Markov Decision Processes with Multiple Objectives *

Krishnendu Chatterjee[1], Rupak Majumdar[2], and Thomas A. Henzinger[1,3]

[1]UC Berkeley [2]UC Los Angeles [3]EPFL
{c_krish,tah}@eecs.berkeley.edu rupak@cs.ucla.edu

Abstract. We consider Markov decision processes (MDPs) with multiple discounted reward 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 achieved by a memoryless strategy; however, unlike the single-objective case, the memoryless strategy may require randomization. Moreover, we show that the Pareto curve can be approximated in polynomial time in the size of the MDP. 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; but the question whether it is realizable by a deterministic memoryless strategy is NP-complete. These results provide efficient algorithms for design exploration in MDP models with multiple objectives.

1 Introduction

Markov decision processes (MDPs) are a widely studied model for dynamic and stochastic systems [8, 2]. An MDP models a dynamic system that evolves through stages. In each stage, a controller chooses one of several actions, 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 state and 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, where the rewards are combined using a discounted sum. 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 product, and in AI planning, one wishes to find a plan that optimizes several distinct goals. The usual MDP model is insufficient to express these natural problems.

* This research was supported in part by the AFOSR MURI grant F49620-00-1-0327, and the NSF grants CCR-0225610, CCR-0234690, and CCR-0427202.

We study MDPs with multiple objectives, an extension of the MDP model where there are several reward functions [4, 10]. 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 [6] and in multi-criterion optimization and decision making in both economics and engineering [5, 11, 9]. 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") [4]. 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. Thus, previous work 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 [1, 2]. Unfortunately, this does not provide any guarantee about the quality of the solutions obtained.

Instead, we study the *approximate* version of the problem, the ϵ-approximate Pareto curve [7] for MDPs with multiple discounted reward criteria. Informally, the ϵ-approximate Pareto curve for $\epsilon > 0$ 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 $1 + \epsilon$. Surprisingly, a polynomial-sized ϵ-approximate Pareto curve always exists. Moreover, we show that such an approximate Pareto curve may be computed efficiently (in polynomial time) in the size of the MDP. Our proof is based on the following characterization of Pareto optimal points: every Pareto optimal value profile can be realized by a memoryless (but possibly randomized) strategy. This enables the reduction of the problem to multi-objective linear programs, and we can apply the methods of [7].

We also study the related *Pareto realizability* decision problem: given a profile of values, is there a Pareto optimal strategy that dominates it? We show that the Pareto realizability problem can be solved in polynomial time. However, if we restrict the set of strategies to be pure (i.e., no randomization), then the problem becomes NP-hard. Our complexities are comparable to the single discounted reward case, where linear programming provides a polynomial-time solution [8]. However, unlike the single reward case, where pure and memoryless optimal strategies always exist, here, checking pure and memoryless realizability is hard.

The results of this paper provide polynomial-time algorithms for both the decision problem and the optimization problem for MDPs with multiple discounted reward objectives. Since the Pareto curve forms a useful "user interface" for desirable solutions, we believe that these results will lead to efficient design space exploration algorithms in multi-criterion design.

The rest of the paper is organized as follows. In Section 2, we give the basic definitions, and show in Section 3 the sufficiency of memoryless strategies. Section 4 gives a polynomial-time algorithm to construct the ϵ-approximate Pareto curve. Section 5 studies the decision version of the problem. Finally, in Section 6, we discuss the extension to MDPs with average (not discounted) reward objectives, and mention some open problems.

2 Discounted Reward Markov Decision Processes

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

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

We denote by $\mathrm{Dest}(s,a) = \mathrm{Supp}(\delta(s,a))$ the set of possible successors of s when the move 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 moves. 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., $\mathrm{Outcome}(s,\sigma) = \{ \langle s_0, s_1, \ldots, s_k, \ldots \rangle \mid \forall k \geq 0. \exists a_k \in A. \sigma(\langle s_0, s_1, \ldots, s_k \rangle)(a_k) > 0.$ 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 move 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 move a real-valued reward. For a reward function r the discounted reward objective is defined as follows. Given a discount factor $0 \leq \beta < 1$, the discounted reward for a strategy σ and an initial state s is $Val_{dis}^\sigma(r,s,\beta) = \sum_{t=0}^{\infty} \beta^t \cdot \mathbb{E}_s^\sigma[r(X_t,\theta_t)]$.

We consider MDPs with k-different reward functions $r_1, r_2, \ldots, r_k$. Given an initial state s, a strategy σ, $0 \leq \beta < 1$, the discounted reward value vector at s for σ, for $\boldsymbol{r} = \langle r_1, r_2, \ldots, r_k \rangle$ is defined as $Val_{dis}^\sigma(\boldsymbol{r},s,\beta) = \langle Val_{dis}^\sigma(r_1,s,\beta), Val_{dis}^\sigma(r_2,s,\beta), \ldots, Val_{dis}^\sigma(r_k,s,\beta) \rangle$.

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 [6] from a state s, if there is no $\sigma' \in \Sigma$ such that $Val_{dis}^\sigma(\boldsymbol{r},s,\beta) \leq Val_{dis}^{\sigma'}(\boldsymbol{r},s,\beta)$, and $Val_{dis}^\sigma(\boldsymbol{r},s,\beta) \neq Val_{dis}^{\sigma'}(\boldsymbol{r},s,\beta)$, i.e., there is no strategy σ' such that for all $1 \leq j \leq r$, $Val_{dis}^\sigma(r_j,s,\beta) \leq Val_{dis}^{\sigma'}(r_j,s,\beta)$ and exists $1 \leq j \leq k$, $Val_{dis}^\sigma(r_j,s,\beta) < Val_{dis}^{\sigma'}(r_j,s,\beta)$. 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 Σ^C of strategies suffices for Pareto optimality for discounted reward objective if for every Pareto-optimal strategy $\sigma \in \Sigma$, there is a strategy $\sigma_c \in \Sigma^C$ such that $Val_{dis}^\sigma(\boldsymbol{r},s,\beta) \leq Val_{dis}^{\sigma_c}(\boldsymbol{r},s,\beta)$.

Theorem 1 (Strategies for optimality [2]). *In MDPs with one reward function r_1, the family Σ^{PM} of pure memoryless strategies suffices for optimality for discounted reward objective, i.e., for all $0 \leq \beta < 1$, there exists a pure memoryless strategy $\sigma^* \in \Sigma^{PM}$, such that for all strategies $\sigma \in \Sigma$, we have $Val_{dis}^\sigma(r_1,s,\beta) \leq Val_{dis}^{\sigma^*}(r_1,s,\beta)$.*

3 Memoryless Strategies Suffice for Pareto Optimality

In the sequel, fix β such that $0 \leq \beta < 1$. Proposition 1 shows the existence of pure memoryless Pareto-optimal strategies.

Proposition 1. *Given reward functions $r_1, r_2, \ldots, r_k$, there exist pure memoryless Pareto-optimal strategies for discounted reward 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 $s \in S$, $r_+(s) = r_1(s) + r_2(s) + \ldots + r_k(s)$. Let $\sigma^* \in \Sigma^{PM}$ be a pure memoryless optimal strategy for the reward function r_+

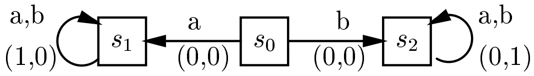

Fig. 1. MDP for Example 1

with the discounted reward objective with discount β (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 $Val_{dis}^{\sigma^*}(\boldsymbol{r}, s, \beta) \leq Val_{dis}^{\sigma}(\boldsymbol{r}, s, \beta)$, and for some j, $Val_{dis}^{\sigma^*}(r_j, s, \beta) < Val_{dis}^{\sigma}(r_j, s, \beta)$. Then we have $Val_{dis}^{\sigma^*}(r_+, s, \beta) = \sum_{j=1}^{k} Val_{dis}^{\sigma^*}(r_j, s, \beta) < \sum_{j=1}^{k} Val_{dis}^{\sigma}(r_j, s, \beta) = Val_{dis}^{\sigma}(r_+, s, \beta)$. This contradicts that σ^* is optimal for r_+. ∎

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 optimal strategy for the single objective $\sum_i w_i r_i$ is Pareto-optimal. This technique is called the *weighted factor method* [4, 10], and used commonly in engineering practice to find subsets of the Pareto set [5]. However, not all Pareto-optimal points are obtained in this fashion, as the following example shows.

Example 1. Consider the MDP from Fig. 1, with two moves 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 discounted reward 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 $(\frac{\beta}{1-\beta}, 0)$ and $(0, \frac{\beta}{1-\beta})$. 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 $Val_{dis}^{\sigma_m}(\boldsymbol{r}, s_0, \beta) = (\frac{\beta}{2\cdot(1-\beta)}, \frac{\beta}{2\cdot(1-\beta)})$. 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 $(\frac{x\cdot\beta}{1-\beta}, \frac{(1-x)\cdot\beta}{1-\beta})$. Hence the set of Pareto-optimal value vectors can be uncountable and value vectors may have irrational values. ∎

We now show that the family of memoryless strategies suffices for Pareto optimality. We assume the state space S is enumerated as $S = \{ 1, 2, \ldots, n \}$. For $t \in S$, we define reward functions r_t as $r_t(s, a) = 1$ if $s = t$ and 0 otherwise, i.e., a reward of value 1 is gained whenever state t is visited. Similarly, we define reward function $r_{t,b}$ for $t \in S$ and $b \in A$ as $r_{t,b}(s, a) = 1$ if $s = t$ and $b = a$, and 0 otherwise, i.e., a reward of value 1 is gained whenever state t is visited and action b is chosen. Given a strategy $\sigma \in \Sigma$ and a state $s \in S$ we define the discounted frequency of the state action pair (t, a):

$$freq_s^{\sigma}(t, a, \beta) = Val_{dis}^{\sigma}(r_{t,a}, s, \beta) = \sum_{k=0}^{\infty} \beta^k \cdot \mathbb{E}_s^{\sigma}[\mathbf{1}_{(X_k = t, \theta_k = a)}],$$

and the discounted frequency of the state t:

$$\mathit{freq}_s^{\sigma}(t,\beta) = \mathit{Val}_{dis}^{\sigma}(r_t,s,\beta) = \sum_{k=0}^{\infty} \beta^k \cdot \mathbb{E}_s^{\sigma}[\mathbf{1}_{(X_k=t)}].$$

Observe that $\sum_{a\in A} \mathit{freq}_s^{\sigma}(t,a,\beta) = \mathit{freq}_s^{\sigma}(t,\beta)$, for all $s,t \in S$ and $\sigma \in \Sigma$. Given a memoryless strategy $\sigma_m \in \Sigma^M$, and a transition function δ we denote by $\delta_{\sigma_m}(s,t) = \sum_{a\in A} \sigma_m(s)(a) \cdot \delta(s,a)(t)$ the probability of transition from s to t given the transition function δ and strategy σ_m.

Proposition 2. *Given a memoryless strategy σ_m, consider a vector $z = (z_1,z_2,\ldots,z_n)$ of variables. Consider the set of equations:*

$$z_i = r_s(i) + \beta \cdot \sum_{j\in S} \delta_{\sigma_m}(j,i) \cdot z_j; \quad \forall i \in S.$$

Then $z_i = \mathit{freq}_s^{\sigma_m}(i,\beta)$ is the unique solution to the above set of equations.

Proof. To establish the desired claim we show that for all $i \in S$, we have $\mathit{freq}_s^{\sigma_m}(i,\beta) = r_s(i) + \beta \cdot \sum_{j\in S} \delta_{\sigma_m}(j,i) \cdot \mathit{freq}_s^{\sigma_m}(j,\beta)$. The uniqueness follows from arguments similar to the uniqueness of values under memoryless strategies (see [2]).

$$\begin{aligned}
\mathit{freq}_s^{\sigma_m}(i,\beta) &= \textstyle\sum_{k=0}^{\infty} \beta^k \cdot \mathbb{E}_s^{\sigma_m}[\mathbf{1}_{(X_k=i)}] \\
&= \textstyle\sum_{k=0}^{\infty} \beta^k \cdot \mathbb{E}_s^{\sigma_m}[\sum_{j\in S} \mathbf{1}_{(X_{k-1}=j)}\ \delta_{\sigma_m}(j,i)] \\
&= r_s(i) + \textstyle\sum_{k=1}^{\infty} \beta^k \cdot \sum_{j\in S} \mathbb{E}_s^{\sigma_m}[\mathbf{1}_{(X_{k-1}=j)}] \cdot \delta_{\sigma_m}(j,i) \\
&= r_s(i) + \beta \cdot \textstyle\sum_{j\in S} \left(\sum_{k=1}^{\infty} \beta^{k-1} \mathbb{E}_s^{\sigma_m}[\mathbf{1}_{(X_{k-1}=j)}]\right) \cdot \delta_{\sigma_m}(j,i) \\
&= r_s(i) + \beta \cdot \textstyle\sum_{j\in S} \left(\sum_{z=0}^{\infty} \beta^{z} \cdot \mathbb{E}_s^{\sigma_m}[\mathbf{1}_{(X_z=j)}]\right) \cdot \delta_{\sigma_m}(j,i) \\
&= r_s(i) + \beta \cdot \textstyle\sum_{j\in S} \mathit{freq}_s^{\sigma_m}(j,\beta) \cdot \delta_{\sigma_m}(j,i). \qquad \blacksquare
\end{aligned}$$

Given a strategy $\sigma \in \Sigma$ and an initial state s, we define a memoryless strategy $\sigma_{f,s}$ from the discounted frequency of the strategy σ as follows:

$$\sigma_{f,s}(t)(a) = \frac{\mathit{freq}_s^{\sigma}(t,a,\beta)}{\mathit{freq}_s^{\sigma}(t,\beta)}; \quad \forall t \in S, \forall a \in A.$$

Since $\sum_{a\in A} \mathit{freq}_s^{\sigma}(t,a,\beta) = \mathit{freq}_s^{\sigma}(t,\beta)$ and $\mathit{freq}_s^{\sigma}(t,a,\beta) \geq 0$, it follows that $\sigma_{f,s}(t)$ is a probability distribution. Thus $\sigma_{f,s}$ is a memoryless strategy. From Proposition 2, and the identity $\mathit{freq}_s^{\sigma}(i,\beta) = r_s(i) + \beta \cdot \sum_{j\in S} \mathit{freq}_s^{\sigma}(j,\beta) \cdot \delta_{\sigma_{f,s}}(j)$, we get the following lemma.

Lemma 1. *For all strategies $\sigma \in \Sigma$, for all $i \in S$, for all $s \in S$, we have $\mathit{freq}_s^{\sigma}(i,\beta) = \mathit{freq}_s^{\sigma_{f,s}}(i,\beta)$.*

Proof. We show that $\mathit{freq}_s^\sigma(i,\beta) = r_s(i) + \beta \cdot \sum_{j\in S} \mathit{freq}_s^\sigma(j,\beta)\cdot \delta_{\sigma_{f,s}}(j)$. The result then follows from Proposition 2.

$$\begin{aligned}
\mathit{freq}_s^\sigma(i,\beta) &= \textstyle\sum_{k=0}^{\infty} \beta^k \cdot \mathbb{E}_s^\sigma[\mathbf{1}_{(X_k=i)}] \\
&= \textstyle\sum_{k=0}^{\infty} \beta^k \cdot \mathbb{E}_s^\sigma[\sum_{j\in S}\sum_{a\in A} \mathbf{1}_{(X_{k-1}=j,\theta_{k-1}=a)}\delta(j,a)(i)] \\
&= r_s(i) + \textstyle\sum_{k=1}^{\infty} \beta^k \cdot \sum_{j\in S}\sum_{a\in A} \mathbb{E}_s^\sigma[\mathbf{1}_{(X_{k-1}=j,\theta_{k-1}=a)}]\cdot\delta(j,a)(i) \\
&= r_s(i) + \beta\cdot \textstyle\sum_{j\in S}\sum_{a\in A}(\sum_{k=1}^{\infty}\beta^{k-1}\mathbb{E}_s^\sigma[\mathbf{1}_{(X_{k-1}=j,\theta_{k-1}=a)}])\cdot\delta(j,a)(i) \\
&= r_s(i) + \beta\cdot \textstyle\sum_{j\in S}\sum_{a\in A}\mathit{freq}_s^\sigma(j,a,\beta)\cdot\delta(j,a)(i) \\
&= r_s(i) + \beta\cdot \textstyle\sum_{j\in S}\sum_{a\in A}\big(\mathit{freq}_s^\sigma(j,\beta)\cdot\frac{\mathit{freq}_s^\sigma(j,a,\beta)}{\mathit{freq}_s^\sigma(j,\beta)}\big)\cdot\delta(j,a)(i) \\
&= r_s(i) + \beta\cdot \textstyle\sum_{j\in S}\mathit{freq}_s^\sigma(j,\beta)\cdot\big(\sum_{a\in A}\frac{\mathit{freq}_s^\sigma(j,a,\beta)}{\mathit{freq}_s^\sigma(j,\beta)}\big)\cdot\delta(j,a)(i) \\
&= r_s(i) + \beta\cdot \textstyle\sum_{j\in S}\mathit{freq}_s^\sigma(j,\beta)\sum_{a\in A}\sigma_{f,s}(j)(a)\cdot\delta(j,a)(i) \\
&= r_s(i) + \beta\cdot \textstyle\sum_{j\in S}\mathit{freq}_s^\sigma(j,\beta)\cdot\delta_{\sigma_{f,s}}(j,i).
\end{aligned}$$
∎

Corollary 1. *For all strategies $\sigma\in\Sigma$, for all $i\in S$ and $a\in A$, for all $s\in S$, we have $\mathit{freq}_s^\sigma(i,a,\beta) = \mathit{freq}_s^{\sigma_{f,s}}(i,a,\beta)$.*

Proof. The following equality follows from definition and Lemma 1.

$$\mathit{freq}_s^{\sigma_{f,s}}(i,a,\beta) = \mathit{freq}_s^{\sigma_{f,s}}(i,\beta)\cdot\sigma_{f,s}(i)(a) = \mathit{freq}_s^\sigma(i,\beta)\cdot\frac{\mathit{freq}_s^\sigma(i,a,\beta)}{\mathit{freq}_s^\sigma(i,\beta)} = \mathit{freq}_s^\sigma(i,a,\beta).$$
∎

Theorem 2. *For all reward functions r, for every strategy $\sigma\in\Sigma$, for all states s we have $\mathit{Val}_{dis}^\sigma(r,s,\beta) = \mathit{Val}_{dis}^{\sigma_{f,s}}(r,s,\beta)$.*

Proof. The result is proved as follows:

$$\begin{aligned}
\mathit{Val}_{dis}^\sigma(r,s,\beta) &= \textstyle\sum_{k=0}^{\infty}\beta^k\cdot\mathbb{E}_s^\sigma[r(X_k,\theta_k)] \\
&= \textstyle\sum_{k=0}^{\infty}\beta^k\cdot\mathbb{E}_s^\sigma[\sum_{i\in S}\sum_{a\in A} r(i,a)\cdot\mathbf{1}_{(X_k=i,\theta_k=a)}] \\
&= \textstyle\sum_{i\in S}\sum_{a\in A}(\sum_{k=0}^{\infty}\beta^k\cdot\mathbb{E}_s^\sigma[\mathbf{1}_{(X_k=i,\theta_k=a)}])\cdot r(i,a) \\
&= \textstyle\sum_{i\in S}\sum_{a\in A}\mathit{freq}_s^\sigma(i,a,\beta)\cdot r(i,a) \\
&= \textstyle\sum_{i\in S}\sum_{a\in A}\mathit{freq}_s^{\sigma_{f,s}}(i,a,\beta)\cdot r(i,a) \quad \text{(by Corollary 1).}
\end{aligned}$$

Similarly, it follows that $\mathit{Val}_{dis}^{\sigma_{f,s}}(r,s,\beta) = \sum_{i\in S}\sum_{a\in A}\mathit{freq}_s^{\sigma_{f,s}}(i,a,\beta)\cdot r(i,a)$. This establishes the result. ∎

Theorem 2 yields Theorem 3, and since the set of memoryless strategies is convex it also shows the set of Pareto optimal points is convex.

Theorem 3. *Given reward functions $\boldsymbol{r} = \langle r_1, r_2, \ldots, r_k\rangle$, for all strategies $\sigma\in\Sigma$ and for all states $s\in S$, the memoryless strategy $\sigma_{f,s}$ satisfies that $\mathit{Val}_{dis}^\sigma(\boldsymbol{r},s,\beta) = \mathit{Val}_{dis}^{\sigma_{f,s}}(\boldsymbol{r},s,\beta)$. The family of Σ^M of memoryless strategies suffices for Pareto optimality for discounted reward objectives.*

4 Approximating the Pareto Curve

Pareto curve. Let M be an MDP with reward functions $\boldsymbol{r} = \langle r_1,\ldots,r_k\rangle$. The *Pareto curve* $P^{dis}(M,s,\beta,\boldsymbol{r})$ of the MDP M at state s with respect to discounted reward objectives is the set of all k-vector of values such that for each $\boldsymbol{v} \in P^{dis}(M, s, \beta, \boldsymbol{r})$, there is a Pareto-optimal strategy σ such that $Val^{\sigma}_{dis}(\boldsymbol{r}, s, \beta) = \boldsymbol{v}$. 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 $P^{dis}(M, s, \beta, \boldsymbol{r})$ as a set of strategies which achieve the Pareto-optimal values (if there is more than one strategy that achieves the same value vector, $P^{dis}(M, s, \beta, \boldsymbol{r})$ contains at least one of them). For an MDP M, and $\epsilon > 0$, an ϵ-approximate Pareto curve, denoted $P^{dis}_{\epsilon}(M, s, \beta, \boldsymbol{r})$, is a set of strategies σ such that there is no other strategy σ' such that for all $\sigma \in P^{dis}_{\epsilon}(M, s, \beta, \boldsymbol{r})$, we have $Val^{\sigma'}_{dis}(r_i, s, \beta) \geq (1 + \epsilon)\, Val^{\sigma}_{dis}(r_i, s, \beta)$ 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 $P^{dis}_{\epsilon}(M, s, \beta, \boldsymbol{r})$.

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 $P_{\epsilon}(L)$ 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 [7].

Theorem 4 ([7]). *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 $P_{\epsilon}(L)$ 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 $P_{\epsilon}(L)$ in time polynomial in $|L|$ and $\frac{1}{\epsilon}$ and exponential in the number of objective functions.*

Proof. The result of part 1 is a direct consequence of Theorem 1 of [7]. The result of part 2 follows from Theorem 3 of [7] and the fact that linear programming can be solved in polynomial time. ∎

Solving MDPs by linear-programming. Given an MDP $M = (S, A, \delta)$ with $S = \{1, \ldots, n\}$, a reward function r, and a discount factor β, the discounted reward objective can be computed as the optimal solution of the linear program [2]. For multi-objective MDPs, we extend the standard linear programming formulation as follows. Given MDP M as before, an initial state s, and reward functions $r_1, \ldots, r_k$, the multi-objective linear program has the set of variables $\{ x(t, a) \mid t \in S, a \in A \}$. Intuitvely $x(t, a)$ is a variable that represents the discounted frequency of the state action pair (t, a), when the starting state is s.

The constraints of the multi-objective linear program over the variables $x(t,a)$ for $t \in S$ and $a \in A$ are given by:

$$\begin{array}{ll} \sum_{a \in A} x(t,a) = r_s(t) + \beta \sum_{u \in S} \sum_{a_1 \in A} \delta(u,a_1)(t) \cdot x(u,a_1) & t \in S \\ x(t,a) \geq 0 & t \in S, a \in A \end{array} \qquad (1)$$

Equation 1 provides constraints on the discounted frequencies. The k objective functions are

$$\max \sum_{t \in S} \sum_{a \in A} r_i(t,a) \cdot x(t,a), \qquad \text{for } i \in \{1,2,\ldots,k\}.$$

Consider any feasible solution $x(t,a)$ for $a \in A, t \in S$. Let $x(t) = \sum_{a \in A} x(t,a)$, then the feasible solution derives a memoryless strategy that plays action a at state t with probability $\frac{x(t,a)}{x(t)}$. The linear program with the ith objective function asks to maximize the discounted reward for reward function r_i over the set of all memoryless strategies. The optimal solution for the linear program with only the ith objective also derives a optimal memoryless strategy for reward function r_i. Further, given a feasible solution $\mathbf{x}$, or equivalently, the memoryless strategy derived from the solution, we can compute the value profile for each reward in polynomial time (since the MDP reduces to a Markov chain when the strategy is fixed).

Consider the above multi-objective linear program for MDP M, state s of M, discount β, and rewards $\boldsymbol{r} = \langle r_1, \ldots, r_k \rangle$, denoted as $L^{dis}(M,s,\beta,\boldsymbol{r})$. Let $P(L^{dis}(M,s,\beta,\boldsymbol{r}))$ denote the Pareto curve for the multi-objective linear program. With abuse of notation, we write $P(L^{dis}(M,s,\beta,\boldsymbol{r}))$ to be the set of memoryless strategies derived from the Pareto-optimal solutions of the multi-objective linear program. It follows that the Pareto curve $P(L^{dis}(M,s,\beta,\boldsymbol{r}))$ characterizes the set of memoryless Pareto-optimal points for the MDP with k-discounted reward objectives. Since memoryless strategies suffices of Pareto optimality for discounted reward objectives (Theorem 3), the following lemma is immediate. Theorem 5 follows from Theorem 4 and Lemma 2.

Lemma 2. *Given MDP M with k-reward functions $\boldsymbol{r}$, state s of M, and discount factor β, let $L^{dis}(M,s,\beta,\boldsymbol{r})$ be the corresponding linear-program for memoryless strategies. The following assertions hold:*

1. *$P(L^{dis}(M,s,\beta,\boldsymbol{r})) = P^{dis}(M,s,\beta,\boldsymbol{r})$, i.e., the Pareto curve for the linear program and the Pareto curve for the discounted reward coincide.*
2. *For all $\varepsilon > 0$, for all ϵ-approximate Pareto curve $P_\epsilon(L^{dis}(M,s,\beta,\boldsymbol{r}))$ of $L^{dis}(M,s,\beta,\boldsymbol{r})$, there is an ϵ-approximate Pareto curve $P_\epsilon^{dis}(M,s,\beta,\boldsymbol{r})$ such that $P_\epsilon(L^{dis}(M,s,\beta,\boldsymbol{r})) = P_\epsilon^{dis}(M,s,\beta,\boldsymbol{r})$.*

Theorem 5. *Given a MDP M with k-reward functions $\boldsymbol{r}$ and discount factor β, the following assertions hold:*

1. *For all $\varepsilon > 0$, there exists an approximate Pareto curve $P_\epsilon^{dis}(M,s,\beta,\boldsymbol{r})$ consisting of a number of feasible solution that is polynomial in $|M|$ and $\frac{1}{\varepsilon}$, but exponential in the number of reward functions.*

2. *For all $\varepsilon > 0$, there is an algorithm to construct a $P_{\epsilon}^{dis}(M, s, \beta, \boldsymbol{r})$ in time polynomial in $|M|$ and $\frac{1}{\epsilon}$ and exponential in the number of reward functions.*

Theorem 5 shows that the Pareto curve can be efficiently ε-approximated. Recall that it follows from Example 1 that the set of Pareto-optimal points can be uncountable and the values can be irrational. Hence the ε-approximation of the Pareto curve is a useful finite approximation.

The approximate Pareto curve constructed allows us to answer trade-off queries about multi-objective MDPs. Specifically, given a multi-objective MDP M with k reward functions $\boldsymbol{r}$ and discount factor β, and a profile $\mathbf{w} = (w_1, \ldots, w_k)$ of values, we can check whether $\mathbf{w}$ is ϵ-close to a Pareto-optimal profile at state s by constructing $P_{\epsilon}^{dis}(M, s, \beta, \boldsymbol{r})$ in polynomial time, and checking that there is some strategy in $P_{\epsilon}^{dis}(M, s, \beta, \boldsymbol{r})$ whose value profile is ϵ-close to $\mathbf{v}$.

5 Pareto Realizability

In this section we study two related aspects of multi-objective MDPs: Pareto realizability, and pure memoryless Pareto realizability. The *Pareto realizability problem* asks, given a multi-objective MDP M with rewards $r_1, \ldots, r_k$ (collectively, $\boldsymbol{r}$) and discount factor β, a state s of M, and a value profile $\mathbf{w} = (w_1, \ldots w_k)$ of k rational values, whether there exists a strategy σ such that $Val_{dis}^{\sigma}(\boldsymbol{r}, s, \beta) \geq \mathbf{w}$. Observe that such a strategy exists if and only if there is a Pareto optimal strategy σ' such that $Val_{dis}^{\sigma'}(\boldsymbol{r}, s, \beta) \geq \mathbf{w}$. Also observe that it follows from Lemma 2 that a value profile $\mathbf{w}$ is realizable if and only if it is realizable by a memoryless strategy. The *pure memoryless* Pareto realizability problem further requires this strategy to be pure and memoryless.

The Pareto realizability problem arises when certain target behaviors are required, and one wishes to check if they can be attained on the model. Pure Pareto realizability arises in situations, such as circuit implementations, where the implemented strategy does not have access to randomization.

Theorem 6. *The Pareto realizability problem for multi-objective MDPs can be solved in polynomial time. The pure memoryless Pareto realizability problem for multi-objective MDPs is NP-complete.*

Proof. We show Pareto realizability is in polynomial time by reduction to linear programming. The reduction is obtained as follows: along with the constraints defined by Equation 1 we add the following set of constraints,

$$w_i \leq \sum_{t \in S} \sum_{a \in A} x(t, a) \cdot r_i(t, a) \qquad i \in \{ 1, \ldots, k \}.$$

The constraints from Equation 1 provide constraints on the discounted frequencies. The last set of constraints ensures that the value for each reward function

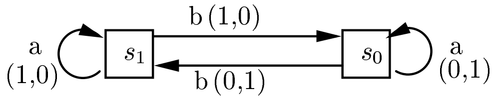

Fig. 2. MDP for Example 2

is greater than or equal to the corresponding profile w_i. Thus, if the set is consistent, then the answer to the Pareto realizability problem is "yes," and if inconsistent, the answer is "no." Consistency of this set can be checked in polynomial time using linear programming.

Pure and memoryless Pareto realizability is in NP since we can guess a pure memoryless strategy and compute its payoffs in polynomial time. We can then check that each payoff is greater than or equal to the given profile. It is NP-hard by reduction from subset sum. The subset sum problem takes as input natural numbers $\{ a_1, \ldots, a_n \}$, and a natural number p, and asks if there exist v_1, v_2, $\ldots$, v_n in $\{ 0, 1 \}$ such that $a_1 v_1 + \ldots + a_n v_n = p$. It is NP-complete [3].

For an instance of the subset sum problem, we construct an MDP with two reward functions as follows. We assume for clarity that $\beta = 1$. The construction can be adapted for any fixed discount factor by suitably scaling the rewards. The MDP has $n+1$ states, numbered from 1 to $n+1$. We fix the start state to be 1. There are two actions, L and R. The transition relation is deterministic, for state $i \in \{ 1, \ldots, n \}$, we have $\delta(i, L)(i+1) = \delta(i, R)(i+1) = 1$. For state $n+1$, we have $\delta(n+1, L)(n+1) = \delta(n+1, R)(n+1) = 1$. The reward function r_1 is defined as $r_1(i, L) = a_i$, and $r_1(i, R) = 0$ for $i \in \{ 1, \ldots, n \}$, and $r_1(n+1, L) = r_1(n+1, R) = 0$. Similarly, the reward function r_2 is defined as $r_2(i, R) = a_i$, and $r_2(i, L) = 0$ for $i \in \{ 1, \ldots, n \}$, and $r_2(n+1, L) = r_2(n+1, R) = 0$. We now ask if the profile $(p, \sum_i a_i - p)$ is pure Pareto realizable for this MDP. From the construction, it is clear that this profile is pure memoryless Pareto realizable iff the answer to the subset sum problem is "yes". In fact, the pure strategy that realizes this profile provides the required v_i's: if move L is played at state i, then $v_i = 1$, else $v_i = 0$. Since the MDP is a DAG, the hardness construction holds if we require the realizing strategy to be pure (not necessarily memoryless). ∎

The *pure Pareto realizability* problem requires the realizing strategy to be pure, but not necessarily memoryless. It follows from the reduction above that the Pareto realizability problem is NP-hard; however, we do not have a characterization of the exact complexity of the problem.

6 Extensions

We now discuss the class of limit-average objectives that is widely studied in context of MDPs. The class of *limit-average objectives* is defined as follows. Given a reward function r, the *limit-average reward* for a strategy σ at an initial state s is $Val^{\sigma}_{avg}(r, s) = \limsup_{k \to \infty} \frac{1}{k} \sum_{t=0}^{k} \mathbb{E}^{\sigma}_{s}[r(X_t, \theta_t)]$. With this definition, Theorem 1 holds for limit-average objectives, and Proposition 1 also extends in a straightforward way. Unfortunately, we now show that Theorem 3 does not generalize. Example 2 below shows that for limit-average objectives the family of memoryless strategies does not capture all Pareto-optimal strategies. However, it is still possible that the Pareto curve for limit-average objectives can be approximated in polynomial time. This remains an open problem.

Example 2. Fig. 2 shows an MDP with two moves 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 limit-average objectives for r_1 and r_2. Given any memoryless strategy σ_m we have at s_0, $Val_{avg}^{\sigma_m}(s_0, r_1) + Val_{avg}^{\sigma_m}(s_0, r_2) = 1$. 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 $Val_{avg}^{\sigma_m}(s_0, r_1) = 1$ and $Val_{avg}^{\sigma_m}(s_0, r_2) = 1$. There is no memoryless Pareto-optimal strategy to achieve this value vector. A simple adaptation of Example 1 shows that the family of pure strategies also does not suffice for Pareto optimality for limit-average objectives. ∎

References

1. 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.
2. J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer, 1997.
3. M.R. Garey and D.S. Johnson. *Computers and Intractability*. W.H. Freeman, 1979.
4. R. Hartley. Finite discounted, vector Markov decision processes. Technical report, Department of Decision Theory, Manchester University, 1979.
5. J. Koski. Multicriteria truss optimization. In *Multicriteria Optimization in Engineering and in the Sciences*. 1988.
6. G. Owen. *Game Theory*. Academic Press, 1995.
7. 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.
8. M.L. Puterman. *Markov Decision Processes*. John Wiley and Sons, 1994.
9. R. Szymanek, F. Catthoor, and K. Kuchcinski. Time-energy design space exploration for multi-layer memory architectures. In *DATE 04*. IEEE, 2004.
10. D.J. White. Multi-objective infinite-horizon discounted Markov decision processes. *Journal of Mathematical Analysis and Applications*, 89(2):639–647, 1982.
11. P. Yang and F. Catthoor. Pareto-optimization based run time task scheduling for embedded systems. In *CODES-ISSS 03*, pages 120–125. ACM, 2003.