Markov Decision Processes with Multiple Long-run Average Objectives

Krishnendu Chatterjee

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2007-105

http://www.eecs.berkeley.edu/Pubs/TechRpts/2007/EECS-2007-105.html

August 22, 2007

Acknowledgement

This research was supported in part by the NSF grants CCR-0225610 and CCR-0234690.

Markov Decision Processes with Multiple Long-run Average Objectives[⋆]

Krishnendu Chatterjee

UC Berkeley
c_krish@eecs.berkeley.edu

Abstract. We consider Markov decision processes (MDPs) with multiple long-run average objectives. Such MDPs occur in design problems where one wishes to simultaneously optimize several criteria, for example, latency and power. The possible trade-offs between the different objectives are characterized by the Pareto curve. We show that every Pareto optimal point can be ε-approximated by a memoryless strategy, for all $\varepsilon > 0$. In contrast to the single-objective case, the memoryless strategy may require randomization. We show that the Pareto curve can be approximated (a) in polynomial time in the size of the MDP for irreducible MDPs; and (b) in polynomial space in the size of the MDP for all MDPs. Additionally, we study the problem if a given value vector is realizable by any strategy, and show that it can be decided in polynomial time for irreducible MDPs and in NP for all MDPs. These results provide algorithms for design exploration in MDP models with multiple long-run average objectives.

1 Introduction

Markov decision processes (MDPs) are standard models for dynamic systems that exhibit both probabilistic and nondeterministic behaviors [10, 4]. An MDP models a dynamic system that evolves through stages. In each stage, a controller chooses one of several actions (the nondeterministic choices), and the system stochastically evolves to a new state based on the current state and the chosen action. In addition, one associates a cost or reward with each transition, and the central question is to find a strategy of choosing the actions that optimizes the rewards obtained over the run of the system. The two classical ways of combing the rewards over the run of the system is are follows: (a) the discounted sum of the rewards and (b) the long-run average of the rewards. In many modeling domains, however, there is no unique objective to be optimized, but multiple, potentially dependent and conflicting objectives. For example, in designing a computer system, one is interested not only in maximizing performance but also in minimizing power. Similarly, in an inventory management system, one wishes to optimize several potentially dependent costs for maintaining each kind of

[⋆] This research was supported in part by the NSF grants CCR-0225610 and CCR-0234690.

product, and in AI planning, one wishes to find a plan that optimizes several distinct goals. These motivate the study of MDPs with multiple objectives.

We study MDPs with multiple long-run average objectives, an extension of the MDP model where there are several reward functions [6, 12]. In MDPs with multiple objectives, we are interested not in a single solution that is simultaneously optimal in all objectives (which may not exist), but in a notion of "trade-offs" called the *Pareto curve*. Informally, the Pareto curve consists of the set of realizable value profiles (or dually, the strategies that realize them) that are not dominated (in every dimension) by any other value profile. Pareto optimality has been studied in co-operative game theory [8] and in multi-criterion optimization and decision making in both economics and engineering [7, 13, 11]. Finding *some* Pareto optimal point can be reduced to optimizing a single objective: optimize a convex combination of objectives using a set of positive weights; the optimal strategy must be Pareto optimal as well (the "weighted factor method") [6]. In design space exploration, however, we want to find not one, but *all* Pareto optimal points in order to better understand the trade-offs in the design. Unfortunately, even with just two rewards, the Pareto curve may have infinitely many points, and also contain irrational payoffs. Many previous works has focused on constructing a sampling of the Pareto curve, either by choosing a variety of weights in the weighted factor method, or by imposing a lexicographic ordering on the objectives and sequentially optimizing each objective according to the order [3, 4]. Unfortunately, this does not provide any guarantee about the quality of the solutions obtained.

The study of the *approximate* version of the problem, the ε-approximate Pareto curve [9] for MDPs with multiple objectives is recent: the problem was studied for discounted sum objectives in [1] and for qualitative ω-regular objectives in [2]. Informally, the ε-approximate Pareto curve for $\varepsilon > 0$ contains a set of strategies (or dually, their payoff values) such that there is no other strategy whose value dominates the values in the Pareto curve by a factor of $1 + \varepsilon$.

Our results. In this work we study the complexity of approximating the Pareto curve for MDPs with multiple long-run average objectives. For a long-run average objective, given an infinite sequence $\langle v_0, v_1, v_2, \ldots \rangle$ of finite reward values the payoff is $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} v_t$. We summarize our results below.

1. We show that for all $\varepsilon > 0$, the value vector of a Pareto-optimal strategy can be ε-approximated by a memoryless strategy. In the case of single objective the definition of long-run average objective can be also alternatively defined as $\limsup$ instead of $\liminf$, and the optimal values coincide. In contrast, in the case of multiple objectives we show that if the long-run average objectives are defined as $\limsup$, then the Pareto-optimal strategies cannot be ε-approximated by memoryless strategies.
2. We show that an approximate Pareto curve can be computed in polynomial time for *irreducible* MDPs [4]; and in polynomial space for general MDPs. The algorithms are obtained by reduction to multi-objective linear-programming and applying the results of [9].

3. We also study the related *realizability* decision problem: given a profile of values, is there a Pareto-optimal strategy that dominates it? We show that the realizability problem can be decided in polynomial time for irreducible MDPs and in NP for general MDPs.

Our work is closely related to the works of [1, 2]. In [1] MDPs with multiple discounted reward objectives was studied. It was shown that memoryless strategies suffices for Pareto optimal strategies, and polynomial time algorithm was given to approximate the Pareto curve by reduction to multi-objective linear-programming and using the results of [9]. In [2] MDPs with multiple qualitative ω-regular objectives was studied. It was shown that the Pareto curve can be approximated in polynomial time: the algorithm first reduces the problem to MDPs with multiple reachability objectives, and then MDPs with multiple reachability objectives can be solved by multi-objective linear-programming. In our case we have the undiscounted setting as well as quantitative objectives and there are new obstacles in the proofs. For example, the notion of "discounted frequencies" used in [1] need not be well defined in the undiscounted setting. Our proof technique uses the results of [1] and a celebrated result Hardy-Littlewood to obtain the result on sufficiency of memoryless strategies for Pareto optimal strategies. Also our reduction to multi-objective linear-programming is more involved: we require several multi-objective linear-programs in the general case, it uses techniques of [2] for transient states and approaches similar to [1] for recurrent states.

2 Markov Decision Processes with Multiple Long-run Average Objectives

We denote the set of probability distributions on a set U by $\mathcal{D}(U)$.

Markov decision processes (MDPs). A Markov decision process (MDP) $G = (S, A, p)$ consists of a finite, non-empty set S of states and a finite, non-empty set A of actions; and a probabilistic transition function $p : S \times A \to \mathcal{D}(S)$, that given a state $s \in S$ and an action $a \in A$ gives the probability $p(s,a)(t)$ of the next state t.

We denote by $\mathrm{Dest}(s,a) = \mathrm{Supp}(p(s,a))$ the set of possible successors of s when the action a is chosen. Given an MDP G we define the set of edges $E = \{ (s,t) \mid \exists a \in A.\ t \in \mathrm{Dest}(s,a) \}$ and use $E(s) = \{ t \mid (s,t) \in E \}$ for the set of possible successors of s in G.

Plays and strategies. A *play* of G is an infinite sequence $\langle s_0, s_1, \ldots \rangle$ of states such that for all $i \geq 0$, $(s_i, s_{i+1}) \in E$. A strategy σ is a recipe that specifies how to extend a play. Formally, a strategy σ is a function $\sigma : S^+ \to \mathcal{D}(A)$ that, given a finite and non-empty sequence of states representing the history of the play so far, chooses a probability distribution over the set A of actions. In general, a strategy depends on the history and uses randomization. A strategy that depends only on the current state is a *memoryless or stationary* strategy, and can be represented as a function $\sigma : S \to \mathcal{D}(A)$. A strategy that does not use randomization is a *pure* strategy, i.e., for all histories $\langle s_0, s_1, \ldots, s_k \rangle$ there exists $a \in A$ such that $\sigma(\langle s_0, s_1, \ldots, s_k\rangle)(a) = 1$. A *pure memoryless* strategy is both pure and memoryless and can be represented as a function $\sigma : S \to A$. We denote by Σ, Σ^M, Σ^P and Σ^{PM} the set of all strategies, all memoryless strategies, all pure strategies and all pure memoryless strategies, respectively.

Outcomes. Given a strategy σ and an initial state s, we denote by $\mathrm{Outcome}(s, \sigma)$ the set of possible plays that start from s, given strategy σ, i.e., $\mathrm{Outcome}(s, \sigma) = \{ \langle s_0, s_1, \ldots, s_k, \ldots\rangle \mid \forall k \geq 0. \exists a_k \in A. \sigma(\langle s_0, s_1, \ldots, s_k\rangle)(a_k) > 0;\ \text{ and } s_{k+1} \in \mathrm{Dest}(s_k, a_k) \}$. Once the initial state and a strategy is chosen, the MDP is reduced to a stochastic process. We denote by X_i and θ_i random variables for the i-th state and the i-th chosen action in this stochastic process. An event is a measurable subset of $\mathrm{Outcome}(s, \sigma)$, and the probabilities of the events are uniquely defined. Given a strategy σ, an initial state s, and an event $\mathcal{A}$, we denote by $\mathrm{Pr}_s^\sigma(\mathcal{A})$ the probability that a path belongs to $\mathcal{A}$, when the MDP starts in state s and the strategy σ is used. For a measurable function f that maps paths to reals, we write $\mathbb{E}_s^\sigma[f]$ for the expected value of f when the MDP starts in state s and the strategy σ is used.

Rewards and objectives. Let $r : S \times A \to \mathbb{R}$ be a reward function that associates with every state and action a real-valued reward. For a reward function r the *inf-long-run* average value is defined as follows: for a strategy σ and an initial state s we have

$$Val_{inf}^{\sigma}(r, s) = \lim\inf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_s^\sigma[r(X_t, \theta_t)].$$

We will also consider the *sup-long-run* average value that is defined as follows: for a strategy σ and an initial state s we have

$$Val_{sup}^{\sigma}(r, s) = \lim\sup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}_s^\sigma[r(X_t, \theta_t)].$$

We consider MDPs with k-different reward functions $r_1, r_2, \ldots, r_k$. Given an initial state s, a strategy σ, the inf-long-run average value vector at s for σ, for $\boldsymbol{r} = \langle r_1, r_2, \ldots, r_k\rangle$ is defined as $Val_{inf}^{\sigma}(\boldsymbol{r}, s) = \langle Val_{inf}^{\sigma}(r_1, s), Val_{inf}^{\sigma}(r_2, s), \ldots, Val_{inf}^{\sigma}(r_k, s)\rangle$. The notation for sup-long-run average objectives is similar.

Comparison operators on vectors are interpreted in a point-wise fashion, i.e., given two real-valued vectors $\boldsymbol{v}_1 = \langle v_1^1, v_1^2, \ldots, v_1^k\rangle$ and $\boldsymbol{v}_2 = \langle v_2^1, v_2^2, \ldots, v_2^k\rangle$, and $\bowtie \in \{<, \leq, =\}$ we write $\boldsymbol{v}_1 \bowtie \boldsymbol{v}_2$ if and only if for all $1 \leq i \leq k$ we have $v_1^i \bowtie v_2^i$. We write $\boldsymbol{v}_1 \neq \boldsymbol{v}_2$ to denote that vector $\boldsymbol{v}_1$ is not equal to $\boldsymbol{v}_2$, i.e., it is not the case that $\boldsymbol{v}_1 = \boldsymbol{v}_2$.

Pareto-optimal strategies. Given an MDP G and reward functions $r_1, r_2, \ldots, r_k$, a strategy σ is a *Pareto-optimal* strategy [8] for inf-long-run average objective from a state s, if there is no $\sigma' \in \Sigma$ such that $Val_{inf}^{\sigma}(\boldsymbol{r}, s) \leq Val_{inf}^{\sigma'}(\boldsymbol{r}, s)$, and $Val_{inf}^{\sigma}(\boldsymbol{r}, s) \neq Val_{inf}^{\sigma'}(\boldsymbol{r}, s)$, i.e., there is no strategy σ' such that for all $1 \leq j \leq k$, we have $Val_{inf}^{\sigma}(r_j, s) \leq Val_{inf}^{\sigma'}(r_j, s)$ and exists $1 \leq j \leq k$, with $\mathit{Val}^{\sigma}_{\mathit{inf}}(r_j, s) < \mathit{Val}^{\sigma'}_{\mathit{inf}}(r_j, s)$. The definition for sup-long-run average objectives is similar. In case $k = 1$, the class of Pareto-optimal strategies are called optimal strategies.

Sufficiency of strategies. Given reward functions $r_1, r_2, \ldots, r_k$, a family $\Sigma^{\mathcal{C}}$ of strategies suffices for ε-Pareto optimality for inf-long-run average objectives if for all $\varepsilon > 0$, for every Pareto-optimal strategy $\sigma \in \Sigma$, there is a strategy $\sigma_c \in \Sigma^{\mathcal{C}}$ such that for all $j = 1, 2, \ldots, k$ and all $s \in S$ we have $\mathit{Val}^{\sigma}_{\mathit{inf}}(r_j, s) \leq \mathit{Val}^{\sigma_c}_{\mathit{inf}}(r_j, s) + \varepsilon$. The notion of sufficiency for Pareto optimality is obtained if the above inequality is satisfied for $\varepsilon = 0$. The definition is similar for sup-long-run average objectives.

Theorem 1 (Strategies for optimality [4]). *In MDPs with one reward function r_1, the family Σ^{PM} of pure memoryless strategies suffices for optimality for inf-long-run average and sup-long-run average objectives, i.e., there exists a pure memoryless strategy $\sigma^* \in \Sigma^{PM}$, such that for all strategies $\sigma \in \Sigma$, the following conditions hold:*

$$\mathit{Val}^{\sigma}_{\mathit{inf}}(r_1, s) \leq \mathit{Val}^{\sigma^*}_{\mathit{inf}}(r_1, s); \qquad \mathit{Val}^{\sigma}_{\mathit{sup}}(r_1, s) \leq \mathit{Val}^{\sigma^*}_{\mathit{sup}}(r_1, s);$$

and moreover, $\mathit{Val}^{\sigma^*}_{\mathit{inf}}(r_1, s) = \mathit{Val}^{\sigma^*}_{\mathit{sup}}(r_1, s)$.

3 Memoryless Strategies Suffice for Pareto Optimality

In this section we study the properties of the family of strategies that suffices for Pareto optimality. We start with a simple result that pure memoryless Pareto-optimal strategies exist, and then show that pure strategies do not capture all Pareto-optimal strategies.

Proposition 1. *Given an MDP G with reward functions $r_1, r_2, \ldots, r_k$, there exist pure memoryless Pareto-optimal strategies for inf-long-run average and sup-long-run average objectives.*

Proof. Given reward functions $r_1, r_2, \ldots, r_k$, consider a reward function $r_+ = r_1 + r_2 + \ldots + r_k$, i.e., for all $s \in S$ and $a \in A$ we have $r_+(s, a) = r_1(s, a) + r_2(s, a) + \ldots + r_k(s, a)$. Let $\sigma^* \in \Sigma^{PM}$ be a pure memoryless optimal strategy for the reward function r_+ with the inf-long-run average objective (such a strategy exists by Theorem 1). We show that σ^* is Pareto-optimal. Assume towards contradiction that σ^* is not a Pareto-optimal strategy, then let $\sigma \in \Sigma$ be such that $\mathit{Val}^{\sigma^*}_{\mathit{inf}}(\boldsymbol{r}, s) \leq \mathit{Val}^{\sigma}_{\mathit{inf}}(\boldsymbol{r}, s)$, and for some j, $\mathit{Val}^{\sigma^*}_{\mathit{inf}}(r_j, s) < \mathit{Val}^{\sigma}_{\mathit{inf}}(r_j, s)$. Then we have $\mathit{Val}^{\sigma^*}_{\mathit{inf}}(r_+, s) = \sum_{j=1}^{k} \mathit{Val}^{\sigma^*}_{\mathit{inf}}(r_j, s) < \sum_{j=1}^{k} \mathit{Val}^{\sigma}_{\mathit{inf}}(r_j, s) = \mathit{Val}^{\sigma}_{\mathit{inf}}(r_+, s)$. This contradicts that σ^* is optimal for r_+. The proof for sup-long-run average objectives is similar. ∎

The above proof can be generalized to any convex combination of the multiple objectives, that is, for positive weights $w_1, \ldots, w_k$, the pure memoryless optimal strategy for the single objective $\sum_i w_i r_i$ is Pareto-optimal. This technique

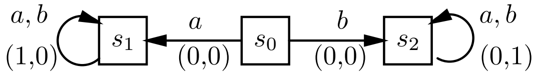

Fig. 1. MDP for Example 1

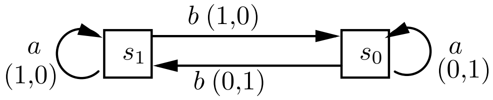

Fig. 2. MDP for Example 2

is called the *weighted factor method* [6, 12], and used commonly in engineering practice to find subsets of the Pareto set [7]. However, not all Pareto-optimal points are obtained in this fashion, as the following example shows that randomization is necessary.

Example 1. Consider the MDP from Fig. 1, with two actions a and b, and two reward functions r_1 and r_2. The transitions and the respective rewards are shown as labeled edges in the figure. Consider the inf-long-run average objectives for reward functions r_1 and r_2. For the pure memoryless strategies (and also the pure strategies) in this MDP, the possible value vectors are $(1,0)$ and $(0,1)$. However, consider a memoryless strategy σ_m that at state s_0 plays action a and b each with probability $1/2$. For $\boldsymbol{r} = (r_1, r_2)$ we have $Val_{inf}^{\sigma_m}(\boldsymbol{r}, s_0) = (\frac{1}{2}, \frac{1}{2})$. The strategy σ_m is Pareto-optimal and no pure memoryless strategy can achieve the corresponding value vector. Hence it follows that the family of pure strategies and pure memoryless strategies do not suffice for Pareto optimality. Note that for all $0 < x < 1$, the memoryless strategy that plays a with probability x is a Pareto-optimal strategy, with value vector $(x, 1-x)$. Hence the set of Pareto-optimal value vectors can be uncountable and may have irrational values. ∎

The following example shows that for sup-long-run average objectives the family of memoryless strategies does not suffice for ε-Pareto optimality. Then we present the main result of this section that shows the family of memoryless strategies suffices for ε-Pareto optimality for inf-long-run average objectives.

Example 2. Fig. 2 shows an MDP with two actions a and b, and two reward functions r_1 and r_2. The transitions and the respective rewards are shown as labeled edges in the figure. Consider the sup-long-run average objectives for r_1 and r_2. Given any memoryless strategy σ_m we have $Val_{sup}^{\sigma_m}(r_1, s_0) + Val_{sup}^{\sigma_m}(r_2, s_0) = 1$. We now consider a strategy σ as follows: the strategy σ is played in rounds. In round j, it first goes to state s_1, plays action a (i.e., stays in s_1), unless the average for reward r_1 is at least $1 - \frac{1}{j}$, then it goes to state s_0, plays action a unless the average reward for reward r_2 is at least $1 - \frac{1}{j}$, then it proceeds to round $j+1$. Given σ, we have $Val_{sup}^{\sigma}(r_1, s_0) = 1$ and $Val_{sup}^{\sigma}(r_2, s_0) = 1$. There is no memoryless Pareto-optimal strategy to achieve this value vector, or even approximate by $0 < \varepsilon < \frac{1}{2}$. ∎

Markov chains. A *Markov chain* $G = (S, p)$ consists of a finite set S of states, and a stochastic transition matrix p, i.e., $p(s, t) \geq 0$ denotes the transition probability from s to t, and for all $s \in S$ we have $\sum_{t \in S} p(s, t) = 1$. Given an MDP $G = (S, A, p)$ and a memoryless strategy $\sigma \in \Sigma^M$ we obtain a Markov chain $G_\sigma = (S, p_\sigma)$ obtained as follows: $p_\sigma(s, t) = \sum_{a \in A} p(s, a)(t) \cdot \sigma(s)(a)$. From Theorem 1 it follows that the values for inf-long-run average and sup-long-run average objectives coincide for Markov chains. Thus we have the following corollary.

Corollary 1. *For all MDPs $G = (S, A, p)$, for all reward functions r_1, for all memoryless strategies $\sigma \in \Sigma^M$, and for all $s \in S$, we have $Val^\sigma_{inf}(r_1, s) = Val^\sigma_{sup}(r_1, s)$.*

We now state a result of Hardy-Littlewood from analysis (see Appendix H of [4] for proof).

Lemma 1 (Hardy-Littlewood result). *Let $\{ d_t \}_{t=0}^\infty$ be an arbitrary sequence of bounded real-numbers. Then the following assertions hold:*

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} d_t \leq \liminf_{\beta \to 1^-} (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot d_t$$
$$\leq \limsup_{\beta \to 1^-} (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot d_t \leq \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} d_t.$$

Lemma 2. *Let $G = (S, A, p)$ be an MDP with k reward functions $r_1, r_2, \ldots, r_k$. For all $\varepsilon > 0$, for all $s \in S$, for all $\sigma \in \Sigma$, there exists a memoryless strategy $\overline{\sigma} \in \Sigma^M$ such that for all $i = 1, 2, \ldots, k$, we have $Val^\sigma_{inf}(r_i, s) \leq Val^{\overline{\sigma}}_{inf}(r_i, s) + \varepsilon$.*

Proof. Given a strategy σ and an initial state s, for $j = 1, 2, \ldots, k$ define a sequence $\{ d^j_t \}_{t=0}^\infty$ as follows: $d^j_t = \mathbb{E}^\sigma_s[r_j(X_t, \theta_t)]$; i.e., d^j_t is the expected reward of the t-th stage for the reward function r_j. The sequence $\{ d^j_t \}_{t=0}^\infty$ is bounded as follows: $\min_{s \in S, a \in A} r_j(s, a) \leq d^j_t \leq \max_{s \in S, a \in A} r_j(s, a)$, for all $t \geq 0$ and for all $j = 1, 2, \ldots, k$. By Lemma 1 we obtain that for all $\varepsilon > 0$, there exists $0 < \beta < 1$ such that for all $j = 1, 2, \ldots, k$ we have

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} d^j_t \leq (1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot d^j_t + \varepsilon;$$

i.e., in other words, for all $j = 1, 2, \ldots, k$ we have

$$Val^\sigma_{inf}(r_j, s) \leq \mathbb{E}^\sigma_s[\sum_{t=0}^{\infty} (1 - \beta) \cdot \beta^t \cdot r_j(X_t, \theta_t)] + \varepsilon.$$

By Theorem 2 of [1] for every strategy σ, there is a memoryless strategy $\overline{\sigma} \in \Sigma^M$ such that for all $j = 1, 2, \ldots, k$ we have

$$\mathbb{E}^\sigma_s[\sum_{t=0}^{\infty} (1 - \beta) \cdot \beta^t \cdot r_j(X_t, \theta_t)] = \mathbb{E}^{\overline{\sigma}}_s[\sum_{t=0}^{\infty} (1 - \beta) \cdot \beta^t \cdot r_j(X_t, \theta_t)].$$

Consider a memoryless strategy $\overline{\sigma}$ that satisfies the above equalities for $j = 1, 2, \ldots, k$. For $j = 1, 2, \ldots, k$ define a sequence $\{ \overline{d}_t^j \}_{t=0}^{\infty}$ as follows: $\overline{d}_t^j = \mathbb{E}_s^{\overline{\sigma}}[r_j(X_t, \theta_t)]$. Again the sequence $\{ \overline{d}_t^j \}_{t=0}^{\infty}$ is bounded as follows: $\min_{s \in S, a \in A} r_j(s, a) \leq \overline{d}_t^j \leq \max_{s \in S, a \in A} r_j(s, a)$, for all $t \geq 0$ and for all $j = 1, 2, \ldots, k$. By Lemma 1 for all $j = 1, 2, \ldots, k$ we obtain that

$$(1 - \beta) \cdot \sum_{t=0}^{\infty} \beta^t \cdot \overline{d}_t^j \leq \lim \sup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \overline{d}_t^j;$$

i.e., for all $j = 1, 2, \ldots, k$ we have $\mathbb{E}_s^{\overline{\sigma}}[\sum_{t=0}^{\infty}(1-\beta) \cdot \beta^t \cdot r_j(X_t, \theta_t)] \leq Val_{sup}^{\overline{\sigma}}(r_j, s)$. Since $\overline{\sigma}$ is a memoryless strategy, by Corollary 1 we obtain that for all $j = 1, 2, \ldots, k$ we have $Val_{sup}^{\overline{\sigma}}(r_j, s) = Val_{inf}^{\overline{\sigma}}(r_j, s)$. Hence it follows that for all $j = 1, 2, \ldots, k$ we have $Val_{inf}^{\sigma}(r_j, s) \leq Val_{inf}^{\overline{\sigma}}(r_j, s) + \varepsilon$. The desired result follows. ∎

Theorem 2 follows from Lemma 2.

Theorem 2. *The family of Σ^M of memoryless strategies suffices for ε-Pareto optimality for inf-long-run average objectives.*

It follows from Example 2 that Lemma 2 does not extend to sup-long-run average objectives. The witness counter-example strategy in Example 2 is an infinite-memory strategy. The result of Lemma 2 extends to sup-long-run average objectives for finite-memory strategies. This is shown in the following lemma.

Lemma 3. *Let $G = (S, A, p)$ be an MDP with k reward functions $r_1, r_2, \ldots, r_k$. For all $\varepsilon > 0$, for all $s \in S$, for all finite-memory strategies $\sigma \in \Sigma$, there exists a memoryless strategy $\overline{\sigma} \in \Sigma^M$ such that for all $i = 1, 2, \ldots, k$, we have $Val_{sup}^{\sigma}(r_i, s) \leq Val_{sup}^{\overline{\sigma}}(r_i, s) + \varepsilon$.*

Proof. (Sketch). Given a finite-memory strategy σ, for all $j = 1, 2, \ldots, k$, and for all $s \in S$, we have $Val_{sup}^{\sigma}(r_j, s) = Val_{inf}^{\sigma}(r_j, s)$. This follows by considering the finite state Markov chain G_σ arising from fixing the strategy σ in G. The result then follows from Lemma 2. ∎

4 Approximating the Pareto Curve

Pareto curve. Let G be an MDP with reward functions $\boldsymbol{r} = \langle r_1, \ldots, r_k \rangle$. The *Pareto curve* $P^{\inf}(G, s, \boldsymbol{r})$ of the MDP G at state s with respect to inf-long-run average objectives is the set of all k-vectors of values such that for each $\boldsymbol{v} \in P^{\inf}(G, s, \boldsymbol{r})$, there is a Pareto-optimal strategy σ such that $Val_{inf}^{\sigma}(\boldsymbol{r}, s) = \boldsymbol{v}$. We are interested not only in the values, but also the Pareto-optimal strategies. We often blur the distinction and refer to the Pareto curve $P^{\inf}(G, s, \boldsymbol{r})$ as a set of strategies which achieve the Pareto-optimal values (if there is more than one strategy that achieves the same value vector, $P^{\inf}(G, s, \boldsymbol{r})$ contains at least one of them). For an MDP G, and $\varepsilon > 0$, an ε-approximate Pareto curve, denoted $P^{\inf}_{\varepsilon}(G, s, \boldsymbol{r})$, is a set of strategies σ such that there is no other strategy σ' such that for all $\sigma \in P^{\inf}_{\varepsilon}(G, s, \boldsymbol{r})$, we have $\mathit{Val}^{\sigma'}_{\mathit{inf}}(r_i, s) \geq (1 + \varepsilon)\, \mathit{Val}^{\sigma}_{\mathit{inf}}(r_i, s)$, for all rewards r_i. That is, the ε-approximate Pareto curve contains strategies such that any Pareto-optimal strategy is "almost" dominated by some strategy in $P^{\inf}_{\varepsilon}(G, s, \boldsymbol{r})$.

Multi-objective linear programming and Pareto curve. A multi-objective linear program L consists of a set k of objective functions $o_1, o_2, \ldots, o_k$, where $o_i(x) = c_i^T \cdot x$, for a vector c_i and a vector x of variables; and a set of linear constraints specified as $A \cdot x \geq b$, for a matrix A and a vector b. A valuation of x is feasible if it satisfies the set of linear constraints. A feasible solution x is a Pareto-optimal point if there is no other feasible solution x' such that $(o_1(x), o_2(x), \ldots, o_k(x)) \leq (o_1(x'), o_2(x'), \ldots, o_k(x'))$ and $(o_1(x), o_2(x), \ldots, o_k(x)) \neq (o_1(x'), o_2(x'), \ldots, o_k(x'))$. Given a multi-objective linear program L, the Pareto curve for L consists of the k-vector of values such that for each $\boldsymbol{v} \in P(L)$ there is a Pareto-optimal point x such that $\boldsymbol{v} = (o_1(x), o_2(x), \ldots, o_k(x))$. The definition of ε-approximate Pareto curve $P_{\varepsilon}(L)$ for L is similar to the definitions of the curves as defined above. The following theorem is a direct consequence of the corresponding theorems in [9].

Theorem 3 ([9]). *Given a multi-objective linear program L with k-objective functions, the following assertions hold:*

1. *For all $\varepsilon > 0$, there exists an approximate Pareto curve $P_{\varepsilon}(L)$ consisting of a number of feasible solution that is polynomial in $|L|$ and $\frac{1}{\varepsilon}$, but exponential in the number of objective functions.*
2. *For all $\varepsilon > 0$, there is an algorithm to construct $P_{\varepsilon}(L)$ in time polynomial in $|L|$ and $\frac{1}{\varepsilon}$ and exponential in the number of objective functions.*

Proof. The result of part 1 is a direct consequence of Theorem 1 of [9]. The result of part 2 follows from Theorem 3 of [9] and the fact that linear programming can be solved in polynomial time. ∎

4.1 Irreducible MDPs

In this subsection we consider a special class of MDPs, namely, *irreducible* MDPs[1] and present algorithm to approximate the Pareto curve by reduction to multi-objective linear-programming.

Irreducible MDPs. An MDP G is *irreducible* if for every pure memoryless strategy $\sigma \in \Sigma^{PM}$ the Markov chain G_σ is completely ergodic (or irreducible), i.e., the graph of G_σ is a strongly connected component. Observe that if G is an irreducible MDP, then for all memoryless strategy $\sigma \in \Sigma^{M}$, the Markov chain G_σ is completely ergodic.

[1] see section 2.4 of [4] for irreducible MDPs with a single reward function.

Long-run frequency. Let G be an irreducible MDP, and $\sigma \in \Sigma^M$ be a memoryless strategy. Let

$$q(s,\sigma)(u) = \lim_{T\to\infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}_s^{\sigma}[\mathbf{1}_{X_t=u}],$$

where $\mathbf{1}_{X_t=u}$ is the indicator function denoting if the t-th state is u, denote the "long-run average frequency" of state u, and let $x_{ua} = q(s,\sigma)(u) \cdot \sigma(u)(a)$ be the "long-run average frequency" of the state action pair (u,a). It follows from the results of [4] (see section 2.4) that $q(s,\sigma)(u)$ exists and is positive for all states $u \in S$, and x_{ua} satisfies the following set of linear-constraints:

$$\begin{aligned}
&(i) \sum_{u\in S}\sum_{a\in A} \big(\delta(u,u') - p(u,a)(u')\big) \cdot x_{ua} = 0; \qquad u' \in S;\\
&(ii) \sum_{u\in S}\sum_{a\in A} x_{ua} = 1;\\
&(iii)\ x_{ua} \geq 0; \qquad a \in A, u \in S,
\end{aligned}$$

where $\delta(u,u')$ is the Kronecker delta. We denote the above set of constraints by $C_{\mathsf{irr}}(G)$.

Multi-objective linear-program. Let $G = (S,A,p)$ be an irreducible MDP with k reward functions $r_1, r_2, \ldots, r_k$. We consider the following multi-objective linear-program over the variables x_{ua} for $u \in S$ and $a \in A$. The k-objectives are as follows:

$$\max \sum_{u\in S}\sum_{a\in A} r_j(u,a) \cdot x_{ua}; \qquad \text{for } j = 1,2,\ldots,k;$$

and the set of linear-constraints are specified as $C_{\mathsf{irr}}(G)$. We denote the above multi-objective linear-program as $L_{\mathsf{irr}}(G,\boldsymbol{r})$.

Lemma 4. *Let $G = (S,A,p)$ be an irreducible MDP, with k reward functions $r_1, r_2, \ldots, r_k$. Let $\boldsymbol{v} \in \mathbb{R}^k$ be a vector of real-values. Then the following statements are equivalent.*

1. *There is a memoryless strategy $\sigma \in \Sigma^M$ such that*

$$\wedge_{j=1}^{k}\big(\mathit{Val}_{\mathit{inf}}^{\sigma}(r_j,s) \geq v_j\big).$$

2. *There is a feasible solution x_{ua} for multi-objective linear-program $L_{\mathsf{irr}}(G,\boldsymbol{r})$ such that*

$$\wedge_{j=1}^{k}\big(\sum_{u\in S}\sum_{a\in A} r_j(u,a) \cdot x_{ua} \geq v_j\big).$$

Proof. We prove both the directions as follows.

1. (1). ⇒ (2). Given a memoryless strategy σ, let

$$x_{ua} = \sigma(u)(a) \cdot \lim_{T\to\infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}_s^{\sigma}[\mathbf{1}_{X_t=u}].$$

Then x_{ua} is a feasible solution to $L_{\mathsf{irr}}(G, \boldsymbol{r})$. Moreover, the value for the inf-long-run average objective can be expressed as follows:

$$Val^{\sigma}_{inf}(r_j, s) = \sum_{u \in S} \sum_{a \in A} \sigma(u)(a) \cdot r_j(u,a) \cdot \lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}^{\sigma}_{s}[\mathbf{1}_{X_t = u}].$$

The desired result follows.

2. (2). ⇒ (1). Let x_{ua} be a feasible solution to $L_{\mathsf{irr}}(G, \boldsymbol{r})$. Consider the memoryless strategy σ defined as follows:

$$\sigma(u)(a) = \frac{x_{ua}}{\sum_{a' \in A} x_{ua'}}.$$

Given the memoryless strategy σ, it follows from Lemma 2.4.2 and Theorem 2.4.3 of [4] that

$$x_{ua} = \sigma(u)(a) \cdot \lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}^{\sigma}_{s}[\mathbf{1}_{X_t = u}].$$

The desired result follows. ∎

It follows from Lemma 4 that the Pareto curve $P(L_{\mathsf{irr}}(G, \boldsymbol{r}))$ characterizes the set of memoryless Pareto-optimal points for the MDP with k inf-long-run average objectives. Since memoryless strategies suffices of ε-Pareto optimality for inf-long-run average objectives (Theorem 2), the following result follows from Theorem 3.

Theorem 4. *Given an irreducible MDP G with k reward functions $\boldsymbol{r}$, for all $\varepsilon > 0$, there is an algorithm to construct a $P^{\inf}_{\varepsilon}(G, s, \boldsymbol{r})$ in time polynomial in $|G|$ and $\frac{1}{\varepsilon}$ and exponential in the number of reward functions.*

4.2 General MDPs

In the case of general MDPs, if we fix a memoryless strategy $\sigma \in \Sigma^M$, then in the resulting Markov chain G_σ, in general, we have both recurrent states and transient states. For recurrent states the "long-run-average frequency" is positive and for transient states the "long-run-average frequency" is zero. For the transient states the strategy determines the probabilities to reach the various closed connected set of recurrent states. We will obtain several multi-objective linear-programs to approximate the Pareto curve: the set of constraints for recurrent states will be obtained similar to the one of $C_{\mathsf{irr}}(G)$, and the set of constraints for the transient states will be obtained from the results of [2] on multi-objective reachability objectives. We first define a partition of the set Σ^M of memoryless strategies.

Partition of strategies. Given an MDP $G = (S, A, p)$, consider the following set of functions: $\mathcal{F} = \{ f : S \to 2^A \setminus \emptyset \}$. The set $\mathcal{F}$ is finite, since $|\mathcal{F}| \leq 2^{|A| \cdot |S|}$.

Given $f \in \mathcal{F}$ we denote by $\Sigma^M \restriction f = \{ \sigma \in \Sigma^M \mid f(s) = \mathrm{Supp}(\sigma(s)), \forall s \in S \}$ the set of memoryless strategies σ such that support of $\sigma(s)$ is $f(s)$ for all states $s \in S$.

Multi-objective linear program for $f \in \mathcal{F}$. Let G be an MDP with reward functions $r_1, r_2, \ldots, r_k$. Let $f \in \mathcal{F}$, and we will present a multi-objective linear-program for memoryless strategies in $\Sigma^M \restriction f$. We first observe that for all $\sigma_1, \sigma_2 \in \Sigma^M \restriction f$, the underlying graph structure of the Markov chains G_{σ_1} and G_{σ_2} is the same, i.e., the recurrent set of states and transient set of states in G_{σ_1} and G_{σ_2} is the same. Hence the computation of the recurrent states and transient states for all strategies in $\Sigma^M \restriction f$ can be achieved by computing it for an arbitrary strategy in $\Sigma^M \restriction f$. Given G, the reward functions, an initial state s, and $f \in \mathcal{F}$, the multi-objective linear program is obtained by applying the following steps.

1. Consider the memoryless strategy $\overline{\sigma} \in \Sigma^M \restriction f$ that plays at u all actions in $f(u)$ uniformly at random, for all $u \in S$. Let U be the reachable subset of states in $G_{\overline{\sigma}}$ from s, and let $\mathcal{R} = \{ R_1, R_2, \ldots, R_l \}$ be the set of closed connected recurrent set of states in $G_{\overline{\sigma}}$, i.e., R_i is a bottom strongly connected component in the graph of $G_{\overline{\sigma}}$. The set U and $\mathcal{R}$ can be computed in linear-time. Let $R = \bigcup_{i=1}^{l} R_i$, and the set $U \setminus R$ consists of transient states.

2. If $s \in R$, then consider R_i such that $s \in R_i$. In the present case, consider the multi-objective linear-program of subsection 4.1 with the additional constraint that $x_{ua} > 0$, for all $u \in R_i$ and $a \in f(u)$, and $x_{ua} = 0$ for all $u \in R_i$ and $a \notin f(u)$. The Pareto curve of the above multi-objective linear-program coincides with the Pareto curve for memoryless strategies in $\Sigma^M \restriction f$. The proof essentially mimics the proof of Lemma 4 restricted to the set R_i.

3. We now consider the case when $s \in U \setminus R$. In this case we will have three kinds of variables: (a) variables x_{ua} for $u \in R$ and $a \in A$; (b) variables y_{ua} for $u \in U \setminus R$ and $a \in A$ (c) variables y_u for $u \in R$. Intuitively, the variables x_{ua} will denote the "long-run average frequency" of the state action pair x_{ua}, and the variables y_{ua} and y_u will play the same role as the variables of the multi-objective linear-program of [2] for reachability objectives (see Fig 3 of [2]). We now specify the multi-objective linear-program

$$\textbf{Objectives } (j = 1, 2, \ldots, k): \qquad \max \sum_{u \in S} \sum_{a \in A} r_j(u, a) \cdot x_{ua};$$

Subject to

$(i)\ \sum_{u\in R_i}\sum_{a\in A}\big(\delta(u,u')-p(u,a)(u')\big)\cdot x_{ua}=0; \qquad u'\in R_i;$

$(ii)\ \sum_{u\in R}\sum_{a\in A}x_{ua}=1;$

$(iii)\ x_{ua}\geq 0; \qquad a\in A, u\in R;$

$(iv)\ x_{ua}>0; \qquad a\in f(u), u\in R;$

$(v)\ x_{ua}=0; \qquad a\not\in f(u), u\in R;$

$(vi)\ \sum_{a\in A}y_{ua}-\sum_{u'\in U}\sum_{a'\in A}p(u',a')(u)\cdot y_{u'a'}=\alpha(u); \qquad u\in U\setminus R;$

$(vii)\ y_u-\sum_{u'\in U\setminus R}\sum_{a'\in A}p(u',a')(u)\cdot y_{u'a'}=0; \qquad u\in R;$

$(viii)\ y_{ua}\geq 0; \qquad u\in U\setminus R, a\in A;$

$(ix)\ y_u\geq 0; \qquad u\in R;$

$(x)\ y_{ua}>0; \qquad u\in U\setminus R, a\in f(u);$

$(xi)\ y_{ua}=0; \qquad u\in U\setminus R, a\not\in f(u);$

$(xii)\ \sum_{u\in R_i}\sum_{a\in A}x_{ua}=\sum_{u\in R_i}y_u; \qquad i=1,2,\ldots,l;$

where $\alpha(u)=1$ if $u=s$ and 0 otherwise. We refer the above set of constraints as $C_{\mathsf{gen}}(G,\boldsymbol{r},f)$ and the above multi-objective linear-program as $L_{\mathsf{gen}}(G,\boldsymbol{r},f)$. We now explain the role of each constraint: the constraints $(i)-(iii)$ coincides with constraints $C_{\mathsf{irr}}(G)$ for the subset R_i, and the additional constraints $(iv)-(v)$ are required to ensure that we have witness strategies such that they belong to $\Sigma^M\restriction f$. The constraints $(vi)-(ix)$ are essentially the constraints of the multi-objective linear-program for reachability objectives defined in Fig 3 of [2]. The additional constraints $(x)-(xi)$ are again required to ensure that witness strategies satisfy that they belong to $\Sigma^M\restriction f$. Intuitively, for $u\in R_i$, the variables y_u stands for the probability to hit u before hitting any other state in R_i. The last constraint specify that the sum total of "long-run average frequency" in a closed connected recurrent set R_i coincides with the probability to reach R_i. We remark that the above constraints can be simplified; e.g., the (iv) and (v) implies (iii), but we present the set constraints in a way such that it can be understood that what new constraints are introduced.

Lemma 5. *Let $G=(S,A,p)$ be an MDP, with k reward functions $r_1,r_2,\ldots,r_k$. Let $\boldsymbol{v}\in\mathbb{R}^k$ be a vector of real-values. Then the following statements are equivalent.*

1. *There is a memoryless strategy $\sigma\in\Sigma^M\restriction f$ such that*

$$\wedge_{j=1}^k\big(\mathit{Val}_{\mathit{inf}}^{\sigma}(r_j,s)\geq v_j\big).$$

2. *There is a feasible solution for the multi-objective linear-program* $L_{\mathsf{gen}}(G, \boldsymbol{r}, f)$ *such that*

$$\wedge_{j=1}^{k} \big(\sum_{u \in S} \sum_{a \in A} r_j(u,a) \cdot x_{ua} \geq v_j \big).$$

Proof. The case when the starting s is a member of the set R of recurrent states, the result follows from Lemma 4. We consider the case when $s \in U \setminus R$. We prove both the directions as follows.

1. (1). ⇒ (2). Let $\sigma \in \Sigma^M \restriction f$ be a memoryless strategy. We now construct a feasible solution for $L_{\mathsf{gen}}(G, \boldsymbol{r}, f)$. For $u \in R$, let

   $$x'_{ua} = \sigma(u)(a) \cdot \lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}_s^{\sigma}[\mathbf{1}_{X_t = u}].$$

   Consider a square matrix P^σ of size $|U \setminus R| \times |U \setminus R|$, defined as follows: $P^\sigma_{u,u'} = \sum_{a \in A} \sigma(u)(a) \cdot p(u,a)(u')$, i.e., P^σ is the one-step transition matrix under p and σ. For all $u \in U \setminus R$, let $y'_{ua} = \sigma(u)(a) \cdot \sum_{n=0}^{\infty} (P^\sigma)^n_{s,u}$. In other words, y'_{ua} denotes "the expected number of times of visiting u and upon doing so choosing action a, given the strategy σ and starting state s". Since states in $U \setminus R$ are transient states, the values y'_{ua} are finite (see Lemma 1 of [2]). For $u \in R$, let $y'_u = \sum_{u' \in U \setminus R} \sum_{a' \in A} p(u',a')(u) \cdot y'_{u'a'}$, i.e., y'_u is the "expected number of times that we will transition into state u for the first time". It follows from arguments similar to Lemma 4 and the results in [2] that above solution is feasible solution to the linear-program $L_{\mathsf{gen}}(G, \boldsymbol{r}, f)$. Moreover, $\sum_{u \in R_i} y'_u = \Pr_s^\sigma(\Diamond R_i)$, for all R_i, where $\Diamond R_i$ denotes the event of reaching R_i. It follows that for all $j = 1, 2, \ldots, k$ we have $Val^{\sigma}_{inf}(r_j, s) = \sum_{u \in R} \sum_{a \in A} r_j(u,a) \cdot x'_{ua}$. The desired result follows.
2. (2). ⇒ (1). Given a feasible solution to $L_{\mathsf{gen}}(G, \boldsymbol{r}, f)$ we construct a memoryless strategy $\sigma \in \Sigma^M \restriction f$ as follows:

   $$\sigma(u)(a) = \begin{cases} \frac{x_{ua}}{\sum_{a' \in A} x_{ua'}} & u \in R; \\ \frac{y_{ua}}{\sum_{a' \in A} y_{ua'}} & u \in U \setminus R; \end{cases}$$

   Observe the constraints $(iv) - (v)$ and $(x) - (xi)$ ensure that the strategy $\sigma \in \Sigma^M \restriction f$. The strategy constructed satisfies the following equalities: for all R_i we have $\Pr_s^\sigma(\Diamond R_i) = \sum_{u \in R_i} y_u$ (this follows from Lemma 2 of [2]); and for all $u \in R_i$ we have

   $$x_{ua} = \sigma(u)(a) \cdot \lim_{T \to \infty} \frac{1}{T} \cdot \sum_{t=0}^{T-1} \mathbb{E}_s^{\sigma}[\mathbf{1}_{X_t = u}].$$

   The above equality follows from arguments similar to Lemma 4. The desired result follows. ∎

Theorem 5. *Given an MDP G with k reward functions $\boldsymbol{r}$, for all $\varepsilon > 0$, there is an algorithm to construct a $P_{\varepsilon}^{\inf}(G, s, \boldsymbol{r})$ in (a) time polynomial in $\frac{1}{\varepsilon}$, and exponential in $|G|$ and the number of reward functions; (b) using space polynomial in $\frac{1}{\varepsilon}$ and $|G|$, and exponential in the number of reward functions.*

Proof. It follows from Lemma 5 that the Pareto curve $P(L_{\mathsf{gen}}(G, \boldsymbol{r}, f))$ characterizes the set of memoryless Pareto-optimal points for the MDP with k inf-long-run average objectives for all memoryless strategies in $\Sigma^M \restriction f$. We can generate all $f \in \mathcal{F}$ in space polynomial in $|G|$ and time exponential in $|G|$. Since memoryless strategies suffices of ε-Pareto optimality for inf-long-run average objectives (Theorem 2), the desired result follows from Theorem 3. ∎

4.3 Realizability

In this section we study the realizability problem for multi-objective MDPs: the *realizability problem* asks, given a multi-objective MDP G with rewards r_1, …, r_k (collectively, $\boldsymbol{r}$) and a state s of G, and a value profile $\mathbf{w} = (w_1, \ldots w_k)$ of k rational values, whether there exists a strategy σ such that $\mathit{Val}_{\mathit{inf}}^{\sigma}(\boldsymbol{r}, s) \geq \mathbf{w}$. Observe that such a strategy exists if and only if there is a Pareto-optimal strategy σ' such that $\mathit{Val}_{\mathit{inf}}^{\sigma'}(\boldsymbol{r}, s) \geq \mathbf{w}$. Also observe that it follows from Theorem 2 that if a value profile $\mathbf{w}$ is realizable, then it is realizable within ε by a memoryless strategy, for all $\varepsilon > 0$. Hence we study the *memoryless realizability* problem that asks, given a multi-objective MDP G with rewards r_1, …, r_k (collectively, $\boldsymbol{r}$) and a state s of G, and a value profile $\mathbf{w} = (w_1, \ldots w_k)$ of k rational values, whether there exists a memoryless strategy σ such that $\mathit{Val}_{\mathit{inf}}^{\sigma}(\boldsymbol{r}, s) \geq \mathbf{w}$. The realizability problem arises when certain target behaviors are required, and one wishes to check if they can be attained on the model.

Theorem 6. *The memoryless realizability problem for multi-objective MDPs with inf-long-run average objectives can be (a) decided in polynomial time for irreducible MDPs; (b) decided in NP for MDPs.*

Proof. The result is obtained as follows.

1. For an irreducible MDP G with k reward functions $r_1, r_2, \ldots, r_k$, the answer to the memoryless realizability problem is "Yes" iff the following set of linear constraints has a solution. The set of constraints consists of the constraints $C_{\mathsf{irr}}(G)$ along with the constraints $\wedge_{j=1}^{k} \big(\sum_{s \in S} \sum_{a \in A} r_j(s, a) \cdot x_{ua} \geq w_j \big)$. Hence we obtain a polynomial time algorithm for the memoryless realizability problem.
2. For an MDP G with k reward functions $r_1, r_2, \ldots, r_k$, the answer to the memoryless realizability problem is "Yes" iff there exists $f \in \mathcal{F}$ such that the following set of linear constraints has a solution. The set of constraints consists of the constraints $C_{\mathsf{gen}}(G, \boldsymbol{r}, f)$ along with the constraints $\wedge_{j=1}^{k} \big(\sum_{s \in S} \sum_{a \in A} r_j(s, a) \cdot x_{ua} \geq w_j \big)$. Hence given the guess f, we have a polynomial time algorithm for verification. Hence the result follows. ∎

Concluding remarks. In this work we studied MDPs with multiple long-run average objectives: we proved ε-Pareto optimality of memoryless strategies for inf-long-run average objectives, and presented algorithms to approximate the Pareto-curve and decide realizability for MDPs with multiple inf-long-run average objectives. By Lemma 3 it follows that the algorithms for inf-long-run average objectives extend to sup-long-run average objectives if we restrict the set of strategies to finite-memory strategies. Recall that by Example 2 it follows that memoryless strategies do not suffice for ε-Pareto optimality for sup-long-run average objectives. The problem of approximating the Pareto curve and deciding the realizability problem for sup-long-run average objectives remain open. The other interesting open problems are as follows: (a) whether memoryless strategies suffices for Pareto optimality, rather than ε-Pareto optimality, for inf-long-run average objectives; (b) whether the problem of approximating the Pareto curve and deciding the realizability problem for general MDPs with inf-long-run average objectives can be solved in polynomial time.

References

1. K. Chatterjee, R. Majumdar and T.A. Henzinger. Markov decision processes with multiple objectives. In *STACS'06* LNCS 3884, pages 325–336, Springer, 2006.
2. K. Etessami, M. Kwiatkowska, M.Y. Vardi and M.Yannakakis. Multi-objective model checking of Markov decision processes. In *TACAS'07* LNCS 4424, Springer, 2007.
3. O. Etzioni, S. Hanks, T. Jiang, R.M. Karp, O. Madari, and O. Waarts. Efficient information gathering on the internet. In *FOCS 96*, pages 234–243. IEEE, 1996.
4. J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer, 1997.
5. M.R. Garey and D.S. Johnson. *Computers and Intractability*. W.H. Freeman, 1979.
6. R. Hartley. Finite discounted, vector Markov decision processes. Technical report, Department of Decision Theory, Manchester University, 1979.
7. J. Koski. Multicriteria truss optimization. In *Multicriteria Optimization in Engineering and in the Sciences*. 1988.
8. G. Owen. *Game Theory*. Academic Press, 1995.
9. C.H. Papadimitriou and M. Yannakakis. On the approximability of trade-offs and optimal access of web sources. In *FOCS 00*, pages 86–92. IEEE Press, 2000.
10. M.L. Puterman. *Markov Decision Processes*. John Wiley and Sons, 1994.
11. R. Szymanek, F. Catthoor, and K. Kuchcinski. Time-energy design space exploration for multi-layer memory architectures. In *DATE 04*. IEEE, 2004.
12. D.J. White. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 89(2):639–647, 1982.
13. P. Yang and F. Catthoor. Pareto-optimization based run time task scheduling for embedded systems. In *CODES-ISSS 03*, pages 120–125. ACM, 2003.