Robust Markov Decision Processes

Wolfram Wiesemann, Daniel Kuhn and Berç Rustem

February 9, 2012

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.

Keywords Robust Optimization; Markov Decision Processes; Semidefinite Programming.

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 the 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 behavioral ecology [5, 20].

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 $s \in \mathcal{S}$, then 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 $\pi_t : (\mathcal{S} \times \mathcal{A})^t \times \mathcal{S} \mapsto \mathcal{M}(\mathcal{A})$. $\pi_t(\cdot|s_0, a_0, \ldots, s_{t-1}, a_{t-1}; s_t)$ 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 maximizes (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 [16]. 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 [16]. 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 [8, 16].

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 [12] and Nilim and El Ghaoui [18] 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 maximizes 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 ambiguity 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 ambiguity 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 maximizes (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 ambiguity 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 ambiguity set $\mathcal{P}$ with this property, [12] and [18] assume that independent transition samples are available for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. Under this assumption, one can employ standard results on the asymptotic properties of the maximum likelihood estimator 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 ambiguity 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. It turns out that many of the results derived for transition sampling in [12] and [18] remain valid in the new setting where the transition probabilities are estimated from observation histories.

The restriction to rectangular ambiguity sets has been introduced in [12] and [18] to facilitate computational tractability. Under the assumption of rectangularity, the robust policy evaluation and improve-

ment problems can be solved efficiently with a modified value or policy iteration. This implies, however, that non-rectangular ambiguity sets have to be projected onto the marginal sets $\mathcal{P}_{sa}$. Not only does this 'rectangularization' 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 ambiguity 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 ambiguity 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 generalized to a hierarchy of increasingly accurate approximations.

The contributions of this paper can be summarized as follows.

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

2. It is stated in [18] that the robust policy evaluation and improvement problems "seem to be hard to solve" for non-rectangular ambiguity sets. We prove that these problems cannot be approximated to any constant factor in polynomial time unless $\mathcal{P} = \mathcal{NP}$. We develop a hierarchy of increasingly accurate conservative approximations, together with ex post bounds on the incurred optimality gap.

3. We present a method to construct ambiguity sets from observation histories. Our approach allows to account for different types of a priori information about the transition kernel, which helps to reduce the size of the ambiguity set. We also investigate the convergence behavior of our ambiguity set when the length of the observation history increases.

The study of robust MDPs with rectangular ambiguity sets dates back to the seventies, see [3, 10, 22, 26] and the surveys in [12, 18]. However, most of the early contributions do not address the construction of suitable ambiguity sets. In [16], 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 [8] 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 [19, 27] and the references cited therein. Note that external variation could be addressed by encoding the unknown model parameters into the states of a partially observable MDP (POMDP) [17]. However, the optimization of POMDPs becomes

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

The remainder of the paper is organized as follows. Section 2 defines and analyzes the classes of robust MDPs that we consider. Sections 3 and 4 study the robust policy evaluation and improvement problems, respectively. Section 5 constructs ambiguity 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. In addition to the transition rewards* $r(s, a, s')$*, an expected reward of* $\mathbf{r}_s \in \mathbb{R}_+$ *is received if the MDP reaches the terminal state* $s \in \mathcal{S}_T$*. We assume that* $p_0(s) = 0$ *for* $s \notin \mathcal{S}_0$*.*

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 ambiguity 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 ambiguity 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

In this paper, we consider ambiguity sets $\mathcal{P}$ of the following type.

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

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 $K_{sa} \in \mathbb{R}^{S\times q}$. The distinction between the sets $\mathcal{P}$ and Ξ allows us to condense all ambiguous parameters in the set Ξ. This will enable

us to simplify notation in Section 5 when we construct ambiguity sets $\mathcal{P}$ from observation histories. We 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 $O_l \preceq 0$. Hence, Ξ results from the finite intersection of closed halfspaces and ellipsoids, which will allow us to solve the policy evaluation and improvement problems efficiently as second-order cone programs and semidefinite programs. 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. This implies that Ξ has a nonempty interior, that is, none of the parameters in Ξ is fully explained by the others. As the following example shows, this is not the case for the transition probabilities $p^\xi(\cdot|s,a)$ in $\mathcal{P}$.

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 ambiguity set $\mathcal{P}$ is (s,a)-*rectangular* if

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

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

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

For any ambiguity set $\mathcal{P}$, we call $\mathcal{P}_{sa}$ and $\mathcal{P}_s$ the *marginal ambiguity 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 ambiguity 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 ambiguity set. Such a dependence

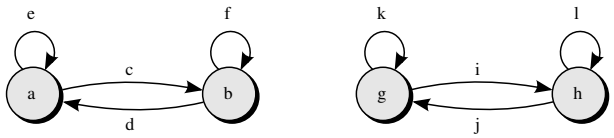

Figure 1: *MDP with two states and two actions. The left and right charts present the transition probabilities for actions 1 and 2, respectively. In both diagrams, nodes correspond to states and arcs to transitions. We label each arc with the probability of the associated transition. We suppress p_0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

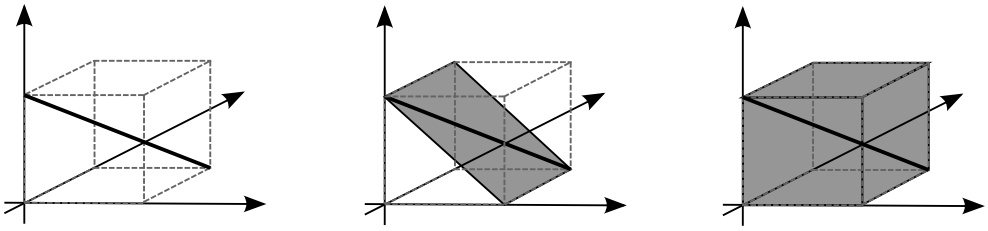

Figure 2: *Illustration of $\mathcal{P}$ (left chart) and the smallest s-rectangular (middle chart) and (s,a)-rectangular (right chart) ambiguity sets that contain $\mathcal{P}$. The charts show three-dimensional projections of $\mathcal{P} \subset \mathbb{R}^8$. The thick line represents $\mathcal{P}$, while the shaded areas visualize the corresponding rectangular ambiguity 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 $\mathbb{R}^3$.*

may arise, for example, when the actions of an MDP relate to varying degrees of intensity with which a task is executed. In Section 6, we will consider a machine replacement problem in which the condition of a machine is influenced by the actions 'repair' and 'wait'. We could imagine an extension of this problem in which there are various types of maintenance actions. In such a variant, the precise probabilities for the evolution of the machine's condition may be unknown, but it may be known that more intensive maintenance actions keep the machine in a better condition than less intensive ones. By definition, (s,a)-rectangularity implies s-rectangularity. (s,a)-rectangular ambiguity sets have been introduced in [12, 18], 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 ambiguity set in Example 2.1 is not s-rectangular. The following example shows that rectangular ambiguity 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 ambiguity set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realizations $\xi \in [0,1]$. This MDP can be assigned an ambiguity set of the form (3). Figure 2 visualizes $\mathcal{P}$ and the smallest s-rectangular and (s,a)-rectangular ambiguity sets that contain $\mathcal{P}$.*

In Section 5, we will construct ambiguity sets from observation histories. The resulting ambiguity sets turn out to be *non-rectangular*, that is, they are neither s- nor (s,a)-rectangular. Unfortunately, the robust policy evaluation and improvement problems over non-rectangular ambiguity sets are intractable

(see Section 2.3), and we will only be able to obtain approximate solutions via semidefinite programming. This is in stark contrast to the robust policy evaluation and improvement problems over s-rectangular and (s,a)-rectangular ambiguity sets, which can be solved efficiently through a sequence of second-order cone programs (see Sections 3.1 and 4). Hence, it may sometimes be beneficial to follow the approach in Example 2.2 and replace a non-rectangular ambiguity set with a larger rectangular set.

2.2 The Robust Policy Improvement Problem

We now consider the policy improvement problem, which asks for a policy that maximizes the worst-case expected total reward (2) over an ambiguity 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 $(s_0, a_0, \ldots, s_{t-1}, a_{t-1}; s_t)$. Throughout this paper, we restrict ourselves to *stationary policies* where π_t is solely determined by s_t for all $t \in \mathcal{T}$.

It is well-known that non-robust finite and infinite horizon MDPs always allow for a deterministic stationary policy that maximizes 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 [5, 20, 23].

To date, the literature on robust MDPs has focused on (s,a)-rectangular ambiguity sets. For this class of ambiguity sets, it is shown in [12, 18] that the worst-case expected total reward (2) is maximized by a deterministic stationary policy for finite and infinite horizon MDPs. Optimal policies can be determined via extensions of the value and policy iteration. For some ambiguity 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:

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

We prove a generalized version of condition (4) in Proposition A.1. A similar result for robust finite horizon MDPs is discussed in [18].

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

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

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

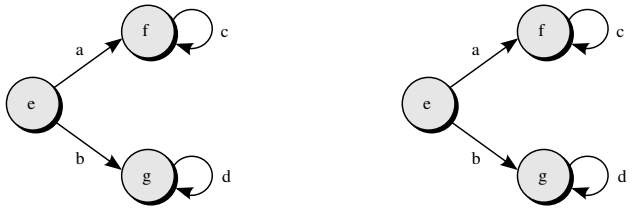

Figure 3: *MDP with three states and two actions. The left and right figures present the transition probabilities and expected rewards for actions* 1 *and* 2*, respectively. The first and second expression in an arc label corresponds to the probability and the expected reward of the associated transition, 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*.*

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

Proof As for claim (a), consider a finite horizon MDP with an s-rectangular ambiguity 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,see Remark 1.1. 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 ambiguity set. Appendix A shows that the saddle point condition (4) extends to this setting. 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.

As for claim (b), consider the robust infinite horizon MDP that is visualized in Figure 3. The ambiguity set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realizations $\xi \in [0,1]$. This MDP can be assigned an s-rectangular ambiguity 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

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

Over $\beta \in [0,1]$, this expression has its unique maximum at $\beta^* = 1/2$, that is, the optimal policy is randomized. 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 randomized. ∎

Figure 3 illustrates the counterintuitive result that randomization is superfluous for (s,a)-rectangular

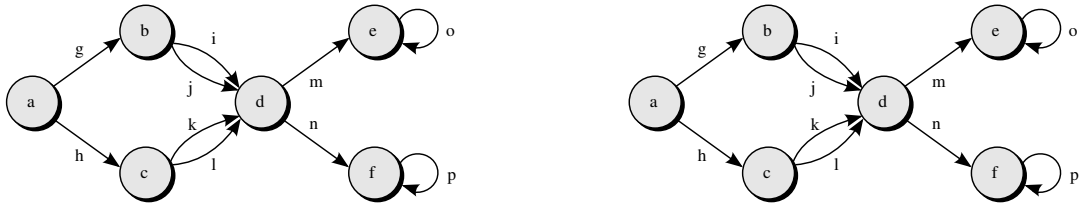

Figure 4: *MDP with six states and two actions. The initial state distribution p_0 places unit mass on state* 1*. The same drawing conventions as in Figure 3 are used.*

ambiguity sets. If we project the ambiguity 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 randomization becomes ineffective.

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

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

Proof Consider the robust infinite horizon MDP with six states and two actions that is visualized in Figure 4. The ambiguity set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realizations $\xi \in [0,1]$. This MDP can be assigned an ambiguity 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 $\beta_2 = \pi_2(1|1, a_0, 3, a_1; 4)$.

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 strictly concave in ξ for all $\beta \in [0,1]^2$ if $M > \lambda^2/(1-\lambda)$. Thus, the 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 maximizer 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 $\mathbf{r}_5 := \lambda^3/(1-\lambda)$ and $\mathbf{r}_6 := 0$, then we can extend the result to robust finite horizon MDPs. ∎

Although the policy improvement problem over non-rectangular ambiguity 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 randomized stationary policies for robust MDPs.

2.3 Complexity of the Robust Policy Evaluation Problem

We show that unless $\mathcal{P} = \mathcal{NP}$, the worst-case expected total reward (2) over non-rectangular ambiguity sets cannot be approximated in polynomial time. To this end, we will reduce the $\mathcal{NP}$-hard 0/1 Integer Programming (IP) problem to the approximate evaluation of (2):

0/1 Integer Programming.
Instance. Given are a matrix $F \in \mathbb{Z}^{m \times n}$ and a vector $g \in \mathbb{Z}^m$.
Question. Is there a vector $x \in \{0,1\}^n$ such that $Fx \leq g$?

The IP problem predominantly studied in the literature also contains a linear objective function $c^\top x$, and it asks whether there a vector $x \in \{0,1\}^n$ such that $Fx \leq g$ and $c^\top x \leq \zeta$ for some $\zeta \in \mathbb{Z}$, see [9]. We can easily transform this problem into an instance of our IP problem by adding the constraint $c^\top x \leq \zeta$.

Assume that $x \in [0,1]^n$ constitutes a fractional solution to the linear inequality system $Fx \leq g$. The following lemma shows that we can obtain an integral vector $y \in \{0,1\}^n$ that satisfies $Fy \leq g$ by rounding x if its components are 'close enough' to zero or one.

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

Proof By construction, $F_{i\cdot}^\top y \leq F_{i\cdot}^\top x + \sum_j |F_{ij}|\, \epsilon < F_{i\cdot}^\top x + 1 \leq g_i + 1$ for all $i \in \{1, \ldots, m\}$. Due to the integrality of F, g and y, we therefore conclude that $Fy \leq g$. ∎

We now show that the robust policy evaluation problem is hard to approximate. We say that the approximation z to the worst-case expected total reward z^* defined in (2) has a *relative error* β if

$$\begin{cases} \dfrac{|z - z^*|}{\min\{|z|, |z^*|\}} \leq \beta & \text{if } \min\{|z|, |z^*|\} > 0, \\ |z - z^*| \leq \beta & \text{otherwise.} \end{cases}$$

Theorem 2.6 *Unless $\mathcal{P} = \mathcal{NP}$, there is no algorithm that approximates the worst-case expected total reward (2) over an ambiguity set of the form (3) with any relative error β in polynomial time for deterministic or randomized stationary policies over a finite or infinite time horizon.*[1]

[1] Here and in the proof, 'polynomial' refers to the size of the problem instance in a suitably chosen encoding [9].

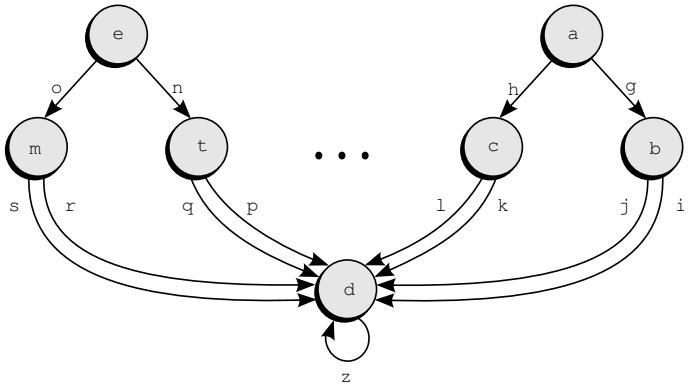

Figure 5: *MDP with $3n + 1$ states and one action. The distribution p_0 places a probability mass of $1/n$ on each state b_j, $j = 1, \ldots, n$. The drawing conventions from Figure 3 are used.*

Proof Assume that there was a polynomial time algorithm that approximates z^* with a relative error β. In the following, we will use this algorithm to decide the IP problem in polynomial time. Since the IP problem is $\mathcal{NP}$-hard, this would imply that $\mathcal{P} = \mathcal{NP}$.

Fix an IP instance specified through F and g. We construct a robust MDP with states $\mathcal{S} = \left\{b_j, b_j^0, b_j^1 : j = 1, \ldots, n\right\} \cup \{\tau\}$, a single action and a discount factor $\lambda \in (0, 1)$ that can be chosen freely. The state transitions and expected rewards are illustrated in Figure 5. We choose $M > (\beta[1 + \beta]n) / \left(\lambda\epsilon^2\right)$, where ϵ is defined as in Lemma 2.5. The ambiguity set $\mathcal{P}$ contains all transition kernels associated with $\xi \in \Xi := \{\xi \in [0, 1]^n : F\xi \leq g\}$. If Ξ is empty, which can be decided in polynomial time, then the IP instance is infeasible. Otherwise, we can decide in polynomial time whether Ξ contains a Slater point, and—if this is not the case—transform Ξ in polynomial time to a lower-dimensional set that contains a Slater point. Note that this requires adaptation of the linear mapping from Ξ to $\mathcal{P}$, which can be achieved in polynomial time as well.

We now show that the IP instance has a feasible solution if and only if the approximate worst-case expected total reward z of the robust MDP from Figure 5 satisfies $|z| \leq \beta$.

Assume first that $|z| > \beta$. If there was a feasible solution $x \in \{0, 1\}^n$ to the IP instance such that $Fx \leq g$, then the expected total reward under the transition kernel associated with $\xi = x$ would be zero. This would imply, however, that $z^* = 0$, and hence the relative error of our approximation algorithm would be $|z - z^*| = |z| > \beta$, which is a contradiction. We therefore conclude that IP is infeasible if the approximate worst-case expected total reward z satisfies $|z| > \beta$.

Assume now that $|z| \leq \beta$. We distinguish the two cases $z^* = 0$ and $z^* \neq 0$. In the first case, there is a transition kernel associated with $\xi \in \Xi$ that results in an expected total reward of zero. This implies that $\xi \in \{0, 1\}^n$, and therefore the IP instance has a feasible solution $x = \xi$. If $z^* \neq 0$, on the other hand, there is no $\xi \in \Xi$ that satisfies $\xi \in \{0, 1\}^n$. We can strengthen this result to conclude that there

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

Table 1: *Properties of infinite horizon MDPs with different ambiguity sets. From left to right, the columns describe the structure of the ambiguity set, the structure of the optimal policy, and the complexity of the policy evaluation and improvement problems over randomized stationary policies. Each ambiguity set is of the form (3). For nonconvex ambiguity 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 ambiguity sets are optimized by randomized stationary policies.*

is no $\xi \in \Xi$ that satisfies $\xi \in ([0,\epsilon] \cup [1-\epsilon,1])^n$, for otherwise we could use Lemma 2.5 to round such a ξ to a vector $\xi' \in \Xi$ that satisfies $\xi' \in \{0,1\}^n$ and $F\xi' \leq g$. This implies, however, that for every $\xi \in \Xi$ there is a component $q \in \{1,\ldots,n\}$ such that $\xi_q \notin ([0,\epsilon] \cup [1-\epsilon,1])$, and therefore the worst-case expected total reward of the robust MDP satisfies

$$z^* \;\geq\; \frac{1}{n}\xi_q(1-\xi_q)\lambda M \;\geq\; \frac{\lambda\epsilon^2 M}{n} \;>\; \beta(1+\beta).$$

If we substitute z^* into the relative error formula, then we obtain

$$\frac{|z-z^*|}{\min\{|z|,|z^*|\}} \;\geq\; \frac{z^*-\beta}{\beta} \;>\; \frac{\beta(1+\beta)-\beta}{\beta} \;=\; \beta,$$

which violates our assumption that the relative error of z does not exceed β. We thus conclude that if $|z| \leq \beta$, then $z^* = 0$ and the IP instance has a feasible solution.

We have shown that unless $\mathcal{P} = \mathcal{NP}$, the robust policy evaluation problem (2) cannot be approximated in polynomial time with any relative error β. Since the policy space of the constructed MDP constitutes a singleton, our proof applies to robust MDPs with deterministic or randomized stationary policies. If we remove the self-loop emanating from state τ, introduce a terminal reward $\mathfrak{r}_\tau := 0$ and multiply the transition rewards with λ, then our proof also applies to robust finite horizon MDPs. ∎

Remark 2.7 *Throughout this section we assumed that $\mathcal{P}$ is a convex set of the form (3). If we extend our analysis to nonconvex ambiguity sets, then we obtain the results in Table 1. Note that the complexity of some of the policy evaluation and improvement problems will be discussed in Sections 3 and 4.*

3 Robust Policy Evaluation

It is shown in [12, 18] that the worst-case expected total reward (2) can be calculated in polynomial time for certain types of (s,a)-rectangular ambiguity sets. We extend this result to the broader class of s-rectangular ambiguity sets in Section 3.1. On the other hand, Theorem 2.6 shows that the evaluation of (2) is difficult for non-rectangular ambiguity sets. We therefore develop conservative approximations for the policy evaluation problem over general ambiguity 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 labor economics, industrial organization and marketing [16].

Our solution approaches for s-rectangular and non-rectangular ambiguity 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

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

$v_s(\pi;\xi)$ 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

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

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 $\widehat{r} : \Pi \times \Xi \overset{\text{a}}{\mapsto} \mathbb{R}^S$ 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 [20]. 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

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

Here, the first identity follows from the matrix inversion lemma, see e.g. Theorem C.2 in [20], 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 $\Pi \times \Xi$. 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 proved assertion (a).

Assertions (b) and (c) follow directly from Theorems 6.1.1 and 6.2.2 in [20], respectively. ∎

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

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

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 ambiguity 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 Ambiguity Sets

We show that the policy evaluation problem (10) is optimized by a constant reward to-go function if the ambiguity 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 ambiguity set $\mathcal{P}$, the policy evaluation problem (10) is optimized 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*

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

Proof 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:

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

Afterwards, we prove that the optimal values of (10) and (12) coincide for s-rectangular ambiguity 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 parametrized 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 [12, 18] 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 [20], this implies that $w^* \geq w$ for every feasible solution w to (13). By non-negativity of p_0, w^* must therefore maximize (13). Since (12) and (13) are equivalent, we have thus shown that w^* maximizes (12).

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

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

Since p_0 is non-negative, 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 ambiguity 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 3.3 *If the ambiguity set $\mathcal{P}$ is s-rectangular, then problem (10) can be solved to any accuracy ϵ in polynomial time $\mathcal{O}\left(q^3L^{3/2}S\log^2\epsilon^{-1} + qAS^2\log\epsilon^{-1}\right)$.*

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 [20]. By construction, w^{N+1} is feasible for the policy evaluation problem (10), see [20]. We can adapt Theorem 5 from [18] 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

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

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 [15]. For $w\in\mathbb{R}^S$, we can evaluate $\phi(\pi; w)$ by solving the following

second-order cone program:

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

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 $\mathcal{O}\left(qAS+q^2L\right)$, and we can solve (16) to accuracy δ in time $\mathcal{O}\left(q^3L^{3/2}\log\delta^{-1}\right)$. 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 $\mathcal{O}\left(q^3L^{3/2}S\log\delta^{-1}+qAS^2\right)$. 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 [12, 18] for a discussion of different numerical schemes.

Remark 3.4 (Finite Horizon MDPs) *For a finite horizon MDP, we can solve the policy evaluation problem (10) over an s-rectangular ambiguity 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 3.3 shows that we obtain an ϵ-optimal solution to the policy evaluation problem (10) in time $\mathcal{O}\left(q^3L^{3/2}S\log\epsilon^{-1}+qAS^2\right)$ if we evaluate $\widehat{\phi}$ to the accuracy $\epsilon/(T-1)$.

Remark 3.5 (Generalized s-Rectangularity) *Consider a robust infinite horizon MDP whose state space $\mathcal{S}$ can be partitioned into nonempty disjoint sets $\mathcal{S}_i$, $i\in\mathcal{I}$, such that the ambiguity set $\mathcal{P}$ satisfies the following* generalized s-rectangularity condition*:*

$$\mathcal{P}=\bigtimes_{i\in\mathcal{I}}\mathcal{P}(\mathcal{S}_i), \qquad \text{where} \qquad \mathcal{P}(\mathcal{S}_i) := \left\{(P_{sa})_{s,a} \,:\, s\in\mathcal{S}_i,\ a\in\mathcal{A},\ P\in\mathcal{P}\right\} \ \text{for } i\in\mathcal{I}.$$

Assume further that under policy π, each subset of states $\mathcal{S}_i$, $i\in\mathcal{I}$, contains a designated entrance state $\sigma_i\in\mathcal{S}_i$ such that $p^\xi(s'|s,a)=0$ for all $s\in\mathcal{S}\setminus\mathcal{S}_i$, $s'\in\mathcal{S}_i\setminus\{\sigma_i\}$, $\xi\in\Xi$ and $a\in\mathcal{A}$ with $\pi(a|s)>0$. Each robust MDP with an s-rectangular ambiguity set satisfies these requirements for $\mathcal{S}_i:=\{i\}$, $i\in\mathcal{I}:=\mathcal{S}$.

Theorem 3.2 extends to generalized s-rectangular ambiguity sets if we replace the contraction ϕ with

$$\phi'_s(\pi; w) := \min_{\xi^s \in \Xi} \left\{ \sum_{s' \in \mathcal{S}'} \sum_{T \in \mathcal{T}(s,s')} \left[\prod_{t=0}^{|T|-1} \pi(a_t|s_t)\, p^{\xi}(s_{t+1}|s_t, a_t) \right] \left[\sum_{t=0}^{|T|-1} \lambda^t\, r(s_t, a_t, s_{t+1}) + \lambda^{|T|}\, w_{s'} \right] \right\} \quad \forall\, s \in \mathcal{S}',$$

where $\mathcal{S}' := \{\sigma_i : i \in \mathcal{I}\}$ denotes the set of entrance states and $\mathcal{T}(s, s')$ is the set of all state-action sequences $T = (s_0 = s, a_0, s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t = s')$, $|T| = t$, that lead from state $s \in \mathcal{S}'$ to state $s' \in \mathcal{S}'$ and whose intermediate states $s_1, \ldots, s_{t-1}$ are elements of $\mathcal{S}_i \setminus \{\sigma_i\}$, where $i \in \mathcal{I}$ satisfies $s = \sigma_i$.

This result can be interpreted as follows. We replace the robust MRP from Theorem 3.2 with a semi-Markov reward process defined over the states $\mathcal{S}'$, where the random holding time in any state $s \in \mathcal{S}'$ depends on both s and the consecutive state $s' \in \mathcal{S}'$ and is determined by the realized state-action sequence $T \in \mathcal{T}(s, s')$. One readily shows that the ambiguity set of this new process is s-rectangular. Note that the new process does not satisfy condition (3) *in general, but this is not required for Theorem 3.2.*

3.2 Robust Policy Evaluation over Non-Rectangular Ambiguity Sets

If the ambiguity set $\mathcal{P}$ is non-rectangular, then Theorem 2.6 implies that constant reward to-go functions are no longer guaranteed to optimize 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 ambiguity set, and define $\overline{\mathcal{P}} := \bigtimes_{s \in \mathcal{S}} \mathcal{P}_s$ as the smallest s-rectangular ambiguity 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 ambiguity set $\mathcal{P}$. As for the second property, the proof of Theorem 3.2 shows that w^* minimizes (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 characterization of the robust value iteration. On one hand, the robust value iteration determines the exact worst-case expected total reward over the rectangularized ambiguity 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 ambiguity set $\mathcal{P}$. Hence, rectangularizing the ambiguity 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 optimization 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*

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

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 [14]).

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 [21] ensures that the system

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

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

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

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

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

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,*

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

as well as the semidefinite program

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

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

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

We first show that (21) is solvable. Since $p_0^\top (w + W\xi)$ is linear in (w, W) and continuous in ξ while Ξ is compact, $\inf_{\xi\in\Xi} \left\{ p_0^\top (w + W\xi) \right\}$ 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 parametrized by $\xi \in \Xi$. However, the feasible region of (21) is *not* bounded because any reward to-go function of the form (we, 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 $w + W\xi \geq 0$ 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-neighborhood of $\overline{\xi}$ that is contained in Ξ. Hence, W must be bounded because all points ξ in this neighborhood 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

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

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, 15] 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) optimizes 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 ambiguity 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 the 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

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

where $F_i \in \mathbb{S}^m$ for $i = 0,\ldots,n$, to accuracy ϵ in time $\mathcal{O}\big(n^2m^{\frac{5}{2}}\log\epsilon^{-1}\big)$, see [24]. 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 $\mathcal{O}\big(n^2m^{\frac{1}{2}}\sum_j m_j^2\big)$. 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 ambiguity sets from observation histories. Asymptotically, this method generates an ambiguity set Ξ that is described by a single quadratic inequality ($L = 1$), which means that problem (20) can be solved in time $\mathcal{O}\big(q^{\frac{9}{2}}S^{\frac{7}{2}}\log\epsilon^{-1} + q^2AS^2\big)$. 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 ambiguity 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 ambiguity 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 ambiguity 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 $\tau \le p_0^\top(w + W\xi)$ 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.

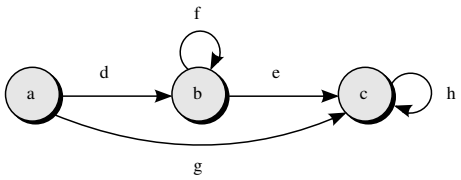

Figure 6: *MDP with three states and one action. p_0 places unit probability mass on state 1. The same drawing conventions as in Figure 3 are used.*

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

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

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 visualized in Figure 6. The ambiguity set $\mathcal{P}$ encompasses all transition kernels that correspond to parameter realizations $\xi \in [0, 1]$. This MDP can be assigned an ambiguity 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 realizations 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 ambiguity set is non-rectangular. Moreover, from Theorem 2.6 we know that this lower bound can be of poor quality. We would therefore like to estimate the approximation error incurred by solving (20). 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 optimization 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 [20]. 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 [4, 11] to reduce the approximation error via piecewise affine approximations of the reward to-go function. Since the resulting generalization 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 discount 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

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

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 ambiguity set $\mathcal{P}$ is s-rectangular. We now extend this result to the policy improvement problem.

Theorem 4.1 *For an s-rectangular ambiguity set $\mathcal{P}$, the policy improvement problem (24) is optimized 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*

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

where ϕ is defined in (11). For each $s \in \mathcal{S}$, let $\pi^s \in \arg\max_{\pi \in \Pi} \{\phi_s(\pi; w^*)\}$ 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 [12, 18] 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 [20] 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 analyzes 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 optimization problem:

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

Second-order cone duality [1, 15] allows us to replace the semi-infinite constraint (27b) with the following linear and conic quadratic constraints:

$$\begin{array}{lll} \exists\, Y \in \mathbb{R}^{q\times L},\, z \in \mathbb{R}^L,\, t \in \mathbb{R}^L \;: & \tau - \sum_{a\in\mathcal{A}} \pi_a k_{sa}^\top (r_{sa} + \lambda w) \leq -\sum_{l=1}^{L} \left(\frac{1-\omega_l}{2} z_l + \frac{\omega_l + 1}{2} t_l \right) & \text{(27b.1)} \\ & \sum_{l=1}^{L} \left(\Omega_l^\top Y_{\cdot l} - \frac{1}{2} o_l \left[z_l - t_l \right] \right) = \sum_{a\in\mathcal{A}} \pi_a K_{sa}^\top (r_{sa} + \lambda w) & \text{(27b.2)} \\ & \left\| \begin{bmatrix} Y_{\cdot l} \\ z_l \end{bmatrix} \right\|_2 \leq t_l \quad \forall\, l = 1, \ldots, L. & \text{(27b.3)} \end{array}$$

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 3.4, 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 ambiguity 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 ambiguity 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 optimized 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 $\bigtimes_s \mathcal{P}_s$. We can try to obtain better solutions to (24) over non-rectangular

ambiguity sets by replacing the constant reward to-go approximations with affine or piecewise affine approximations. The associated optimization 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 optimization procedure.

1. *Initialization.* Choose $\pi^1 \in \Pi$ (best policy found) and set $i := 1$ (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 optimization problem

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

   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. For $L = 1$, Algorithm 4.1 converges to a partial optimum of the policy improvement problem (24). We refer to [13] for a detailed convergence analysis.

5 Constructing Ambiguity 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 denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the probability space for the Markov chain of state-action pairs induced by π^0 and the unknown true transition kernel P^0. The sample space Ω represents the set of all state-action sequences in $(\mathcal{S} \times \mathcal{A})^\infty$, while $\mathcal{F}$ is defined as the product σ-field Σ^∞, where Σ denotes the power set of $\mathcal{S} \times \mathcal{A}$. Moreover, $\mathbb{P}$ is the product measure induced by π^0 an $\mathcal{P}^0$, see e.g. [2, Theorem 4.11.2]. We can use the observation (30) to construct an ambiguity 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 ambiguity 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 ambiguity set which incorporates all available a priori information about P^0. We then combine this structural information with 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 ambiguity set that satisfies the requirements from Sections 3 and 4. We close with an asymptotic analysis of our approach.

5.1 Structural Ambiguity Set

Traditionally, ambiguity 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. In [12] and [18], such a priori knowledge is incorporated through maximum a posteriori models and moment information about the transition probabilities, respectively. In this paper, we follow a different approach and condense all available a priori information about the MDP into the *structural ambiguity set* $\mathcal{P}^0$. The use of structural information excludes irrelevant transition kernels and therefore leads to a smaller ambiguity 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 ambiguity 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 [6], 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 [6]. For a sufficiently long observation, we can therefore assign zero probability to unobserved transitions.

We illustrate the construction of the structural ambiguity 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 modeled 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 centered 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 derive the log-likelihood function for the observation (30) and calculate 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 [25]. Our approach differs in two main aspects. Firstly, due to the nature of the newsvendor problem, the observation history in [25] 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 [25] 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 behavior 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 maximizing 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.*

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

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. Note that ξ^n coincides with the empirical transition probabilities. This is an artefact of the structural ambiguity sets defined in Example 5.1, and it does not generalize to other classes of structural ambiguity sets.

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

Assumption (A2) 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 ambiguity set for the unknown true transition kernel P^0 would be the structural ambiguity 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.

Proof 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 $s \in \mathcal{S}_0$ 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 $(s^i_1, a^i_1, \ldots, s^i_{T-1}, a^i_{T-1}, s^i_T)$, $i = 1, \ldots, m$, of the finite horizon MDP as one observation

$$(s^1_1, a^1_1, \ldots, s^1_{T-1}, a^1_{T-1}, s^1_T, a^1_T;\;\; \ldots \;\; ; s^m_1, a^m_1, \ldots, s^m_{T-1}, a^m_{T-1}, s^m_T, a^m_T)$$

of the corresponding infinite horizon MDP. In this concatenated observation, the terminal actions $a^i_T \in \mathcal{A}$ 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 $s \in \mathcal{S}_0$ with $p_0(s) > 0$.

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

Remark 5.4 *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

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

with

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

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

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 ambiguity 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 ambiguity set as n tends to infinity. In Theorem 5.5 below we establish that $\mathcal{P}^n$ converges to the unknown true transition kernel P^0 of the MDP and analyze 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 column 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 ambiguity set $\mathcal{P}^n$ will not converge to a singleton as the observation length n tends to infinity.

In the following proposition, we analyze the Hausdorff distance between the two sets Ξ^n and $\{\xi^0\}$. 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 5.5 *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.

Proof See Appendix C. ∎

We now show that under the assumptions of Theorem 5.5, 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 ambiguity set.

Proposition 5.6 *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 5.7 *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 5.6 Fix any $\alpha < 1/2$. By Theorem 5.5, 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 5.6 immediately extends to the affine reward to-go approximations obtained from the semidefinite program (20).

Corollary 5.8 *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 5.5 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 $\tau^n \leq p_0^\top v(\pi; \xi)$ 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 $\tau^n \geq p_0^\top w^n$ for the vector w^n defined in Proposition 5.6. Hence, the assertion follows from the convergence of $p_0^\top w^n$, see Proposition 5.6. ∎

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.

Proposition 5.9 *For $\Xi = \Xi^n$, let π^n denote an optimal policy determined by Algorithm 4.1 or the robust*

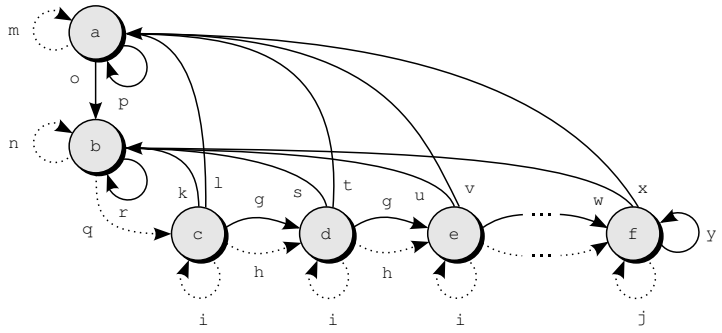

Figure 7: *MDP for the machine replacement problem. Shown are the transition probabilities for the two actions 'do nothing' (dashed arcs) and 'repair' (solid arcs). The states 8, R1 and R2 pay an expected reward of -20, -2 and -10, respectively, while no reward is received in the other states. We use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

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 ambiguity set is non-rectangular.

6 Numerical Example

We apply the robust policy evaluation and improvement methods from Sections 3 and 4 to the machine replacement problem presented in [8]. This problem concerns a single machine whose condition is described by eight 'operative' states $1,\ldots,8$ and two 'repair' states R1 and R2. The initial state is selected according to a uniform distribution over all states. 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 maximizes the expected total reward under the discount factor $\lambda = 0.8$. If all transition probabilities are known, then we can model this problem as an MDP, see Figure 7. One can readily transform this MDP into an equivalent one that satisfies the definitions in Section 1.

We first investigate the statistical properties of the confidence regions presented in Section 5. These confidence regions require us to select a structural ambiguity set $\mathcal{P}^0$ that describes the available a priori

knowledge about the true transition probabilities P^0 from Figure 7. We start with the naive ambiguity set presented in Example 5.1, which only requires us to specify which transitions can occur. For this choice of $\mathcal{P}^0$, we need to estimate a parameter vector ξ^0 with 25 components: three for each of the states $1, \ldots, 7$, two for state 8, and one for state R1 and R2, respectively. To derive a confidence region for ξ^0, we assume that we have access to an observation history that is generated by a historical policy π^0. In our experiments, we assume that π^0 chooses the actions 'do nothing' and 'repair' in each operative state $1, \ldots, 7$ with probability 0.8 and 0.2, respectively. In the states 8 and R2, the policy always chooses the action 'repair', while the action 'do nothing' is chosen deterministically in state R1. Using this policy, we randomly generate 10,000 observation histories of length $n \in \{1{,}000,\ 2{,}000,\ \ldots,\ 50{,}000\}$. For these histories, Figure 8 reports the empirical frequencies with which the true transition probabilities P^0 from Figure 7 are contained in 80%, 90%, 95% and 99% confidence regions (Section 5.2), as well as their quadratic approximations (Section 5.3). While the exact confidence regions are overly optimistic for short observation histories, the figure shows that the empirical frequencies converge to the selected confidence levels as the length of the observation history increases. The quadratic approximations are even more optimistic, and observation histories of length $n \geq 25{,}000$ are required to obtain a reasonably accurate match between empirical frequencies and the selected confidence levels. Both the exact and the approximated confidence regions are less optimistic if larger confidence levels are selected. We also recorded the empirical frequencies with which the true transition probabilities P^0 are contained in the s-rectangular and (s,a)-rectangular projections of the exact and quadratically approximated confidence regions. For observation histories of length $n \in \{1{,}000,\ 2{,}000,\ \ldots,\ 50{,}000\}$, these frequencies were always above 98.85% (80% confidence region), 99.15% (90% confidence region), 99.35% (95% confidence region) and 99.65% (99% confidence region) for s-rectangular projections, and even higher for (s,a)-rectangular projections. Hence, the rectangular projections result in ambiguity sets that are much more conservative than desired, given a selected confidence level. We conclude that for the naive structural ambiguity set from Example 5.1, neither the quadratic approximation nor the rectangular projections provide completely satisfying confidence regions. Nevertheless, the quadratic approximation seems to perform better than the (s,a)-rectangular and s-rectangular projections in our tests.

We now consider the robust policy evaluation problem. Under the true transition probabilities from Figure 7, the expected total reward of the policy π^0 is -11.43. We want to estimate this value using the information provided by the structural ambiguity set $\mathcal{P}^0$ and an observation history. To this end, we can apply the robust value iteration from Section 3 to (s,a)-rectangular or s-rectangular projections of the exact confidence region derived in Section 5.2, or we can solve the semidefinite program (20), which uses the quadratically approximated confidence region from Section 5.3. The left part of Table 2 compares the worst-case expected total rewards obtained with these three approaches for various lengths n of the

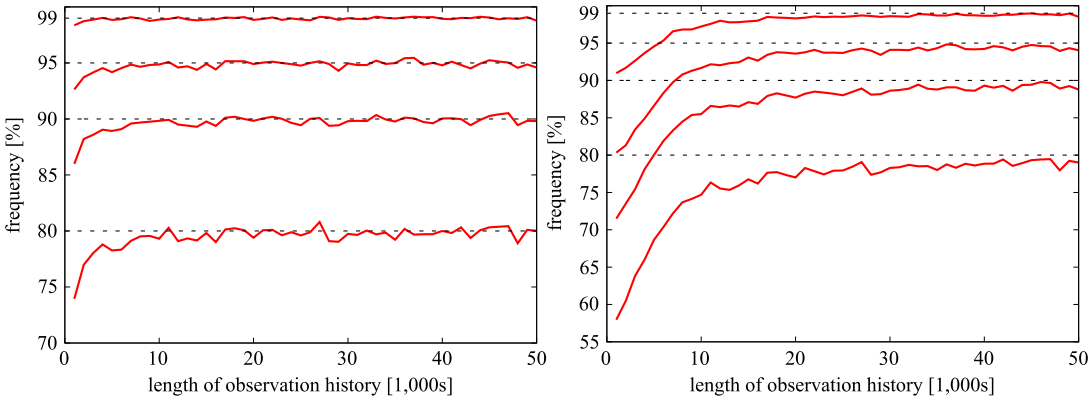

Figure 8: *Statistical properties of the exact and approximated confidence regions using the naive structural ambiguity set described in Example 5.1. The left (right) graph describes the empirical frequencies with which the true transition probabilities P^0 are contained in the exact (quadratically approximated) confidence regions, respectively. From bottom to top, the solid lines refer to the* 80%, 90%, 95% *and* 99% *confidence regions.*

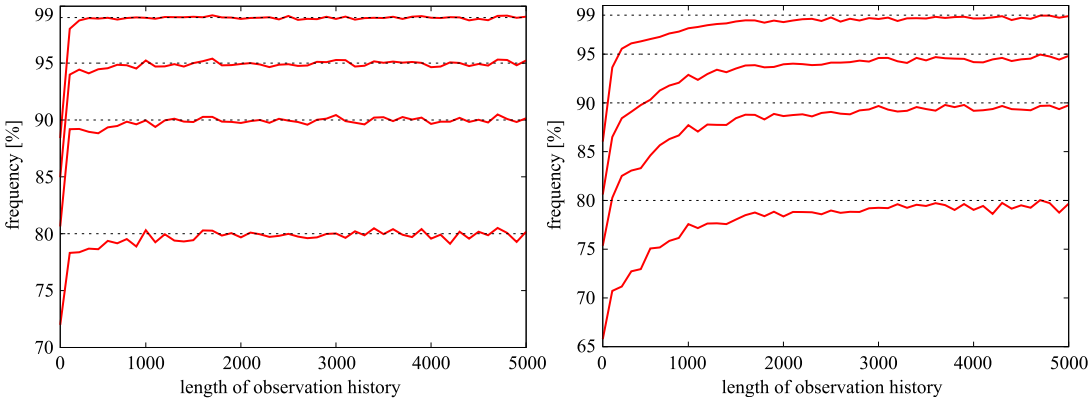

Figure 9: *Statistical properties of the exact and approximated confidence regions using the refined structural ambiguity set described in the text. The interpretation of the graphs is similar to Figure 8.*

observation history. As expected, the s-rectangular policy evaluation problem provides better estimates than its (s,a)-rectangular counterpart, and the semidefinite programming approximation outperforms both variants of the robust value iteration. In all cases, the approximation quality increases with the length of the observation history. The table also shows that the linear approximation of the reward to-go function is surprisingly accurate. Indeed, the gap between the lower and upper bounds is negligible in all of our experiments.

The transition probabilities in Figure 7 are highly structured. In particular, the probabilities associated with the transitions emanating from state s under either action are identical for the states $s \in \{1,\ldots,7\}$. We now assume that although these probabilities are unknown, they are known to be identical for the states $s \in \{1,\ldots,7\}$. This additional information can be incorporated into the struc-

	no structural knowledge					structural knowledge			
n	80%	90%	95%	99%	n	80%	90%	95%	99%
1,000	-29.60	-30.94	-32.31	-33.85	500	-24.81	-26.42	-27.93	-31.08
10,000	-16.05	-16.38	-16.65	-17.17	1,000	-19.99	-20.82	-22.01	-24.02
25,000	-14.17	-14.39	-14.53	-14.85	2,500	-16.34	-16.95	-17.57	-18.69
50,000	-13.35	-13.45	-13.54	-13.74	5,000	-14.80	-15.12	-15.55	-16.23
(s,a)-rect. 1,000	-28.45	-29.93	-31.11	-33.04	500	-23.28	-25.31	-27.22	-29.73
10,000	-15.51	-15.82	-16.03	-16.49	1,000	-19.08	-20.26	-21.21	-23.20
25,000	-13.89	-14.03	-14.16	-14.41	2,500	-15.88	-16.48	-17.08	-17.91
s-rect. 50,000	-13.09	-13.20	-13.26	-13.42	5,000	-14.40	-14.80	-15.01	-15.81
1,000	-20.46	-21.61	-22.47	-24.11	500	-14.25	-14.78	-15.19	-16.60
	-20.46	*-21.61*	*-22.46*	*-24.10*		*-14.22*	*-14.76*	*-15.18*	*-16.57*
10,000	-12.95	-13.05	-13.12	-13.27	1,000	-13.29	-13.56	-13.77	-14.29
	-12.94	*-13.04*	*-13.12*	*-13.27*		*-13.26*	*-13.53*	*-13.75*	*-14.27*
25,000	-12.36	-12.41	-12.47	-12.57	2,500	-12.52	-12.67	-12.70	-13.00
nonrect.	*-12.36*	*-12.41*	*-12.46*	*-12.56*		*-12.50*	*-12.65*	*-12.67*	*-12.99*
50,000	-12.07	-12.11	-12.14	-12.22	5,000	-12.05	-12.20	-12.22	-12.52
	-12.07	*-12.11*	*-12.13*	*-12.21*		*-12.03*	*-12.19*	*-12.21*	*-12.50*

Table 2: *Robust policy evaluation results for 100 randomly generated observation histories of different length n. The left part of the table documents the average expected total reward estimates for the 80%, 90%, 95% and 99% confidence regions with the naive structural ambiguity set from Example 5.1, and the right part of the table reports the same values for the refined ambiguity set described in the text. For the semidefinite programming approximation ('nonrect.'), the table also presents the associated upper bounds (italicized values).*

tural ambiguity set $\mathcal{P}^0$ to reduce the dimension of ξ^0 from 25 to 5. The impact of this refined ambiguity set on the statistical properties of the confidence regions and the robust policy evaluation problem is documented in Figure 9 and the right part of Table 2, respectively. With the refined ambiguity set, observation histories of length 2,500 are sufficient to obtain a reasonably accurate match between empirical frequencies and the selected confidence levels, and both the robust value iterations and the semidefinite programming approximation provide significantly better estimates for the expected total reward of π^0.

We now use the observation histories generated by the historical policy π^0 to solve the robust policy improvement problem from Section 4. While we can use the exact confidence region from Section 5.2 for the (s,a)-rectangular value iteration, we use the quadratic approximation from Section 5.3 to obtain a tractable reformulation for the contraction mapping underlying the s-rectangular value iteration. The optimal policy π^* for the true transition probabilities P^0 yields an expected total reward of -5.98, and it chooses the 'repair' action in states 6, 7, 8 and R2, whereas the 'do nothing' action is selected in all other states. Table 3 reports the estimated and the true expected total rewards of the policies determined by the (s,a)-rectangular and s-rectangular robust value iteration, as well as for the sequential convex optimization algorithm from Section 4. In all three cases, we employ the refined structural ambiguity set from the previous paragraph. The table shows that in this example, the true expected total rewards (evaluated under P^0) of the policies determined by each of the three approaches is close to the expected

	(s,a)-rectangular				s-rectangular				nonrectangular			
n	80%	90%	95%	99%	80%	90%	95%	99%	80%	90%	95%	99%
500	-20.19	-21.70	-23.18	-25.93	-10.59	-11.41	-12.32	-14.37	-8.53	-8.86	-9.34	-10.03
	-6.45	*-6.65*	*-6.84*	*-7.57*	*-6.13*	*-6.06*	*-6.20*	*-7.04*	*-6.04*	*-6.04*	*-6.04*	*-6.06*
1,000	-15.05	-16.02	-17.12	-19.49	-8.98	-9.18	-9.81	-10.65	-7.67	-8.03	-8.07	-8.63
	-6.20	*-6.18*	*-6.14*	*-6.25*	*-5.99*	*-5.99*	*-5.99*	*-5.99*	*-6.02*	*-6.02*	*-6.02*	*-6.02*
2,500	-10.85	-11.84	-12.21	-13,35	-7.56	-7.90	-8.19	-8.52	-6.99	-7.10	-7.25	-7.41
	-6.05	*-6.06*	*-6.05*	*-6.04*	*-5.98*	*-5.99*	*-5.99*	*-5.98*	*-6.01*	*-6.00*	*-6.00*	*-6.01*
5,000	-9.29	-9.68	-10.06	-10.79	-7.16	-7.27	-7.46	-7.75	-6.69	-6.75	-6.87	-7.02
	-6.03	*-6.02*	*-6.03*	*-6.02*	*-5.98*	*-5.98*	*-5.98*	*-5.98*	*-5.99*	*-5.99*	*-5.99*	*-5.99*

Table 3: *Robust policy improvement results for 100 randomly generated observation histories of different length n. From left to right, the table documents the average estimated (roman type) and true (italicized) expected total reward for the (s,a)-rectangular and s-rectangular robust value iteration, as well as the semidefinite programming approximation.*

total reward of π^*. However, the (s,a)-rectangular and s-rectangular robust value iteration significantly underestimate the expected total reward of the determined policies. It is interesting to note that for short observation histories, the suboptimality of the s-rectangular robust value iteration and the semidefinite programming approximation stems from the fact that both methods suggest to randomize between the 'repair' and the 'do nothing' actions in state 5. The policies determined by the (s,a)-rectangular robust value iteration, on the other hand, tend to choose the 'repair' action in both state 4 and state 5.

We close with a comparison of the runtimes required by the (s,a)-rectangular and s-rectangular value iterations, and by the semidefinite programming approximations for the policy evaluation and improvement problems. To this end, we employ the refined structural ambiguity set and measure the runtimes over 100 observation histories (95% confidence, $n = 2{,}500$) generated by the historical policy π^0 for variants of the machine replacement problem with 10, 20, ..., 200 states. All results were generated on an Intel Core 2 Duo CPU with 2.80GHz clock speed and 4GB RAM. We solve the (s,a)-rectangular and s-rectangular value iterations with KNITRO 7.0.0 using the GAMS modeling software, and we solve the semidefinite programming approximations with SDPT3 4.0 using the YALMIP modeling package.[2]

Figure 10 (left) shows that for the policy evaluation problem, the (s,a)-rectangular value iteration requires almost twice as much time as its s-rectangular counterpart. This is not surprising as the evaluation of the contraction mapping associated with the (s,a)-rectangular value iteration requires the solution of $S \cdot A$ optimization problems, whereas the contraction mapping ϕ underlying the s-rectangular value iteration is evaluated through the solution of S optimization problems. Both variants of the value iteration require about 50 iterations on average, independent of the problem size. The semidefinite programming approximation requires the least computation time across all problem sizes considered. This is due to the fact that the semidefinite program (20) needs to be solved only once, and the size

[2]More information about software packages can be found at `www.ziena.com/knitro.htm` (KNITRO), `www.gams.com` (GAMS), `www.math.nus.edu.sg/~mattohkc/sdpt3.html` (SDPT3) and `users.isy.liu.se/johanl/yalmip` (YALMIP).

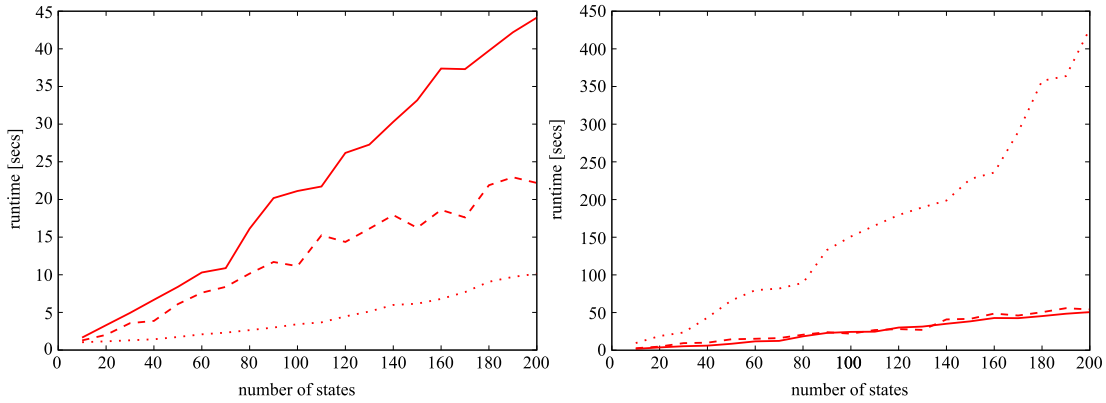

Figure 10: *Scalability of the robust policy evaluation and improvement methods. Shown are the runtimes of the (s,a)-rectangular value iteration (solid line), the s-rectangular value iteration (dashed line) and the semidefinite programming approximation (dotted line) for the policy evaluation problem (left chart) and the policy improvement problem (right chart).*

of (20) grows linearly in the number of states.

Figure 10 (right) shows that for the policy improvement problem, both variants of the robust value iteration require a comparable amount of time. This is due to the fact that the policy evaluation and the policy improvement problems are equally time-consuming for the (s,a)-rectangular value iteration, whereas the contraction mapping φ underlying the s-rectangular value iteration requires the solution of larger optimization problems, see (25). The semidefinite programming approximation, on the other hand, requires significantly higher computation times than both variants of the robust value iteration. This is due to the fact that the sequential convex optimization algorithm from Section 4 requires the solution of one policy evaluation problem and S policy improvement problems per iteration, all of which constitute semidefinite programs. On average, both variants of the robust value iteration require about 50 iterations, whereas the sequential convex optimization algorithm terminates after 3 iterations on average. We remark that the (s,a)-rectangular policy evaluation and improvement problems could be solved more efficiently with the solution techniques presented in [12, 18].

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 optimization 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 optimization 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 a machine replacement problem, and we demonstrated that our approach scales to nontrivial problem sizes.

Acknowledgments

The authors wish to express their gratitude to the referees for their constructive criticism which led to substantial improvements of the paper. Also, financial support from the EPSRC grants EP/H0204554/1 and EP/I014640/1 is gratefully acknowledged.

References

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

[2] R. B. Ash and C. A. Doléans-Dade. *Probability and Measure Theory*. Harcourt Academic Press, 2009.

[3] 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.

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

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

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

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

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

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

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

[11] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58(4):902–917, 2010.

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

[13] 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.

[14] D. Kuhn, W. Wiesemann, and A. Georghiou. Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming*, 130(1):177–209, 2011.

[15] 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.

[16] 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.

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

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

[19] 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.

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

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

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

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

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

[25] 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.

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

[27] 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 Ambiguity Sets

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

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

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

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

The constraints in this problem are equivalent to $w\le\varphi(w)$, see (25). Since φ is a contraction mapping, see Theorem 4.1, non-negativity of p_0 and Theorem 6.2.2 in [20] allow us to reexpress the problem as

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

The max-min expressions in the constraints satisfy the conditions of Corollary 37.3.2 in [21]. Hence, we can interchange the order of the operators in the constraints to obtain the following reformulation.

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

The ambiguity 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.

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

Since p_0 is non-negative, Theorems 6.1.1 and 6.2.2 in [20] 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 [6], which establish asymptotic properties of maximum likelihood estimators of ordinary MCs. To keep the paper self-contained, we summarize these results in Theorem B.1.

Theorem B.1 *Consider a finite MC with state set $\mathcal{X} = \{1, \dots, X\}$ 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 $(x_1, \dots, x_m)$ 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

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

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

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 maximizer 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 34. 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 $p^\theta(\cdot|s,a) = k_{sa} + K_{sa}\theta$. We claim that for any $\theta \in \Theta$, the set $D(\theta)$ equals

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

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 $[k_{sa}\; K_{sa}]^\top_{s'\cdot} \neq 0$ 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 neighborhood of θ that is contained in Θ, and all points θ' in this neighborhood have to satisfy $p^{\theta'}(s'|s,a) \geq 0$. Since $p^\theta(s'|s,a) = 0$, this implies that $[K_{sa}]^\top_{s'\cdot} = 0$, and hence $[k_{sa}]_{s'} = 0$ as well. This contradicts our assumption that $[k_{sa}\; K_{sa}]^\top_{s'\cdot} \neq 0$. 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 $J(\theta) \in \mathbb{R}^{|\mathcal{X}|^2 \times U}$ 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 maximizer 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 maximizer 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

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

where $\psi := \log[p_0(s_1)] + \log\big[\pi^0(a_1|s_1)\big]$. 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 maximizer 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 34. 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 5.5

We first investigate the convergence behavior 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*

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

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 [7].

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 [7].

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 $\pi^0(a|s)$ [7]. 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*

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

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

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

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

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

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

Proof The proof of Theorem 5.3 shows that the unique global maximizer 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 $[k_{sa}\ K_{sa}]^\top_{s'\cdot} \neq 0$. 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 $[K_{sa}]^\top_{s'\cdot}\, x = 0$ 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 5.5.

Proof of Theorem 5.5 Let $\mathbb{B}$ denote the closed unit ball centered 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,

$x^\top \Sigma\, x \geq \nu \|x\|^2$ 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. ∎