Reinforcement learning for discretized Euclidean MDPs

Manfred Jaeger and Kim Guldstrand Larsen

Department of Computer Science, Aalborg University, Denmark

Abstract. Modern model checking tools like UPPAAL Stratego provide a rich framework for modeling cyber-physical systems involving non-determinism as well as time, stochastic and continuous state descriptors. A key objective is to design controllers for such systems that optimize a given objective, e.g., minimizing energy consumption. At an abstract level, the controller design problem can be cast as optimizing a strategy in a continuous (Euclidean) Markov decision process. Partitioning the continuous state space is a simple yet effective strategy to solve this optimization problem in a flexible, non-parametric manner. In previous work we have introduced a reinforcement learning strategy under an undiscounted cost objective on dynamically refined partitions, and we have analyzed at the semantic level approximations of Euclidean MDPs by Imprecise MDPs. In this paper we are extending the approximation analysis to discounted and average cost objectives, and we are moving to close the gap between the theoretical analysis and the practical reinforcement learning approach. We introduce several alternative simulation strategies that on the one hand maintain approximation guarantees as the granularity of the partitioning increases, and on the other hand turns our learning scenario into a standard Q-learning procedure.

1 Introduction

Modern model-checkers (e.g. PRISM [22], STORM [9, 14], MODEST [15], KeYmaeraX [27], UPPAAL Stratego [23, 8]) are increasingly having a focus on integration of state-of-the-art reinforcement learning (RL) and model checking (MC) [25]. In this effort RL is leveraged to efficiently construct near-optimal control strategies while model checking techniques are used to give absolute [7], probabilistic [3, 2] or statistical guarantees [13] of crucial safety properties.

Most importantly, the existing model checkers offer a variety of rich and mature modelling formalisms for defining Markov decision processes (MDPs) that may be used to run simulations for the off-line training of RL policies. Compared to traditional RL scenarios this allows for a number of additional capabilities, e.g. in terms of "targeted sampling" of initial states, rare configurations, etc. The modelling formalisms range from finite state MDPs (STORM, PRISM) to continuous-space (Euclidean) Markov decision processes (EMDPs) (MODEST, UPPAAL Stratego) and Simulink (KeYmaeraX). For continuous-space models abstractions are often used in order to obtain the required guarantees [17, 7].

Most of the above integrations of RL and model checking rely on external components for the RL training (e.g. Open Gym [14] and Simulinks RL Toolbox [17]). In contrast, the tool UPPAAL Stratego offers its own RL method for continuous-space MDPs [20]. Here the learning method is based on a dynamic partition-refinement approach for function approximations providing high flexibility regarding the types of functions that can be approximated, but also closely aligns with continuous-time model-checking techniques, so-called zones.

The RL method of UPPAAL Stratego has already been applied for the construction of near-optimal controllers in a number of industrial applications including traffic-lights [10], water management [11, 12], floor heating systems [24], heat-pumps [16], as well as distributed fleets of autonomous mobile robots [5].

However, the proven practical usefulness of the system is not fully complemented by theoretical guarantees. The RL approach in UPPAAL Stratego is based on sampled runs in the continuous state space. This, and the interleaving with partition-refinement steps, means that classic convergence guarantees for RL [19] in finite state spaces are not directly applicable to this approach. In [21] we have started to develop theoretical underpinnings of the UPPAAL Stratego approach by using imprecise Markov decision processes (IMDPs) to formalize partition-based abstractions, and to approximate EMDPs by standard finite state MDPs that provide upper and lower bounds on the cost function of the EMDP.

This paper extends this work with two main contributions:

- we extend our earlier approximation analysis to average cost objective, which are particularly pertinent for many cyber-physical system applications, yet require a treatment that is very different from discounted and undiscounted cost objectives;
- we show how strategy synthesis for EMDPs can be performed by standard Q-learning on finite state approximations; in particular, we establish for different cost objectives different types of near-optimality guarantees for learned strategies, and we introduce a hierarchy of simulation capabilities of system models, which enable different types of learning and optimization approaches.

Related Work

Numerous works consider abstraction of MDPs over infinite (continuous) state spaces by finite state systems, both in the context of formal verification [26] and reinforcement learning[31, 30]. While the construction of abstractions is quite similar to ours, the underlying purpose, and hence the theoretical analysis, of the abstractions is quite different: in [26] the main concern is the approximation of probabilities of path properties expressed in temporal logic. Moreover, the focus is on bounded time horizons. This is quite different from our focus on minimizing costs over unbounded horizons. The problem of using learning approaches for strategy synthesis is described as an area for future research in [26].

In contrast, [31, 30] are concerned with reinforcement learning in finite state abstractions of MDPs over metric spaces. In [30] this includes an approach for adaptive refinement of the partitions during learning, which is closely related to our approach presented in [20]. A main difference between our work and [31, 30] is a different learning setting reflected in a different objective: [31, 30] focus on minimizing *regret* over a fixed time horizon, i.e., the difference between the expected rewards received during training over a fixed period of time, and the expected rewards under an optimal policy. This is the most appropriate criterion in online learning settings. In our context, we assume that we can learn from unlimited simulation data in an offline setting, and the goal is to learn an optimal strategy for use in later deployment of the system.

2 Euclidean MDPs

We start by introducing our continuous state space system model. The definitions in this and the following section mostly follow [21], but with a few simplifications that only incur a loss of non-essential generality. The following definition is closely related to the *stochastic hybrid systems* of [26], and is a special case of MDPs over Borel spaces[1]. In the following we assume familiarity with some basic concepts of measure theory and Markov processes. The most relevant fundamental concepts are also reviewed in Appendix A.

Definition 1 (Euclidean Markov Decision Processes). *A Euclidean Markov decision process (EMDP) is a tuple $\mathcal{M} = (\mathcal{S}, Act, T, \mathcal{C})$ where:*

- *$\mathcal{S} \subseteq \mathbb{R}^K$ is a compact subset of the K-dimensional Euclidean space equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{S})$.*
- *Act is a finite set of actions,*
- *$T : \mathcal{S} \times Act \times \mathcal{B}(\mathcal{S}) \to [0, 1]$ defines for every $a \in Act$ a transition kernel on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, i.e., $T(s, a, \cdot)$ is a probability distribution on $\mathcal{B}(\mathcal{S})$ for all $s \in \mathcal{S}$, and $T(\cdot, a, B)$ is measurable for all $B \in \mathcal{B}(\mathcal{S})$.*
- *$\mathcal{C} : \mathcal{S} \times Act \to [0, c_{\max}]$ is a cost-function for state-action pairs, such that for all $a \in Act$: $\mathcal{C}(\cdot, a)$ is measurable, and $c_{\max}$ is a global upper bound on costs.*

Example 1. The following toy example describes a moving agent in a square in the 2d-plane. Let $\mathcal{S} = [-1, 1] \times [-1, 1]$, $Act = \{\mathit{right}, \mathit{left}, \mathit{up}, \mathit{down}\}$. We associate with the actions expected transitions of length 0.3 in the corresponding directions, given by vectors $v(\mathit{right}) = (0.3, 0), \ldots, v(\mathit{down}) = (0, -0.3)$. For $(x, y) \in \mathcal{S}$ and $a \in Act$, let $T((x, y), a, \cdot)$ be the distribution on $\mathcal{B}(\mathcal{S})$ defined as $\mathit{clip}(N((x, y) + v(a), 0.003))$, where $N(\ldots)$ is a Gaussian distribution with mean $(x, y) + v(a)$ and diagonal covariance matrix with values 0.003, and $\mathit{clip}((x, y)) := (\max\{-1, \min\{1, x\}\}, \max\{-1, \min\{1, y\}\})$ ensures that the result stays in $\mathcal{S}$. Figure 1 on the left shows for the state $s = (0.8, 0.7)$ (marked by a black $+$) samples of 500 successor states according to $T(s, a, \cdot)$ for each $a \in Act$. The cost function

$$\mathcal{C}((x, y), \mathit{right}) = 2 * (x^2 + y^2) + 0.6(1 - x) \tag{1}$$

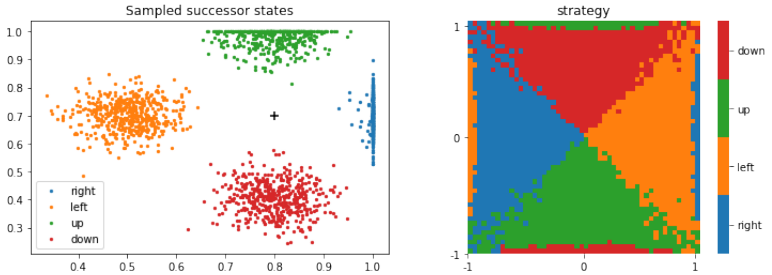

Fig. 1: Sampled successor states and strategy in Example 1 and Example 2

consists of two elements: a cost proportional to the squared Euclidean distance to the origin, and a cost that is inversely proportional to the effectiveness of the action *right*: at $x = 1$ it is impossible to move to the right, and the cost component $0.6(1-x)$ here is zero. The cost then linearly increases for decreasing x. The costs for actions *left*,*up*,*down* are defined analogously with the same $2 * (x^2 + y^2)$ term, but with $(1 - x)$ replaced by $(1 + x), (1 - y)$ and $(1 + y)$, respectively.

We are mostly concerned with EMDPs that satisfy the following continuity conditions. In this definition we denote with d_{tv} the total variation distance between distributions.

Definition 2 (Continuous EMDP). *A Euclidean MDP $\mathcal{M}$ is* continuous *if*

- *For each $\epsilon > 0$ there exists $\delta > 0$, such that for all $s, s' \in \mathcal{S}$, $a \in Act$: $\parallel s - s' \parallel < \delta \Rightarrow d_{\mathrm{tv}}(T(s, a, \cdot), T(s', a, \cdot)) \leq \epsilon$.*
- *$\mathcal{C}(\cdot, a)$ is continuous on $\mathcal{S}$ for all $a \in Act$.*

The EMDP of Example 1 is continuous. A *run* π of an MDP is a sequence of alternating states and actions $s_0 a_0 s_1 a_1 s_2 a_2 \ldots$. A *state-run* is a sequence of states $s_0, s_1, \ldots$. An initial segment $s_0 a_0 \ldots s_t a_t$ of a run is denoted $\pi_{0:t}$.

Definition 3 (Strategy). *A (memoryless, stationary, deterministic) strategy for an MDP $\mathcal{M}$ is a function $\sigma : \mathcal{S} \rightarrow Act$, mapping states to actions, such that for every $a \in Act$ the set $\{s \in \mathcal{S} | \sigma(s) = a\}$ is measurable.*

A fixed strategy σ turns an EMDP into a Markov process whose transition kernels we denote by $T_\sigma(s, \cdot)$. Together with an initial state $s_0 = s$ the transition kernel defines a probability distribution $P_{s,\sigma}$ over state runs (see Appendix A for details). The strategy being deterministic, a state-run induces a unique run. Depending on context, we therefore also view $P_{s,\sigma}$ as a distribution over runs. Finally, for $k \geq 1$, we denote by T_σ^k the transition kernel representing k successive transitions according to T_σ, and by $P_{s,\sigma}^k$ the distribution of the k'th state in a run.

Definition 4 ((Expected) Cost). *We consider four different types of expected costs. The first three types are defined by first defining the* undiscounted, discounted *and* path-average *costs of individual runs* π*:*

$$\mathcal{C}_u(\pi) = \sum_{i\geq 0} \mathcal{C}(s_i, a_i), \tag{2}$$

$$\mathcal{C}_\lambda(\pi) = \sum_{i\geq 0} \lambda^i \mathcal{C}(s_i, a_i) \ (\lambda \in (0,1)), \tag{3}$$

$$\mathcal{C}_{p-avg}(\pi) = \limsup_{N\to\infty} 1/N \sum_{i=0}^{N} \mathcal{C}(s_i, a_i). \tag{4}$$

Given a strategy σ *and initial state* s*, the* expected cost $\mathbb{E}_{s,\sigma}[\mathcal{C}]$ *then is the expectation of the cost of* π *under the distribution* $P_{s,\sigma}$*. A common alternative definition of expected average cost is given by*

$$\mathbb{E}_{s,\sigma}[\mathcal{C}_{e-avg}] := \limsup_{N\to\infty} \mathbb{E}_{s,\sigma}[1/N \sum_{i=0}^{N} \mathcal{C}(s_i, a_i)]. \tag{5}$$

For any type of cost $\mathcal{C}$*, when* $\min_\sigma \mathbb{E}_{s,\sigma}[\mathcal{C}]$ *exists, then this is referred to as the expected cost of* s*, denoted* $\mathbb{E}_s[\mathcal{C}]$*.*

Example 2. (Example 1 continued). Figure 1 on the right shows a strategy for the EMDP of Example 1. This strategy was learned using the partitioning approach described in Section 3, and is constant on small (measurable) squares that partition the state space $\mathcal{S}$ (see Example 6 below for more details). The strategy is optimized for $\mathcal{C}_\lambda$ with $\lambda = 0.6$. It consists of trying to move into the middle of the state space, except for positions very close to the boundaries, where it is preferable to minimize the short-term action-related costs over the long-term benefit of minimizing the state-related cost in the future.

Of the two versions of expected average cost the more commonly used version is $\mathcal{C}_{e-avg}$ (e.g. used in [28]). The following example illustrates possible discrepancies between the two versions, and motivates our preference for the path-average version. However, as we shall see below, in important classes of models that we will focus on in the sequel, the two versions actually coincide. For ease of exposition, the example is based on a MDP with an infinite countable state space, not an EMDP in our sense.

Example 3. Let $\mathcal{S} = \mathbb{Z}$, $Act = \{z, m\}$. Transition probabilities are defined as follows: $T(0, z, 0) = 1$, $T(0, m, 1) = T(0, m, -1) = 0.5$. For all states other than 0, both actions z and m have the same, deterministic effect: for $i \geq 1$: $T(i, z, i+1) = T(i, m, i+1) = 1$; for $i \leq -1$: $T(i, z, i-1) = T(i, m, i-1) = 1$. The cost is defined as follows: $\mathcal{C}(0, z) = 1$, $\mathcal{C}(0, m) = 0$. For $i \geq 1$ the cost is independent of the action: $\mathcal{C}(i) = 10$ if $2^k \leq i < 2^{k+1}$ with k odd, and $\mathcal{C}(i) = -10$ if $2^k \leq i < 2^{k+1}$ with k even. Finally, for $i \leq -1$: $\mathcal{C}(i) = -\mathcal{C}(-i)$.

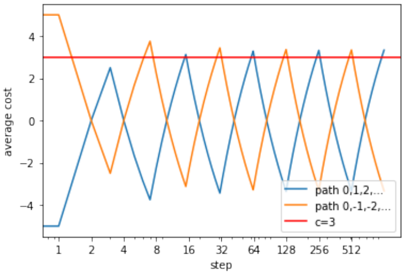

Fig. 2: Cost paths under the m strategy in Example 3.

If one takes action m at state 0, then with probability 0.5 the resulting state path is $0, 1, 2, 3, \dots$ (regardless of which actions are taken from the second step onwards), and with probability 0.5 it is $0, -1, -2, -3, \dots$. Figure 2 shows the development of the average cost $1/N \sum_{i=0}^{N} \mathcal{C}(s_i)$ along these two paths. At $N = 2^{k+1} - 1$ with k odd, then one easily obtains a (crude) lower bound of 3 (marked by the horizontal red line in Figure 2) for the average cost on the $0, 1, 2, \dots$ path, and similarly at even k for the $0, -1, -2, \dots$ path. Thus, with probability 1, $\limsup_N \mathcal{C}_{p-avg}(\pi) > 3$, and hence $\mathbb{E}_{0,m}[\mathcal{C}_{p-avg}] \geq 3$, where m denotes any policy that selects action m at state 0. On the other hand, by the symmetries of the cost and transition probability definitions: $\mathbb{E}_{0,m}[1/N \sum_{i=0}^{N} \mathcal{C}(s_i)] = 0$ for all N, and therefore $\limsup_N \mathbb{E}_{0,m}[\sum_{i=0}^{N} \mathcal{C}(s_i, a_i)] = 0$. For any policy that takes action z at 0, on the other hand, both versions of average cost coincide, and yield a cost of 1. Thus, under $\mathcal{C}_{p-avg}$ the strategy is preferred that takes action z at 0, whereas under $\mathcal{C}_{e-avg}$ action m is preferred.

We would argue that $\mathcal{C}_{e-avg}$ is sensible when one is interested in optimizing a strategy that is to be implemented for multiple, independently running agents or systems, whose costs are in some sense shared or pooled at all times (modeled by the inner expectation in the average cost definition). If, on the other hand, we are thinking of a single agent or system that needs to minimize cost over the single run that it will execute, then the expectation can only be over the different average costs incurred by different infinite runs, i.e., modeled by the outer expectation of $\mathcal{C}_{p-avg}$.

In the preceding example the strategy m led to a process in which all states are transient. In the context of average cost objectives, one typically requires assumptions about recurrence that ensure, at the very least, that average cost optimal strategies exist [28, Section 8.3], [18, 1, 32]. In the case of continuous state spaces, these assumptions are expressed in terms of mixing properties of the transition kernels, which take the place of irreducibility and aperiodicity assumptions for discrete state Markov chains. In the following, we introduce a version of such assumptions for our EMDPs that follows the terminology and presentation of [4].

Definition 5 (Small EMDP). *An EMDP is* small *if there exist a measure ξ on $\mathcal{S}$ that is not identically 0, such that for all $s \in \mathcal{S}$ and all strategies σ: $T_\sigma(s, \cdot) \geq \xi(\cdot)$.*

Example 4. The EMDP of Example 1 is small: we can define ξ by the density function for $(x, y) \in \mathcal{S}$:

$$f_\xi(x, y) = min_{(x',y') \in \mathcal{S}, a \in Act} N((x', y') + v(a), 0.003)(x, y),$$

where on the right we now identify $N(\ldots)$ with its density function. Due to compactness of $\mathcal{S}$, the minimum on the right exists, and is > 0 for all (x, y).

In the preceding example 'smallness' was obtained by the ability to transition in a single step from any point of the state space to all parts of the state space (even though with potentially very small probability). Another way to achieve the small property is to allow the system at every step to "re-initialize" with a small probability $q > 0$ to a designated state s_0. Then definition 5 is satisfied with ξ the pointmass of q on s_0.

The following is an adaptation to our special context of general fundamental results in the ergodic theory of Markov processes.

Theorem 1. *Let $\mathcal{M}$ be a small EMDP, and σ any strategy. Then the transition kernel T_σ has a unique stationary distribution $\tilde{P}_\sigma$. Denoting expectations w.r.t. $\tilde{P}_\sigma$ by $\tilde{\mathbb{E}}_\sigma$, then for all $s \in \mathcal{S}$:*

$$\lim_{k \to \infty} d_{\mathrm{tv}}(P^k_{s,\sigma}, \tilde{P}_\sigma) = 0, \tag{6}$$

$$\lim_{k \to \infty} |\mathbb{E}_{s,\sigma}[\mathcal{C}(s_k, \sigma(s_k))] - \tilde{\mathbb{E}}_\sigma[\mathcal{C}(\cdot, \sigma(\cdot))]| = 0. \tag{7}$$

and

$$P_{s,\sigma}(\lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \mathcal{C}(s_k, \sigma(s_k)) = \tilde{\mathbb{E}}_\sigma[\mathcal{C}(\cdot, \sigma(\cdot))]) = 1. \tag{8}$$

Proof: In a small EMDP in our sense, $\mathcal{S}$ is a recurrent petite set in the sense of [4, Section 6.1, Section 7.2.1]. Our theorem then follows from Proposition 7.11,Theorem 7.8, and Theorem 8.7 of [4]. □

3 Approximations Induced by Partitions

We approximate an EMDP by a finite state MDP whose states are subsets of $\mathcal{S}$: let $\mathcal{A} = \{\nu_1, \ldots, \nu_{|\mathcal{A}|}\} \subset 2^{\mathcal{S}}$ be a finite partition of $\mathcal{S}$. We call an element $\nu \in \mathcal{A}$ a *region* and shall assume that each such ν is Borel measurable. For $s \in \mathcal{S}$ we denote by $[s]_\mathcal{A}$ the unique region $\nu \in \mathcal{A}$ such that $s \in \nu$. We say that a partition $\mathcal{B}$ *refines* a partition $\mathcal{A}$ if for every $\nu \in \mathcal{B}$ there exists $\mu \in \mathcal{A}$ with $\nu \subseteq \mu$. We write $\mathcal{A} \sqsubseteq \mathcal{B}$ in this case. Given $\mathcal{A}$, we define a standard MDP by means of an *adversary*:

Definition 6 (Adversary, Induced MDP). *Let $\mathcal{M}$ be an EMDP with action set Act, $\mathcal{A}$ a partition of its state space. An adversary α is a mapping that assigns to pairs $(\nu, a) \in \mathcal{A} \times Act$ a probability distributions $\alpha(\nu, a)$ over ν. $A[\mathcal{A}]$ denotes the space of possible adversaries.*

The finite state MDP $\mathcal{M}_\alpha^{\mathcal{A}} = (\mathcal{A}, Act, T_\alpha^{\mathcal{A}}, \mathcal{C}_\alpha^{\mathcal{A}})$ then is defined by the transition probabilities

$$T_\alpha^{\mathcal{A}}(\nu, a, \nu') = \int_\nu T(s, a, \nu') d\alpha(\nu, a)(s) \tag{9}$$

and costs

$$\mathcal{C}_\alpha^{\mathcal{A}}(\nu, a) = \int_\nu \mathcal{C}(s, a) d\alpha(\nu, a)(s). \tag{10}$$

We denote with $\mathit{\Pi}[\mathcal{S}]$ and $\mathit{\Pi}[\mathcal{A}]$ the space of possible strategies on the EMDP $\mathcal{M}$, and on the induced $\mathcal{M}_\cdot^{\mathcal{A}}$, respectively. For $\mathcal{M}_\alpha^{\mathcal{A}}$, $\sigma \in \mathit{\Pi}[\mathcal{A}]$ and $\nu \in \mathcal{A}$ then expected costs $\mathbb{E}_{\nu,\sigma,\alpha}^{\mathcal{A}}[\mathcal{C}]$ and $\mathbb{E}_{\nu,\alpha}^{\mathcal{A}}[\mathcal{C}]$ (for optimal σ) are defined as before for all cost types. Our main question is how well $\mathbb{E}_{[s]_\mathcal{A},\alpha}^{\mathcal{A}}[\mathcal{C}]$ is guaranteed to approximate $\mathbb{E}_s[\mathcal{C}]$. Our first result relates expected costs for partitions of different granularities.

Theorem 2. *Let $\mathcal{M}$ be a continuous EMDP and $\mathcal{C} \in \{\mathcal{C}_u, \mathcal{C}_\lambda\}$. Let $\mathcal{A} \sqsubseteq \mathcal{B}$ be partitions. For all $\alpha \in A[\mathcal{B}]$ there exist $\alpha^-, \alpha^+ \in A[\mathcal{A}]$, such that*

$$\mathbb{E}_{[s_0]_\mathcal{A},\alpha^-}^{\mathcal{A}}[\mathcal{C}] \le \mathbb{E}_{[s_0]_\mathcal{B},\alpha}^{\mathcal{B}}[\mathcal{C}] \le \mathbb{E}_{[s_0]_\mathcal{A},\alpha^+}^{\mathcal{A}}[\mathcal{C}] \tag{11}$$

If $\mathcal{M}$ is small, then the same holds for $\mathcal{C} \in \{\mathcal{C}_{p-avg}, \mathcal{C}_{e-avg}\}$.

The proof of this and the following theorems is given in Appendix B. For $\mathcal{C} = \mathcal{C}_u$ this result was essentially already given in [21]. However, there a slightly more general class of adversaries was considered, so that also the $\mathcal{C} = \mathcal{C}_u$ case needs to be addressed again in the proof of this theorem. The same applies to our next theorem.

Theorem 3. *Let $\mathcal{M}$ be a continuous EMDP, $\mathcal{A}$ a partition of $\mathcal{S}$. Let $\mathcal{C} \in \{\mathcal{C}_u, \mathcal{C}_\lambda\}$. Then there exist adversaries α^-, α^+, such that for all $s_0 \in \mathcal{S}$:*

$$\mathbb{E}_{[s_0]_\mathcal{A},\alpha^-}^{\mathcal{A}}[\mathcal{C}] \le \mathbb{E}_{s_0}^{\mathcal{M}}[\mathcal{C}] \le \mathbb{E}_{[s_0]_\mathcal{A},\alpha^+}^{\mathcal{A}}[\mathcal{C}]. \tag{12}$$

If $\mathcal{M}$ is small, then (12) also holds for $\mathcal{C} \in \{\mathcal{C}_{p-avg}, \mathcal{C}_{e-avg}\}$.

A question of key interest is whether the bounds (12) become arbitrarily tight (for all possible α^-, α^+) if $\mathcal{A}$ is a sufficiently fine partition. The following definition provides the precise formulation of this question.

Definition 7. *A sequence $\mathcal{A}_0 \sqsubseteq \mathcal{A}_1 \sqsubseteq \cdots \sqsubseteq \mathcal{A}_k \sqsubseteq \cdots$ is called* refining *if $\lim_{k\to\infty} \max_{\nu\in\mathcal{A}_k} \sup_{s,s'\in\nu} ||s - s'|| = 0$. A refining sequence* approximates *$\mathbb{E}_s^{\mathcal{M}}[\mathcal{C}]$ if*

$$\lim_{k\to\infty} \Big(\sup_{\alpha\in A[\mathcal{A}_k]} \mathbb{E}_{[s]_{\mathcal{A}_k},\alpha}^{\mathcal{A}_k}[\mathcal{C}] - \inf_{\alpha\in A[\mathcal{A}_k]} \mathbb{E}_{[s]_{\mathcal{A}_k},\alpha}^{\mathcal{A}_k}[\mathcal{C}]\Big) = 0. \tag{13}$$

In [21] it was conjectured that for all continuous EMDPs, all refining sequences approximate $\mathbb{E}_s^{\mathcal{M}}[\mathcal{C}_u]$. This conjecture turns out to be false, however. The following example describes a continuous (but not small!) EMDP that gives a counterexample to the conjecture both for $\mathcal{C}_u$ and $\mathcal{C}_{p-avg}$.

Example 5. Let $\mathcal{S} = [-1/2, 3/2]$. Let Act contain only a single action. Thus, choice of actions and strategies are vacuous, and we therefore drop the action argument in all notation. The transition probabilities $T(x, \cdot)$ are defined for all $x \in \mathcal{S}$ by uniform distribution over an interval of length $1/2$. For $x \in (0.5, 1)$ this interval is centered on a point to the right of x, and for $x \in (0, 0.5)$ on a point to the left of x. Precisely, using $U_{[a,b]}$ to denote the uniform distribution on the interval $[a, b]$, we define for $x \in (0, 1)$:

$$T(x, \cdot) = U_{[\frac{3}{2}x - \frac{1}{2}, \frac{3}{2}x]}.$$

For all $x \in [1, 3/2]$ define $T(x, \cdot) = U_{[1,3/2]}$, and for $x \in [-1/2, 0]$ define $T(x, \cdot) = U_{[-1/2,0]}$. Thus, the transition model represents a random walk with a drift towards the right for $x > 0.5$, a drift to the left for $x < 0.5$, and the intervals $[-1/2, 0]$ and $[1, 3/2]$ as absorbing sets. Let the cost be defined as

$$\mathcal{C}(x) = \begin{cases} 0 & x \leq 0 \\ x & x \in (0, 1) \\ 1 & x \geq 1 \end{cases}$$

This EMDP is continuous. It is not small, because e.g. the transition kernels $T(0, \cdot)$ and $T(1, \cdot)$ assign probability 1 to the disjoint sets $[-1/2, 0]$ and $[1, 3/2]$, respectively. With probability 1, runs π will eventually reach one of the absorbing intervals. $\mathcal{C}_{p-avg}(\pi) = 1$ if π is absorbed in $[1, 3/2]$, and $\mathcal{C}_{p-avg}(\pi) = 0$ if π is absorbed in $[-1/2, 0]$. Thus, for any $x \in \mathcal{S}$, $\mathbb{E}_x[\mathcal{C}_{p-avg}]$ is equal to the probability that a process started at x is absorbed in $[1, 3/2]$, which is nonzero for all $x > 0$.

Let $\mathcal{A}$ be any partition of $\mathcal{S}$ formed by half-open intervals $\nu = [l, u[$ such that both 0 and 1 are interior points of their respective partition elements. Let α be the adversary that assigns probability 1 to the left boundary point l of ν. In the resulting MDP $\mathcal{M}_\alpha^{\mathcal{A}}$ then the set A of partition elements ν with non-empty intersections with $[-1/2, 0]$ is an absorbing set with zero cost. From all other partition elements there is a non-zero probability of reaching A. Note that for ν contained in $[1, 3/2]$ there is a non-zero probability to transition to the element containing 1 as an interior point. The left boundary point of this element is < 1, and therefore in $\mathcal{M}_\alpha^{\mathcal{A}}$ it is possible to "break out" of $[1, 3/2]$, in contrast to the underlying EMDP. By standard Markov chain theory, thus, with probability 1 a run of $\mathcal{M}_\alpha^{\mathcal{A}}$ will end up in A, and therefore $\mathbb{E}_{\nu,\alpha}^{\mathcal{A}}[\mathcal{C}_{p-avg}] = 0$ for all ν.

Now consider the undiscounted cost. Here, in the original EMDP we have $\mathbb{E}_x[\mathcal{C}_u] = 0$ for all $x \in [-1/2, 0]$. Now let α be the adversary defined by the uniform distribution over ν. For any ν contained in $[-1/2, 0]$ there is a nonzero probability of transitioning to the region ν_0 containing 0, and similarly as above, to "break out" of the zero cost component. With probability 1 this will happen infinitely often in a run of $\mathcal{M}_\alpha^{\mathcal{A}}$, and so $\mathbb{E}_{\nu,\alpha}^{\mathcal{A}}[\mathcal{C}_u] = \infty$ for all $\nu \in \mathcal{A}$.

For undiscounted cost, and average costs for small EMDPs, on the other hand, we obtain the desired result:

Theorem 4. *Let $\mathcal{M}$ be a continuous EMDP, and $\lambda < 1$. Any refining sequence approximates $\mathbb{E}_s^{\mathcal{M}}[\mathcal{C}_\lambda]$. If $\mathcal{M}$ is small, then any refining sequence approximates $\mathbb{E}_s^{\mathcal{M}}[\mathcal{C}_{p-avg}]$ and $\mathbb{E}_s^{\mathcal{M}}[\mathcal{C}_{e-avg}]$.*

4 Learning in $\mathcal{M}_\alpha^\mathcal{A}$

We now turn to learning strategies for the standard finite state MDPs $\mathcal{M}_\alpha^\mathcal{A}$. When $\mathbb{E}_{[s_0]_\mathcal{A},\alpha}^\mathcal{A}[\mathcal{C}]$ is close to $\mathbb{E}_{s_0}^{\mathcal{M}}[\mathcal{C}]$, then optimality guarantees for Q-learning [19, 32] ensure that learning in $\mathcal{M}_\alpha^\mathcal{A}$ yields a near-optimal strategy for the underlying EMDP. In the cases covered by Theorem 4 we obtain these guarantees, albeit with the caveat that the theorem does not give a constructive rule for how fine the partition $\mathcal{A}$ needs to be in order to support a desired approximation error bound.

4.1 Q-learning

We first briefly review Q-learning under the $\mathcal{C}_\lambda$ objective, and introduce notation and terminology we need in the sequel. For average cost objectives similar Q-learning approaches exist [32], and most of the discussion and results of this section carries over to learning under an average cost objective. A detailed analysis of this is outside the scope of this paper, however.

Q-learning aims to learn for each region-action pair (ν, a) the expected cost of performing action a in state ν, and following an optimal strategy thereafter. The Q-values thus defined are initialized as $Q_0(\nu, a) = 0$ for all ν, a. Based on an observed run $\pi = \nu_0 a_0 \nu_1 a_1 \ldots \nu_t a_t \nu_{t+1} \ldots$ the Q-values are iteratively updated as

$$Q_{t+1}(\nu_t, a_t) = (1 - \beta_t) Q_t(\nu_t, a_t) + \beta_t(\mathcal{C}(\nu_t, a_t) + \lambda \min_{a \in Act} Q_t(\nu_{t+1}, a)), \tag{14}$$

where $\beta_t \in (0, 1)$ is the *learning rate* at iteration t. In order to ensure convergence with high probability, β_t is defined as a decreasing function in the number of times the pair (ν_t, a_t) has been updated at previous time points $s < t$ [19]. Thus, β_t is a function of $\pi_{0:t}$. Moreover, during learning, actions a_t are typically not selected according to a fixed, stationary strategy, but according to a strategy that also is designed to *explore* new actions, and, in online learning scenarios, to already *exploit* actions that promise low costs according to the current Q-values. Since according to (14) the current Q-values are a function of $\pi_{0:t}$, this means that both aspects are covered by defining action selection based on a history-dependent strategy $a_t = \sigma_t(\pi_{0:t})$. We write $\boldsymbol{\sigma} = (\sigma_t)_t$ for such a history dependent strategy. The data may also consist of multiple runs (a.k.a. *episodes*) starting at the same or at different initial state. For notational simplicity we here take the data to consist of a single long sequence indexed by t, even though in our examples and experiments data usually consists of multiple episodes.

4.2 Adversary and Sampling Design

Ideally, given an EMDP $\mathcal{M}$ with a partition $\mathcal{A}$ one would define adversaries α^-, α^+ satisfying (12), such that simulating $\mathcal{M}^{\mathcal{A}}_{\alpha^-}, \mathcal{M}^{\mathcal{A}}_{\alpha^+}$ is computationally feasible. Then, the learned cost functions $\mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\alpha^-}[\mathcal{C}], \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\alpha^+}[\mathcal{C}]$ will provide guaranteed lower and upper bound on the true cost in $\mathcal{M}$. Note that even when Theorem 4 does not apply, a good choice of α^-, α^+ will lead to tight bounds.

The construction of α^-, α^+ in the proof of Theorem 3 is based on the true distribution $P^{\mathcal{M}}$ and expected costs $\mathbb{E}^{\mathcal{M}}_{s,\sigma}$ in $\mathcal{M}$, which are not known in practice. We therefore have to design α^-, α^+ that are likely to provide valid lower and upper bounds, and that are amenable for use in simulations. With regard to the latter issue, we now introduce classes of adversaries C1,C2,C3 essentially defined by the tractability of their defining integrals (9), (10). We always assume that we can compute costs $\mathcal{C}(s,a)$, and sample transitions according to $T(s,a,\cdot)$ in the underlying EMDP.

C1 For all ν, a, one can sample states $s \in \nu$ according to $\alpha(\nu, a)$.
C2 In addition to C1, the cost values (10) can be computed for all ν, a.
C3 In addition to C2, the transition probabilities (9) can be computed for all ν, a, ν'.

For MDPs $\mathcal{M}^{\mathcal{A}}_{\alpha}$ defined by C3 adversaries, one can, in principle, compute standard matrix representations of the transition probabilities and cost function, and apply all standard tools for solving finite state MDPs. Class C2 is important because it is sufficient to support Q-learning: we can sample runs according to the distribution defined by $\mathcal{M}^{\mathcal{A}}_{\boldsymbol{\sigma},\alpha}$ for any (possibly non-stationary) strategy $\boldsymbol{\sigma}$: given a current state-action pair (ν, a), we sample a successor state ν' by randomly sampling $s \in \nu$ according to α, then sampling $s' \in \mathcal{S}$ according to $T(s,a,\cdot)$, and finally setting $\nu' := [s']_{\mathcal{A}}$. If, according to C2, we can at each step also compute the cost value $\mathcal{C}^{\mathcal{A}}_{\alpha}(\nu, a)$, then we obtain exactly the data needed for Q-learning.

A simple type of adversaries is defined by *representative points*: $\alpha(\nu, a)$ is defined as a pointmass on one designated element $s \in \nu$. Then α belongs at least to C2. The adversaries of Example 5 were of this type. In this example even C3 holds, due to the simple transition model defined by uniform distributions in the underlying $\mathcal{M}$. Natural candidates for representative points are those that minimize or maximize the cost for an action in a region (as in Example 5), and that thereby are good candidates for defining α^-, α^+ that satisfy (12). Another canonical construction is to let $\alpha(\nu, a)$ be the uniform distribution on ν (for all a). We refer to this as the *mean adversary*, denoted α^{mean}.

Example 6. (Example 2 continued). Similar to Example 5 we consider partitions $\mathcal{A}_i$ in the form of uniform grids with region dimensions $1/i \times 1/i$. For any region ν and all $a \in \mathit{Act}$ it is easy to identify the states $s \in \nu$ that minimize or maximize the cost $\mathcal{C}(s,a)$. We denote the resulting representative state adversaries as α^-, α^+. These adversaries then are of type C2. The transition probabilities $T^{\mathcal{A}}_{\alpha^-}, T^{\mathcal{A}}_{\alpha^+}$ require integrals over Gaussian densities, for which no closed-form solutions (but good numerical approximations) exist. Thus, C3 does not hold in a strict sense, but is satisfied up to numerical approximations.

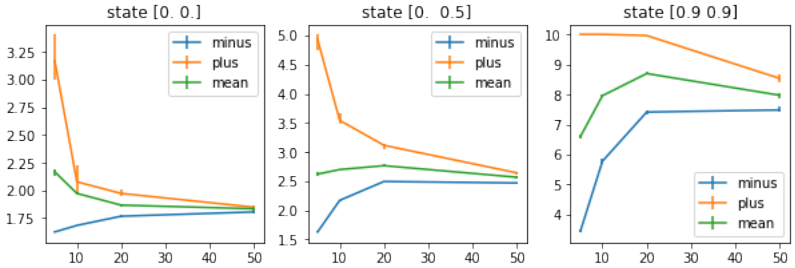

Fig. 3: Cost values (y-axis) at selected states learned from partitions of granularities 5,10,20,50, (x-axis) and the $\alpha^-, \alpha^+, \alpha^{mean}$ adversaries

Considering the α^{mean} adversary, we find that this, too, satisfies **C2**: clearly we can sample states uniformly in a grid cell ν. Also, due to the simple polynomial cost function (1), the integrals defining $\mathcal{C}^{\mathcal{A}}_{\alpha^{mean}}$ can be easily computed.

Figure 3 shows for three selected states in $\mathcal{S}$ the learned discounted cost values ($\lambda = 0.6$) of the regions $[s]_{\mathcal{A}_i}$ for $i = 5, 10, 20, 50$ under the $\alpha^-, \alpha^+, \alpha^{mean}$ adversaries. For learning, we use a strategy $\boldsymbol{\sigma}$ that always selects the next action uniformly at random, and where the learning rate β_t is $1/\sqrt{n}$, with n the number of times (ν_t, a_t) has already been updated. Learning was repeated 10 times. The curves show the average cost values of the 10 learning runs, and the error bars indicate the absolute minimum and maximum values obtained in the 10 runs. Notice that there is very little variation from granularity $i = 20$ onwards.

Figure 1 on the right shows the strategy learned for $\mathcal{M}^{\mathcal{A}_{50}}_{\alpha^-}$ (the strategies learned for the α^+ and α^{mean} adversaries look very similar).

4.3 Learning with C1 adversaries

While satisfied in Example 6, even C2 can easily be out of reach. We now consider an approach to approximate Q-learning for $\mathcal{M}^{\mathcal{A}}_{\alpha}$ when only C1 is true. For any adversary α with C1, we can approximate simulations of $\mathcal{M}^{\mathcal{A}}_{\alpha}$ in a Q-learning scenario with a non-stationary strategy $\boldsymbol{\sigma}$ as follows:

- Given the current history $\pi_{0:t}$ and selected action $a_t = \sigma_t(\pi_{0:t})$:
  - sample a state $s \in \nu_t$ according to $\alpha(\nu_t, a_t)$
  - return the cost value $\mathcal{C}_t = \mathcal{C}(s, a_t)$, and sample the next state ν' according to $T(s, a, \nu')_{\nu' \in \mathcal{A}}$.

This simulation generates runs $\nu_0 a_0 \nu_1 a_1 \ldots$ according to the distribution $P_{\nu_0, \boldsymbol{\sigma}}$ defined by $\mathcal{M}^{\mathcal{A}}_{\alpha}$, $\boldsymbol{\sigma}$, and initial state ν_0. The simulations differ from exact simulations of $\mathcal{M}^{\mathcal{A}}_{\alpha}$ (or any MDP) in that the observed cost $\mathcal{C}_t$ at step t no longer is a function of the state-action pair (ν_t, a_t). However, one can still perform the Q-learning updates (14) with $\mathcal{C}_t$ instead of $\mathcal{C}(\nu_t, a_t)$. We denote the function defined by these updates at time t as $\tilde{Q}_t$. We now show that in expectation, we obtain the same results as with standard Q-learning from proper simulations of

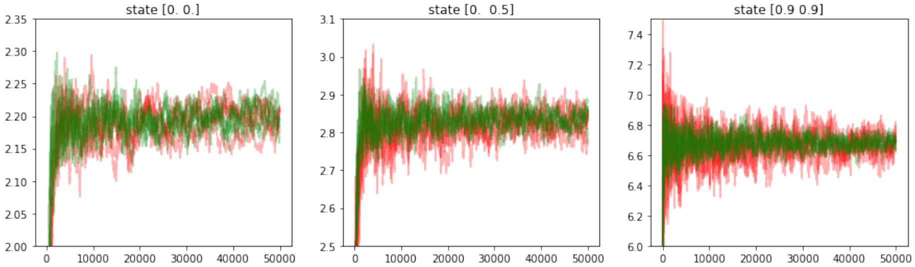

Fig. 4: Learned Q (green) vs. $\tilde{Q}$-values (red) for selected regions

$\mathcal{M}^{\mathcal{A}}_{\alpha}$. Before stating this result, we have to reconsider the strategy $\boldsymbol{\sigma}$ used in learning. We have said earlier that σ_t only depends on $\pi_{0:t}$, even if σ_t is defined in terms of the current Q-values. This is no longer the case for $\tilde{Q}$, where the cost values $\mathcal{C}_t$, and hence the current $\tilde{Q}$-values, are no longer fully determined by $\pi_{0:t}$. We therefore have to restrict the following theorem explicitly to training strategies that are a function only of $\pi_{0:t}$. This is not a serious limitation, since the dependence on current Q-values is mostly desirable in online training scenarios.

In the following we assume a single fixed adversary α. To reduce clutter in the notation, we omit the α index from expectations and probabilities.

Theorem 5. *For all $(\nu, a) \in \mathcal{A} \times Act$, $\nu_0 \in \mathcal{A}$, strategies $\boldsymbol{\sigma}$ such that σ_t is a function of $\pi_{0:t}$, and $t \geq 0$:*

$$\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\tilde{Q}_t(\nu, a)) = \mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(Q_t(\nu, a)). \tag{15}$$

Example 7. (Example 6 continued). We now fix the granularity at $i = 5$, and consider the α^{mean} adversary. We compare the Q-values learned during proper Q-learning with the $\tilde{Q}$ values obtained during our approximation of the Q learning process. For both exact and approximate Q-learning we perform 5 learning runs. Each learning run consists of 50000 episodes of length 10. We record the Q and $\tilde{Q}$ values at the end of each episode.

Figure 4 shows for the regions containing the points $(0, 0)$, $(0, 0.5)$ and $(0.9, 0.9)$ the developments of the Q (green) and $\tilde{Q}$ (red) values over the course of the 50000 episodes. One can see that in expectation the learned Q and $\tilde{Q}$ coincide, but that the $\tilde{Q}$ values exhibit a larger variance (especially for the $(0.9, 0.9)$ region).

5 Conclusion

We have developed a general approach to approximate an EMDP by standard finite state MDPs defined by partitions of the continuous state space. We have shown that under suitable conditions on the EMDP, this approximation becomes more and more precise for discounted and average cost objectives as the granularity of the partition increases. We have further developed conditions on the adversaries that allow us to effectively sample system runs on which standard Q-learning methods with their known convergence guarantees can be applied. In case where our system does not allow to implement adversaries with these properties, we find that under much weaker (usually satisfied) conditions, we can still simulate system runs, such that applying Q-learning on these runs yields in expectation the same results we would get if the stronger conditions were satisfied.

In this paper we have focused on the static scenario where the partition is fixed during learning. A question for future work is how to best interleave learning steps on a given partition with refinement steps of the partition, so that an overall convergence to the cost function of the EMDP is guaranteed.

6 Acknowledgment

This research has been partially funded by the Villum Investigator Grant S4OS (37819) from Villum Foundation.

Bibliography

[1] A. Arapostathis, V. S. Borkar, E. Fernández-Gaucherand, M. K. Ghosh, and S. I. Marcus. Discrete-time controlled markov processes with average cost criterion: A survey. *SIAM Journal on Control and Optimization*, 31 (2):282–344, 1993.

[2] E. Bacci and D. Parker. Probabilistic guarantees for safe deep reinforcement learning. In N. Bertrand and N. Jansen, editors, *Formal Modeling and Analysis of Timed Systems - 18th International Conference, FORMATS 2020, Vienna, Austria, September 1-3, 2020, Proceedings*, volume 12288 of *Lecture Notes in Computer Science*, pages 231–248. Springer, 2020. doi: 10.1007/978-3-030-57628-8_14. URL `https://doi.org/10.1007/978-3-030-57628-8_14`.

[3] E. Bacci and D. Parker. Verified probabilistic policies for deep reinforcement learning. In J. V. Deshmukh, K. Havelund, and I. Perez, editors, *NASA Formal Methods - 14th International Symposium, NFM 2022, Pasadena, CA, USA, May 24-27, 2022, Proceedings*, volume 13260 of *Lecture Notes in Computer Science*, pages 193–212. Springer, 2022. doi: 10.1007/978-3-031-06773-0_10. URL `https://doi.org/10.1007/978-3-031-06773-0_10`.

[4] M. Benaïm and T. Hurth. *Markov Chains on Metric Spaces*. Springer, 2022.

[5] S. Bøgh, P. G. Jensen, M. Kristjansen, K. G. Larsen, and U. Nyman. Distributed fleet management in noisy environments via model-predictive control. In A. Kumar, S. Thiébaux, P. Varakantham, and W. Yeoh, editors, *Proceedings of the Thirty-Second International Conference on Automated Planning and Scheduling, ICAPS 2022, Singapore (virtual), June 13-24,*

2022, pages 565–573. AAAI Press, 2022. URL https://ojs.aaai.org/index.php/ICAPS/article/view/19843.

[6] G. E. Cho and C. D. Meyer. Comparison of perturbation bounds for the stationary distribution of a markov chain. *Linear Algebra and its Applications*, 335(1-3):137–150, 2001.

[7] A. David, P. G. Jensen, K. G. Larsen, A. Legay, D. Lime, M. G. Sørensen, and J. H. Taankvist. On time with minimal expected cost! In F. Cassez and J. Raskin, editors, *Automated Technology for Verification and Analysis - 12th International Symposium, ATVA 2014, Sydney, NSW, Australia, November 3-7, 2014, Proceedings*, volume 8837 of *Lecture Notes in Computer Science*, pages 129–145. Springer, 2014. doi: 10.1007/978-3-319-11936-6_10. URL https://doi.org/10.1007/978-3-319-11936-6_10.

[8] A. David, P. G. Jensen, K. G. Larsen, M. Mikucionis, and J. H. Taankvist. Uppaal stratego. In C. Baier and C. Tinelli, editors, *Tools and Algorithms for the Construction and Analysis of Systems - 21st International Conference, TACAS 2015, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2015, London, UK, April 11-18, 2015. Proceedings*, volume 9035 of *Lecture Notes in Computer Science*, pages 206–211. Springer, 2015. doi: 10.1007/978-3-662-46681-0_16. URL https://doi.org/10.1007/978-3-662-46681-0_16.

[9] C. Dehnert, S. Junges, J. Katoen, and M. Volk. A storm is coming: A modern probabilistic model checker. In R. Majumdar and V. Kuncak, editors, *Computer Aided Verification - 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part II*, volume 10427 of *Lecture Notes in Computer Science*, pages 592–600. Springer, 2017. doi: 10.1007/978-3-319-63390-9_31. URL https://doi.org/10.1007/978-3-319-63390-9_31.

[10] A. B. Eriksen, H. Lahrmann, K. G. Larsen, and J. H. Taankvist. Controlling signalized intersections using machine learning. In *World Conference on Transport Research*, volume 48 of *Transportation Research Procedia*, pages 987–997. Science Direct, 2020.

[11] M. A. Goorden, K. G. Larsen, J. E. Nielsen, T. D. Nielsen, M. R. Rasmussen, and J. Srba. Learning safe and optimal control strategies for storm water detention ponds. In R. M. Jungers, N. Ozay, and A. Abate, editors, *7th IFAC Conference on Analysis and Design of Hybrid Systems, ADHS 2021, Brussels, Belgium, July 7-9, 2021*, volume 54 of *IFAC-PapersOnLine*, pages 13–18. Elsevier, 2021. doi: 10.1016/j.ifacol.2021.08.467. URL https://doi.org/10.1016/j.ifacol.2021.08.467.

[12] M. A. Goorden, P. G. Jensen, K. G. Larsen, M. Samusev, J. Srba, and G. Zhao. STOMPC: stochastic model-predictive control with uppaal stratego. In A. Bouajjani, L. Holík, and Z. Wu, editors, *Automated Technology for Verification and Analysis - 20th International Symposium, ATVA 2022, Virtual Event, October 25-28, 2022, Proceedings*, volume 13505 of *Lecture Notes in Computer Science*, pages 327–333. Springer,

2022. doi: 10.1007/978-3-031-19992-9_21. URL https://doi.org/10.1007/978-3-031-19992-9_21.

[13] T. P. Gros, H. Hermanns, J. Hoffmann, M. Klauck, and M. Steinmetz. Deep statistical model checking. In A. Gotsman and A. Sokolova, editors, *Formal Techniques for Distributed Objects, Components, and Systems - 40th IFIP WG 6.1 International Conference, FORTE 2020, Held as Part of the 15th International Federated Conference on Distributed Computing Techniques, DisCoTec 2020, Valletta, Malta, June 15-19, 2020, Proceedings*, volume 12136 of *Lecture Notes in Computer Science*, pages 96–114. Springer, 2020. doi: 10.1007/978-3-030-50086-3_6. URL https://doi.org/10.1007/978-3-030-50086-3_6.

[14] D. Gross, N. Jansen, S. Junges, and G. A. Pérez. COOL-MC: A comprehensive tool for reinforcement learning and model checking. In W. Dong and J. Talpin, editors, *Dependable Software Engineering. Theories, Tools, and Applications - 8th International Symposium, SETTA 2022, Beijing, China, October 27-29, 2022, Proceedings*, volume 13649 of *Lecture Notes in Computer Science*, pages 41–49. Springer, 2022. doi: 10.1007/978-3-031-21213-0_3. URL https://doi.org/10.1007/978-3-031-21213-0_3.

[15] A. Hartmanns and H. Hermanns. The modest toolset: An integrated environment for quantitative modelling and verification. In E. Ábrahám and K. Havelund, editors, *Tools and Algorithms for the Construction and Analysis of Systems - 20th International Conference, TACAS 2014, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2014, Grenoble, France, April 5-13, 2014. Proceedings*, volume 8413 of *Lecture Notes in Computer Science*, pages 593–598. Springer, 2014. doi: 10.1007/978-3-642-54862-8_51. URL https://doi.org/10.1007/978-3-642-54862-8_51.

[16] I. R. Hasrat, P. G. Jensen, K. G. Larsen, and J. Srba. End-to-end heat-pump control using continuous time stochastic modelling and uppaal stratego. In Y. A. Ameur and F. Craciun, editors, *Theoretical Aspects of Software Engineering - 16th International Symposium, TASE 2022, Cluj-Napoca, Romania, July 8-10, 2022, Proceedings*, volume 13299 of *Lecture Notes in Computer Science*, pages 363–380. Springer, 2022. doi: 10.1007/978-3-031-10363-6_24. URL https://doi.org/10.1007/978-3-031-10363-6_24.

[17] P. Herber, J. Adelt, and T. Liebrenz. Formal verification of intelligent cyber-physical systems with the interactive theorem prover keymaera X. In S. Götz, L. Linsbauer, I. Schaefer, and A. Wortmann, editors, *Proceedings of the Software Engineering 2021 Satellite Events, Braunschweig/Virtual, Germany, February 22 - 26, 2021*, volume 2814 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2814/short-A3-2.pdf.

[18] O. Hernandez-Lerma, J. Hennet, and J. Lasserre. Average cost markov decision processes: Optimality conditions. *Journal of mathematical analysis and applications*, 158(2):396–406, 1991.

[19] T. Jaakkola, M. Jordan, and S. Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

[20] M. Jaeger, P. G. Jensen, K. G. Larsen, A. Legay, S. Sedwards, and J. H. Taankvist. Teaching stratego to play ball: Optimal synthesis for continuous space mdps. In Y. Chen, C. Cheng, and J. Esparza, editors, *Automated Technology for Verification and Analysis - 17th International Symposium, ATVA 2019, Taipei, Taiwan, October 28-31, 2019, Proceedings*, volume 11781 of *Lecture Notes in Computer Science*, pages 81–97. Springer, 2019. doi: 10.1007/978-3-030-31784-3_5. URL `https://doi.org/10.1007/978-3-030-31784-3_5`.

[21] M. Jaeger, G. Bacci, G. Bacci, K. G. Larsen, and P. G. Jensen. Approximating euclidean by imprecise markov decision processes. In *Leveraging Applications of Formal Methods, Verification and Validation: Verification Principles: 9th International Symposium on Leveraging Applications of Formal Methods, ISoLA 2020, Rhodes, Greece, October 20–30, 2020, Proceedings, Part I 9*, pages 275–289. Springer, 2020.

[22] M. Z. Kwiatkowska, G. Norman, and D. Parker. PRISM: probabilistic symbolic model checker. In T. Field, P. G. Harrison, J. T. Bradley, and U. Harder, editors, *Computer Performance Evaluation, Modelling Techniques and Tools 12th International Conference, TOOLS 2002, London, UK, April 14-17, 2002, Proceedings*, volume 2324 of *Lecture Notes in Computer Science*, pages 200–204. Springer, 2002. doi: 10.1007/3-540-46029-2_13. URL `https://doi.org/10.1007/3-540-46029-2_13`.

[23] K. G. Larsen, P. Pettersson, and W. Yi. UPPAAL in a nutshell. *Int. J. Softw. Tools Technol. Transf.*, 1(1-2):134–152, 1997. doi: 10.1007/s100090050010. URL `https://doi.org/10.1007/s100090050010`.

[24] K. G. Larsen, M. Mikucionis, M. Muñiz, J. Srba, and J. H. Taankvist. Online and compositional learning of controllers with application to floor heating. In M. Chechik and J. Raskin, editors, *Tools and Algorithms for the Construction and Analysis of Systems - 22nd International Conference, TACAS 2016, Held as Part of the European Joint Conferences on Theory and Practice of Software, ETAPS 2016, Eindhoven, The Netherlands, April 2-8, 2016, Proceedings*, volume 9636 of *Lecture Notes in Computer Science*, pages 244–259. Springer, 2016. doi: 10.1007/978-3-662-49674-9_14. URL `https://doi.org/10.1007/978-3-662-49674-9_14`.

[25] K. G. Larsen, A. Legay, G. Nolte, M. Schlüter, M. Stoelinga, and B. Steffen. Formal methods meet machine learning (F3ML). In T. Margaria and B. Steffen, editors, *Leveraging Applications of Formal Methods, Verification and Validation. Adaptation and Learning - 11th International Symposium, ISoLA 2022, Rhodes, Greece, October 22-30, 2022, Proceedings, Part III*, volume 13703 of *Lecture Notes in Computer Science*, pages 393–405. Springer, 2022. doi: 10.1007/978-3-031-19759-8_24. URL `https://doi.org/10.1007/978-3-031-19759-8_24`.

[26] A. Lavaei, S. Soudjani, A. Abate, and M. Zamani. Automated verification and synthesis of stochastic hybrid systems: A survey. *Automatica*, 146:

110617, 2022.

[27] A. Platzer and J. Quesel. Keymaera: A hybrid theorem prover for hybrid systems (system description). In A. Armando, P. Baumgartner, and G. Dowek, editors, *Automated Reasoning, 4th International Joint Conference, IJCAR 2008, Sydney, Australia, August 12-15, 2008, Proceedings*, volume 5195 of *Lecture Notes in Computer Science*, pages 171–178. Springer, 2008. doi: 10.1007/978-3-540-71070-7_15. URL `https://doi.org/10.1007/978-3-540-71070-7_15`.

[28] M. L. Puterman. *Markov Decision Processes*. Wiley, 2005.

[29] E. Seneta. Perturbation of the stationary distribution measured by ergodicity coefficients. *Advances in Applied Probability*, 20(1):228–230, 1988.

[30] S. R. Sinclair, S. Banerjee, and C. L. Yu. Adaptive discretization for episodic reinforcement learning in metric spaces. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(3):1–44, 2019.

[31] Z. Song and W. Sun. Efficient model-free reinforcement learning in metric spaces. *arXiv preprint arXiv:1905.00475*, 2019.

[32] Y. Wan, A. Naik, and R. S. Sutton. Learning and planning in average-reward markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR, 2021.

A Mathematical Background

The K-dimensional real space $\mathbb{R}^K$ is equipped with the standard Euclidean metric $d_{eucl}((x_1, \ldots, x_K), (y_1, \ldots, y_K)) := \sqrt{\sum_{i=1}^{K}(x_i - y_i)^2}$. Probability distributions on $\mathbb{R}^K$ are defined on systems of subsets of $\mathbb{R}^K$ that form a *σ-algebra*, i.e., that include the empty set, and are closed under countable unions and complementation. The elements of the σ-algebra are also referred to as the *measurable* sets. The *Borel σ-algebra* $\mathcal{B}(\mathbb{R}^K)$ is the smallest *σ-algebra* that contains all sets that are open with respect to d_{eucl}. For a measurable subset $\mathcal{S} \subseteq \mathbb{R}^K$, the Borel σ-algebra $\mathcal{B}(\mathcal{S})$ consists of all elements of $\mathcal{B}(\mathbb{R}^K)$ that are subsets of $\mathcal{S}$.

A *transition kernel* models a single probabilistic transition step in a continuous state space. It corresponds to a transition probability matrix in finite state Markov chains. Formally, a transition kernel on $\mathcal{S}$ is a function $T : \mathcal{S} \times \mathcal{B}(\mathcal{S}) \to [0, 1]$, such that $T(s, \cdot)$ is a probability distribution on $\mathcal{B}(\mathcal{S})$ for all $s \in \mathcal{S}$, and $T(\cdot, B)$ is measurable for all $B \in \mathcal{B}(\mathcal{S})$. The first of these conditions is the obvious requirement to define the probability distribution for the successor state s' given current state s. The second condition is needed in order to make integrals well-defined that define probabilities of events of interest. For example, given a distribution P_0 for the initial state s_0, the probability that the successor state s_1 lies in the measurable set B is

$$P(s_1 \in B) = \int_{\mathcal{S}} T(s_0, B) P_0(ds_0). \tag{16}$$

To generalize from Markov processes to Markov decision processes, the transition kernels are parameterized by the actions $a \in Act$, i.e., are functions $T(s, a, B)$. For a given strategy $\sigma : \mathcal{S} \to Act$, $T_\sigma(s, B) := T(s, \sigma(s), B)$ then is a standard transition kernel, under the condition that the sets $\{s \in \mathcal{S} | \sigma(s) = a\}$ are measurable. A transition kernel T representing a single transition step induces for all $k \geq 1$ a transition kernel T^k representing k successive steps taken according to T. Formally, $T^1 = T$, and

$$T^{k+1}(s_0, B) = \int_{\mathcal{S}} T(s_k, B) T^k(s_0, ds_k).$$

Generalizing (16), a transition kernel T in conjunction with an initial state distribution P_0 defines the probabilities of *cylinder sets* of the form $s_0 \in B_0, s_1 \in B_1, ..., s_k \in B_k$ (B_i measurable sets), which, in turn, induce a unique probability distribution on the space $\mathcal{S}^\infty$ of infinite state sequences (equipped with a canonical σ-algebra). When the transition kernel is defined by a strategy σ, and the initial distribution P_0 is given by a pointmass on a fixed initial state s, then we denote this distribution by $P_{s,\sigma}$, and its marginal for the state at fixed step k by $P^k_{s,\sigma}$.

For two probability distributions P, Q defined on $\mathcal{B}(\mathcal{S})$ the *total variation distance* is defined as

$$d_{tv}(P, Q) := \sup_{B \in \mathcal{B}(\mathcal{S})} |P(B) - Q(B)|.$$

B Proofs

Proof of Theorem 2: We consider the special case where $\mathcal{B}$ is equal to $\mathcal{A}$, except that a single $\nu_0 \in \mathcal{A}$ is sub-divided into $\nu_0 = \nu_0' \cup \nu_0''$ in $\mathcal{B}$. The general case follows directly from this. Let $\alpha \in A[\mathcal{B}]$ be given. For the left inequality of (11) we show that

$$\forall \sigma \in \Pi[\mathcal{B}],\ \exists \hat{\sigma} \in \Pi[\mathcal{A}],\ \exists \alpha^- \in A[\mathcal{A}] : \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\hat{\sigma},\alpha^-}[\mathcal{C}] \leq \mathbb{E}^{\mathcal{B}}_{[s_0]_{\mathcal{B}},\sigma,\alpha}[\mathcal{C}] \tag{17}$$

This we show for all cost types, and without the restriction to small $\mathcal{M}$. Let σ be given. Let $a' := \sigma(\nu_0')$, $a'' := \sigma(\nu_0'')$, and assume $a' \neq a''$ (the case $a' = a''$ can be treated by a simpler variation of the same arguments). We first construct a history-dependent, randomized strategy $\tilde{\sigma}$. For a run π of $\mathcal{M}^{\mathcal{A}}$ and $t \geq 0$ let $\pi_{[0:t-1]}$ denote the prefix of π containing the first t state-action pairs. A run π in $\mathcal{M}^{\mathcal{A}}$ defines a run $\hat{\pi}$ in $\mathcal{M}^{\mathcal{B}}$ by replacing state-action pairs ν_0, a' with $\nu_0' a'$, and ν_0, a'' with $\nu_0'' a''$. Define for $t \geq 0$ and $a \in Act$:

$$\tilde{\sigma}(\nu, \pi_{[0:t-1]})(a) := \begin{cases} P^{\mathcal{B}}_{\sigma,\alpha}(s_t = \nu_0' | \hat{\pi}_{[0:t-1]}) & \nu = \nu_0', a = a' \\ P^{\mathcal{B}}_{\sigma,\alpha}(s_t = \nu_0'' | \hat{\pi}_{[0:t-1]}) & \nu = \nu_0'', a = a'' \\ 1 & \nu \notin \{n_0', n_0''\}, a = \sigma(\nu) \\ 0 & \text{otherwise} \end{cases}$$

Define α^- as

$$\alpha^-(\nu, a) := \begin{cases} \alpha(\nu_0', a') & \nu = \nu_0, a = a' \\ \alpha(\nu_0'', a'') & \nu = \nu_0, a = a'' \\ \alpha(\nu, a) & \nu \neq \nu_0 \end{cases}$$

We now obtain that for all $\pi_{[0:t]}$ in $\mathcal{M}^{\mathcal{A}}$:

$$P^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\tilde{\sigma},\alpha^-}(\pi_{[0:t]}) = P^{\mathcal{B}}_{[s_0]_{\mathcal{B}},\sigma,\alpha}(\hat{\pi}_{[0:t]}).$$

Since $\mathcal{C}^{\mathcal{A}}_{\alpha^-}(\pi) = \mathcal{C}^{\mathcal{B}}_{\alpha}(\hat{\pi})$, this implies that $\mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\tilde{\sigma},\alpha^-}[\mathcal{C}] = \mathbb{E}^{\mathcal{B}}_{[s_0]_{\mathcal{B}},\sigma,\alpha}[\mathcal{C}]$. Since $\mathcal{M}^{\mathcal{A}}_{\alpha^-}$ is a finite state MDP, there exist for all cost definitions optimal strategies that are deterministic and stationary, i.e., there exists $\hat{\sigma} \in \Pi[\mathcal{A}]$ with $\mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\hat{\sigma},\alpha^-}[\mathcal{C}] \leq \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\tilde{\sigma},\alpha^-}[\mathcal{C}]$.

For the right inequality we show

$$\forall \sigma \in \Pi[\mathcal{A}]\ \exists \hat{\sigma} \in \Pi[\mathcal{B}]\ \exists \alpha^+ \in A[\mathcal{A}] : \mathbb{E}^{\mathcal{B}}_{[s_0]_{\mathcal{B}},\hat{\sigma},\alpha}[\mathcal{C}] \leq \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\sigma,\alpha^+}[\mathcal{C}]. \tag{18}$$

Let $\sigma \in \Pi[\mathcal{A}]$, and $\hat{\sigma}$ the induced strategy in $\mathcal{M}^{\mathcal{B}}$ defined by $\hat{\sigma}(\nu_0') = \hat{\sigma}(\nu_0'') = \sigma(\nu_0)$ and $\hat{\sigma}(\nu) = \sigma(\nu)$ for all $\nu \neq \nu_0$. We now first consider the case $\mathcal{C} \in \{\mathcal{C}_{p-avg}, \mathcal{C}_{e-avg}\}$. Since $\mathcal{M}$ is small, the induced Markov chain $\mathcal{M}^{\mathcal{B}}_{\hat{\sigma},\alpha}$ is irreducible and aperiodic with unique stationary distribution $\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}$. Define

$$\alpha^+(\nu_0, \sigma(\nu_0)) := \frac{\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0')\alpha(\nu_0', \sigma(\nu_0)) + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'')\alpha(\nu_0'', \sigma(\nu_0))}{\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0') + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'')},$$

and $\alpha^+(\nu, \sigma(\nu)) = \alpha(\nu, \sigma(\nu))$ for all $\nu \neq \nu_0$. Then

$$T^{\mathcal{A}}(\nu_0, \sigma(\nu_0), \cdot) = \frac{\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0')T^{\mathcal{B}}(\nu_0', \sigma(\nu_0), \cdot) + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'')T^{\mathcal{B}}(\nu_0'', \sigma(\nu_0)), \cdot}{\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0') + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'')},$$

and for $\nu \neq \nu_0$

$$T^{\mathcal{A}}(\nu, \sigma(\nu), \nu_0) = T^{\mathcal{B}}(\nu, \sigma(\nu), \nu_0') + T^{\mathcal{B}}(\nu, \sigma(\nu), \nu_0'').$$

Comparing the stationarity equations for $\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}$ and $\tilde{P}^{\mathcal{A}}_{\sigma,\alpha^+}$ one then finds

$$\tilde{P}^{\mathcal{A}}_{\sigma,\alpha^+}(\nu_0) = \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0') + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'').$$

Together with

$$\mathcal{C}^{\mathcal{A}}_{\alpha^+}(\nu_0, \sigma(\nu_0)) = \frac{\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0')\mathcal{C}^{\mathcal{B}}(\nu_0', \sigma(\nu_0)) + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'')\mathcal{C}^{\mathcal{B}}(\nu_0'', \sigma(\nu_0))}{\tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0') + \tilde{P}^{\mathcal{B}}_{\hat{\sigma},\alpha}(\nu_0'')}$$

this yields $\mathbb{E}^{\mathcal{B}}_{[s_0]_{\mathcal{B}},\hat{\sigma},\alpha}[\mathcal{C}] = \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\sigma,\alpha^+}[\mathcal{C}]$.

Finally, we turn to $\mathcal{C} \in \{\mathcal{C}_u, \mathcal{C}_\lambda\}$. Assume without loss of generality that $\mathbb{E}^{\mathcal{B}}_{\nu_0',\hat{\sigma},\alpha}[\mathcal{C}] \leq \mathbb{E}^{\mathcal{B}}_{\nu_0'',\hat{\sigma},\alpha}[\mathcal{C}]$. Define $\alpha^+(\nu_0, \sigma(\nu_0)) = \alpha(\nu_0'', \sigma(\nu_0))$. Then the inequality in (18) is verified by comparing the cost equations for $\mathcal{M}^{\mathcal{B}}_{\hat{\sigma},\alpha}$ and $\mathcal{M}^{\mathcal{A}}_{\sigma,\alpha^+}$.

□

Proof of Theorem 3:

We start with the right inequality of (12) for $\mathcal{C} \in \{\mathcal{C}_u, \mathcal{C}_\lambda\}$. Let $\sigma \in \Pi[\mathcal{A}]$, and $\hat{\sigma} \in \Pi[\mathcal{S}]$ the induced strategy on $\mathcal{M}$. Let $s_0 \in \mathcal{S}$ be given. We define a non-stationary adversary α^* that takes the transition step $k \geq 1$ as an additional input by letting $\alpha^*(\nu, \sigma(\nu), k)$ be equal to $P^k_{s_0,\hat{\sigma}}$ conditioned on ν, denoted $P^k_{s,\hat{\sigma}}|\nu$ (given the fixed strategy σ, the adversary need only be defined for state/action pairs $(\nu, \sigma(\nu))$).

We first show by induction that for all $k \geq 0$ and $\nu \in \mathcal{A}$:

$$P^{\mathcal{M},k}_{s_0,\hat{\sigma}}(\nu) = P^{\mathcal{A},k}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}(\nu). \tag{19}$$

For $k = 0$ we have that the probabilities on both sides are 1 for $\nu = [s_0]_{\mathcal{A}}$. For the induction step we have

$$P^{\mathcal{M},k}_{s_0,\hat{\sigma}}(\nu) = \sum_{\nu' \in \mathcal{A}} P^{\mathcal{M},k-1}_{s_0,\hat{\sigma}}(\nu') \int_{\nu'} T(s, \hat{\sigma}(s), \nu) dP^{\mathcal{M},k-1}_{s_0,\hat{\sigma}}|\nu'(s) =$$

$$\sum_{\nu' \in \mathcal{A}} P^{\mathcal{A},k-1}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}(\nu')\alpha(\nu', \sigma(\nu'), \nu) = P^{\mathcal{A},k}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}(\nu). \tag{20}$$

We now consider the expected cost at step ($k \geq 0$), and obtain:

$$\mathbb{E}^{\mathcal{M}}_{s_0,\hat{\sigma}}[\mathcal{C}(s_k,\hat{\sigma}(s_k))] = \int_{\mathcal{S}} \mathcal{C}(s,\hat{\sigma}(s))dP^{\mathcal{M},k}_{s_0,\hat{\sigma}}(s) = \\ \sum_{\nu\in\mathcal{A}} P^{\mathcal{M},k}_{s_0,\hat{\sigma}}(\nu)\int_{\nu} \mathcal{C}(s,\hat{\sigma}(s))dP^{\mathcal{M},k}_{s_0,\hat{\sigma}}|\nu(s) = \sum_{\nu\in\mathcal{A}} P^{\mathcal{A},k}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}(\nu)\int_{\nu} \mathcal{C}^{\mathcal{A}}(\nu,\sigma(\nu)) = \\ \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}[\mathcal{C}(\nu_k,\sigma(\nu_k))]. \tag{21}$$

For $\mathcal{C} \in \{\mathcal{C}_u, \mathcal{C}_\lambda\}$ the pointwise equality (21) for each k implies the equality of expectations over paths, and hence

$$\mathbb{E}^{\mathcal{M}}_{s_0}[\mathcal{C}] \leq \mathbb{E}^{\mathcal{M}}_{s_0,\hat{\sigma}}[\mathcal{C}] = \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}[\mathcal{C}]. \tag{22}$$

For the fixed stationary strategy σ, an adversary α that maximizes $\mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\sigma,\alpha}[\mathcal{C}]$ can always be chosen to be stationary. Thus, even though α^* was defined as a non-stationary adversary, we have

$$\mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\alpha^*,\sigma}[\mathcal{C}] \leq \max_{\alpha} \mathbb{E}^{\mathcal{A}}_{[s_0]_{\mathcal{A}},\sigma,\alpha}, \tag{23}$$

where now the maximum is only taken over stationary adversaries. Combining (22) and (23), and observing that σ was an arbitrary strategy on $\mathcal{M}^{\mathcal{A}}$, we obtain the right-hand inequality of (12).

We next prove the left inequality of (12). Let $\mathcal{A}$ be given, and $\sigma \in \varPi[\mathcal{S}]$. Let $\mathcal{A} \sqcap \sigma$ denote the refinement of $\mathcal{A}$ that is defined by intersecting $\mathcal{A}$ with the partition $\{\{s \in \mathcal{S} | \sigma(s) = a\} | a \in \mathit{Act}\}$. Then σ directly induces a strategy $\hat{\sigma}$ on $\mathcal{M}^{\mathcal{A}\sqcap\sigma}$. As above, we define a non-stationary adversary $\alpha^* \in A[\mathcal{A} \sqcap \sigma]$ by letting $\alpha^*(\nu, \sigma(\nu), k)$ be equal to $P^k_{s_0,\sigma}|\nu$. In analogy to (21) we obtain

$$\mathbb{E}^{\mathcal{A}\sqcap\sigma}_{[s_0]_{\mathcal{A}\sqcap\sigma},\hat{\sigma},\alpha^*}[\mathcal{C}(\nu_k,\hat{\sigma}(s_k))] = \mathbb{E}^{\mathcal{M}}_{[s_0]_{\mathcal{A}},\sigma}[\mathcal{C}(s_k,\sigma(\nu_k))]. \tag{24}$$

Together with the left inequality of (11) (with $\mathcal{B} = \mathcal{A} \sqcap \sigma$) and again the fact that a minimizing adversary can be chosen to be stationary, this shows the left side of (12).

The proof for $\mathcal{C} \in \{\mathcal{C}_{p-avg}, \mathcal{C}_{e-avg}\}$ is mostly an application of Theorem 1, which immediately shows that in this case for all $s_0 \in \mathcal{S}$ and strategies σ

$$\mathbb{E}^{\mathcal{M}}_{s_0,\sigma}[\mathcal{C}_{e-avg}] = \mathbb{E}^{\mathcal{M}}_{s_0,\sigma}[\mathcal{C}_{p-avg}] = \tilde{\mathbb{E}}^{\mathcal{M}}_{\sigma}[\mathcal{C}(s,\sigma(s)]. \tag{25}$$

The smallness of $\mathcal{M}$ implies that for all partitions $\mathcal{A}$ and $\sigma \in \varPi[\mathcal{A}], \alpha \in A[\mathcal{A}]$ the Markov chain $\mathcal{M}^{\mathcal{A}}_{\alpha,\sigma}$ is irreducible and aperiodic with a unique stationary distribution $\tilde{P}^{\mathcal{A}}_{\alpha,\sigma}$, and therefore in analogy to (25) for all $\sigma \in \varPi[\mathcal{A}], \nu_0 \in \mathcal{A}$:

$$\mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma,\alpha}[\mathcal{C}_{e-avg}] = \mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma,\alpha}[\mathcal{C}_{p-avg}] = \tilde{\mathbb{E}}^{\mathcal{A}}_{\sigma,\alpha}[\mathcal{C}_\alpha(\nu,\sigma(\nu)]. \tag{26}$$

We prove (12) by the same lines of argument as above. However, the step-wise equalities (21), (24) are now replaced by analogous properties of the stationary limit distributions.

Beginning again with the right inequality of (12), let $\sigma \in \Pi[\mathcal{A}]$. The induced $\hat{\sigma} \in \Pi[\mathcal{S}]$ has a stationary limit distribution $\tilde{P}^{\mathcal{M}}_{\hat{\sigma}}$. Define $\alpha \in A[\mathcal{A}]$ as $\alpha(\nu, \sigma(\nu)) = \tilde{P}^{\mathcal{M}}_{\hat{\sigma}}|\nu$ (actions other than $\sigma(\nu)$ are irrelevant). We now have that the stationary distribution $\tilde{P}^{\mathcal{A}}_{\sigma,\alpha}$ is equal to the marginal of $\tilde{P}^{\mathcal{M}}_{\hat{\sigma}}$ on $\mathcal{A}$, and hence

$$\tilde{\mathbb{E}}^{\mathcal{M}}_{\hat{\sigma}}[\mathcal{C}(s, \hat{\sigma}(s)] = \sum_{\nu \in \mathcal{A}} \tilde{P}^{\mathcal{M}}_{\hat{\sigma}}(\nu) \int_{\nu} \mathcal{C}(s, \hat{\sigma}(s)) d(\tilde{P}^{\mathcal{M}}_{\hat{\sigma}}|\nu)(s) =$$
$$\sum_{\nu \in \mathcal{A}} \tilde{P}^{\mathcal{A}}_{\sigma,\alpha}(\nu) \mathcal{C}^{\mathcal{A}}(\nu, \sigma(\nu)) = \tilde{\mathbb{E}}^{\mathcal{A}}_{\sigma,\alpha}[\mathcal{C}^{\mathcal{A}}_{\alpha}(\nu, \sigma(\nu))] \quad (27)$$

For the left inequality of (12) let $\mathcal{A} \sqcap \sigma$ and $\hat{\sigma} \in \Pi[\mathcal{A} \sqcap \sigma]$ be defined as before. Define $\alpha^{-}(\nu, \sigma(\nu)) = \tilde{P}^{\mathcal{M}}_{\sigma}|\nu$. Again, the stationary distribution $\tilde{P}^{\mathcal{A} \sqcap \sigma}_{\alpha^{-},\hat{\sigma}}$ is equal to the marginal of $\tilde{P}^{\mathcal{M}}_{\sigma}$ on $\mathcal{A} \sqcap \sigma$, and we obtain (27) for $\mathcal{A} \sqcap \sigma$ in place of $\mathcal{A}$. Together with (11) this yields the desired result.

□

Proof of Theorem 4: For any partition $\mathcal{A}$ let $\delta(\mathcal{A}) = \max_{\nu \in \mathcal{A}} \sup_{s,s' \in \nu} ||s - s'||$ denote the *diameter* of $\mathcal{A}$. For both parts of the theorem it is sufficient to show that for all $\epsilon > 0$: when the diameter of $\mathcal{A}_k$ is sufficiently small, then for all $\sigma \in \Pi[\mathcal{A}_k]$, $\alpha^{-}, \alpha^{+} \in A[\mathcal{A}_k]$, and $\nu \in \mathcal{A}_k$:

$$|\mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma,\alpha^{-}}[\mathcal{C}] - \mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma,\alpha^{+}}[\mathcal{C}]| < \epsilon. \quad (28)$$

Then, assuming without loss of generality that

$$\mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma^{*},\alpha}[\mathcal{C}_{\lambda}] \leq \mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma^{*\prime},\alpha'}[\mathcal{C}_{\lambda}]$$

where $\sigma^{*\prime}, \sigma^{*}$ are the optimal strategies for $\mathcal{M}^{\mathcal{A}_k}_{\alpha'}, \mathcal{M}^{\mathcal{A}_k}_{\alpha}$, we obtain

$$\mathbb{E}^{\mathcal{A}_k}_{\nu,\alpha}[\mathcal{C}_{\lambda}] = \mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma^{*},\alpha}[\mathcal{C}_{\lambda}] \leq \mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma^{*\prime},\alpha'}[\mathcal{C}_{\lambda}] \leq \mathbb{E}^{\mathcal{A}_k}_{\nu,\sigma^{*},\alpha'}[\mathcal{C}_{\lambda}] \leq \mathbb{E}^{\mathcal{A}_k}_{\nu,\alpha}[\mathcal{C}_{\lambda}] + \epsilon.$$

We first show (28) for $\mathcal{C} = \mathcal{C}_{\lambda}$. For $N \geq 0$ we define the *truncated expected cost* $\mathbb{E}^{N}$ by taking the sum in (3) only over $i = 0, \ldots, N$. For each $\lambda < 1$ and each $\epsilon > 0$ there then exists an $N \geq 0$ such that

$$0 \leq \mathbb{E}_{\cdot}[\mathcal{C}_{\lambda}] - \mathbb{E}^{N}_{\cdot}[\mathcal{C}_{\lambda}] < \epsilon/2, \quad (29)$$

where these bounds apply uniformly to all expectations both in the EMDP $\mathcal{M}$, the induced IMPDPs $\mathcal{M}^{\mathcal{A}_i}$, for all strategies and adversaries, and for all states, respectively regions. According to Theorem 4 of [21] there exists a $\delta > 0$, such that for all partitions $\mathcal{A}_k$ with $\delta(\mathcal{A}_k) \leq \delta$, all strategies σ defined on $\mathcal{M}^{\mathcal{A}_k}$, all pairs of adversaries α^{-}, α^{+}, and all $\nu \in \mathcal{A}$:

$$|\mathbb{E}^{\mathcal{A}_k,N}_{\nu,\sigma,\alpha^{-}}[\mathcal{C}_{\lambda}] - \mathbb{E}^{N,\mathcal{A}_k}_{\nu,\sigma,\alpha^{+}}[\mathcal{C}_{\lambda}]| < \epsilon/2 \quad (30)$$

(the theorem and proof in [21] are for undiscounted cost, but the case of discounted costs is directly implied by this). In conjunction with (29), (30) implies (28).

We now turn to average cost $\mathcal{C}_{p-avg}$ (which, under the smallness assumption, is equal to $\mathcal{C}_{e-avg}$). Let $\sigma, \alpha^-, \alpha^+, \nu$ as above in (28). Let $\tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}, \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^+}$ be the stationary distributions of $\mathcal{M}^{\mathcal{A}_k}_{\sigma,\alpha^-}, \mathcal{M}^{\mathcal{A}_k}_{\sigma,\alpha^+}$. Our proof relies on perturbation bounds for Markov chains [29, 6], which in our context can be written as

$$d_{tv}(\tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}, \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^+}) \leq \frac{1}{1-\tau_1(\mathcal{M}^{\mathcal{A}_k}_{\sigma,\alpha^-})} \max_{\nu\in\mathcal{A}_k} d_{tv}(T^{\mathcal{A}_k}_{\alpha^-}(\nu,\sigma(\nu),\cdot), T^{\mathcal{A}_k}_{\alpha^+}(\nu,\sigma(\nu),\cdot)), \quad (31)$$

where

$$\tau_1(\mathcal{M}^{\mathcal{A}_k}_{\sigma,\alpha}) := \max_{\nu,\nu'\in\mathcal{A}_k} d_{tv}(T^{\mathcal{A}_k}_{\alpha}(\nu,\sigma(\nu),\cdot), T^{\mathcal{A}_k}_{\alpha}(\nu',\sigma(\nu),\cdot)) \quad (32)$$

is the *ergodicity coefficient*. We first obtain a bound on $\tau_1(\mathcal{M}^{\mathcal{A}_k}_{\sigma,\alpha})$ that is uniform for k, σ, α. Let $\nu, \nu' \in \mathcal{A}_k$.

$$\begin{aligned} d_{tv}(T^{\mathcal{A}_k}_{\alpha}(\nu,\sigma(\nu),\cdot), T^{\mathcal{A}_k}_{\alpha}(\nu',\sigma(\nu),\cdot)) = \\ \frac{1}{2}\sum_{\bar{\nu}} |T^{\mathcal{A}_k}_{\alpha}(\nu,\sigma(\nu),\bar{\nu}) - T^{\mathcal{A}_k}_{\alpha}(\nu',\sigma(\nu),\bar{\nu})| \\ \leq \frac{1}{2}\sum_{\bar{\nu}} |T^{\mathcal{A}_k}_{\alpha}(\nu,\sigma(\nu),\bar{\nu}) - \xi(\bar{\nu})| + \frac{1}{2}\sum_{\bar{\nu}} |T^{\mathcal{A}_k}_{\alpha}(\nu',\sigma(\nu),\bar{\nu}) - \xi(\bar{\nu})| \\ = 1 - \xi(\mathcal{S}). \quad (33) \end{aligned}$$

For the last equality observe that the smallness of $\mathcal{M}$ implies that for all $\mathcal{A}, \sigma, \alpha, \nu, \bar{\nu}$: $T^{\mathcal{A}}_{\alpha}(\nu,\sigma(\nu),\bar{\nu}) \geq \xi(\bar{\nu})$

Due to the continuity of $\mathcal{M}$ we have that for $\epsilon > 0$ there exists δ such that for partitions $\mathcal{A}_k$ with $\delta(\mathcal{A}_k) \leq \delta$ and all $\nu \in \mathcal{A}_k$

$$d_{tv}(T^{\mathcal{A}_k}_{\alpha^-}(\nu,\sigma(\nu),\cdot), T^{\mathcal{A}_k}_{\alpha^+}(\nu,\sigma(\nu),\cdot)) < \epsilon/2 \quad (34)$$

and

$$|\mathcal{C}_{\alpha^-}(\nu) - \mathcal{C}_{\alpha^+}(\nu)| < \epsilon/2. \quad (35)$$

(34) and (33) imply that for a given ϵ there exists $\delta > 0$, such that for all $\sigma, \alpha^-, \alpha^+$:

$$d_{tv}(\tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}, \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^+}) < \epsilon. \quad (36)$$

We therefore obtain that for sufficiently fine $\mathcal{A}_k$

$$
\begin{aligned}
&|\mathbb{E}^{\mathcal{A}_k}_{\sigma,\alpha^-}[\mathcal{C}_{p-avg}] - \mathbb{E}^{\mathcal{A}_k}_{\sigma,\alpha^+}[\mathcal{C}_{p-avg}]| = \\
&\qquad |\sum_{\nu\in\mathcal{A}_k} \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}(\nu)\mathcal{C}_{\alpha^-}(\nu) - \sum_{\nu\in\mathcal{A}_k} \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^+}(\nu)\mathcal{C}_{\alpha^+}(\nu)| \leq \\
&\qquad |\sum_{\nu\in\mathcal{A}_k} \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}(\nu)\mathcal{C}_{\alpha^-}(\nu) - \sum_{\nu\in\mathcal{A}_k} \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}(\nu)\mathcal{C}_{\alpha^+}(\nu)| + \\
&\qquad |\sum_{\nu\in\mathcal{A}_k} \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^+}(\nu)\mathcal{C}_{\alpha^+}(\nu) - \sum_{\nu\in\mathcal{A}_k} \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}(\nu)\mathcal{C}_{\alpha^+}(\nu)| \leq \\
&\qquad \epsilon/2 + \sum_{\nu\in\mathcal{A}_k} \mathcal{C}_{\alpha^+}(\nu)|\tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^+}(\nu) - \tilde{P}^{\mathcal{A}_k}_{\sigma,\alpha^-}(\nu)\mathcal{C}_{\alpha^+}(\nu)| \leq \\
&\qquad\qquad\qquad\qquad\qquad\qquad \epsilon/2 + c_{max}\cdot 2\cdot\epsilon.
\end{aligned}
$$

where the $\epsilon/2$ bound of the first term is due to (35), and the $c_{max}\cdot 2\cdot\epsilon$ bound for the second term is due to (34). □

Proof of Theorem 5: By induction on t. For $t = 0$ we have $\tilde{Q}_0 \equiv Q_0 \equiv 0$. Assume (15) holds for t. Then we first write

$$
\mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma}(\tilde{Q}_{t+1}(\nu,a)) = \sum_{\pi_{0:t}\in(\mathcal{A}\times Act)^t} P^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\pi_{0:t})\mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma}(\tilde{Q}_{t+1}(\nu,a)|\pi_{0:t}), \tag{37}
$$

and similarly for $\mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma}(Q_{t+1}(\nu,a))$. Since the probabilities $P^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\pi_{0:t})$ are the same for $\tilde{Q}_t$ and Q_t, it is sufficient to show that for all $\pi_{0:t}$

$$
\mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma}(\tilde{Q}_{t+1}(\nu,a)|\pi_{0:t}) = \mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma}(Q_{t+1}(\nu,a)|\pi_{0:t}). \tag{38}
$$

If $\nu_t \neq \nu$ in $\pi_{0:t}$, or $a \neq \sigma_t(\pi_{0:t})$, then the Q and $\tilde{Q}$ values of (ν,a) are not updated in the $t+1$'st iteration, and (15) holds by induction hypothesis.

Assume, then, that $(\nu_t, a_t) = (\nu, a)$. We obtain:

$$
\begin{aligned}
\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\tilde{Q}_{t+1}(\nu,a)|\pi_{0:t}) = (1-\beta_t(\pi_{0:t}))\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\tilde{Q}_t(\nu,a)|\pi_{0:t}) + \\
\beta_t(\pi_{0:t})(\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\mathcal{C}_t|\pi_{0:t}) + \lambda\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\min_{a'\in Act}\tilde{Q}_t(\nu_{t+1},a)|\pi_{0:t})),
\end{aligned} \tag{39}
$$

where in the rightmost term the $\tilde{Q}_t(\nu_{t+1},a)$ now are to be understood as random variables defined by the random next state ν_{t+1}. The distribution of $\mathcal{C}_t$ conditional on $\pi_{0:t}$ only depends on $\nu_t = \nu$, and the expectation is

$$
\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\mathcal{C}_t|\pi_{0:t}) = \int_{\nu}\mathcal{C}(s,a)d\alpha(\nu,a)(s) = \mathcal{C}^{\mathcal{A}}_{\alpha}(\nu,a). \tag{40}
$$

The distribution for the random ν_{t+1} given $\pi_{0:t}$ only depends on $(\nu_t, a_t) = (\nu, a)$. We can therefore write:

$$\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(\min_{a'\in Act}\tilde{Q}_t(\nu_{t+1},a)|\pi_{0:t}) = \int_\nu \sum_{\nu'\in\mathcal{A}} T(s,a,\nu')\min_{a'\in Act}\tilde{Q}_t(\nu',a')d\rho(s) = \sum_{\nu'\in\mathcal{A}}\min_{a'\in Act}\tilde{Q}_t(\nu',a')\int_\nu T(s,a,\nu')d\rho(s) = \sum_{\nu'\in\mathcal{A}}\alpha^\rho_T(\nu,a)(\nu')\min_{a'\in Act}\tilde{Q}_t(\nu',a'). \tag{41}$$

Substituting the right-hand sides of (40) and (40) into the right-hand side of (39), and replacing by induction hypothesis $\tilde{Q}_t$ with Q_t everywhere, we obtain

$$(1-\beta_t(\pi_{0:t}))\mathbb{E}^{\mathcal{A}}_{\nu_0,\boldsymbol{\sigma}}(Q_t(\nu,a)|\pi_{0:t})+ \beta_t(\pi_{0:t})(\alpha^\rho_C(\nu,a)+\lambda\sum_{\nu'\in\mathcal{A}}\alpha^\rho_T(\nu,a)(\nu')\min_{a'\in Act}\tilde{Q}_t(\nu_{t+1},a) = \mathbb{E}^{\mathcal{A}}_{\nu_0,\sigma}(Q_{t+1}(\nu,a)|\pi_{0:t}). \tag{42}$$

□