arXiv:2611.12663v1 [math.OC] 9 Nov 2026

# Multi-objective Markov Decision Process: Set-valued Bellman's principle and fixed point perspective

Zachary Feinstein* Gabriela Kováčová† Ariel Neufeld‡ Birgit Rudloff§

November 10, 2026

**Abstract**

We study infinite-horizon discounted Markov decision processes with multiple cost objectives and finite state and action spaces. We consider the set of value functions generated by randomized Markov policies, where each function describes the expected discounted costs from all initial states under a single policy. We establish a set-valued Bellman equation and characterize this set as the greatest fixed point of the associated Bellman operator. We also show that choosing an attainable cost vector separately for each initial state can produce a function that no single Markov policy realizes. Nevertheless, the attainable costs at individual states satisfy their own Bellman recursion. These results provide a dynamic programming and fixed point framework for multi-objective Markov decision processes.

## 1 Introduction

An *infinite-horizon Markov decision problem* (MDP) provides a mathematical framework for sequential decision-making in a stochastic system that operates indefinitely. Let $\mathcal{S}$ denote the state space and $\mathcal{A}$ the action space. At decision time $t = 0, 1, \ldots$, the system occupies a state $S_t \in \mathcal{S}$, and the decision-maker selects an admissible action $A_t \in \mathcal{A}(S_t)$. Conditional on $(S_t, A_t) = (s, a)$, the next state $S_{t+1}$ is generated according to the controlled transition probability $P(s'|s, a)$, and the transition incurs the immediate cost $c(S_t, A_t, S_{t+1})$. The Markov property requires that, conditional on the current state and action, the distribution of $S_{t+1}$ be independent of the preceding history.

A policy $\pi \in \Pi$ specifies how actions are selected over time, where $\Pi$ denotes the set of admissible policies. Given an initial state $S_0 = s$, the expected infinite-horizon discounted cost

*Stevens Institute of Technology, zfeinste@stevens.edu
†Reykjavík University, gabrielak@ru.is
‡Nanyang Technological University (NTU), ariel.neufeld@ntu.edu.sg
§Vienna University of Economics and Business, birgit.rudloff@wu.ac.at

under policy $\pi$ is

$$J^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t c(S_t, A_t, S_{t+1}) \,\middle|\, S_0 = s\right], \tag{1.1}$$

where $\gamma \in (0,1)$ denotes the given discount factor which assigns progressively less weight to costs incurred in the distant future and ensures that the total discounted cost is finite. The objective is to determine the optimal value $V^*(s) := \inf_{\pi\in\Pi} J^\pi(s)$ as well as the corresponding optimal policy. We refer to [Put14] and [BR11] for books on (standard, single-objective) Markov decision problems.

Infinite-horizon MDPs arise naturally in continuing operational systems, including inventory and production control, queueing and service operations, maintenance and replacement, transportation, energy management, and healthcare. Brown and Haugh [BH17], for example, study discounted infinite-horizon models for inventory control with autoregressive demand and for dynamic service allocation in a multiclass queue. Financial applications include consumption-investment decisions with a random lifetime, dividend-payment policies for insurance companies, and utility maximization in financial markets; infinite-horizon formulations of these problems are developed by Bäuerle and Rieder [BR11, Chapter 9, Sections 9.1–9.3]. Credit management provides another example: Liebman [Lie72] formulates the selection of credit-control policies as an infinite-horizon MDP and solves an equivalent linear-programming representation.

The principal solution approach is based on the *dynamic programming principle* (DPP). This allows us to identify the value function $V^*(s)$ as the unique solution (fixed point) of the equation

$$V^*(s) = \inf_{a\in\mathcal{A}(s)} \sum_{s'\in\mathcal{S}} \big[c(s,a,s') + \gamma V^*(s')\big] P(s'|s,a), \tag{1.2}$$

known as the *Bellman optimality equation*. Moreover, the optimal policy can be derived from the optimal action $a^*(s)$ which minimizes the right-hand side of the Bellman equation. This then enables the use of numerical methods such as value iteration or policy iteration to find the optimal value and policy [Put14].

In this paper, we study *multi-objective* infinite-horizon Markov decision problems. In contrast to the standard setting discussed above, the decision maker incurs a vector-valued cost $C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^n$ in each time period. The goal is to derive the corresponding dynamic programming results, where we pursue an appropriate counterpart of the Bellman's optimality equation for multi-objective MDP and its interpretation in terms of a fixed point.

A number of multi-objective and set-valued dynamic problems have been studied in recent years. These include (finite-horizon) multi-objective optimal control problems, with applications such as mean-risk portfolio selection [KR21]. The validity of dynamic programming is the central

question: [KR21] establish that a value function of a multi-objective control problem is a *set-valued* object and the counterpart of dynamic programming principle is a *set-valued Bellman's principle*, a recursive relationship for the (set-valued) value function. For continuous-time multi-objective problems, the corresponding counterpart, the set-valued Hamilton-Jacobi-Bellman equations, have been studied [IZ23, Vis23, HV20]. A set-valued Bellman's principle has also been derived in the context of robust multi-objective control [CK26], multivariate and set-valued dynamic risk measures [FR13, FR14, FR17, AF21], Nash equilibria of dynamic games [FRZ21] and mean-field games [IZ24].

Despite the growing body of work on multi-objective control and dynamic programming, to the best knowledge of the authors, these have not been explored in the context of multi-objective Markov decision processes and infinite horizon. The current work aims to fill this gap by establishing the validity of an appropriate set-valued Bellman's principle and providing an interpretation in terms of a fixed-point framework. By establishing the DPP, this paper provides a solid theoretical framework for future work on reinforcement learning for multi-objective control.

## 2 Setting

### 2.1 Preliminaries

Let $\mathcal{T} = \{0, 1, \dots\}$ and let $\mathcal{S}$ be a finite set. The set of probability measures (distributions) over $\mathcal{S}$ is denoted by $\mathcal{M}(\mathcal{S})$, the power set of $\mathcal{S}$ is denoted by $\mathcal{P}(\mathcal{S}) = 2^{\mathcal{S}}$. For a finite space $\mathcal{S}$ the corresponding $\sigma$-algebra is $\mathcal{P}(\mathcal{S})$, the $\sigma$-algebras of product spaces are obtained as product $\sigma$-algebras. In view of the finiteness assumption, we omit mentions of the $\sigma$-algebras and measurable spaces.

We use the supremum norm $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$ for $x \in \mathbb{R}^n$ and denote by $\mathbb{B}_K^n = \{x \in \mathbb{R}^n \mid \|x\|_\infty \leq K\}$ the closed ball around the origin with radius $K > 0$. We denote the set of measurable functions $v : \mathcal{S} \to \mathbb{R}^n$ by $\mathscr{L}(\mathcal{S}, \mathbb{R}^n)$, and the notation $\mathscr{L}(\mathcal{S}, \mathbb{B}_K^n)$ will be used for the subset of those functions with values in $\mathbb{B}_K^n$. Due to the finiteness of the space $\mathcal{S}$, the notion of measurability is trivial and, therefore, not addressed.

### 2.2 Multi-objective Markov decision process

Throughout, we consider a state process $(S_t)_{t\in\mathcal{T}}$ with values in a finite state space $\mathcal{S}$ and an action (control) process $(A_t)_{t\in\mathcal{T}}$ with values in a finite action space $\mathcal{A}$. The transition of the state process fulfills the controlled Markov property with transition probabilities (stochastic

kernels) $P : \mathcal{S} \times \mathcal{A} \to \mathcal{M}(\mathcal{S})$,

$$\text{Prob}(S_{t+1} = s' \,|\, S_t = s, A_t = a) = P(s'|s, a).$$

Moreover, an initial state distribution $S_0 \sim \nu \in \mathcal{M}(S)$ is given.

In combination with a scalar-valued cost (or reward) function, we would obtain a standard (single-objective) Markov decision process. In this work, we consider its multi-objective counterpart, a multi-objective Markov decision process. The cost accrued is multivariate, given by a multi-cost function $C : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^n$. The aim is to minimize the expected discounted (multi-)cost on an infinite time horizon

$$\mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \right],$$

where $\gamma \in (0, 1)$ is a given discount factor. This leads to a multi-objective control problem.

We denote by $\mathcal{H}_t = \underbrace{(\mathcal{S} \times \mathcal{A}) \times \cdots \times (\mathcal{S} \times \mathcal{A})}_{t \text{ times}} \times \mathcal{S}$ the space of all (state) histories until time $t \in \mathcal{T}$, and by $\mathcal{H} = (\mathcal{S} \times \mathcal{A}) \times (\mathcal{S} \times \mathcal{A}) \times \cdots$ the space of all (state) histories on the infinite time horizon. The (stochastic) process of histories up to time $t$ will be denoted by $H_t = (S_0, A_0, \ldots, S_t)$ and its realizations (plays) will be denoted by $h_t = (s_0, a_0, \ldots, s_t)$.

### 2.3 Multi-objective optimization

A multi-objective optimization problem is a problem of the form

$$\text{minimize } F(x) \text{ with respect to } \leq \text{ subject to } x \in \mathbb{X}, \tag{MOP}$$

where $\mathbb{X}$ is a non-empty feasible set, $F : \mathbb{X} \to \mathbb{R}^n$ is a vector-valued objective mapping, and $\leq$ is the standard (component-wise) partial order on $\mathbb{R}^n$ induced by the *ordering cone* $\mathbb{R}^n_+$. The *image of the feasible set* of problem (MOP) is the set $F[\mathbb{X}] = \{F(x)|x \in \mathbb{X}\}$ and the *upper image* of (MOP) is defined as $\mathcal{U} = \text{cl}\,(F[\mathbb{X}] + \mathbb{R}^n_+)$.

The prior literature, see e.g. [KR21, CK26], has established that a Bellman-type recursion for (finite-horizon) multi-objective control holds for a value function that is a set-valued object corresponding to the image of the feasible set or the upper image. While the upper image is usually preferred due to its set optimization interpretation as an infimum of the multi-objective problem, see [HHL+15], in this work, we opt to utilize the image of the feasible set since boundedness simplifies the fixed point argument. Note that the close relationship between $F[\mathbb{X}]$ and $\mathcal{U}$ makes it possible to easily extend a recursion for the image of the feasible set into a recursion for the upper image, see e.g. [CK26, Theorem 3.3 or Theorem 4.4].

A feasible point $\bar{x} \in \mathbb{X}$ is called a *minimizer* of (MOP) if $(F[\bar{x}] - \mathbb{R}^n_+ \setminus \{0\}) \cap F[\mathbb{X}] = \emptyset$ and it is called a *weak minimizer* if $(F[\bar{x}] - \operatorname{int} \mathbb{R}^n_+) \cap F[\mathbb{X}] = \emptyset$. A (weak) minimizer is in the literature also referred to as *(weakly) Pareto optimal point* or a *(weakly) efficient point*. The set of images of all (weak) minimizers of (MOP) is often known as the *(weak) Pareto frontier* or *(weakly) efficient frontier*.

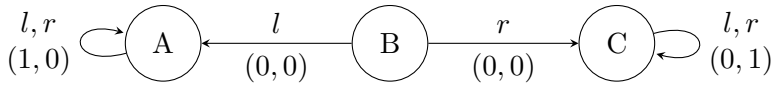
### 2.4 Policy classes

The decision maker controls the MDP through the choice of a policy $\pi = (\pi_0, \pi_1, \ldots)$. At time $t$, the observed history is $h_t = (s_0, a_0, s_1, \ldots, a_{t-1}, s_t)$. A general history-dependent randomized decision rule is a stochastic kernel $\pi_t(\cdot \mid h_t)$ on the admissible action set $\mathcal{A}(s_t)$. We distinguish between *history-dependent* policies, for which the decision rule $\pi_t$ may depend on the full history $h_t$, and *Markov* policies, for which $\pi_t$ depends only on the current state $s_t$. We also distinguish between *deterministic* policies, where the decision rule selects a single admissible action, and *randomized* policies, where the decision rule specifies a probability distribution over admissible actions. Thus, a randomized Markov decision rule is of the form $\pi_t(\cdot \mid s_t)$, whereas a deterministic Markov decision rule is induced by a measurable selector $A_t$ satisfying $A_t(s_t) \in \mathcal{A}(s_t)$.

For single-objective discounted MDPs, restricting attention to *Markov* policies is without loss of optimality under the standard assumptions of infinite-horizon MDP theory. In particular, the optimal value obtained over history-dependent policies coincides with the optimal value obtained over Markov policies [BR11, Chapter 7]. Moreover, if for every state the infimum in the corresponding Bellman equation is attained by an admissible action and a measurable minimizing selector can be chosen, then an optimal deterministic Markov policy exists. In finite-state and finite-action discounted MDPs, this attainment condition is automatic; in this setting, there exists an optimal deterministic stationary policy [Put14, Theorem 6.2.7]. For the multi-objective setting considered in this paper, [CMH06, Theorem 3] provides the corresponding reduction at the level of Pareto images: the image generated by Pareto-optimal history-dependent policies coincides with the image generated by Pareto-optimal Markov policies. Hence, restricting attention to Markov policies does not remove any Pareto-optimal objective vector.

The role of *randomization* is different. In a single-objective MDP with an expected discounted-cost criterion, randomization does not improve the optimal value. More precisely, optimizing the corresponding Bellman operator over randomized decision rules yields the same value as optimizing over deterministic decision rules [Put14, Proposition 6.2.1]. Intuitively, for a fixed state, the expected one-step cost plus continuation value under a randomized action is a convex combination of the corresponding quantities under deterministic actions; hence no randomized

Figure 1: Multi-objective MDP of Example 2.1




action can improve upon the best deterministic action. Consequently, if the infimum in the Bellman equation is attained by an admissible action and a measurable minimizer can be selected, an optimal randomized Markov policy can be replaced by a deterministic Markov policy with the same scalar value. If no minimizing action exists, this is a nonattainment issue for the scalar optimization problem rather than a strict improvement obtained through randomization.

In multi-objective MDPs, however, the object of interest is not a single scalar value but the attainable set of a vector-valued criteria. Since the underlying order is only partial, randomization can generate objective vectors that are not attainable by deterministic policies, for instance through convex combinations of deterministic objective vectors. Such randomized objective vectors may be Pareto optimal and therefore relevant for the efficient frontier, as illustrated in Example 2.1. This is a genuine distinction from the scalar case: the issue is not merely whether an infimum is attained, but whether all Pareto-relevant objective vectors are attainable within the chosen policy class. Since deterministic policies may fail to generate the full Pareto-relevant image, we work throughout this paper with randomized Markov policies.

**Example 2.1.** [CMH06, Example 1] Consider the multi-objective MDP given in Figure 1 with initial state $S_0 = s_0 = B$ and multi-cost function $C(s, a, s') = C(s)$. In this case, the expected discounted multi-cost is fully determined by the action $A_0$. Only two expected discounted costs can be achieved across all deterministic policies, $\left(\frac{\gamma}{1-\gamma}, 0\right)$ for $\pi_0(B) = l$ and $\left(0, \frac{\gamma}{1-\gamma}\right)$ for $\pi_0(B) = r$. However, consider a randomized policy where each action $l$ and $r$ is chosen with equal probability in stage $t = 0$. This is a Pareto optimal policy and its expected cost $\left(\frac{\gamma}{2(1-\gamma)}, \frac{\gamma}{2(1-\gamma)}\right)$ cannot be achieved by any deterministic policy. □

In the remainder of this subsection, we formalize admissible policies. Let $\mathcal{A}(\cdot) : \mathcal{S} \rightrightarrows \mathcal{A}$ be a set-valued mapping with non-empty values, where $\mathcal{A}(s)$ represents the admissible actions for state $s \in \mathcal{S}$. We denote by

$$\mathcal{D} = \{\varpi : \mathcal{S} \to \mathcal{M}(\mathcal{A}) \quad | \quad \forall s \in \mathcal{S} : \ \varpi(\mathcal{A}(s)|s) = 1\}$$

the set of admissible randomized (Markov) decision rules and by

$$\Pi = \{\pi = (\pi_0, \pi_1, \dots) \quad | \quad \forall t \in \mathcal{T} : \ \pi_t \in \mathcal{D}\}$$

the set of all admissible randomized Markov policies. For a policy $\pi \in \Pi$ and a decision rule $\varpi \in \mathcal{D}$ we define a forward shift operator $\theta(\pi) = (\pi_1, \pi_2, \dots)$ and a concatenation $(\varpi, \pi) = (\varpi, \pi_0, \pi_1, \dots)$.

Observe that these operators preserve admissibility. Namely, for any policy $\pi \in \Pi$ and decision rule $\varpi \in \mathcal{D}$, it holds

$$\{\theta(\pi) \mid \pi \in \Pi\} = \Pi = \{(\varpi, \pi) \mid \varpi \in \mathcal{D},\ \pi \in \Pi\}. \tag{2.1}$$

## 2.5 Probability measure

We follow [HLL12, Section 2.2] to establish the probability measure used by the decision maker who implements policy $\pi$. Since the results are standard, we provide them without proofs.

By the Ionescu-Tulcea theorem [HLL12, Proposition C.10], for each policy $\pi \in \Pi$ there exists a unique probability measure $\mathbb{P}_\nu^\pi$ on $\mathcal{H} = (\mathcal{S} \times \mathcal{A}) \times (\mathcal{S} \times \mathcal{A}) \times \cdots$ such that

$$\begin{aligned}
&\mathbb{P}_\nu^\pi(S_0 = s) = \nu(s),\\
&\mathbb{P}_\nu^\pi(A_t = a | H_t = h_t) = \pi_t(a|s_t),\\
&\mathbb{P}_\nu^\pi(S_{t+1} = s | H_t = h_t, A_t = a_t) = P(s|s_t, a_t),
\end{aligned}$$

where $h_t = (s_0, a_0, \dots, s_t)$. The expectation under $\mathbb{P}_\nu^\pi$ will be denoted $\mathbb{E}_\nu^\pi$. In case of a Dirac measure $\nu = \delta_s$, we will use the notation $\mathbb{E}^\pi[\,\cdot\,|S_0 = s]$.

With a slight abuse of notation, we use $\theta : \mathcal{H} \to \mathcal{H}$ to denote a forward shift operator given by $\theta((s_0, a_0, s_1, a_1, \dots)) = (s_1, a_1, s_2, a_2, \dots)$ as well as a forward shift operator for the policies. In the following lemma we summarize the decomposition and Markov type properties on the level of probability measures.

**Lemma 2.2.** *Let $\mathbb{P}_\nu^\pi$ be the probability measure on $\mathcal{H}$ induced by policy $\pi \in \Pi$, prior $\nu \in \mathcal{M}(\mathcal{S})$ and transition kernel $P$. The following holds for the distribution of $H = (S_0, A_0, S_1, \dots)$:*

*(i) The marginal law of $(S_0, A_0, S_1)$ is given by*

$$\mathbb{P}_\nu^\pi(S_0 = s_0, A_0 = a_0, S_1 = s_1) = \nu(s_0)\pi_0(a_0|s_0)P(s_1|s_0, a_0)$$

*for all $(s_0, a_0, s_1) \in \mathcal{H}_1$.*

*(ii) The conditional tail distribution of $\theta(H)|S_1 = s$ is given by*

$$\mathbb{P}_\nu^\pi\big(\theta(H) \in \cdot \,\big|\, S_1 = s\big) = \mathbb{P}_{\delta_s}^{\theta(\pi)}(\cdot).$$

*(iii) For any $h_1 = (s_0, a_0, s_1) \in \mathcal{H}_1$ and any $B \subseteq \mathcal{H}$, a decomposition is given by*

$$\mathbb{P}_\nu^\pi\big(H_1 = h_1, \theta(H) \in B\big) = \mathbb{P}_\nu^\pi\big(H_1 = h_1\big)\mathbb{P}_{\delta_{s_1}}^{\theta(\pi)}(B).$$

*(iv) Given a function $f : \mathcal{S} \times (\mathcal{A} \times \mathcal{S})^k \to \mathbb{R}^n$ it holds*

$$\begin{aligned}&\mathbb{E}^\pi\left[f(S_0, A_0, \ldots, S_k) \,|\, S_0 = s_0\right] \\ &= \int_{(\mathcal{A}\times\mathcal{S})^k} f(s_0, a_0, \ldots, s_k) P(\mathrm{d}s_k|s_{k-1}, a_{k-1})\pi_{k-1}(\mathrm{d}a_{k-1}|s_{k-1}) \ldots P(\mathrm{d}s_1|s_0, a_0)\pi_0(\mathrm{d}a_0|s_0).\end{aligned}$$

# 3 From policies to values

Throughout this section, let the admissible policy $\pi \in \Pi$ be fixed. We introduce the expected discounted multi-cost associated with policy $\pi$ and derive its recursive form, which will be used to derive a Bellman-type recursion for the multi-objective problem.

For a policy $\pi$ we define a policy (state) value function $V^\pi : \mathcal{S} \to \mathbb{R}^n$ as

$$V^\pi(s) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \;\Big|\; S_0 = s\right] \qquad \text{for all } s \in \mathcal{S}.$$

The policy state value function $V^\pi$ satisfies a recursive relationship known as the *Bellman consistency equation*. Observe that this recursive form for multi-objective MDP is a direct counterpart to the single-objective case.

**Theorem 3.1.** *Given an admissible policy $\pi \in \Pi$, for all states $s \in \mathcal{S}$ it holds*

$$\begin{aligned}V^\pi(s) &= \mathbb{E}^\pi\left[C(S_0, A_0, S_1) + \gamma V^{\theta(\pi)}(S_1) \;\mid\; S_0 = s\right] \\ &= \sum_{a\in\mathcal{A}} \pi_0(a|s) \sum_{s'\in\mathcal{S}} P(s'|s, a)\left(C(s, a, s') + \gamma V^{\theta(\pi)}(s')\right).\end{aligned}$$

Before proving the theorem, we first need the following lemma.

**Lemma 3.2.** *Define a function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ with $\pi_0(a|s) > 0$ as*

$$Q^\pi(s, a) := \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \;\Big|\; S_0 = s, A_0 = a\right],$$

*and set $Q^\pi(s, a) := 0$ otherwise. For the state value function $V^\pi$ and the state-action value function $Q^\pi$ it holds*

$$V^\pi(s) = \sum_{a\in\mathcal{A}} \pi_0(a|s) Q^\pi(s, a) \tag{3.1}$$

*and for* $(s,a) \in \mathcal{S} \times \mathcal{A}$ *satisfying* $\pi_0(a|s) > 0$ *it holds*

$$Q^\pi(s,a) = \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V^{\theta(\pi)}(s') \right). \tag{3.2}$$

*Proof.* The tower property of the expectation is used throughout this proof. The absolute integrability of the series follows from the finiteness of the spaces $\mathcal{S}, \mathcal{A}$. We use Lemma 2.2. For (3.1), let $s \in \mathcal{S}$ and observe that

$$\begin{aligned} V^\pi(s) &= \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \,\middle|\, S_0 = s \right] = \mathbb{E}^\pi \left[ \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \,\middle|\, S_0, A_0 \right] \,\middle|\, S_0 = s \right] \\ &= \mathbb{E}^\pi \left[ Q^\pi(S_0, A_0) \mid S_0 = s \right] = \sum_{a \in \mathcal{A}} \pi_0(a|s) Q^\pi(s,a). \end{aligned}$$

To prove (3.2), observe that, in view of Lemma 2.2(ii), it holds

$$\mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} C(S_t, A_t, S_{t+1}) \mid S_1 = s \right] = \mathbb{E}^{\theta(\pi)} \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \mid S_0 = s \right] = V^{\theta(\pi)}(s).$$

Then by the Markov property, for $(s,a) \in \mathcal{S} \times \mathcal{A}$ with $\pi_0(a|s) > 0$ we obtain

$$\begin{aligned} Q^\pi(s,a) &= \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \,\middle|\, S_0 = s, A_0 = a \right] \\ &= \mathbb{E}^\pi \left[ C(S_0, A_0, S_1) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} C(S_t, A_t, S_{t+1}) \,\middle|\, S_0 = s, A_0 = a \right] \\ &= \mathbb{E}^\pi \left[ C(S_0, A_0, S_1) + \gamma \mathbb{E}^\pi \left[ \sum_{t=1}^{\infty} \gamma^{t-1} C(S_t, A_t, S_{t+1}) \,\middle|\, S_1 \right] \,\middle|\, S_0 = s, A_0 = a \right] \\ &= \mathbb{E}^\pi \left[ C(S_0, A_0, S_1) + \gamma V^{\theta(\pi)}(S_1) \,\middle|\, S_0 = s, A_0 = a \right] \\ &= \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V^{\theta(\pi)}(s') \right). \end{aligned}$$

■

*Proof of Theorem 3.1.* In view of (3.1) and (3.2) from Lemma 3.2, we obtain

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi_0(a|s) Q^\pi(s,a) = \sum_{a \in \mathcal{A}} \pi_0(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V^{\theta(\pi)}(s') \right),$$

respectively

$$\begin{aligned} V^\pi(s) &= \mathbb{E}^\pi \left[ Q^\pi(S_0, A_0) \mid S_0 = s \right] = \mathbb{E}^\pi \left[ \mathbb{E}^\pi \left[ C(S_0, A_0, S_1) + \gamma V^{\theta(\pi)}(S_1) \,\middle|\, S_0, A_0 \right] \mid S_0 = s \right] \\ &= \mathbb{E}^\pi \left[ C(S_0, A_0, S_1) + \gamma V^{\theta(\pi)}(S_1) \,\middle|\, S_0 = s \right]. \end{aligned}$$

■

Before moving to the optimal value function, we provide bounds on the value associated

with an arbitrary policy. Given the finite state and action spaces, there exists a finite bound

$$\bar{C} := \max_{s,s' \in \mathcal{S}, a \in \mathcal{A}} \left\| C(s,a,s') \right\|_\infty < \infty$$

for the supremum norm (as well as each component) of the multi-cost function. We show that the value associated with any policy is bounded (in the supremum norm) by $M := \frac{\bar{C}}{1-\gamma}$.

**Lemma 3.3.** *For an admissible policy $\pi \in \Pi$, it holds $V^\pi \in \mathscr{L}(\mathcal{S}, \mathbb{B}_M^n)$, i.e.,*

$$\left\| V^\pi(s) \right\|_\infty \leq \frac{\bar{C}}{1-\gamma} \quad \text{for all states } s \in \mathcal{S}.$$

*Proof.* For an arbitrary history $(S_0, A_0, S_1, A_1, \dots)$, it holds

$$\left\| \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \right\|_\infty \leq \max_{i=1,\dots,n} \sum_{t=0}^{\infty} \gamma^t \left| C_i(S_t, A_t, S_{t+1}) \right| \leq \frac{\bar{C}}{1-\gamma}.$$

The result follows. ∎

**Remark 3.4.**

- Bounds on the multi-cost would be similarly obtained if the finiteness of the state and action spaces was replaced by a compactness assumption alongside mild continuity of the cost function.
- For the state-action value function $Q^\pi$ we obtain an identical bound $\left\| Q^\pi(s,a) \right\|_\infty \leq \frac{\bar{C}}{1-\gamma}$ across all $(s,a) \in \mathcal{S} \times \mathcal{A}$, that is, $Q^\pi \in \mathscr{L}(\mathcal{S} \times \mathcal{A}, \mathbb{B}_M^n)$.

## 4 Optimal value function

We now return to the multi-objective problem faced by the decision maker, the problem of minimizing the expected discounted multi-cost across all admissible Markov policies,

$$\min_{\pi \in \Pi} \quad V^\pi(S_0) = \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \mid S_0 \right] \quad \text{w.r.t. } \leq_{\mathbb{R}_+^n}. \tag{4.1}$$

The aim of this work is to establish a Bellman-type recursive relationship for this multi-objective control problem. It has been established in the literature in the context of finite-horizon problems that the counterpart of the dynamic programming principle for *multi-objective* controls is a *set-valued Bellman's principle*, see e.g. [KR21].

We define the (optimal) value function of problem (4.1) as the set (collection) of functions providing the policy values

$$\mathbb{V}^* = \{ s \mapsto V^\pi(s) \mid \pi \in \Pi \} \subseteq \mathscr{L}(\mathcal{S}, \mathbb{R}^n)$$

across all admissible policies. This section delivers the main result in two steps. First, we prove

the validity of the *set-valued Bellman optimality equation* for the value function of the infinite horizon problem. Second, we interpret this optimality equation in terms of a *fixed point*. These results show that the structure of dynamic programming results for infinite horizon problems is preserved also in the context of multiple objectives, when the value function is interpreted as a set of functions.

**Remark 4.1.** In the context of multi-objective control, the optimal value function is often treated as a *set-valued function* attaining the upper image of the multi-objective problem, see [KR21, CK26]. In this paper, we instead work with *a set of functions* – specifically, the collection of policy-valued functions across all admissible policies – which enables an easy interpretation of the set-valued Bellman's principle in terms of a fixed point. See Appendix A for a further discussion.

## 4.1 Set-valued Bellman optimality equation

For single-objective control on infinite horizon, the dynamic programming principle has the form of a *Bellman optimality equation*, which shows that the optimal value function can be obtained through one-step recursions. We show that this remains true also in the multi-objective case, when the optimal value function is treated as a set of functions.

**Theorem 4.2.** *It holds*

$$\mathbb{V}^* = \left\{ s \longmapsto \sum_{a\in\mathcal{A}} \varpi(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a)\left(C(s,a,s') + \gamma V(s')\right) \;\middle|\; \varpi \in \mathcal{D},\, V \in \mathbb{V}^* \right\}. \tag{4.2}$$

The proof of Theorem 4.2 will be given after Lemma 4.3.

**Lemma 4.3.** *Let us define the set* $\mathbb{Q}^*$ *of state-action values as*

$$\mathbb{Q}^* := \{(s,a) \mapsto Q^\pi(s,a) \;\mid\; \pi \in \Pi\}.$$

*Then, it holds*

$$\mathbb{Q}^* = \left\{ (s,a) \mapsto \sum_{s'\in\mathcal{S}} P(s'|s,a)\left(C(s,a,s') + \gamma V(s')\right) \;\middle|\; V \in \mathbb{V}^* \right\} \tag{4.3}$$

*and*

$$\mathbb{V}^* = \left\{ s \mapsto \sum_{a\in\mathcal{A}} \varpi(a|s) Q(s,a) \;\middle|\; \varpi \in \mathcal{D},\, Q \in \mathbb{Q}^* \right\}. \tag{4.4}$$

*Proof.* In view of (2.1), it holds $\mathbb{V}^* = \{V^\pi \mid \pi \in \Pi\} = \{V^{\theta(\pi)} \mid \pi \in \Pi\}$. Then, Lemma 3.2 yields

$$\mathbb{Q}^* = \{(s,a) \mapsto Q^\pi(s,a) \;\mid\; \pi \in \Pi\} \overset{(3.2)}{=} \left\{ (s,a) \mapsto \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V^{\theta(\pi)}(s') \right) \;\mid\; \pi \in \Pi \right\}$$
$$= \left\{ (s,a) \mapsto \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V(s') \right) \;\mid\; V \in \mathbb{V}^* \right\}.$$

This proves (4.3). For (4.4), we use Lemma 3.2 to observe

$$\mathbb{V}^* = \{s \mapsto V^\pi(s) \;\mid\; \pi \in \Pi\} \overset{(3.1)}{=} \left\{ s \mapsto \sum_{a \in \mathcal{A}} \pi_0(a|s) Q^\pi(s,a) \;\mid\; \pi \in \Pi \right\}$$
$$\subseteq \left\{ \; s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) Q(s,a) \;\middle|\; \varpi \in \mathcal{D},\, Q \in \mathbb{Q}^* \; \right\}.$$

For the reverse inclusion, let $\varpi \in \mathcal{D}$ and $Q \in \mathbb{Q}^*$. By definition of the collection $\mathbb{Q}^*$, there exists a policy $\pi \in \Pi$ such that $Q = Q^\pi$. In view of (2.1), also policy $(\varpi, \theta(\pi)) \in \Pi$ is admissible and since $Q^\pi$ does not depend on the decision rule $\pi_0$, it follows that $Q^\pi = Q^{(\varpi,\theta(\pi))}$. Finally, the above together with Lemma 3.2 implies

$$\left\{ \; s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) Q(s,a) \;\middle|\; \varpi \in \mathcal{D},\, Q \in \mathbb{Q}^* \; \right\} = \left\{ \; s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) Q^\pi(s,a) \;\middle|\; \varpi \in \mathcal{D},\, \pi \in \Pi \; \right\}$$
$$= \left\{ \; s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) Q^{(\varpi,\theta(\pi))}(s,a) \;\middle|\; \varpi \in \mathcal{D},\, \pi \in \Pi \; \right\}$$
$$\overset{(3.1)}{=} \left\{ \; s \mapsto V^{(\varpi,\theta(\pi))}(s) \;\middle|\; \varpi \in \mathcal{D},\, \pi \in \Pi \; \right\} \subseteq \mathbb{V}^*.$$

■

*Proof of Theorem 4.2.* The set-valued Bellman optimality equation follows from Lemma 4.3,

$$\mathbb{V}^* \overset{(4.4)}{=} \left\{ \; s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) Q(s,a) \;\middle|\; \varpi \in \mathcal{D},\, Q \in \mathbb{Q}^* \; \right\}$$
$$\overset{(4.3)}{=} \left\{ \; s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V(s') \right) \;\middle|\; \varpi \in \mathcal{D},\, V \in \mathbb{V}^* \; \right\}.$$

■

In the next subsection, we will interpret the set-valued Bellman optimality equation in terms of a *fixed point*. For that purpose, we define a mapping $f : \mathcal{D} \times \mathscr{L}(\mathcal{S}, \mathbb{R}^n) \to \mathscr{L}(\mathcal{S}, \mathbb{R}^n)$ as

$$f(\varpi, v)(s) = \int_{\mathcal{A}} \int_{\mathcal{S}} \left( C(s,a,s') + \gamma v(s') \right) P(\mathrm{d}s'|s,a) \varpi(\mathrm{d}a|s). \tag{4.5}$$

We note that the recursion (4.2) can be expressed in terms of the mapping $f$ as

$$\mathbb{V}^* = \{ \; f(\varpi, v) \mid \varpi \in \mathcal{D},\, v \in \mathbb{V}^* \; \}.$$

In Lemma 3.3 we have shown that a policy value $V^\pi$ corresponding to an admissible policy $\pi$ is bounded, i.e., $V^\pi \in \mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)$. Now we show that the mapping $f$ preserves this bound.

**Lemma 4.4.** *If $\varpi \in \mathcal{D}$ and $v \in \mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)$, then $f(\varpi, v) \in \mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)$.*

*Proof.* For $v \in \mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)$ it holds $|v_i(s')| \leq M$ for all $s' \in \mathcal{S}$ and all $i = 1, \dots, n$. Then,

$$\begin{aligned}
\|f(\varpi, v)(s)\|_\infty &= \left\| \sum_{a \in \mathcal{A}} \varpi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma v(s') \right) \right\|_\infty \\
&\leq \max_{i=1,\dots,n} \sum_{a \in \mathcal{A}} \varpi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a) \cdot \left| C_i(s,a,s') + \gamma v_i(s') \right| \\
&\leq \sum_{a \in \mathcal{A}} \varpi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a) \cdot (\bar{C} + \gamma M) = M
\end{aligned}$$

■

## 4.2 Fixed point

For single-objective infinite horizon control problems, the Bellman optimality equation allows an interpretation as a *fixed point*. We show that the same is true also for the set-valued counterpart in case of a multi-objective infinite horizon control problem.

In light of the set-valued Bellman optimality equation (4.2) and Lemma 4.4, we define a mapping $\mathbb{F} : \mathcal{P}(\mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)) \to \mathcal{P}(\mathscr{L}(\mathcal{S}, \mathbb{B}^n_M))$ as

$$\begin{aligned}
\mathbb{F}(\mathbb{V}) &:= \left\{ s \mapsto \sum_{a \in \mathcal{A}} \varpi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s,a) \left( C(s,a,s') + \gamma V(s') \right) \;\middle|\; \varpi \in \mathcal{D},\, V \in \mathbb{V} \right\} \\
&= \left\{ f(\varpi, v) \;\middle|\; \varpi \in \mathcal{D},\, v \in \mathbb{V} \right\}.
\end{aligned}$$

Observe that $\mathbb{F}$ maps a set of functions onto a set of functions and Lemma 4.4 guarantees that if $\mathbb{V} \subseteq \mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)$, then $\mathbb{F}(\mathbb{V}) \subseteq \mathscr{L}(\mathcal{S}, \mathbb{B}^n_M)$.

The set-valued Bellman optimality equation (4.2) stated in terms of the mapping $\mathbb{F}$ corresponds to $\mathbb{V}^* = \mathbb{F}(\mathbb{V}^*)$. The *Tarski fixed point framework* allows us to interpret the value function $\mathbb{V}^*$ as a fixed point of the mapping $\mathbb{F}$. In contrast to the scalar setting, herein we do not obtain uniqueness of the fixed point. However, we prove below that the optimal value function $\mathbb{V}^*$ corresponds to the *largest* fixed point of the mapping $\mathbb{F}$.

**Corollary 4.5.** *Let $g : \mathbb{X} \times \mathbb{Y} \to \mathbb{X}$ and define $G : \mathcal{P}(\mathbb{X}) \to \mathcal{P}(\mathbb{X})$ by $G(X) := \{g(x,y) \,|\, x \in X, y \in Y_0\}$ for any $X \in \mathcal{P}(\mathbb{X})$ and some fixed $Y_0 \subseteq \mathbb{Y}$. The set of fixed points of $G$ (i.e., $\{X^* \subseteq \mathbb{X} \mid X^* = G(X^*)\}$) forms a complete lattice under the subset order relation $\subseteq$. In particular, there exists a largest fixed point $\bar{X} = G(\bar{X})$ such that $X^* \subseteq \bar{X}$ for any $X^* = G(X^*)$.*

*Proof.* Note that $(\mathcal{P}(\mathbb{X}), \subseteq)$ forms a complete lattice where the infimum is defined by the intersection and the supremum by the union (as the largest subset of a collection of sets is the intersection and the smallest superset is the union). Additionally, trivially, $G(X) \subseteq G(X')$ for any $X \subseteq X' \in \mathcal{P}(\mathbb{X})$. The result trivially follows by the Tarski fixed point theorem [AB06, Theorem 1.11]. ■

**Remark 4.6.** The smallest fixed points for $G$ as constructed in Corollary 4.5 is, trivially, the empty set $\emptyset$.

In order to prove that the optimal value function $\mathbb{V}^*$ is the largest fixed point of the mapping $\mathbb{F}$, we rewrite the mapping as $\mathbb{F}(\mathbb{V}) = \{f(\varpi, v) \,|\, \varpi \in \mathcal{D}, v \in \mathbb{V}\}$.

**Theorem 4.7.** *Let $\mathbb{V} \subseteq \mathscr{L}(\mathcal{S}, \mathbb{B}_M^n)$ be a fixed point of $\mathbb{F}$, i.e., it holds $\mathbb{V} = \mathbb{F}(\mathbb{V})$. Then $\mathbb{V} \subseteq \mathbb{V}^*$. Hence, $\mathbb{V}^*$ is the largest fixed point of $\mathbb{F}$.*

*Proof.* Consider an arbitrary $v \in \mathbb{V}$, where $\mathbb{V}$ is a fixed point of $\mathbb{F}$. Then, there exists $\varpi_0 \in \mathcal{D}$ and $v^{(1)} \in \mathbb{V}$ such that $v = f(\varpi_0, v^{(1)})$. By the same argument, $v^{(1)} = f(\varpi_1, v^{(2)})$ for some $\varpi_1 \in \mathcal{D}$ and $v^{(2)} \in \mathbb{V}$. We can repeat this argument $k$ times for any $k \in \mathbb{N}$ and obtain

$$\begin{aligned} v(s_0) &= f(\varpi_0, f(\varpi_1, \dots f(\varpi_{k-1}, v^{(k)})))(s_0) \\ &= \int_{\mathcal{A}\times\mathcal{S}} \cdots \int_{\mathcal{A}\times\mathcal{S}} \left( \sum_{t=0}^{k-1} \gamma^t C(s_t, a_t, s_{t+1}) + \gamma^k v^{(k)}(s_k) \right) P(\mathrm{d}s_k | s_{k-1}, a_{k-1}) \varpi_{k-1}(\mathrm{d}a_{k-1} | s_{k-1}) \\ &\qquad \dots P(\mathrm{d}s_1 | s_0, a_0) \varpi_0(\mathrm{d}a_0 | s_0) \end{aligned}$$

for all $s_0 \in \mathcal{S}$. In view of Lemma 2.2(iv), it holds

$$v(s_0) = \mathbb{E}^\pi \left[ \sum_{t=0}^{k-1} \gamma^t C(S_t, A_t, S_{t+1}) + \gamma^k v^{(k)}(S_k) \;\middle|\; S_0 = s_0 \right]$$

for a policy $\pi \in \Pi$ satisfying $(\pi_0, \dots, \pi_{k-1}) = (\varpi_0, \dots, \varpi_{k-1})$. In particular, we obtain that for $v \in \mathbb{V}$, there exists a policy $\pi \in \Pi$, such that for all $s_0 \in \mathcal{S}$, it holds

$$v(s_0) = \lim_{k\to\infty} \mathbb{E}^\pi \left[ \sum_{t=0}^{k-1} \gamma^t C(S_t, A_t, S_{t+1}) + \gamma^k v^{(k)}(S_k) \;\middle|\; S_0 = s_0 \right]. \tag{4.6}$$

Now, let $|\mathcal{S}| \in \mathbb{N}$ denote the cardinality of the finite state space $\mathcal{S}$ and let $\mathcal{S} = \{s_1, \dots, s_{|\mathcal{S}|}\}$ denote its elements. Then, for each $k \in \mathbb{N}$, we can define an $n|\mathcal{S}|$-dimensional vector $\mathfrak{v}^k := (v^{(k)}(s_1), \dots, v^{(k)}(s_{|\mathcal{S}|})) \in \mathbb{R}^{n|\mathcal{S}|}$. Note that the sequence of vectors $(\mathfrak{v}^k)_{k\in\mathbb{N}}$ is uniformly bounded since each element of the sequence $(v^{(k)})_{k\in\mathbb{N}} \subseteq \mathbb{V}$ is uniformly bounded by $M$. Therefore, by Heine-Borel, there exists a subsequence $(\mathfrak{v}^{k_j})_{j\in\mathbb{N}}$, and an element $\mathfrak{v}_\infty \in \mathbb{R}^{n|\mathcal{S}|}$ such that

$\lim_j \mathfrak{v}^{k_j} = \mathfrak{v}_\infty$. In particular, this implies for every $j \in \mathbb{N}$ and $s \in \mathcal{S}$ that

$$\left\| v^{(k_j)}(s) \right\|_\infty \leq \left\| \mathfrak{v}^{k_j} \right\|_\infty \leq \sup_\ell \left\| \mathfrak{v}^{k_\ell} \right\|_\infty < \infty.$$

This, together with dominated convergence and the assumption that $0 < \gamma < 1$, ensures that

$$\lim_{j\to\infty} \mathbb{E}^\pi \left[ \gamma^{k_j} v^{(k_j)}(S_{k_j}) \mid S_0 = s_0 \right] = 0.$$

Therefore, by (4.6) and using that the multi-cost function $C$ is uniformly bounded by $\bar{C} < \infty$, we obtain by dominated convergence that

$$\begin{aligned} v(s_0) &= \lim_{j\to\infty} \mathbb{E}^\pi \left[ \sum_{t=0}^{k_j-1} \gamma^t C(S_t, A_t, S_{t+1}) + \gamma^{k_j} v^{(k_j)}(S_{k_j}) \mid S_0 = s_0 \right] \\ &= \mathbb{E}^\pi \left[ \sum_{t=0}^{\infty} \gamma^t C(S_t, A_t, S_{t+1}) \mid S_0 = s_0 \right] = V^\pi(s_0). \end{aligned}$$

Hence, $v = V^\pi \in \mathbb{V}^*$ and $\mathbb{V} \subseteq \mathbb{V}^*$. Since $\mathbb{V}$ was an arbitrary fixed point of $\mathbb{F}$, it follows that $\mathbb{V}^*$ is the largest set being a fixed point of $\mathbb{F}$ and contains all other fixed points of $\mathbb{F}$, which means by the subset relation that $\mathbb{V}^*$ is the largest fixed point of $\mathbb{F}$. ■

# A Optimal value function: Set of functions versus set-valued function

We define the optimal value function as the *set of all policy value functions*

$$\mathbb{V}^* = \{ s \mapsto V^\pi(s) \mid \pi \in \Pi \} \subseteq \mathscr{L}(\mathcal{S}, \mathbb{R}^n).$$

In contrast, other works (e.g. [KR21, CK26]) define the optimal value function of a multi-objective control problem as a *set-valued function*. In the context of the present problem, such (alternative) optimal value function would be $\mathcal{V}^* : \mathcal{S} \rightrightarrows \mathbb{R}^n$ given by

$$\mathcal{V}^*(s) = \{ V^\pi(s) \mid \pi \in \Pi \} = \{ v(s) \mid v \in \mathbb{V}^* \}.$$

Clearly, all elements of $\mathbb{V}^*$ are *selectors* of the set-valued function $\mathcal{V}^*$, i.e. $\mathbb{V}^* \subseteq \operatorname{Sel} \mathcal{V}^*$, where $\operatorname{Sel} \mathcal{V}^*$ denotes the set of all selectors of $\mathcal{V}^*$. However, Example A.1 illustrates that not all selectors of $\mathcal{V}^*$ belong to $\mathbb{V}^*$, i.e. not all selectors of $\mathcal{V}^*$ are attainable by a single admissible policy across all states. Thus, $\mathbb{V}^* \subsetneq \operatorname{Sel} \mathcal{V}^*$ can hold. Following, e.g., [Mol05, Theorem 2.1.6], this can only happen if $\mathbb{V}^*$ is not a *decomposable* set, i.e., there exists a pasting $s \mapsto V(s) := V^\pi(s)\mathbf{1}_{\{s\in D\}} + V^{\bar{\pi}}(s)\mathbf{1}_{\{s\in D^c\}}$ for some $\pi, \bar{\pi} \in \Pi$ and $D \subseteq \mathcal{S}$ such that $V \notin \mathbb{V}^*$. We leave the study of when we recover decomposability of $\mathbb{V}^*$ for future work. Here $\mathbf{1}_{\{\cdot\}}$ denotes the indicator

function.

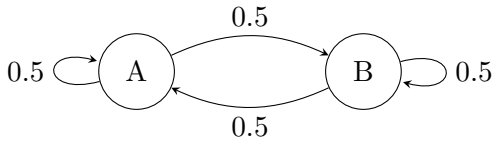



Figure 2: Transition probabilities of MDP in Example A.1.

**Example A.1.** Consider a simple MDP with two states $\mathcal{S} = \{A, B\}$ and two actions $\mathcal{A} = \{l, r\}$. The transition probabilities are independent of the chosen action and are depicted in Figure 2, the cost function depends only on the action and is given by $C(s, a, s') = C(a) = \binom{1}{0}\mathbf{1}_{\{a=l\}} + \binom{0}{1}\mathbf{1}_{\{a=r\}}$.

The policy $\pi^l$ of always choosing action $l$ yields the value $V^{\pi^l}(A) = V^{\pi^l}(B) = \sum_{t=0}^{\infty}\gamma^t\binom{1}{0} = \binom{\frac{1}{1-\gamma}}{0}$. Similarly, the policy $\pi^r$ of always choosing action $r$ yields $V^{\pi^r}(A) = V^{\pi^r}(B) = \binom{0}{\frac{1}{1-\gamma}}$. Then, a function $v : \mathcal{S} \to \mathbb{R}^2$ given by

$$v(A) = \begin{pmatrix} \frac{1}{1-\gamma} \\ 0 \end{pmatrix}, \qquad v(B) = \begin{pmatrix} 0 \\ \frac{1}{1-\gamma} \end{pmatrix}$$

is a *selector* of the set-valued function $\mathcal{V}^*$, i.e., $v(s) \in \mathcal{V}^*(s)$ for all $s \in \mathcal{S}$. But $v$ is not produced by any admissible policy, i.e., $v \notin \mathbb{V}^*$.

The main body of this paper derived the appropriate dynamic programming principle for the value function $\mathbb{V}^*$ in Theorem 4.2. A natural follow up question is whether also $\mathcal{V}^*$ admits a recursive form. As a consequence of the Bellman optimality equation (4.2) for $\mathbb{V}^*$, for all states $s \in \mathcal{S}$, it holds

$$\mathcal{V}^*(s) = \left\{ \sum_{a\in\mathcal{A}} \varpi(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a)\left(C(s,a,s') + \gamma V(s')\right) \;\middle|\; \varpi \in \mathcal{D},\, V \in \mathbb{V}^* \right\}. \tag{A.1}$$

Note, however, that (A.1) is not a full recursion for the set-valued function $\mathcal{V}^*$ since it involves only a subset $\mathbb{V}^* \subsetneq \operatorname{Sel}\mathcal{V}^*$ of the selectors of $\mathcal{V}^*$. However, a true recursion, and thus a set-valued Bellman's principle for the set-valued function $\mathcal{V}^*$ can be obtained.

**Theorem A.2.** *For all states $s \in \mathcal{S}$, it holds*

$$\mathcal{V}^*(s) = \left\{ \sum_{a\in\mathcal{A}} \varpi(a|s) \sum_{s'\in\mathcal{S}} P(s'|s,a)\left(C(s,a,s') + \gamma V(s')\right) \;\middle|\; \varpi \in \mathcal{D},\, V \in \mathit{Sel}\,\mathcal{V}^* \right\}. \tag{A.2}$$

We provide a *sketch of the proof.* Consider a pair $\varpi \in \mathcal{D}$, $V \in \operatorname{Sel}\mathcal{V}^*$ generating an element of the right-hand side of equation (A.2). Given a selector, for every state $s_1 \in \mathcal{S}$ there exists

a (Markov randomized) policy $\pi^{(s_1)} \in \Pi$ such that $V(s_1) = V^{\pi^{(s_1)}}(s_1)$. Consequently, the considered right-hand side element of equation (A.2) is generated as an expectation under a *history-dependent* randomized policy $\bar{\pi}$ given by $\bar{\pi}_0(\cdot|s_0) = \varpi(\cdot|s_0)$ and $\bar{\pi}_t(\cdot|s_1, s_t) = \pi_{t-1}^{(s_1)}(\cdot|s_t)$ for $t > 0$.

To show that the right-hand side element belongs to the set $\mathcal{V}^*(s)$, it needs to be shown that an objective value arising from a history-dependent randomized policy can also be obtained with a Markov randomized policy. [CMH06, Theorem 3] provides this result in a setting with multi-cost (more precisely, reward) functions $\tilde{C} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$. Defining $\tilde{C}(s, a) = \int_{\mathcal{S}} C(s, a, s')P(\mathrm{d}s'|s, a)$ for all pairs $(s, a)$ reduces our problem to the setting of [CMH06]. We omit the remaining details for the sake of brevity. ■

Note that it is quite common in the multi-objective case to obtain a set-valued Bellman's principle for several objects related to the multi-objective problem. For instance, in [KR21] is was shown that for the mean-risk problem the *upper images* recurs backwards in time, but also the *image sets* as well as the *weakly efficient frontiers*. The focus in [KR21] was then on the upper image as it is a convex set, and thus the most convenient to work with and to approximate and the best choice for a numerical implementation of the recursion.

The above Bellman's principle (A.2) for the set-valued function $\mathcal{V}^*$ corresponds to the recursion of the image sets. However, as the focus in the present paper is on deriving for the first time a Bellman's principle for a multi-objective infinite-horizon Markov decision problem and provide in total analogy to the scalar case an interpretation of it as a fixed point, the set-valued Bellman optimality equation (4.2) for the set $\mathbb{V}^*$ of all policy value functions is the right choice, as it allows for a direct application of Tarski fixed point theorem in Section 4.2. However, for future work it might be the case that other formulations of the Bellman's principle for a multi-objective infinite-horizon Markov decision problem like the one in terms of the set-valued function $\mathcal{V}^*$ derived in Theorem A.2 might be relevant as well.

## References


[AB06] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, Berlin, 2006.

[AF21] Çağın Ararat and Zachary Feinstein. Set-valued risk measures as backward stochastic difference inclusions and equations. *Finance and Stochastics*, 25(1):43–76, 2021.

[BH17] David B Brown and Martin B Haugh. Information relaxation bounds for infinite horizon markov decision processes. *Operations Research*, 65(5):1355–1379, 2017.

[BR11] Nicole Bäuerle and Ulrich Rieder. *Markov Decision Processes with Applications to Finance*. Springer Berlin, Heidelberg, 2011.

[CK26] Igor Cialenco and Gabriela Kováčová. Vector-valued robust stochastic control. *SIAM Journal on Financial Mathematics*, Forthcoming, 2026.

[CMH06] Krishnendu Chatterjee, Rupak Majumdar, and Thomas A. Henzinger. Markov decision processes with multiple objectives. In Bruno Durand and Wolfgang Thomas, editors, *STACS 2006*, pages 325–336, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[FR13] Zachary Feinstein and Birgit Rudloff. Time consistency of dynamic risk measures in markets with transaction costs. *Quantitative Finance*, 13(9):1473–1489, 2013.

[FR14] Zachary Feinstein and Birgit Rudloff. Multi-portfolio time consistency for set-valued convex and coherent risk measures. *Finance and Stochastics*, 19(1):67–107, October 2014.

[FR17] Zachary Feinstein and Birgit Rudloff. A recursive algorithm for multivariate risk measures and a set-valued Bellman's principle. *Journal of Global Optimization*, 68(1):47–69, August 2017.

[FRZ21] Zachary Feinstein, Birgit Rudloff, and Jianfeng Zhang. Dynamic set values for nonzero-sum games with multiple equilibriums. *Mathematics of Operations Research*, 47(1):616–642, 2021.

[HHL+15] Andreas H. Hamel, Frank Heyde, Andreas Löhne, Birgit Rudloff, and Carola Schrage. Set optimization – a rather short introduction. In Andreas H. Hamel, Frank Heyde, Andreas Löhne, Birgit Rudloff, and Carola Schrage, editors, *Set Optimization and Applications - The State of the Art*, pages 65–141, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[HLL12] Onésimo Hernández-Lerma and Jean Bernard Lasserre. *Discrete-Time Markov Control Processes*. Springer New York, 2012.

[HV20] A.H. Hamel and D. Visetti. The value functions approach and hopf-lax formula for multiobjective costs via set optimization. *Journal of Mathematical Analysis and Applications*, 483(1):123605, 2020.

[IZ23] Melih İşeri and Jianfeng Zhang. Set-valued Hamilton–Jacobi–Bellman equations. *arXiv preprint arXiv:2311.05727*, 2023.

[IZ24] Melih İşeri and Jianfeng Zhang. Set values for mean field games. *Transactions of the American Mathematical Society*, 377:7117–7174, 2024.

[KR21] Gabriela Kováčová and Birgit Rudloff. Time consistency of the mean-risk problem. *Operations Research*, 69(4):1100–1117, jul 2021.

[Lie72] Leon H Liebman. A markov decision model for selecting optimal credit control policies. *Management Science*, 18(10):B–519, 1972.

[Mol05] I. Molchanov. *Theory of Random Sets*. Probability and Its Applications. Springer London, 2005.

[Put14] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[Vis23] Daniela Visetti. Characteristic curves for set-valued Hamilton–Jacobi equations. *Set-Valued and Variational Analysis*, 31(3):21, 2023.