(27b)}\\
& && \pi \geq 0, \quad \mathrm{e}^\top \pi = 1. && \text{(27c)}
\end{aligned}$$

Second-order cone duality [1, 14] allows us to replace the semi-infinite constraint (27b) with the following linear and conic quadratic constraints:

$$\exists\, Y \in \mathbb{R}^{q \times L},\, z \in \mathbb{R}^L,\, t \in \mathbb{R}^L \; : \qquad \tau - \sum_{a \in \mathcal{A}} \pi_a k_{sa}^\top (r_{sa} + \lambda w) \leq -\sum_{l=1}^{L} \left(\frac{1-\omega_l}{2} z_l + \frac{\omega_l + 1}{2} t_l \right) \tag{27b.1}$$

$$\sum_{l=1}^{L} \left(\Omega_l^\top Y_{\cdot l} - \frac{1}{2} o_l \left[t_l - z_l \right] \right) = \sum_{a \in \mathcal{A}} \pi_a K_{sa}^\top (r_{sa} + \lambda w) \tag{27b.2}$$

$$\left\| \begin{bmatrix} Y_{\cdot l} \\ z_l \end{bmatrix} \right\|_2 \leq t_l \quad \forall\, l = 1, \ldots, L. \tag{27b.3}$$

Here, Ω_l satisfies $\Omega_l^\top \Omega_l = -O_l$. The assertion now follows if we evaluate $\varphi(w^i)$ at iteration i to an accuracy $\delta < \epsilon(1-\lambda)^2/8$ and stop as soon as $\left\| w^{N+1} - w^N \right\|_\infty \leq \epsilon(1-\lambda)/4$ at some iteration N. ∎

In analogy to Remark 3.5, we can solve the policy improvement problem for finite horizon MDPs via robust backward induction in polynomial time $\mathcal{O}\left((q+A+L)^{1/2}(qL+A)^3 S \log \epsilon^{-1} + qAS^2\right)$.

Since the policy improvement problem (24) contains the policy evaluation problem (10) as a special case, Theorem 2.6 implies that (24) is intractable for non-rectangular uncertainty sets. In analogy to Section 3, we can obtain a suboptimal solution to (24) by considering constant approximations of the reward to-go function. The following result is an immediate consequence of Proposition 3.6 and Theorem 4.1.

Corollary 4.3 *For a non-rectangular uncertainty set $\mathcal{P}$, consider the following variant of the policy improvement problem (24), which approximates the reward to-go function by a constant function.*

$$\sup_{\pi \in \Pi} \sup_{w \in \mathbb{R}^S} \left\{ p_0^\top w \; : \; w \leq \widehat{r}(\xi) + \lambda \widehat{P}(\xi)\, w \;\; \forall\, \xi \in \Xi \right\} \tag{28}$$

Problem (28) is optimised by the unique fixed point $w^* \in \mathbb{R}^S$ of the contraction mapping φ defined in (25).

In analogy to Proposition 3.6, the policy improvement problem (24) is equivalent to its approximation (28) if we replace $\mathcal{P}$ with $\times_s \mathcal{P}_s$. We can try to obtain better solutions to (24) over non-rectangular uncertainty sets by replacing the constant reward to-go approximations with affine or piecewise affine approximations. The associated optimisation problems are bilinear semidefinite programs and as such difficult to solve. Nevertheless, we can obtain a suboptimal solution with the following heuristic.

Algorithm 4.1. Sequential convex optimisation procedure.

1. *Initialisation.* Choose $\pi^1 \in \Pi$ (best policy found) and $i := 1$ (iteration counter).

2. *Policy Evaluation.* Solve the semidefinite program (20) for $\pi = \pi^i$ and store the τ-, w- and W-components of the solution in τ^i, w^i and W^i, respectively. Abort if $i > 1$ and $\tau^i = \tau^{i-1}$.

3. *Policy Improvement.* For each $s \in \mathcal{S}$, solve the semi-infinite optimisation problem

$$
\begin{aligned}
\underset{\sigma_s, \pi_s}{\text{maximise}} \quad & \sigma_s && \text{(29a)} \\
\text{subject to} \quad & \sigma_s \in \mathbb{R}, \quad \pi_s \in \mathbb{R}^A \\
& w_s + W_{s\cdot}^\top \xi + \sigma_s \leq \sum_{a \in \mathcal{A}} \pi_{sa} \left(k_{sa} + K_{sa}\xi\right)^\top \left(r_{sa} + \lambda \left[w + W\xi\right]\right) \qquad \forall \xi \in \Xi, && \text{(29b)} \\
& \pi_s \geq 0, \quad \mathrm{e}^\top \pi_s = 1, && \text{(29c)}
\end{aligned}
$$

where $(w, W) = (w^i, W^i)$. Set $\pi^{i+1}(a|s) := \pi^*_{sa}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, where π^*_s denotes the π_s-component of an optimal solution to (29) for state $s \in \mathcal{S}$. Set $i := i + 1$ and go back to Step 2.

Upon termination, the best policy found is stored in π^{i-1}, and τ^i is an estimate for the worst-case expected total reward of π^{i-1}. Depending on the number L of constraints that define Ξ, this estimate is exact (if $L = 1$) or a lower bound (if $L > 1$). We can equivalently reformulate (if $L = 1$) or conservatively approximate (if $L > 1$) the semi-infinite constraint (29b) with a semidefinite constraint. Since this reformulation parallels the proof of Theorem 3.8, we omit the details. Step 3 of the algorithm aims to increase the slack in the constraint (20c) of the policy evaluation problem solved in Step 2. One can show that if $\sigma_s > 0$ for some state $s \in \mathcal{S}$ that can be visited by the MDP, then Step 2 will lead to a better objective value in the next iteration. Algorithm 4.1 converges to a partial optimum of the policy improvement problem (24). We refer to [12] for a detailed convergence analysis.

5 Constructing Uncertainty Sets from Observation Histories

Assume that an observation history

$$
(s_1, a_1, \ldots, s_n, a_n) \in (\mathcal{S} \times \mathcal{A})^n \tag{30}
$$

of the MDP under some known stationary policy π^0 is available. We can use the observation (30) to construct an uncertainty set that contains the MDP's unknown true transition kernel P^0 with a probability of at least $1 - \beta$. The worst-case expected total reward of any policy π over this uncertainty set then provides a valid lower bound on the expected total reward of π under P^0 with a confidence of at least $1 - \beta$.

In the following, we first define the structural uncertainty set which incorporates all available a priori information about P^0. We then combine this structural information with the statistical information in the form of observation (30) to construct a confidence region for P^0. This confidence region will not be of the form (3). Section 5.3 therefore elaborates an approximate uncertainty set that is in line with the methods presented in Sections 3 and 4. We close with an asymptotic analysis of our approach.

5.1 Structural Uncertainty Set

Traditionally, uncertainty sets for the transition kernels of MDPs are constructed under the assumption that all transitions $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ are possible and that no a priori knowledge about the associated transition probabilities is available. In reality, however, one often has structural information about the MDP. For example, some transitions may be impossible, or certain functional relations between the transition probabilities may be known. We condense this kind of information into the *structural uncertainty set* $\mathcal{P}^0$, which captures all available a priori knowledge about the MDP. The use of structural information excludes irrelevant transition kernels and therefore leads to a smaller uncertainty set (and hence a tighter lower bound on the expected total reward). In Section 6, we will exemplify the benefits of this approach.

Formally, we assume that the structural uncertainty set $\mathcal{P}^0$ represents the affine image of a set Ξ^0, and that $\mathcal{P}^0$ and Ξ^0 satisfy our earlier definition (3) of $\mathcal{P}$ and Ξ. In the remainder of the paper, we denote by ξ^0 the parameter vector associated with the unknown true transition kernel P^0 of the MDP, that is, $P^0_{sa} = p^{\xi^0}(\cdot|s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$. We require that

(A1) Ξ^0 contains the parameter vector ξ^0 in its interior: $\xi^0 \in \operatorname{int} \Xi^0$.

Assumption (A1) implies that all vanishing transition probabilities are known a priori. This requirement is standard in the literature on statistical inference for Markov chains [5], and it is naturally satisfied if structural knowledge about the MDP is available. Otherwise, one may use the observation (30) to infer which transitions are possible. Indeed, it can be shown under mild assumptions that the probability to *not* observe a possible transition decreases exponentially with the length n of the observation [5]. For a sufficiently long observation, we can therefore assign zero probability to unobserved transitions.

We illustrate the construction of the structural uncertainty set $\mathcal{P}^0$ in an important special case.

Example 5.1 *For every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, let $\mathcal{S}_{sa} \subseteq \mathcal{S}$ denote the (nonempty) set of possible subsequent states if the MDP is in state s and action a is chosen. Assume that all sets $\mathcal{S}_{sa}$ are known, while no other structural information about the MDP's transition kernel is available. In the following, we define Ξ^0 and $p^\xi(\cdot|s, a)$ for this setting. For $(s, a) \in \mathcal{S} \times \mathcal{A}$, all but one of the probabilities corresponding to transitions (s, a, s'), $s' \in \mathcal{S}_{sa}$, can vary freely within the $(|\mathcal{S}_{sa}| - 1)$-dimensional probability simplex, while the remaining transition probability is uniquely determined through the others. We therefore set the dimension of Ξ^0 to $q := \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}(|\mathcal{S}_{sa}| - 1)$. For each $(s, a) \in \mathcal{S} \times \mathcal{A}$, we define the set $\overline{\mathcal{S}}_{sa}$ of explicitly modelled transition probabilities through $\overline{\mathcal{S}}_{sa} := \mathcal{S}_{sa} \setminus \{\overline{s}_{sa}\}$, where $\overline{s}_{sa} \in \mathcal{S}_{sa}$ can be chosen freely. Let μ be a bijection that maps each triple (s, a, s'), $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $s' \in \overline{\mathcal{S}}_{sa}$, to a* *component $\{1,\ldots,q\}$ of Ξ^0. We identify $\xi_{\mu(s,a,s')}$ with the probability of transition (s,a,s'). We define*

$$\Xi^0 := \left\{ \xi \in \mathbb{R}^q \,:\, \xi \geq 0, \sum_{s' \in \overline{\mathcal{S}}_{sa}} \xi_{\mu(s,a,s')} \leq 1 \;\; \forall \, (s,a) \in \mathcal{S} \times \mathcal{A} \right\} \tag{31}$$

and set $p^\xi(s'|s,a) := \xi_{\mu(s,a,s')}$ for $(s,a) \in \mathcal{S}\times\mathcal{A}$ and $s' \in \overline{\mathcal{S}}_{sa}$, as well as $p^\xi(\overline{s}_{sa}|s,a) := 1-\sum_{s'\in\overline{\mathcal{S}}_{sa}} \xi_{\mu(s,a,s')}$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$. The constraints in (31) ensure that all transition probabilities are non-negative.

5.2 Confidence Regions from Maximum Likelihood Estimation

In the following, we use the observation (30) to construct a confidence region for ξ^0. This confidence region will be centred around the maximum likelihood estimator associated with the observation (30), and its shape will be determined by the statistical properties of the likelihood difference between ξ^0 and its maximum likelihood estimator. To this end, we first calculate the log-likelihood function for the observation (30) and derive the corresponding maximum likelihood estimator. We then use existing statistical results for Markov chains (hereafter MCs) to construct a confidence region for ξ^0.

We remark that maximum likelihood estimation has recently been applied to construct confidence regions for the newsvendor problem [24]. Our approach differs in two main aspects. Firstly, due to the nature of the newsvendor problem, the observation history in [24] constitutes a collection of independent samples from a common distribution. Secondly, the newsvendor problem belongs to the class of single-stage stochastic programs, and the techniques developed in [24] do not readily extend to MDPs.

The probability to observe the state-action sequence (30) under the policy π^0 and some transition kernel associated with $\xi \in \Xi^0$ is given by

$$p_0(s_1)\, \pi^0(a_n|s_n) \prod_{t=1}^{n-1} \left[\pi^0(a_t|s_t)\, p^\xi(s_{t+1}|s_t,a_t) \right]. \tag{32}$$

The log-likelihood function $\ell_n : \Xi^0 \mapsto \mathbb{R} \cup \{-\infty\}$ is given by the logarithm of (32), where we use the convention that $\log(0) := -\infty$. Thus, we set

$$\ell_n(\xi) := \sum_{t=1}^{n-1} \log\left[p^\xi(s_{t+1}|s_t,a_t) \right] + \zeta, \qquad \text{where} \qquad \zeta := \log\left[p_0(s_1)\right] + \sum_{t=1}^{n} \log\left[\pi^0(a_t|s_t)\right]. \tag{33}$$

Note that the remainder term ζ is finite and does not depend on ξ. Due to the monotonicity of the logarithmic transformation, the expressions (32) and (33) attain their maxima over Ξ^0 at the same points. Note also that we index the log-likelihood function with the length n of the observation (30). This will be useful later when we investigate its asymptotic behaviour as n tends to infinity.

The order of the transitions (s_t, a_t, s_{t+1}) in the observation (30) is irrelevant for the log-likelihood function (33). Hence, we can reexpress the log-likelihood function as

$$\ell_n(\xi) = \sum_{(s,a,s') \in N} n_{sas'} \log \left[p^{\xi}(s'|s,a) \right] + \zeta, \tag{33'}$$

where $n_{sas'}$ denotes the number of transitions from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ under action $a \in \mathcal{A}$ in (30), and $N := \{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \, : \, n_{sas'} > 0\}$ represents the set of observed transitions.

We obtain a maximum likelihood estimator ξ^n by maximising the concave log-likelihood function ℓ_n over Ξ^0. Since the observation (30) has strictly positive probability under the transition kernel associated with ξ^0, we conclude that $\ell_n(\xi^n) \geq \ell_n(\xi^0) > -\infty$. Note that the maximum likelihood estimator may not be unique if ℓ_n fails to be strictly concave.

Remark 5.2 (Analytical Solution) *Sometimes the maximum likelihood estimator can be calculated analytically. Consider, for instance, the log-likelihood function associated with Example 5.1.*

$$\ell_n(\xi) = \sum_{\substack{(s,a,s') \in N : \\ s' \in \overline{\mathcal{S}}_{sa}}} n_{sas'} \log \left[\xi_{\mu(s,a,s')} \right] + \sum_{(s,a,\overline{s}_{sa}) \in N} n_{sa\overline{s}_{sa}} \log \Big[1 - \sum_{s' \in \overline{\mathcal{S}}_{sa}} \xi_{\mu(s,a,s')} \Big] + \zeta$$

The gradient of ℓ_n vanishes at ξ^n defined through $\xi^n_{\mu(s,a,s')} := n_{sas'} / \sum_{s'' \in \mathcal{S}} n_{sas''}$ if $\sum_{s'' \in \mathcal{S}} n_{sas''} > 0$ and $\xi^n_{\mu(s,a,s')} := 0$ otherwise. Since $\xi^n \in \Xi^0$, see (31), it constitutes a maximum likelihood estimator.

For $\xi \in \Xi^0$, the log-likelihood $\ell_n(\xi)$ describes the (logarithm of the) probability to observe the state-action sequence (30) under the transition kernel associated with ξ. For a sufficiently long observation, we therefore expect the log-likelihood $\ell_n(\xi^0)$ of the unknown true parameter vector ξ^0 to be 'not much smaller' than the log-likelihood $\ell_n(\xi^n)$ of the maximum likelihood estimator ξ^n. Guided by this intuition, we intersect the set Ξ^0 with a constraint that bounds this log-likelihood difference.

$$\Xi^0 \cap \{\xi \in \mathbb{R}^q \, : \, \ell_n(\xi) \geq \ell_n(\xi^n) - \delta\} \tag{34}$$

Here, $\delta \in \mathbb{R}_+$ determines the upper bound on the anticipated log-likelihood difference between ξ^0 and ξ^n. Expression (34) raises two issues. Firstly, it is not clear how δ should be chosen. Secondly, the intersection does not constitute a valid uncertainty set since it is not of the form (3b). In the following, we address the choice of δ. We postpone the discussion of the second issue to the next section.

Our choice of δ relies on statistical inference and requires two further assumptions:

(A2) The MC with state set $\mathcal{S}$ and transition kernel $\widehat{P}(\pi^0; \xi)$ is irreducible for some $\xi \in \Xi^0$, see (7a).

(A3) The matrix with rows $[K_{sa}]^\top_{s'\cdot}$ for $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ with $\pi^0(a|s) > 0$ has rank $\kappa > 0$.

Remember that a finite MC with state set $\mathcal{S}$ is called *irreducible* if for any pair of states $s, s' \in \mathcal{S}$, there is a strictly positive probability that the MC visits state s' in the future if it is currently in state s. Assumption (A2) therefore guarantees that the MDP visits every state infinitely often as the observation length n tends to infinity. Assumption (A3) ensures that the historical policy π^0 chooses at least one state-action pair with unknown transition probabilities $p^{\xi^0}(\cdot|s,a)$. If this was not the case, then the observation (30) would not allow any inference about ξ^0, and the tightest possible uncertainty set for the unknown true transition kernel P^0 would be the structural uncertainty set $\mathcal{P}^0$.

We can now establish an asymptotic relation between ξ^n and ξ^0.

Theorem 5.3 *Under the assumptions (A1)–(A3), we have*

$$2\left[\ell_n(\xi^n) - \ell_n(\xi^0)\right] \quad \underset{n\to\infty}{\longrightarrow} \quad \chi^2_\kappa, \tag{35}$$

where '$\longrightarrow$' denotes convergence in distribution and χ^2_κ is a χ^2-distribution with κ degrees of freedom.

Remark 5.4 *A sequence of random variables X_i with cumulative distribution functions F_i, $i = 1, 2, \ldots$, is said to* converge in distribution *to a random variable X with cumulative distribution function F if $\lim_{n\to\infty} F_n(x) = F(x)$ at all points $x \in \mathbb{R}$ where F is continuous.*

Proof of Theorem 5.3 See Appendix B. ■

Theorem 5.3 can be interpreted as follows. The observation (30) constitutes a random vector whose true distribution is determined by the expression (32) if we set $\xi = \xi^0$. Since ξ^0 is unknown, the distribution of the observation (30) is unknown as well. Similarly, the maximum likelihood estimator ξ^n depends on the observation (30) and is therefore a random vector with an unknown distribution. Theorem 5.3 shows, however, that the distribution of the random variable $2\left[\ell_n(\xi^n) - \ell_n(\xi^0)\right]$ is asymptotically known: it converges to a χ^2_κ distribution. Thus, under the assumptions (A1)–(A3), we obtain a $(1-\beta)$-confidence region for ξ^0 if we set δ in (34) to one half of the $(1-\beta)$-quantile of the χ^2_κ distribution.

$$\mathbb{P}\left(\xi^0 \in \Xi^0 \cap \{\xi \in \mathbb{R}^q \,:\, \ell_n(\xi) \geq \ell_n(\xi^n) - \delta\}\right) \;\geq\; 1 - \beta$$

The support of the χ^2_κ distribution is unbounded above, and thus δ grows indefinitely if β goes to zero. For a fixed observation length n, the set (34) therefore reduces to Ξ^0 for $\beta \longrightarrow 0$.

Theorem 5.3 provides an asymptotic convergence result for robust *infinite* horizon MDPs. Robust *finite* horizon MDPs, on the other hand, are not directly amenable to an asymptotic analysis since they reach a terminal state after finitely many transitions. The most natural way to estimate the transition kernel of a finite horizon MDP is to assume that the MDP is 'restarted', that is, the same MDP is run several times. Theorem 5.3 can be applied to this situation as follows. We construct an infinite horizon MDP whose state space consists of the states of the finite horizon MDP, together with an auxiliary 'restarting' state τ. Apart from the transitions of the finite horizon MDP, the infinite horizon MDP contains deterministic transitions from all terminal states $s \in \mathcal{S}_T$ to τ, as well as transitions from τ to all initial states $s \in \mathcal{S}_1$ with action-independent transition probabilities $p_0(s)$. We do not specify a discount factor λ or one-step rewards r since they are irrelevant for Theorem 5.3. We interpret m observation histories $(s_1^i, a_1^i, \ldots, s_{T-1}^i, a_{T-1}^i, s_T^i)$, $i = 1, \ldots, m$, of the finite horizon MDP as one observation

$$(s_1^1, a_1^1, \ldots, s_{T-1}^1, a_{T-1}^1, s_T^1, a_T^1; \ \ldots \ ; s_1^m, a_1^m, \ldots, s_{T-1}^m, a_{T-1}^m, s_T^m, a_T^m)$$

of the corresponding infinite horizon MDP. In this concatenated observation, the terminal actions $a_T^i \in \mathcal{A}$ may be chosen freely. We can now apply Theorem 5.3 to the constructed infinite horizon MDP if it satisfies the assumptions (A1)–(A3). This is the case if the finite horizon MDP satisfies the assumptions (A1) and (A3) and if each of its states can be reached from an initial state $s \in \mathcal{S}_1$ with $p_0(s) > 0$.

We close with a variant of Theorem 5.3 that relaxes the assumption (A2).

Remark 5.5 *Even if assumption (A2) is violated, the MDP will eventually enter a set of irreducible states $\overline{\mathcal{S}} \subseteq \mathcal{S}$ from which it cannot escape. If we remove from the observation (30) all state-action pairs $(s_1, a_1, \ldots, s_\tau, a_\tau)$ for which $s_t \notin \overline{\mathcal{S}}$, $t = 1, \ldots, \tau$, then Theorem 5.3 can be applied to the reduced MDP that only consists of the states in $\overline{\mathcal{S}}$.*

5.3 Quadratic Approximation

The confidence region for the unknown parameter vector ξ^0 in (34) is not consistent with the definition (3b) that underlies our computational techniques developed in Sections 3 and 4. We therefore approximate the left-hand side of the constraint $\ell_n(\xi) \geq \ell_n(\xi^n) - \delta$ in (34) by a second-order Taylor expansion around the maximum likelihood estimator ξ^n and set

$$\Xi^n := \Xi^0 \cap \{\xi \in \mathbb{R}^q \, : \, \varphi_n(\xi) \geq 0\}, \tag{36}$$

where

$$\varphi_n(\xi) := [\nabla_\xi \, \ell_n(\xi^n)]^\top (\xi - \xi^n) - \frac{1}{2} (\xi - \xi^n)^\top \left[\nabla_\xi^2 \, \ell_n(\xi^n)\right] (\xi - \xi^n) + \delta \tag{37a}$$

with

$$[\nabla_{\xi}\, \ell_n(\xi^n)]^{\top} = \sum_{(s,a,s')\in N} \frac{n_{sas'}}{p^{\xi^n}(s'|s,a)}\, [K_{sa}]^{\top}_{s'\cdot} \tag{37b}$$

$$\text{and} \qquad \nabla^2_{\xi}\, \ell_n(\xi^n) = \sum_{(s,a,s')\in N} \frac{n_{sas'}}{[p^{\xi^n}(s'|s,a)]^2} \left([K_{sa}]^{\top}_{s'\cdot}\right)^{\top} \left([K_{sa}]^{\top}_{s'\cdot}\right). \tag{37c}$$

Note that the expressions in (37b) and (37c) are well-defined since $p^{\xi^n}(s'|s,a) > 0$ for all $(s,a,s') \in N$, see our discussion surrounding the log-likelihood function (33'). Moreover, Ξ^n is of the form (3b) since it emerges from the intersection of Ξ^0 with an ellipsoid. One can show that Ξ^n contains a Slater point whenever δ is strictly positive.

The set Ξ^n in (36) induces an uncertainty set of the form

$$\mathcal{P}^n := \left\{ P \in [\mathcal{M}(\mathcal{S})]^{S\times A} \ : \ \exists\, \xi \in \Xi^n \ \text{ such that } \ P_{sa} = p^{\xi}(\cdot|s,a) \ \ \forall\, (s,a) \in \mathcal{S}\times\mathcal{A} \right\}.$$

We now investigate the asymptotic properties of this uncertainty set as n tends to infinity. In Theorem 5.6 below we establish that $\mathcal{P}^n$ converges to the unknown true transition kernel P^0 of the MDP and analyse the speed of convergence. Afterwards, we show that the solutions of the robust policy evaluation and improvement problems converge to the solutions of the nominal policy evaluation and improvement problems under the unknown true transition kernel P^0. All subsequent convergence results rely on the following stronger version of assumption (A3).

(A3') The matrix with rows $[K_{sa}]^{\top}_{s'\cdot}$ for $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$ with $\pi^0(a|s) > 0$ has full rank.

Assumption (A3') stipulates that the mapping from ξ to the probabilities of all possible transitions under π^0 is injective. Indeed, if assumption (A3') is violated, then there are different parameter vectors $\xi, \xi' \in \Xi^0$ such that $p^{\xi}(s'|s,a) = p^{\xi'}(s'|s,a)$ for all possible transitions (s,a,s') under the data generating policy π^0. In this case, we cannot distinguish between ξ and ξ' based on the information provided by any observation of the type (30), and the uncertainty set $\mathcal{P}^n$ will not converge to a singleton as the observation length n tends to infinity.

In the following proposition, we analyse the Hausdorff distance between the two sets Ξ^n and $\left\{\xi^0\right\}$. Recall that the Hausdorff distance between two sets $X, Y \subseteq \mathbb{R}^q$ is defined as

$$d^{\mathrm{H}}(X,Y) := \max\left\{ \sup_{x\in X} \inf_{y\in Y} \|x-y\|_{\infty}\,,\ \sup_{y\in Y} \inf_{x\in X} \|x-y\|_{\infty} \right\}.$$

Theorem 5.6 *Under the assumptions (A1), (A2) and (A3'), we have*

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha d^{\mathrm{H}} \left[\Xi^n, \{\xi^0\} \right] \right) = 0 \qquad \forall \, \alpha < 1/2, \tag{38}$$

where 'plim' denotes convergence in probability.

Remark 5.7 *Theorem 5.6 is equivalent to the statement that*

$$\lim_{n \longrightarrow \infty} \mathbb{P} \left(\max_{\xi \in \Xi^n} \left\| \xi - \xi^0 \right\|_\infty \leq \frac{\epsilon}{n^\alpha} \right) = 1$$

for every $\alpha < 1/2$ and $\epsilon > 0$.

Proof of Theorem 5.6 See Appendix C. ∎

We now show that under the assumptions of Theorem 5.6, the solution provided by the constant reward to-go approximation from Proposition 3.6 converges to the expected total reward $p_0^\top v(\xi^0)$ of policy π as n tends to infinity. Note that $\mathcal{P}^n$ constitutes a non-rectangular uncertainty set.

Proposition 5.8 *Let $\vartheta^n(\xi) = w^n$ be the constant reward to-go approximation described in Proposition 3.6 if we set $\Xi = \Xi^n$. Under the assumptions (A1), (A2) and (A3'), we have*

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \left| p_0^\top w^n - p_0^\top v(\pi; \xi^0) \right| \right) = 0 \qquad \forall \, \alpha < 1/2, \tag{39}$$

where $p_0^\top v(\pi; \xi^0)$ denotes the expected total reward under π and the unknown true transition kernel P^0.

Remark 5.9 *Proposition 5.8 is equivalent to the statement that for every $\alpha < 1/2$ and $\epsilon > 0$, we have*

$$\lim_{n \longrightarrow \infty} \mathbb{P} \left(\left| p_0^\top w^n - p_0^\top v(\pi; \xi^0) \right| \leq \frac{\epsilon}{n^\alpha} \right) = 1.$$

While Ξ^n is constructed from the observation (30) under the historical policy π^0, $p_0^\top w^n$ estimates the expected total reward of policy π. Note that π^0 and π can be different.

Proof of Proposition 5.8 Fix any $\alpha < 1/2$. By Theorem 5.6, we have

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \max_{\xi \in \Xi^n} \left\| \xi - \xi^0 \right\|_\infty \right) = 0. \tag{40}$$

The proof of Theorem 3.2 shows that for each w^n, $n \in \mathbb{N}$, there is $\xi^{n,1}, \ldots, \xi^{n,S} \in \Xi^n$ such that

$$w^n = \widehat{r}(\pi; \xi^{n,1}, \ldots, \xi^{n,S}) + \lambda \widehat{P}(\pi; \xi^{n,1}, \ldots, \xi^{n,S}) \, w^n, \tag{41}$$

where for $\xi^1, \ldots, \xi^S \in \Xi^n$, the rectangular rewards $\widehat{r}(\pi; \xi^1, \ldots, \xi^S)$ and the rectangular transition kernel $\widehat{P}(\pi; \xi^1, \ldots, \xi^S)$ are defined through $\left[\widehat{r}(\pi; \xi^1, \ldots, \xi^S)\right]_s := \widehat{r}_s(\pi; \xi^s)$ and $\left[\widehat{P}(\pi; \xi^1, \ldots, \xi^S)\right]_{s\cdot}^\top := \widehat{P}_{s\cdot}^\top(\pi; \xi^s)$ for all $s \in \mathcal{S}$, respectively. Note that the existence of $\xi^{n,1}, \ldots, \xi^{n,S}$ does not depend on the structure of Ξ^n, see (14). By unrolling the recursion (41), we see that

$$w^n = v(\pi; \xi^{n,1}, \ldots, \xi^{n,S}) := \sum_{t=0}^{\infty} \left[\lambda \widehat{P}(\pi; \xi^{n,1}, \ldots, \xi^{n,S})\right]^t \widehat{r}(\pi; \xi^{n,1}, \ldots, \xi^{n,S}),$$

where for $\xi^1, \ldots, \xi^S \in \Xi^n$, $v(\pi; \xi^1, \ldots, \xi^S)$ represents a rectangular variant of the reward to-go function v. One can adapt the proof of Proposition 3.1 (a) to show that this rectangular reward to-go function is Lipschitz continuous on the compact set Ξ^0. Equation (40) therefore implies that

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \left\| v(\pi; \xi^{n,1}, \ldots, \xi^{n,S}) - v(\pi; \xi^0, \ldots, \xi^0) \right\|_\infty\right) = 0.$$

Equation (39) now follows from $w^n = v(\pi; \xi^{n,1}, \ldots, \xi^{n,S})$ and $v(\pi; \xi^0) = v(\pi; \xi^0, \ldots, \xi^0)$. ■

Proposition 5.8 immediately extends to the affine reward to-go approximations obtained from the semidefinite program (20).

Corollary 5.10 *Let τ^n denote the optimal value of τ in the semidefinite program (20) with $\Xi = \Xi^n$. Under the assumptions (A1), (A2) and (A3'), we have*

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \left|\tau^n - p_0^\top v(\pi; \xi^0)\right|\right) = 0 \qquad \forall\, \alpha < 1/2.$$

Proof Fix $\alpha < 1/2$. Theorem 5.6 and the Lipschitz continuity of v, see Proposition 3.1 (a), imply that

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \max_{\xi \in \Xi^n} \left|p_0^\top v(\pi; \xi) - p_0^\top v(\pi; \xi^0)\right|\right) = 0.$$

Proposition 3.1 (c) and Theorem 3.8 ensure that $\tau^n \le p_0^\top v(\pi; \xi)$ for all $\xi \in \Xi^n$, $n \in \mathbb{N}$. We conclude that

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \left[\tau^n - p_0^\top v(\pi; \xi^0)\right]^+\right) = 0,$$

where $[x]^+ := \max\{x, 0\}$ for $x \in \mathbb{R}$. In a probabilistic sense, τ^n therefore underestimates $p_0^\top v(\pi; \xi^0)$. At the same time, Proposition 3.10 guarantees that $\tau^n \ge p_0^\top w^n$ for the vector w^n defined in Proposition 5.8. Hence, the assertion follows from the convergence of $p_0^\top w^n$, see Proposition 5.8. ■

The above convergence results extend to the policy improvement problem discussed in Section 4. Since the derivation of the following result does not require any new ideas, we state it without a proof.

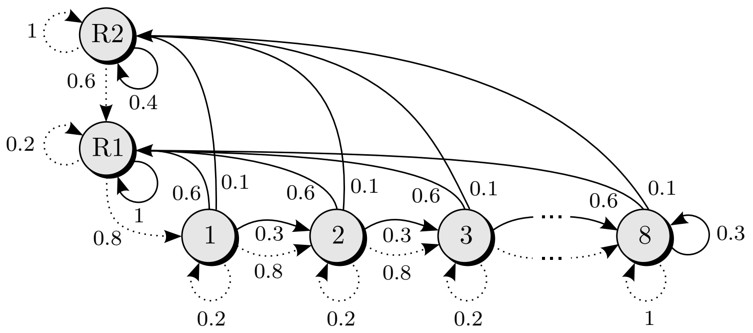

Figure 7: *MDP for the machine replacement problem. Shown are the transition probabilities for the two actions 'do nothing' (dashed arcs) and 'repair' (solid arcs). The states 8, R1 and R2 pay an expected reward of -20, -2 and -10, respectively, while no reward is received in the other states. We use the same drawing conventions as in Figure 1.*

Proposition 5.11 *For $\Xi = \Xi^n$, let π^n denote an optimal policy determined by Algorithm 4.1 or the robust value iteration described in Corollary 4.3. Under the assumptions (A1), (A2) and (A3'), we have*

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(n^\alpha \left| p_0^\top v(\pi^n; \xi^0) - \min_{\pi \in \Pi} \left\{ p_0^\top v(\pi; \xi^0) \right\} \right| \right) = 0 \qquad \forall\, \alpha < 1/2,$$

where the second term in the absolute value represents the expected total reward of the optimal policy under the MDP's unknown true transition kernel P^0.

Note that both the constant and the affine reward to-go approximations guarantee convergence to the nominal solutions of the policy evaluation and improvement problems as n tends to infinity. However, the next section will show that we can expect the affine approximations to convergence faster if the uncertainty set is non-rectangular.

6 Numerical Example

We apply the policy evaluation and improvement methods from Sections 3 and 4 to the machine replacement problem presented in [7]. The problem concerns a single machine whose condition is described by eight 'operative' states $1, \ldots, 8$ and two 'repair' states R1 and R2. At each time period, the decision maker receives an expected reward that depends on the machine's current state. The state in the subsequent time period is random and depends on both the current state and the chosen action ('do nothing' or 'repair'). The goal is to find a policy that maximises the expected total reward under the discount factor $\lambda = 0.8$. If all transition probabilities are known, we can model this problem as an MDP, see Figure 7. It is easy to transform this MDP into an equivalent one that satisfies the definitions in Section 1.

Consider the policy that chooses the actions 'do nothing' and 'repair' with probability 0.8 and 0.2, respectively, in each operative state $1, \ldots, 7$. In states 8 and R2, the policy always chooses the action

n	RVI	SDP (LB)	SDP (UB)	$P^0 \in \mathcal{P}^n$?
500	-43.90	-30.37	-26.97	87%
1000	-32.34	-20.74	-18.81	92%
2500	-20.35	-15.36	-15.32	91%
500	-16.82	-14.95	-14.95	87%
1000	-15.20	-14.00	-13.99	88%
2500	-14.07	-13.31	-13.30	92%

Table 2: *Policy evaluation results for 100 randomly generated observation histories of different observation length n. From left to right, the columns report the observation length, the average lower bound provided by the robust value iteration (RVI), the average lower and upper bounds obtained from the semidefinite program (20), and the percentage of instances in which P^0 is contained in $\mathcal{P}^n$. The first three rows were obtained without a priori knowledge, whereas the last three rows exploit the structural knowledge described in the text.*

'repair', while the action 'do nothing' is chosen in state R1. The expected total reward of this policy is -12.34. Assume now that instead of the transition probabilities, we only have access to an observation history. We can use the structural uncertainty set $\mathcal{P}^0$ described in Example 5.1 and intersect it with a 90% confidence region for the unknown transition probabilities, see Section 5.3. The resulting uncertainty set is non-rectangular, and we can apply the robust value iteration from Proposition 3.6 or solve the semidefinite program (20) to obtain a lower bound on the worst-case expected total reward (2). The results for randomly generated observation histories are presented in the first part of Table 2. Note that the uncertainty set $\mathcal{P}^n$ contains the MDP's true transition kernel P^0 in about 90% of the observation histories. As the observation length n increases, the lower bounds obtained from both the robust value iteration and the semidefinite program (20) converge to the true expected total reward. However, the lower bounds provided by the semidefinite program are significantly tighter. From the optimality gaps we conclude that the semidefinite programming approximation performs well in this example.

The transition kernel in Figure 7 is highly structured. In particular, the probabilities associated with the transitions emanating from state s under either action are identical for $s \in \{1, \ldots, 7\}$. We now assume that although these probabilities are unknown, they are known to be identical for $s \in \{1, \ldots, 7\}$. This additional information can be incorporated into the structural uncertainty set $\mathcal{P}^0$ to reduce the dimension of Ξ^0. The results are presented in the second part of Table 2. As the table shows, the incorporation of the additional structural information leads to significantly tighter bounds.

We now use the random observation histories to solve the robust policy improvement problem. The optimal policy for the unknown true transition kernel P^0 achieves an expected total reward of -7.98. Table 3 reports on the performance of the policies determined by the robust value iteration and the sequential convex optimisation algorithm from Section 4. Both methods perform well in this example. Nevertheless, the sequential convex optimisation algorithm provides tighter worst-case estimates. This is not surprising since the algorithm employs affine approximations of the reward to-go function.

	RVI		SCO	
n	LB	nominal	LB	nominal
500	-12.35	-8.05	-10.45	-8.05
1000	-10.64	-8.00	-9.51	-8.00
2500	-9.50	-7.99	-8.99	-7.99

Table 3: *Policy improvement results for 100 randomly generated observation histories of different observation length n. From left to right, the columns report the observation length, the average lower bound and nominal performance of the robust value iteration (RVI), and the average lower bound and nominal performance of the sequential convex optimisation procedure (SCO). In both cases, the nominal performance describes the expected total reward of the worst-case optimal policy under the unknown true transition kernel P^0.*

We finally remark that we have considered variants of the MDP in Figure 7 with up to 1000 states. On average, the solution of the associated semidefinite program (20) required between 0.38 secs (10 states) and 228.92 secs (1000 states). Numerical results for the robust value iteration are reported in [11, 17].

7 Conclusion

We studied robust Markov decision processes (MDPs) in which the transition kernel is unknown. Traditionally, the policy evaluation and improvement problems for robust MDPs are solved in two steps. In the first step, one constructs a confidence region for the unknown parameters. Afterwards, one solves a robust optimisation problem over this confidence region.

We proposed a variant of this approach that differs in two important aspects. Firstly, existing methods rely on transition sampling to construct the confidence region for the MDP's transition kernel. In contrast, we use observation histories which are much easier to obtain in practice. Secondly, previous approaches solve an unduly conservative approximation of the aforementioned robust optimisation problem. As we pointed out in Section 2, this approximation can destroy vital characteristics of robust MDPs. We developed two novel approximations that retain these characteristics. Moreover, our approximations provide tighter bounds than the existing techniques. We applied our method to the machine replacement problem, and we demonstrated that our approach scales to nontrivial problem sizes.

Acknowledgements

Financial support from EPSRC Grant EP/H0204554/1 is gratefully acknowledged.

References

[1] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Mathematical Programming*, 95(1):3–51, 2003.

[2] J. D. Bagnell, A. Y. Ng, and J. Schneider. Solving uncertain Markov decision problems. Technical Report CMU-RI-TR-01-25, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2001.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization.* Princeton University Press, 2009.

[4] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, 2007.

[5] P. Billingsley. *Statistical Inference for Markov Processes.* The University of Chicago Press, 1961.

[6] P. Billingsley. *Probability and Measure.* Wiley Blackwell, 1995.

[7] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, Accepted for Publication.

[8] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* W. H. Freeman, 1979.

[9] R. Givan, S. Leach, and T. Dean. Bounded-parameter Markov decision processes. *Artificial Intelligence*, 122(1–2):71–109, 2000.

[10] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. Accepted for Publication in Operations Research, 2009.

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

[12] J. Korski, F. Pfeuffer, and K. Klamroth. Biconvex sets and optimization with biconvex functions: A survey and extensions. *Mathematical Methods of Operations Research*, 66(3):373–407, 2007.

[13] D. Kuhn, W. Wiesemann, and A. Georghiou. Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming*, Forthcoming, 2009.

[14] M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284(1–3):193–228, 1998.

[15] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

[16] G. E. Monahan. A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

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

[18] I. C. Paschalidis and S.-C. Kang. A robust approach to Markov decision problems with uncertain transition probabilities. In *Proceedings of the 17th IFAC World Congress*, pages 408–413, 2008.

[19] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, 1994.

[20] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

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

[22] J. N. Tsitsiklis. Computational complexity in Markov decision theory. *HERMIS – An International Journal of Computer Mathematics and its Applications*, 9(1):45–54, 2007.

[23] L. Vandenberghe and S. Boyd. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

[24] Z. Wang, P. W. Glynn, and Y. Ye. Likelihood robust optimization for data-driven newsvendor problems. Working paper, Department of Management Science and Engineering, Stanford University, USA, 2009.

[25] C. C. White and H. K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, 42(4):739–749, 1994.

[26] H. Xu and S. Mannor. The robustness-performance tradeoff in Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 1537–1544, 2006.

A Saddle Point Condition for s-Rectangular Uncertainty Sets

Proposition A.1 *For an infinite horizon MDP with an s-rectangular uncertainty set $\mathcal{P}$, we have*

$$\sup_{\pi \in \Pi} \inf_{P \in \mathcal{P}} \mathbb{E}^{P,\pi}\left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 \sim p_0\right] = \inf_{P \in \mathcal{P}} \sup_{\pi \in \Pi} \mathbb{E}^{P,\pi}\left[\sum_{t=0}^{\infty} \lambda^t r(s_t, a_t, s_{t+1}) \;\middle|\; s_0 \sim p_0\right]. \quad (42)$$

Proof It follows from the proof of Theorem 4.1 that the left-hand side of (42) is equivalent to

$$\max_{w \in \mathbb{R}^S} \left\{ p_0^\top w \,:\, w_s \leq \max_{\pi \in \Pi} \min_{\xi^s \in \Xi} \left\{ \widehat{r}_s(\pi; \xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\pi; \xi^s)\, w \right\} \quad \forall\, s \in \mathcal{S} \right\}.$$

The constraints in this problem are equivalent to $w \leq \varphi(w)$, see (25). Since φ is a contraction mapping, see Theorem 4.1, non-negativity of p_0 and Theorem 6.2.2 in [19] allow us to reexpress the problem as

$$\min_{w \in \mathbb{R}^S} \left\{ p_0^\top w \,:\, w_s \geq \max_{\pi \in \Pi} \min_{\xi^s \in \Xi} \left\{ \widehat{r}_s(\pi; \xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\pi; \xi^s)\, w \right\} \quad \forall\, s \in \mathcal{S} \right\}.$$

The max-min expressions in the constraints satisfy the conditions of Corollary 37.3.2 in [20]. Hence, we can interchange the order of the operators in the constraints to obtain the following reformulation.

$$\min_{w \in \mathbb{R}^S} \left\{ p_0^\top w \,:\, w_s \geq \min_{\xi^s \in \Xi} \max_{\pi \in \Pi} \left\{ \widehat{r}_s(\pi; \xi^s) + \lambda \widehat{P}_{s\cdot}^\top(\pi; \xi^s)\, w \right\} \quad \forall\, s \in \mathcal{S} \right\}.$$

The uncertainty set $\mathcal{P}$ is s-rectangular, and the sth constraint only depends on the components $\pi(\cdot|s)$ of π. Hence, similar transformations as in Theorems 3.2 and 4.1 yield the following reformulation.

$$\min_{w \in \mathbb{R}^S} \min_{\xi \in \Xi} \left\{ p_0^\top w \,:\, w_s \geq \widehat{r}_s(\pi; \xi) + \lambda \widehat{P}_{s\cdot}^\top(\pi; \xi)\, w \;\; \forall\, s \in \mathcal{S},\; \pi \in \Pi \right\}. \quad (43)$$

Since p_0 is non-negative, Theorems 6.1.1 and 6.2.2 in [19] imply that for a given $\xi \in \Xi$, the optimal solution w satisfies $w = \max_{\pi \in \Pi} \{v(\pi; \xi)\}$. The equivalence of (43) and the right-hand side of (42) now follows from the property (6) of the reward to-go function v. ∎

B Proof of Theorem 5.3

The proof of Theorem 5.3 relies on the Theorems 2.1, 2.2 and 5.1 in [5], which establish asymptotic properties of maximum likelihood estimators of ordinary MCs. To keep the paper self-contained, we summarise these results in Theorem B.1.

Theorem B.1 *Consider a finite MC with state set $\mathcal{X} = \{1, \ldots, X\}$ and transition probabilities $p_{xy}(\theta)$, $x, y \in \mathcal{X}$, that depend on an unknown parameter vector θ ranging over an open set $\Theta \subseteq \mathbb{R}^U$. Assume that the following conditions are satisfied:*

(C1) Each function p_{xy} has continuous partial derivatives of third order throughout Θ.

(C2) The set-valued mapping $D(\theta) := \{(x, y) \in \mathcal{X} \times \mathcal{X} : p_{xy}(\theta) > 0\}$ is constant, that is, there is a set $D \subseteq \mathcal{X} \times \mathcal{X}$ such that $D(\theta) = D$ for all $\theta \in \Theta$.

(C3) The Jacobian matrix of the transition kernel $(p_{xy}(\theta))_{x,y}$ has rank U throughout Θ.

(C4) For each $\theta \in \Theta$, the MC is irreducible.

Let $(x_1, \ldots, x_m)$ denote an observation of the MC under its true transition kernel $p_{xy}(\theta^0)$, where $\theta^0 \in \Theta$, and let m_{xy} denote the number of observations of transition $(x, y) \in \mathcal{X} \times \mathcal{X}$. For the sequence of functions $f_m(\theta) := \sum_{(x,y)\in D} m_{xy} \log [p_{xy}(\theta)]$, Θ contains a sequence of random vectors $\overline{\theta}^m$ that satisfy

$$2\left[f_m(\overline{\theta}^m) - f_m(\theta^0)\right] \underset{m\to\infty}{\longrightarrow} \chi^2_U, \tag{44a}$$

$$m^{1/2}\left(\overline{\theta}^m - \theta^0\right) \underset{m\to\infty}{\longrightarrow} \mathcal{N}(0, \Gamma). \tag{44b}$$

Here, $\mathcal{N}(0, \Gamma)$ is a multivariate normal distribution with zero mean and finite covariance matrix $\Gamma \succ 0$. Moreover, $\overline{\theta}^m$ is a strict local maximiser of f_m with probability going to one as m tends to infinity.

In order to apply Theorem B.1 to MDPs, we interpret the state-action sequence (30) as an observation history of an ordinary MC. Theorem 5.3 then follows from (44a). To simplify the exposition, we prove Theorem 5.3 first under assumption (A3') on page 33. At the end of this section, we extend our proof to hold under the weaker assumption (A3).

We interpret the state-action sequence (30) as an observation of n states of an MC with state set

$$\mathcal{X} := \left\{(s, a) \in \mathcal{S} \times \mathcal{A} : \pi^0(a|s) > 0\right\}. \tag{45a}$$

The MC is in state $(s, a) \in \mathcal{X}$ whenever the underlying MDP is in state s and the decision maker chooses action a. Note that we omit state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$ with $\pi^0(a|s) = 0$ in (45a). As we will see, this is a necessary (but not sufficient) condition for the MC to be irreducible, see condition (C4) of Theorem B.1. By construction, the MC starts in state $(s, a) \in \mathcal{X}$ with probability $p_0(s)\, \pi^0(a|s)$, and it moves from state $(s, a) \in \mathcal{X}$ to state $(s', a') \in \mathcal{X}$ with probability $p^{\xi^0}(s'|s, a)\, \pi^0(a'|s')$, where ξ^0 is the unknown true parameter of the underlying MDP. Since the historical policy π^0 is stationary, the MC indeed satisfies the Markov property.

We can establish the following relationship between the MC and the MDP.

$$\Theta := \operatorname{int} \Xi^0 \tag{45b}$$

$$\text{and} \qquad p_{xy}(\theta) := p^\theta(s'|s, a)\, \pi^0(a'|s') \qquad \text{for } \theta \in \Theta \text{ and } x = (s, a),\ y = (s', a') \in \mathcal{X}. \tag{45c}$$

By assumption (A1), we have $\xi^0 \in \operatorname{int} \Xi^0$. Hence, Θ indeed contains the unknown true parameter vector $\theta^0 := \xi^0$ of the MC as required by Theorem B.1.

We now show that the MC defined through (45) satisfies the conditions (C1)–(C4) of Theorem B.1.

Lemma B.2 *If the MDP satisfies assumptions (A2) and (A3'), then the MC defined through (45) satisfies the conditions (C1)–(C4) of Theorem B.1.*

Proof Condition (C1) is satisfied since p_{xy} is affine in θ for all $x, y \in \mathcal{X}$, see definitions (45c) and (3).

As for condition (C2), the definitions (45a) and (45c) imply that

$$D(\theta) = \left\{(x, y) \in \mathcal{X} \times \mathcal{X} \ :\ p^\theta(s'|s, a) > 0 \ \text{ for } x = (s, a) \text{ and } y = (s', a')\right\}.$$

We recall that $p^\theta(s'|s, a) = k_{sa} + K_{sa}\theta$. We claim that for any $\theta \in \Theta$, the set $D(\theta)$ equals

$$D := \Big\{(x, y) \in \mathcal{X} \times \mathcal{X} \ :\ [k_{sa}\ K_{sa}]_{s'\cdot}^\top \neq 0 \ \text{ for } x = (s, a) \text{ and } y = (s', a')\Big\}.$$

By construction, $D(\theta) \subseteq D$ for all $\theta \in \Theta$. It remains to show that $D \subseteq D(\theta)$ for all $\theta \in \Theta$. Assume to the contrary that $[k_{sa}\ K_{sa}]_{s'\cdot}^\top \neq 0$ but $p^\theta(s'|s, a) = 0$ for $x = (s, a)$, $y = (s', a') \in \mathcal{X}$ and $\theta \in \Theta$. Since Θ is an open set, there is a neighbourhood of θ that is contained in Θ, and all points θ' in this neighbourhood have to satisfy $p^{\theta'}(s'|s, a) \geq 0$. Since $p^\theta(s'|s, a) = 0$, this implies that $[K_{sa}]_{s'\cdot}^\top = 0$, and hence $[k_{sa}]_{s'} = 0$ as well. This contradicts our assumption that $[k_{sa}\ K_{sa}]_{s'\cdot}^\top \neq 0$. We therefore conclude that $p^\theta(s'|s, a) > 0$ for all $\theta \in \Theta$, that is, $D \subseteq D(\theta)$ for all $\theta \in \Theta$.

We now consider condition (C3). The Jacobian $J(\theta) \in \mathbb{R}^{X^2 \times U}$ of the MC's transition kernel is defined through $J_{xy,u} := \partial p_{xy}(\theta)/\partial \theta_u$ for $x, y \in \mathcal{X}$ and $u = 1, \ldots, U$. For $x = (s, a)$, $y = (s', a') \in \mathcal{X}$, we have $\partial p_{xy}(\theta)/\partial \theta_u = \pi^0(a'|s')\, [K_{sa}]_{s'u}$. Thus, assumption (A3') ensures that $J(\theta)$ has rank U.

In view of condition (C4), we note that the irreducibility of a finite MC only depends on the structure of the set of transitions with strictly positive probability; the actual probabilities are irrelevant. However, the proof of condition (C2) implies that for all state pairs $(x, y) \in \mathcal{X} \times \mathcal{X}$, either $p_{xy}(\theta) > 0$ for all $\theta \in \Theta$ or $p_{xy}(\theta) = 0$ for all $\theta \in \Theta$. Hence, the set of transitions with strictly positive probability does not depend on θ, and the MC defined through (45) is irreducible for *all* $\theta \in \Theta$ if and only if it is irreducible for *some* $\theta \in \Theta$. Condition (C4) therefore follows from assumption (A2). ∎

We can now apply Theorem B.1 to the MC defined through (45). This allows us to prove Theorem 5.3 under the stronger assumption (A3').

Proof of Theorem 5.3 Under assumption (A3') the assumptions of Lemma B.2 are satisfied, and we can apply Theorem B.1 to the MC defined through (45). Hence, we know that Θ contains a sequence $\overline{\theta}^n$ that satisfies (44a), and each $\overline{\theta}^n$ constitutes a strict local maximiser of f_n with probability going to one as n tends to infinity. By definition (45c) of p, every function f_n is concave, which implies that $\overline{\theta}^n$ is indeed the unique global maximiser of f_n with probability going to one as n tends to infinity.

Let m_{xy} denote the number of observations of transition $(x, y) \in \mathcal{X} \times \mathcal{X}$ in (30). We additionally set $m_{xy} := 0$ for $(x, y) \in (\mathcal{S} \times \mathcal{A})^2 \setminus (\mathcal{X} \times \mathcal{X})$. For any $\theta \in \Theta$, we have

$$\begin{aligned}
\ell_n(\theta) = \sum_{(s,a,s')\in N} n_{sas'} \log\left[p^\theta(s'|s,a)\right] + \zeta \;&=\; \sum_{\substack{x=(s,a)\in\mathcal{X},\\ y=(s',a')\in\mathcal{X}:\\ m_{xy}>0}} m_{xy} \log\left[p^\theta(s'|s,a)\right] + \zeta \\
= \sum_{\substack{x,y\in\mathcal{X}:\\ m_{xy}>0}} m_{xy} \log\left[p_{xy}(\theta)\right] + \psi \;&=\; \sum_{(x,y)\in D} m_{xy} \log\left[p_{xy}(\theta)\right] + \psi = f_n(\theta) + \psi, \qquad (46)
\end{aligned}$$

where $\psi := \log[p_0(s_1)] + \log\left[\pi^0(a_1|s_1)\right]$. The first equality follows from the definition of ℓ_n in (33'). The second equality holds because $n_{sas'} = \sum_{a'\in\mathcal{A}} m_{(s,a),(s',a')}$ and $m_{(s,a),(s',a')} = 0$ if $\pi^0(a|s) = 0$ or $\pi^0(a'|s') = 0$. The third equality follows from the definition (45c) of p and our choice of ψ. As for the fourth equality, note that all $x, y \in \mathcal{X}$ with $m_{xy} > 0$ satisfy $p_{xy}(\theta^0) > 0$ for $\theta^0 = \xi^0$. Lemma B.2 therefore ensures that $(x, y) \in D(\theta^0) = D$. The last equality follows from the definition of f_n in Theorem B.1.

From (46) and the fact that $\theta^0 = \xi^0$ we conclude that $l_n(\xi^0) = f_n(\theta^0) + \psi$. Moreover, (46) implies that $\overline{\theta}^n$ defined in Theorem B.1 represents the unique global maximiser of ℓ_n with probability going to one as n tends to infinity. The assertion of Theorem 5.3 now follows from (44a). ∎

Remark B.3 *Throughout this section, we replaced assumption (A3) with the stronger assumption (A3') from page 33. Under assumption (A3), the Jacobian of the MC's transition kernel may violate condition (C3) of Theorem B.1. We circumvent this problem by decomposing the affine mapping p in (45c) into the composition of a linear surjection, followed by an affine injection. If we replace* Θ *with the image of* $\operatorname{int}\Xi^0$ *under the surjection and p with the injection, all conditions of Theorem B.1 remain satisfied.*

C Proof of Theorem 5.6

We first investigate the convergence behaviour of the sequence φ_n of quadratic functions defined in (37a). To this end, Lemma C.1 investigates the asymptotic properties of the observation frequencies $n_{sas'}$, while Lemma C.2 investigates ξ^n, $\nabla_\xi \ell_n(\xi^n)$ and $\nabla^2_\xi \ell_n(\xi^n)$. These auxiliary results will then allow us to establish the convergence of the sequence of confidence regions Ξ^n defined in (36).

We recall that the *expected return time* of a state s in an MC is defined as the expected number of transitions between two successive visits of state s. We extend this definition to MDPs by defining the expected return time of state s under policy π as the expected return time of s in the MC defined through the state set $\mathcal{S}$ and the transition kernel (7a) with $\xi = \xi^0$.

Lemma C.1 *Under the assumptions (A1) and (A2), we have*

$$\frac{n_{sas'}}{n} \quad \underset{n\to\infty}{\longrightarrow} \quad \frac{\pi^0(a|s)\, p^{\xi^0}(s'|s,a)}{\mu_s} \qquad \textit{almost surely for all } (s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}, \tag{47}$$

where $\mu_s \in [1,\infty)$ denotes the expected return time of state $s \in \mathcal{S}$ under policy π^0.

Proof We first show that the expected return times μ_s are finite. To this end, let $\mathrm{MC}_{\mathcal{S}}(\pi;\xi)$ denote the MC defined through the state set $\mathcal{S}$ and the transition kernel (7a). Due to assumption (A2), $\mathrm{MC}_{\mathcal{S}}(\pi^0;\xi)$ is irreducible for some $\xi \in \Xi^0$. By a similar argument as in the proof of Lemma B.2, we may conclude that $\mathrm{MC}_{\mathcal{S}}(\pi^0;\xi)$ is indeed irreducible for all $\xi \in \operatorname{int} \Xi^0$. Assumption (A1) then guarantees that $\mathrm{MC}_{\mathcal{S}}(\pi^0;\xi^0)$ is irreducible, which implies that its expected return times μ_s are finite.

In view of equation (47), let n_s and n_{sa} denote the numbers of occurrences of state $s \in \mathcal{S}$ and state-action pair $(s,a) \in \mathcal{S} \times \mathcal{A}$ in the observation (30), respectively. As usual, $n_{sas'}$ denotes the number of occurrences of the state-action sequence $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, and n represents the observation length. Note that the random variables n_s, n_{sa} and $n_{sas'}$ depend on n. If $\pi^0(a|s) = 0$, then $n_{sas'} = 0$, and (47) is trivially satisfied. We therefore assume that $\pi^0(a|s) > 0$. We show that

$$\text{(A)}\ \frac{n_s}{n} \underset{n\to\infty}{\longrightarrow} \frac{1}{\mu_s} \ \text{a.s.,} \quad \text{(B)}\ \frac{n_{sa}}{n_s} \underset{n\to\infty}{\longrightarrow} \pi^0(a|s) \ \text{a.s., and} \quad \text{(C)}\ \frac{n_{sas'}}{n_{sa}} \underset{n\to\infty}{\longrightarrow} p^{\xi^0}(s'|s,a) \ \text{a.s.,}$$

where 'a.s.' abbreviates 'almost surely'. Statements (A) and (B) imply that n_s and n_{sa} become nonzero a.s. as n tends to infinity, and therefore the identity $n_{sas'}/n = (n_{sas'}/n_{sa})(n_{sa}/n_s)(n_s/n)$ holds a.s. as n tends to infinity. The assertion of this lemma then follows from the continuous mapping theorem [6].

As for claim (A), note that n_s represents the number of visits of $\mathrm{MC}_{\mathcal{S}}(\pi^0;\xi^0)$ to state $s \in \mathcal{S}$. Since $\mathrm{MC}_{\mathcal{S}}(\pi^0;\xi^0)$ is irreducible, the ergodic theorem ensures that $n_s/n \longrightarrow 1/\mu_s$ a.s. as n tends to infinity [6].

In order to prove claims (B) and (C), we introduce a new MC denoted as $\mathrm{MC}_{\mathcal{SA}}$. By construction,

$\mathrm{MC}_{\mathcal{SA}}$ is in state $s \in \mathcal{S}$ whenever the underlying MDP is in state s and the decision maker has not yet chosen any action, while $\mathrm{MC}_{\mathcal{SA}}$ is in state $(s,a) \in \mathcal{S} \times \mathcal{A}$ whenever the MDP is in state s and the decision maker has chosen action a (but before the MDP moves to a new state s'). We can interpret the state-action sequence (30) as an observation of $2n$ states of $\mathrm{MC}_{\mathcal{SA}}$, where $\mathrm{MC}_{\mathcal{SA}}$ starts in state s_1, then moves to state (s_1, a_1), after which it enters state s_2 and so on. Formally, we define $\mathrm{MC}_{\mathcal{SA}}$ through the state set $\mathcal{S} \cup (\mathcal{S} \times \mathcal{A})$ and the transition probabilities

$$p_{xy} = \begin{cases} \pi^0(a|s) & \text{if } x = s \in \mathcal{S} \text{ and } y = (s,a) \in \mathcal{S} \times \mathcal{A}, \\ p^{\xi^0}(s'|s,a) & \text{if } x = (s,a) \in S \times \mathcal{A} \text{ and } y = s' \in \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

To prove claim (B), fix $(s,a) \in \mathcal{S} \times \mathcal{A}$ and let X_i be a random binary variable that adopts the value 1 if and only if $\mathrm{MC}_{\mathcal{SA}}$ moves to state (s,a) after the ith visit of state s. By the strong Markov property, the random variables X_i are independent and identically distributed with expected value $\pi^0(a|s)$ [6]. Thus, the strong law of large numbers implies that $\sum_{i=1}^{m} X_i/m \longrightarrow \pi^0(a|s)$ a.s. as m tends to infinity. According to claim (A), $n_s \longrightarrow \infty$ a.s. as n tends to infinity. Hence, we obtain that $\sum_{i=1}^{n_s} X_i/n_s \longrightarrow \pi^0(a|s)$ a.s. as n tends to infinity. Claim (B) then follows from the fact that $n_{sa} = \sum_{i=1}^{n_s} X_i$.

The proof of claim (C) widely parallels the above argumentation for claim (B). ∎

Lemma C.2 *Under the assumptions (A1), (A2) and (A3'), observation (30) satisfies*

$$\lim_{n \longrightarrow \infty} \mathbb{P}\big(\nabla_\xi \ell_n(\xi^n) = 0\big) = 1, \tag{48a}$$

$$\operatorname*{plim}_{n \longrightarrow \infty} \big(n^\alpha \left\|\xi^n - \xi^0\right\|\big) = 0 \qquad\qquad \forall\, \alpha < 1/2, \tag{48b}$$

$$\operatorname*{plim}_{n \longrightarrow \infty} \left(\left\|\frac{1}{n}\left[\nabla^2_\xi \ell_n(\xi^n)\right] - \Sigma\right\|\right) = 0, \tag{48c}$$

where $\nabla_\xi \ell_n(\xi^n)$ and $\nabla^2_\xi \ell_n(\xi^n)$ are defined in (37b) and (37c), respectively, and

$$\Sigma := \sum_{(s,a,s') \in N_0} \frac{\pi^0(a|s)}{\mu_s\, p^{\xi^0}(s'|s,a)} \left([K_{sa}]^\top_{s'\cdot}\right)^\top \left([K_{sa}]^\top_{s'\cdot}\right), \tag{48d}$$

where $N_0 := \left\{(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \,:\, [k_{sa}\; K_{sa}]^\top_{s'\cdot} \neq 0\right\}$. Moreover, the matrix Σ is positive definite.

Proof The proof of Theorem 5.3 shows that the unique global maximiser ξ^n of ℓ_n is an element of $\operatorname{int} \Xi^0$ with probability going to one as n tends to infinity. This proves (48a).

In view of (48b), consider any sequence X_n of random variables. One can show that if $n^\alpha X_n$ converges in distribution, then $n^\beta X_n$ converges to zero in probability for all $\beta < \alpha$. Thus, (48b) follows from (44b).

Let us now consider (48c). We can replace the set N in the summation index of $\nabla^2_\xi \ell_n(\xi^n)$ in (37c) with the set N_0 used in (48d). Indeed, $N \subseteq N_0$ holds because $n_{sas'} > 0$ implies that $p^{\xi^0}(s'|s,a) > 0$ and therefore $[k_{sa}\ K_{sa}]^\top_{s'.} \neq 0$. Likewise, the numerator in (37c) vanishes for each index $(s,a,s') \in N_0 \setminus N$. Equation (48c) now follows from Lemma C.1, (48b) and the continuous mapping theorem.

It is clear that Σ is positive semidefinite. Also, $x^\top \Sigma\, x = 0$ if and only if $[K_{sa}]^\top_{s'.}\, x = 0$ for all $(s,a,s') \in N_0$ with $\pi^0(a|s) > 0$. Assumption (A3') implies that this is the case if and only if $x = 0$. Thus, the matrix Σ has full rank and is therefore positive definite. ■

We can now prove Theorem 5.6.

Proof of Theorem 5.6 Let $\mathbb{B}$ denote the closed unit ball centred at the origin of $\mathbb{R}^q$. For fixed $\alpha < 1/2$, (38) is satisfied if and only if for all $\epsilon, \gamma > 0$, there is $m \in \mathbb{N}$ such that for all $n \geq m$,

$$\mathbb{P}\left(n^\alpha\left(\Xi^n - \xi^0\right) \subseteq \epsilon\mathbb{B}\right) \geq 1 - \gamma, \tag{49}$$

where operations on sets are understood in the Minkowski sense. We define $\phi_n(x) := \varphi_n\left(n^{-\alpha}x + \xi^0\right)$. According to the definition (36) of Ξ^n, we have

$$n^\alpha\left(\Xi^n - \xi^0\right) \subseteq \{x \in \mathbb{R}^q \,:\, \phi_n(x) \geq 0\}$$

because the set on the right-hand side ignores the constraints from Ξ^0. Hence, (49) holds if

$$\mathbb{P}\left(\{x \in \mathbb{R}^q \,:\, \phi_n(x) \geq 0\} \subseteq \epsilon\mathbb{B}\right) \geq 1 - \gamma,$$

which is equivalent to

$$\mathbb{P}\left(\{x \in \mathbb{R}^q \,:\, \phi_n(x) < 0\} \supseteq \epsilon\mathbb{B}^{\mathrm{c}}\right) \geq 1 - \gamma, \tag{50}$$

where $\epsilon\mathbb{B}^{\mathrm{c}} := \mathbb{R}^q \setminus \epsilon\mathbb{B}$ denotes the complement of $\epsilon\mathbb{B}$. We prove (50) in two steps. We first show that ϕ_n is negative on $\epsilon\mathbb{B}^{\mathrm{c}} \cap 2\epsilon\mathbb{B}$. Afterwards, we show that $\phi_n(0) > \phi_n(x)$ for all $x \in \epsilon\mathbb{B}^{\mathrm{c}} \cap 2\epsilon\mathbb{B}$. Since ϕ_n is concave, this implies that ϕ_n remains negative on $\mathbb{R}^q \setminus 2\epsilon\mathbb{B}$ with high probability. We can then conclude that ϕ_n is negative on the whole set $\epsilon\mathbb{B}^{\mathrm{c}}$ with high probability, which proves (50).

Using the definition (37a) of φ_n and Lemma C.2, one can show that

$$\operatorname*{plim}_{n \longrightarrow \infty}\left(\sup_{x \in 2\epsilon\mathbb{B}}\left|n^{2\alpha-1}\phi_n(x) - \frac{1}{2}x^\top \Sigma\, x\right|\right) = 0, \tag{51}$$

where Σ is defined in (48d). In a probabilistic sense, $n^{2\alpha-1}\phi_n(x)$ therefore converges uniformly to $x^\top \Sigma\, x/2$ over $2\epsilon\mathbb{B}$. Since Σ is positive definite, see Lemma C.2, there is $\nu > 0$ such that $\Sigma \succeq \nu I$, that is,

$x^\top \Sigma\, x \geq \nu \left\|x\right\|^2$ for all x. We thus obtain that for any $\eta > 0$, we can choose m such that for all $n \geq m$,

$$\mathbb{P}\left(n^{2\alpha-1}\phi_n(0) \geq -\eta,\;\; n^{2\alpha-1}\phi_n(x) \leq -\frac{\nu}{2}\epsilon^2 + \eta \;\; \forall\, x \in \epsilon\mathbb{B}^{\mathrm{c}} \cap 2\epsilon\mathbb{B}\right) \geq 1 - \gamma.$$

For $\eta < \nu\epsilon^2/4$ this is equivalent to

$$\mathbb{P}\left(\phi_n(0) > \phi_n(x),\;\; \{x \in \mathbb{R}^q \,:\, \phi_n(x) < 0\} \supseteq \epsilon\mathbb{B}^{\mathrm{c}} \cap 2\epsilon\mathbb{B}\right) \geq 1 - \gamma.$$

According to our previous discussion, this proves equation (50) and the assertion of the theorem. ∎