Journal of Machine Learning Research 27 (2026) 1-53 Submitted 3/26; Revised 6/26; Published 7/26

Optimising Utility Functions in Multi-Objective Markov Decision Processes

Manel Rodriguez-Soto MANEL.RODRIGUEZ@IIIA.CSIC.ES
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, 08193, Spain

Editor: Marcello Restelli

Abstract

Multi-objective Markov Decision Processes (MOMDPs) are among the most prevalent formal frameworks for addressing sequential decision-making problems involving multiple, potentially conflicting objectives. In most MOMDP approaches, a utility function is employed to aggregate these objectives into a single scalar criterion that encodes user preferences. Despite its widespread adoption, the theoretical foundations of MOMDPs remain incomplete in two main respects: first, there is no general characterisation of the classes of utility functions that guarantee the existence of an optimal policy; second, we do not know which preference relations can be represented by utility functions. This work advances both lines of research through a theoretical analysis of MOMDPs. Specifically, we examine each problem under the two principal formulations of utility functions for MOMDPs: the Scalarised Expected Returns (SER) criterion and the Expected Scalarised Returns (ESR) criterion. Our formal findings allow us to derive formal conditions that describe the families of utility functions and preference relations for which MOMDP algorithms should focus. These analyses can guide the development of new MOMDP algorithms explicitly grounded in our formal results.

Keywords: Markov decision processes, multi-objective Markov decision processes, multi-objective reinforcement learning, utility functions, preference relations

1. Introduction

Sequential decision-making problems are pervasive, influencing diverse domains such as autonomous driving (Elallid et al., 2022), robotics (Zhao et al., 2020), finance (Charpentier et al., 2021), and healthcare (Coronato et al., 2020), among others. In recent years, Reinforcement Learning (RL) has become a fundamental framework for tackling these complex tasks (Kaelbling et al., 1996; Li, 2017). Most RL research has traditionally focused on problems in which an agent is tasked with achieving a single objective (e.g., maximising financial returns in trading or winning a race in a competitive setting). However, real-world applications frequently involve multiple, often conflicting, objectives (Vamplew et al., 2022), such as a self-driving car balancing safety, efficiency, and passenger comfort simultaneously.

Multi-Objective Reinforcement Learning (MORL) has emerged as an up-and-coming framework for addressing decision-making problems that involve multiple objectives (Roijers et al., 2013; Roijers and Whiteson, 2017; Rădulescu et al., 2019). Although MORL is a relatively recent development compared to single-objective RL, its state-of-the-art methods have demonstrated significant potential in tackling inherently multi-objective real-world

challenges (Hayes et al., 2022; Vamplew et al., 2022). The majority of MORL approaches are *utility-based*, and assume that a *utility function* exists to combine all objectives into a single scalar value, thereby enabling the learning agent to weigh trade-offs among competing objectives (Hayes et al., 2022; Vamplew et al., 2024). Thus, the MORL agent's goal becomes learning the optimal policy (behaviour) that maximises the considered utility function in a set environment, called a *Multi-Objective Markov Decision Problem* (MOMDP).

However, determining the most appropriate utility function is a non-trivial and often context-dependent problem. To address this, much of the MORL literature focuses on learning a set of candidate policies, known as the *undominated set*, which optimise all potential utility functions (Hayes et al., 2022; Röpke et al., 2023). This approach ensures that once a utility function is specified, the decision-maker can seamlessly select the policy from the undominated set that maximises the chosen utility, providing flexibility and adaptability in multi-objective decision-making.

Thus, utility functions are widely regarded as a fundamental concept of MOMDPs (Vamplew et al., 2024). Utility functions encode the user's preferences over different objectives and guide learning (Rădulescu et al., 2019). Hence, both concepts (utilities and preferences) are at the core of current MORL approaches. However, the state-of-the-art approach of considering the undominated set as the most general solution concept of MORL relies on two assumptions:

1. **On utilities:** It assumes that we can always optimise any utility function.

2. **On preferences:** It assumes that we can always express any user preference as a utility function.

Unfortunately, these assumptions do not always hold. Numerous examples exist of preferences that cannot be represented by a utility function, such as the lexicographic order (Wray et al., 2015), as formally demonstrated in (Debreu, 1997). Similarly, even in problems with a finite set of states and actions, there are utility functions for which no optimal policy exists for any state. An explicit example of this phenomenon is provided in later sections.

These counterexamples raise the need to answer the following two main theoretical questions: (1) Can we determine the families of utility functions for which optimal policies are guaranteed to exist? (2) Can we determine the family of preferences that can be represented as utility functions? Furthermore, both questions need to be answered at least twice: once for each of the two main criteria (i.e., definitions) of how utility functions are applied in MORL: the *Scalarised Expected Returns* (SER) criterion, and the *Expected Scalarised Returns* (ESR) criterion (Hayes et al., 2022). The state of the art on MORL has only very recently addressed these fundamental research questions for some families of utility functions, and only for the SER criterion (Rodriguez-Soto et al., 2024).

Against this background, in this paper, we advance towards answering both research questions by carrying out an in-depth analysis of utility functions in MOMDPs, for both SER and ESR criteria, leading to the following three contributions:

1. **We provide two novel MOMDP fundamental theoretical concepts**. We introduce the first formal definitions of preferences between policies in MOMDPs and of utility maximisation in MOMDPs for both SER and ESR criteria.

2. Given a utility function in an MOMDP, **we provide sufficient and insufficient conditions to guarantee the existence of an optimal policy** maximising it for both SER and ESR criteria.

3. **We provide sufficient conditions to express preferences between policies as utility functions** for both SER and ESR criteria. These preferences are expressed by a so-called *quasi-representative* utility function, which contains the same set of maximal policies as the preference relation.

We expect our theoretical results to lead to novel MORL algorithms that can exploit the analytical properties of the utility functions in MOMDPs introduced here.

The remainder of the paper is structured as follows. Section 2 reviews the necessary background on multi-objective Markov decision processes. Section 3 then establishes sufficient and insufficient conditions for the existence of utility-maximising policies under the SER criterion, while Section 4 does so for the ESR criterion. Section 5 subsequently presents a family of preferences that can be maximised with utility functions. Section 6 discusses related work, and Section 7 concludes by summarising the main theoretical contributions of this work and outlining directions for future research.

2. Background

The background section is divided into several parts, providing the definitions of all necessary concepts to understand the novel concepts and theorems presented in this work. This section is divided as follows. First, Section 2.1 introduces some probabilistic concepts. Then, Section 2.2 provides a brief background on single-objective Markov decision processes. The remainder of the background section (from Section 2.3 onwards) is devoted to introducing all the relevant and necessary concepts of multi-objective Markov decision processes.

2.1 Probability Theory

It is worthwhile to remind a few basic formal definitions of probabilistic concepts, as they will be required to introduce more complex multi-objective reinforcement learning solution concepts. The definitions provided here can be found in standard probability textbooks such as Bertsekas and Tsitsiklis (2008); McColl (2004).

First of all, a real-valued *random variable* is a function that assigns a real numerical value to each possible outcome of a random experiment. Take, for instance, an experiment in which we randomly select a person and measure their weight. Here, the weight would be a random variable.

Next, the *cumulative distribution function* (CDF) of a random variable describes the probability that the variable takes a value less than or equal to a given real number. The CDF fully characterises the probability distribution of the random variable.

Definition 1 (Cumulative distribution function in random variables) *Let X be a real-valued random variable. The* cumulative distribution function *of X is the function $F_X : \mathbb{R} \to [0, 1]$ defined by*

$$F_X(x) \doteq \mathbb{P}(X \leq x), \qquad \forall x \in \mathbb{R}.$$

Then, a vector of random variables $\vec{X} = (X_1, \dots, X_n)$ is called a *random vector*. Thus, for any random vector, its cumulative distribution function is defined as the joint CDF of each of its random variable components X_i:

Definition 2 (Cumulative distribution function in random vectors) *Let $\vec{X}$ be a real-valued random vector. The* cumulative distribution function *of $\vec{X}$ is the function $F_{\vec{X}} : \mathbb{R}^n \to [0, 1]$ defined by*

$$F_{\vec{X}}(\vec{x}) \doteq \mathbb{P}(X_1 \leq x_1, \dots X_n \leq x_n), \qquad \forall \vec{x} \in \mathbb{R}^n.$$

Another very relevant property of random vectors is their *support*. Intuitively, the support of a random vector is the set of all its possible values (analogous to the image of a function). Formally:

Definition 3 (Support) *Let $\vec{X}$ be a real-valued random vector. The* support *of $\vec{X}$ is the only subset $supp(\vec{X}) \subseteq \mathbb{R}^n$ that satisfies:*

$$\mathbb{P}(\vec{X} \in supp(\vec{X})) = 1 \quad \textit{and} \quad \forall \vec{\mathcal{X}} \subseteq supp(\vec{X}) : \quad \mathbb{P}(\vec{X} \in \vec{\mathcal{X}}) = 1 \implies \vec{\mathcal{X}} = supp(\vec{X}). \quad (1)$$

Now that we have random vectors fully characterised, we need a criterion to order them. The common approach in the MORL literature is to order them with the (first-order) *stochastic dominance* criterion (Hayes et al., 2021; Röpke et al., 2023). Intuitively, a random vector $\vec{X}$ stochastically dominates another one $\vec{Y}$ if, for any threshold value $\vec{x}$, there is a higher or equal probability of obtaining more value in all components x_i of $\vec{x}$ with X than with Y. Formally:

Definition 4 (First-order stochastic dominance of random vectors) *Let $\vec{X}$ and $\vec{Y}$ be real-valued random vectors with cumulative distribution functions $F_{\vec{X}}$ and $F_{\vec{Y}}$, respectively. We say that $\vec{X}$* first-order stochastically dominates *$\vec{Y}$ if and only if*

$$F_{\vec{X}}(\vec{x}) \leq F_{\vec{Y}}(\vec{x}) \quad \forall \vec{x} \in \mathbb{R}^n. \quad (2)$$

2.2 Single-Objective Markov Decision Processes

Markov decision processes (MDPs) (Kaelbling et al., 1996; Sutton and Barto, 1998) formalise sequential decision-making problems. An MDP models an environment in which an agent interacts iteratively, selecting actions that induce transitions between states while receiving an immediate scalar reward after each action, reflecting the objective the agent seeks to optimise:

Definition 5 (Markov Decision Process) *A (single-objective)[1] Markov Decision Process (MDP) is defined as a tuple $\mathcal{M} \doteq \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, T \rangle$ of three sets and one function: the set of states $\mathcal{S}$, the set of actions $\mathcal{A}(s)$ available at each state s, the reward set $\mathcal{R} \subseteq \mathbb{R}$, and the transition function $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathcal{R} \to [0, 1]$ specifying the probability $T(s, a, s', r) =$*

1. Through the paper, we refer to a single-objective MDP simply as an MDP.

$\mathbb{P}(s', r \mid s, a)$ *that the next state is* s' *and next reward is* r *if an action* a *is performed upon the state* s[2].

Following the reinforcement learning literature, we consider the set of rewards $\mathcal{R}$ to be bounded (Puterman, 1998). Moreover, we will use the term *reward function* for any function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ that tells the (expected) reward r an agent receives at state s when applying action a. When needed, we will also use the nomenclature $R(s, a, p) \doteq r$ that specifies with what probability the reward r is obtained at state s applying action a (i.e., $p = \mathbb{P}(r \mid s, a)$)[3].

An agent's behaviour in an MDP is called a *policy*. In its most general form, a policy indicates how likely an agent will perform an action a if the agent is currently in state s, given all the previous transitions visited (Puterman, 1998). This behaviour is referred to as a *non-stationary* policy (also called *history-dependent*, *time-dependent* policy). If the agent's behaviour depends only on the current state s and time t, and is independent of previous transitions, then we say that the agent follows a *stationary* policy. In this work, unless explicitly stated otherwise, we always refer to *stationary* policies. Formally (Sutton and Barto, 1998; Puterman, 1998; Roijers and Whiteson, 2017):

Definition 6 (Stationary policy) *Let* $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, T \rangle$ *be a Markov Decision Process. We define a* (stationary) policy π *of* $\mathcal{M}$ *as any probability distribution of the form* $\pi(s, a) = \mathbb{P}(a \mid s)$.

Interestingly, stationary policies do not depend on which time step we are in. Formally, they are *time-invariant* (also called *Markov* or *Markovian* policies (Puterman, 1998)). A stationary policy $\pi(s, a)$ specifies the probability that an agent adhering to it will execute action a when the current state is s.

The agent aims to learn the policy accumulating the maximum number of expected discounted rewards. Fortunately, the maximum element can be attained within the set of stationary policies (Sutton and Barto, 1998), meaning we only need to compare and evaluate stationary policies. Thus, to evaluate a given (stationary) policy, we need to compute the (expected) discounted sum of rewards that an agent obtains by following it. This operation is formalised by means of the so-called *value function* $V : \mathcal{S} \to \mathbb{R}$, defined as:

$$V^{\pi}(s) \doteq \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R(S_{t+k}, A_{t+k}) \mid S_t = s, \pi], \tag{3}$$

where $\gamma \in [0, 1)$ is the discount factor, indicating how much we care about future rewards, and t is any time-step. To shorten notation, the discounted sum of rewards is called the

2. Some works formalise the transition function T as a probabilistic model of only future states and not of future rewards (i.e., it is assumed that there is a deterministic function $R(s, a, s')$ such that given a triplet $\langle s, a, s' \rangle$ the agent will obtain a fixed reward $R(s, a, s')$). While such a simplification is enough for a large number of MDPs, our analysis necessarily requires considering the most general case in which for any triplet $\langle s, a, s' \rangle$ there might exist multiple associated rewards r with a probability $P(s', r \mid s, a) > 0$.
3. While the notation $R(s, a, p)$ is not common in the literature, it will be very useful to explain some of the examples that we will detail throughout the paper.

discounted return G_t:

$$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R(S_{t+k}, A_{t+k}), \tag{4}$$

which simplifies the value function notation as

$$V^{\pi}(s) \doteq \mathbb{E}[G_t \mid S_t = s, \pi]. \tag{5}$$

Value functions allow us to order policies partially (Sutton and Barto, 1998). Hence, they allow the formalisation of the agent's objective as the problem of learning the stationary policy that maximises the value function. This policy is defined as the *optimal* policy:

Definition 7 (Optimal policy) *Given an MDP* $\mathcal{M}$*, its optimal policy* π_* *is the policy that maximises the value function* V^{π}*. Formally:*

$$V^{\pi_*}(s) \geq V^{\pi}(s), \tag{6}$$

for every state s *of the MDP* $\mathcal{M}$*, and every policy* π *of* $\mathcal{M}$*.*

The optimal policy is the solution concept in single-objective RL. At least one optimal policy exists for any MDP with a finite state and action space, which is also deterministic and stationary (Feinberg, 2011).

2.3 Multi-Objective Markov Decision Processes

Multi-objective reinforcement learning (MORL) focuses on decision-making problems in which an agent must simultaneously optimise multiple, potentially competing objectives (for instance, in healthcare, balancing a patient's clinical outcomes with their autonomy). In standard single-objective RL, the reward function encodes the agent's goal. MORL generalises this framework to environments characterised by multiple reward functions. Such environments are typically formalised as *Multi-Objective Markov Decision Processes* (Roijers et al., 2013; Hayes et al., 2022):

Definition 8 (Multi-Objective MDP) *An* n*-objective Markov Decision Process (MOMDP) is defined as a tuple* $\vec{\mathcal{M}} \doteq \langle \mathcal{S}, \mathcal{A}, \vec{\mathcal{R}}, T \rangle$ *of three sets and one function: the set of states* $\mathcal{S}$*, the set of actions* $\mathcal{A}(s)$ *available at each state* s*, the vector of reward sets* $\vec{\mathcal{R}} \subseteq \mathbb{R}^n$*, providing a reward set* $\mathcal{R}_i$ *for each objective* $i \in \{1, \ldots, n\}$*, and the transition function* $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \vec{\mathcal{R}} \to [0, 1]$ *specifying the probability* $T(s, a, s', \vec{r}) = \mathbb{P}(s', \vec{r} \mid s, a)$ *that the next state is* s' *and next reward is* $\vec{r}$ *if an action* a *is performed upon the state* s*.*

Policies in an MOMDP are assessed through a vectorial value function, denoted $\vec{V}$, which comprises the value functions associated with each objective: $\vec{V}(s) = (V_1(s), \ldots, V_n(s))$.

In scenarios where it is infeasible to simultaneously optimise all objectives, the agent must establish a prioritisation among them. The prevailing assumption in most multi-objective reinforcement learning (MORL) approaches is that the individual value functions corresponding to each objective can be combined into a single scalar quantity that reflects such preferences. Under this framework, the agent's objective is to maximise this aggregated

value function. This aggregation is achieved by means of an utility function u (also referred to as a scalarisation function) as discussed in (Roijers and Whiteson, 2017; Rădulescu et al., 2019; Vamplew et al., 2024). In its most general form, a utility function u is defined as any mapping from a vector space (typically the space of value vectors $\vec{V}$) into the real numbers $\mathbb{R}$. Formally:

Definition 9 (Utility function) *Let $\vec{\mathcal{M}}$ be an MOMDP of n objectives, Let $\mathcal{S}$ be the set of states of $\vec{\mathcal{M}}$, let $\Pi(\vec{\mathcal{M}})$ be the set of policies of $\vec{\mathcal{M}}$, let $\vec{G}_t^{\pi}(s)$ be the random variable of possible returns obtained by following a policy π starting at state $s \in \mathcal{S}$ at time-step t. Let $\vec{\mathcal{X}} \subseteq \mathbb{R}^n$ be any subset of $\mathbb{R}^n$ that satisfies both:*

$$\pi \in \Pi(\vec{\mathcal{M}}) \implies \vec{V}^{\pi}(s) \in \vec{\mathcal{X}} \quad \forall s \in \mathcal{S}, \quad \textit{and} \tag{7}$$

$$\pi \in \Pi(\vec{\mathcal{M}}) \implies \textit{supp}(\vec{G}_t^{\pi}(s)) \subseteq \vec{\mathcal{X}} \quad \forall s \in \mathcal{S}. \tag{8}$$

Then, we say that any function $u : \vec{\mathcal{X}} \to \mathbb{R}$ is a utility function *of $\vec{\mathcal{M}}$.*

Notice that, in particular, for any finite MOMDP with reward set $\vec{\mathcal{R}}$, if we are following discount factor $0 < \gamma < 1$, any function $u : \vec{\mathcal{X}} \to \mathbb{R}$ defined over the set $\vec{\mathcal{X}} = [\frac{1}{1-\gamma}\min_{\mathcal{R}_1}(r), \frac{1}{1-\gamma}\max_{\mathcal{R}_1}(r)] \times \cdots \times [\frac{1}{1-\gamma}\min_{\mathcal{R}_n}(r), \frac{1}{1-\gamma}\max_{\mathcal{R}_n}(r)] \in \mathbb{R}^n$ is always an utility function of the MOMDP.

Throughout this paper, we will study several families of utility functions (such as continuous functions, linear functions, and so on). For all families of utilities, when we say that a given utility function $u : \vec{\mathcal{X}} \to \mathbb{R}^n$ satisfies a given condition (such as being continuous), we always mean that it satisfies such a condition at least over the entirety of its dominion $\vec{\mathcal{X}}$. Importantly, Definition 9 imposes that such a dominion $\vec{\mathcal{X}}$ covers at minimum all possible value vectors and all possible returns of a given MOMDP. Thus, for example, a utility function that is only linear in some closed subset $\vec{\mathcal{X}}$, and another utility function that is linear in the entirety of $\mathbb{R}^n$ will seem equally linear for the agent learning at the MOMDP.

There are two main ways of applying utility functions u in MOMDPs: the *Scalarised Expected Returns* (SER) criterion, and the *Expected Scalarised Returns* (ESR) criterion.

Definition 10 (Utility function criteria) *Given an MOMDP of n objectives and a utility function u, then:*

- ***SER:*** *If u is applied directly to value functions, then we say that u follows the* ***Scalarised Expected Returns*** *criterion. With the SER criterion, the agent's goal can be expressed as learning a policy that maximises the higher-order function*

$$f_V(u, \pi)(s) \doteq u(\mathbb{E}[\vec{G}_t | S_t = s, \pi]) = u(\vec{V}(s)) = (u \circ \vec{V})(s). \tag{9}$$

- ***ESR:*** *If u is applied directly to the vectorial discounted returns $\vec{G}_t$, then we say that u follows the* ***Expected Scalarised Returns*** *criterion. With the ESR criterion, the agent's goal can be expressed as learning a policy that maximises the higher-order function*

$$f_G(u, \pi)(s) \doteq \mathbb{E}[u(\vec{G}_t) | S_t = s, \pi]. \tag{10}$$

The choice of utility criterion has a significant influence on the policies learned by a MORL agent (Roijers et al., 2018; Rădulescu et al., 2020; Hayes et al., 2022). Therefore, it is important to understand when each criterion (SER or ESR) is most appropriate. The SER criterion is most suitable when a user's utility depends on the outcomes of multiple executions of a policy. In such settings, the user is interested in maximising utility across repeated interactions. Thus, here, the SER is the appropriate optimisation objective, as it applies the utility function to the expected (average) returns. Consider, for instance, a navigation system that selects a route for a commuter every day. The commuter is not primarily concerned with the outcome of any single trip, but rather with minimising their average travel time over months or years of commuting. Since utility is derived from the aggregate performance of many trips, rather than from an individual journey, the SER criterion is the appropriate optimisation objective.

By contrast, the ESR criterion is more appropriate when utility depends on the outcome of each particular execution of a policy. In these cases, the user considers the utility of each execution directly, rather than averaging outcomes over many runs. Consequently, the ESR criterion is the correct one here, as it optimises the expected utility of each obtained return. For example, in a medical treatment scenario, a patient may have only one opportunity to choose a treatment. Since the consequences of that single decision cannot be averaged over repeated executions, ESR provides the more appropriate optimisation criterion.

For an in-depth discussion on the matter, we refer to Section 5.3 of (Hayes et al., 2022). Importantly, as recognised in recent surveys such as (Hayes et al., 2022), the ESR criterion has been extensively understudied in the reinforcement learning literature.

The family of linear utility functions is especially notable. Among its most relevant properties, for any linear utility function l, they return the same result for both the ESR and the SER criterion (i.e., $f_G(l, \pi) = f_V(l, \pi)$ for f_G, f_V defined as in Def. 10), as it is well-known in the MORL literature (Roijers et al., 2013).

Recall that, in single-objective MDPs, the solution concept is well defined (a deterministic, stationary optimal policy). In contrast, no universal analogue exists for multi-objective MDPs. Instead, the agent's utility function is typically assumed to be largely unspecified, subject only to mild structural assumptions (for instance, linearity or monotonicity) (Roijers et al., 2013; Rădulescu et al., 2019; Hayes et al., 2022). Consequently, MORL algorithms, rather than identifying a single policy, aim to learn a set of *candidate policies* π, each of which is optimal for some admissible utility function u. The following sections examine the principal solution concepts proposed for MOMDPs.

2.4 Solution Concepts of MOMDPs under the SER Criterion

As explained before, there is a wide array of solution concepts of MOMDPs, which depend on the assumptions imposed on the utility functions. In the absence of any assumption, the objective is to compute the set of maximal policies with respect to *any* utility function. It is therefore essential to clarify what it means for a policy to be *maximal* under a given utility function. A very predominant portion of the MORL literature adopts both the scalarised expected returns (SER) criterion and the so-called *state-independent* (SI) criterion to define optimality (Roijers et al., 2013; Roijers and Whiteson, 2017; Rădulescu et al., 2019; Hayes et al., 2022). Under the SI criterion, a policy π is said to *maximise* a utility function if and

only if it is maximal with respect to the expectation over possible initial states (for a given value function $\vec{V}$ and the random variable of initial states S_0). We denote this expectation by $\vec{V}^{\pi}_{SI}$:

$$\vec{V}^{\pi}_{SI} \doteq \mathbb{E}[\vec{V}^{\pi}(S_0)]. \tag{11}$$

Thus, for any utility function u, under the SER and SI criteria, the optimal policy π_* (if exists) is the one that maximises $u(\vec{V}^{\pi}_{SI})$:

$$\pi_* \doteq \arg\max_{\pi} u(\vec{V}^{\pi}_{SI}). \tag{12}$$

The simplicity of the *state-independent* (SI) criterion has made it become ubiquitous in MORL and RL. While this criterion has no significant side-effects in single-objective RL, it can produce contradictory policies in multi-objective RL, as shown in Theorem 2 further ahead, in Section 3.

Consequently, all solution concepts for MOMDPs have been formulated exclusively under this *state-independent* criterion, In particular, the prevailing general solution, the *undominated set*, is defined as the set of policies that maximise at least one utility function according to the SI criterion. Formally (Hayes et al., 2022):

Definition 11 (Undominated set) *Given an MOMDP $\vec{\mathcal{M}}$, its* undominated set $U(\vec{\mathcal{M}})$ *is defined as the set of policies for which there exists a utility function u with a maximal scalarised value.*

$$U(\vec{\mathcal{M}}) \doteq \{\pi \in \Pi(\vec{\mathcal{M}}) \mid \exists u : \forall \pi' \in \Pi(\vec{\mathcal{M}}),\ u(\vec{V}^{\pi}_{SI}) \geq u(\vec{V}^{\pi'}_{SI})\}, \tag{13}$$

where $\Pi(\vec{\mathcal{M}})$ is the set of all possible policies of an MOMDP $\vec{\mathcal{M}}$.

As previously mentioned, it is typically assumed that the utility function we aim to maximise satisfies at least some properties, such as being strictly monotonically increasing (SMI). In particular, many MORL works only focus on computing the subset of the undominated set restricted to SMI utility functions, known as the *Pareto front*.

Interestingly, in MORL, the Pareto front can be (and usually is) formalised without any mention of utility functions. Instead, the Pareto front definition relies on the concept of *Pareto-dominant* policies[4]. As previously expressed, when facing multiple objectives, it is typically impossible to maximise all of them simultaneously because improvements in one objective may require sacrifices in another. Rather than searching for a single globally optimal solution, one therefore seeks policies whose value vector $\vec{V}^{\pi}$ cannot be improved in any objective without degrading at least one other objective. Such solution policies are said to be Pareto-dominant, and collectively form the Pareto front. Formally:

Definition 12 (Pareto dominance) *Let $\vec{v}, \vec{v}' \in \mathbb{R}^n$ be two n-dimensional vectors. Then, we say that $\vec{v}$* Pareto-dominates *$\vec{v}'$ (and denote it as $\vec{v} \succ_P \vec{v}'$) if and only if:*

$$\forall i \in \{1, \ldots, n\} : v_i \geq v'_i \quad \textit{and} \quad \exists j \in \{1, \ldots, n\} : v_j > v'_j. \tag{14}$$

4. It is easy to prove that both formalisations (with SMI utility functions or with Pareto-dominant policies) are in fact equivalent.

Then, with the concept of Pareto dominance, we formalise the Pareto front as follows:

Definition 13 (Pareto front) *Given an MOMDP $\vec{\mathcal{M}}$, its* Pareto front $PF(\vec{\mathcal{M}})$ *is defined as the set of policies π for which their associated value vector $\vec{V}^{\pi}_{SI}$ is* not *Pareto-dominated by the value vector of any other policy $\vec{V}^{\pi'}_{SI}$:*

$$PF(\vec{\mathcal{M}}) \doteq \{\pi \in \Pi(\vec{\mathcal{M}}) \mid \nexists \pi' \in \Pi(\vec{\mathcal{M}}) : \vec{V}^{\pi'}_{SI} \succ_P \vec{V}^{\pi}_{SI}\}. \tag{15}$$

For any MOMDP, its Pareto front includes all policies that maximise at least one strictly monotonically increasing utility function (SMI) *within the* MOMDP. This distinction is important: in more detail, if the utility function is SMI for at least all possible value vectors or returns of the MOMDP, then every policy maximising it (according to the SI criterion) will be inside the Pareto front.

Next, we consider the subset of the undominated set restricted to linear utility functions, commonly called the *convex hull*. Linear utility functions have several important properties, the most relevant being that they can transform any MOMDP into a single-objective MDP (Roijers and Whiteson, 2017). The convex hull of a MOMDP contains all policies that are maximal for at least one linear utility function (again, according to the SI criterion). Formally (Hayes et al., 2022):

Definition 14 (Convex hull) *Given a MOMDP $\vec{\mathcal{M}}$, and its set of policies Π, its convex hull $CH(\mathcal{M})$ is the subset of policies $\pi \in \Pi$ that are optimal for some weight vector $\vec{w}$:*

$$CH(\mathcal{M}) \doteq \{\pi \in \Pi \mid \exists \vec{w} \in \mathbb{R}^n : \forall \pi' \in \Pi,\ \vec{w} \cdot \vec{V}^{\pi}_{SI} \geq \vec{w} \cdot \vec{V}^{\pi'}_{SI}\}, \tag{16}$$

By definition, the convex hull contains an excessive amount of policies. In practice, it is enough to compute a subset P of the convex hull that contains *at least one policy* that is maximal (according to the SI criterion) for each possible linear utility function. This relevant subset is called the *convex coverage set* (Hayes et al., 2022). The convex coverage set $CCS(\mathcal{M})$ of any MOMDP $\mathcal{M}$ is formalised as:

Definition 15 (Convex coverage set) *Given a MOMDP $\vec{\mathcal{M}}$, and its set of policies Π, a convex coverage set $CCS(\mathcal{M})$ is any set of policies $P \subseteq \Pi$ that contains at least one optimal policy for every weight vector $\vec{w}$:*

$$P \text{ is a } CCS(\mathcal{M}) \iff \forall \vec{w} \in \mathbb{R}^n : \exists \pi \in P : \forall \pi' \in \Pi,\ \vec{w} \cdot \vec{V}^{\pi}_{SI} \geq \vec{w} \cdot \vec{V}^{\pi'}_{SI}. \tag{17}$$

It is clear from Equation 17 that the convex coverage set CCS, in general, is not unique. Moreover, CCS is always guaranteed to exist since, in particular, the complete convex hull of an MOMDP is also a convex coverage set. Furthermore, it is known that, for any MOMDP, it is always possible to build an associated CCS with only deterministic stationary policies (Roijers and Whiteson, 2017). An interesting consequence of this property is that any stationary policy of an MOMDP can be expressed as a linear combination of the finite set of deterministic stationary policies of any CCH. Moreover, all current MORL algorithms focused on obtaining solution sets for linear utility functions compute a convex coverage set (Hayes et al., 2022).

2.5 Solution Concepts of MOMDPs under the ESR Criterion

Some recent works in the literature have tackled the problem of defining optimality under the *expected scalarised returns* (ESR) criterion (Hayes et al., 2021; Röpke et al., 2023). As in the SER case, they also focus exclusively on the *state-independent* (SI) criterion. Thus, given an MOMDP, the goal is to find a policy π that maximises the expectation over possible initial states (for some scalarised discounted return distribution $u(\vec{G}^\pi)$ and the random variable of initial states S_0). Formally:

$$\mathbb{E}[u(\vec{G}^\pi)]_{SI} \doteq \mathbb{E}[f_G(u,\pi)(S_0)] = \mathbb{E}[u(\vec{G}_t)|S_t = S_0, \pi]. \tag{18}$$

Thus, for any utility function u, under the ESR and SI criteria, the optimal policy π_* (if exists) is the one that maximises $\mathbb{E}[u(\vec{G}^\pi)]_{SI}$:

$$\pi_* \doteq \arg\max_\pi \mathbb{E}[u(\vec{G}^\pi)]_{SI}. \tag{19}$$

Computing ESR solution concepts adds difficulty to the SER case, as it requires considering the distribution of all possible returns a policy may obtain, not just the expected return. Hence, all existing ESR solution concepts require a method to compare the returns probability distributions, based on the literature on distributional (single-objective) reinforcement learning (Bellemare et al., 2017).

First of all, notice that the discounted return $\vec{G}_t$ of any policy in an MOMDP is clearly a random vector. Thus, we can establish a stochastic dominance relationship between discounted return distributions.

Interestingly, if the return distribution of some policy π stochastically dominates the return distribution of another policy π', any user with a strictly monotonically increasing utility function would prefer $\vec{X}$ over $\vec{Y}$, as proven in (Röpke et al., 2023). Thus, we are interested in finding policies whose associated value vectors are stochastically dominant. We say that a policy *stochastically dominates* another policy if their associated return distribution do so. Formally:

Definition 16 (First-order stochastic dominance of policies) *Given an MOMDP $\vec{\mathcal{M}}$, let $\pi, \pi' \in \Pi(\vec{\mathcal{M}})$ be two of its policies. Let $\vec{G}^\pi, \vec{G}^{\pi'}$ be their respective associated return distributions. Then, π* first-order stochastically dominates *π' if and only if*

$$F_{\vec{G}^\pi}(\vec{x}) \leq F_{\vec{G}^{\pi'}}(\vec{x}) \quad \forall \vec{x} \in \mathbb{R}^n. \tag{20}$$

Hence, even if the utility function of an MOMDP is unknown, if we assume that such a function is SMI, it suffices to compute a policy that stochastically dominates all the rest. The set of all stochastically dominant policies is called the *distributed undominated set* (DUS) (Röpke et al., 2023). Among its most relevant properties, the DUS is a superset of any Pareto front (Def. 13).

2.6 Beyond Utility Functions

As aforementioned, *most* MORL approaches assume that the solution policy maximises an aggregation of the multiple objectives determined by some utility function. Consequently,

they develop algorithms to compute solution sets, such as the undominated set or the Pareto front.

Nevertheless, there is also a well-established body of MORL literature that does not even consider utility functions (e.g., (Gábor et al., 1998; Van Moffaert et al., 2014; Skalse et al., 2022)). Instead, they aim to find policies whose associated value vectors satisfy specific conditions, which may or may not be represented as a utility function. Probably the most famous example is the subfield of *lexicographic* MORL, in which the goal is to compute a policy that satisfies a given *lexicographic order* (Wray et al., 2015; Vamplew et al., 2018; Li and Czarnecki, 2019; Vamplew et al., 2021).

Lexicographic orders determine a ranking between objectives, and comparisons between policies are performed sequentially following this ranking. The most important objective is considered first, and only if two policies are equivalent with respect to this objective is the next objective used to break the tie, and so on.

This ordering reflects decision scenarios in which certain criteria are considered fundamentally more important than others. For example, a safety-related objective may take absolute precedence over performance metrics, meaning that improvements in lower-priority objectives cannot compensate for losses in higher-priority ones. In the context of MOMDPs, lexicographic orders induce a total ordering over policies by comparing their value vectors. Formally:

Definition 17 (Lexicographic order) *Let $v, v' \in \mathbb{R}^n$ be two n-dimensional vectors, and let $l = \langle o(1), \dots, o(n) \rangle$ be a permutation of $\{1, \dots, n\}$. Then, we say that v dominates v'* according to the lexicographic order induced by l *(and denote it as $v \succ_l v'$) if and only if:*

$$\forall i \in \{1, \dots, k-1\} : v_{\sigma(i)} = v'_{\sigma(i)} \qquad \text{and} \quad v_{\sigma(k)} \geq v'_{\sigma(k)}. \tag{21}$$

Despite the increasing prominence of MORL works that do not consider utility functions, their connection with the rest of *utility-based* MORL literature still has not been formally established, which leads to our interest in formalising preference relations in MOMDPs in the upcoming Section 5. Finally, we discuss this topic in more detail in Section 6.

3. Utility-Optimal Policies under the SER Criterion

In this section, we tackle the first two contributions aforementioned in the Introduction. These contributions are: (1) providing a formal definition of utility maximisation in MORL, and (2) providing the sufficient conditions of utility functions that guarantee the existence of a policy maximising them. A graphical summary of the formal results introduced in this section is presented in Figure 1. In particular, in this section we are going to focus on utility functions that follow the SER criterion (Def. 10), the most common utility function criterion in the MORL literature (Roijers and Whiteson, 2017). Moreover, we already mentioned that the MORL literature formalises the majority of its solution concepts under the *state-independent* criterion. However, for a complete evaluation of utility functions in MOMDPs, we need to formalise solution concepts that account for every state of an MOMDP (i.e., following the *state-dependent* criterion). The MORL research area almost exclusively defines solution concepts under the *state-independent* criterion. However, a complete analysis of utility functions in MOMDPs needs more general optimality definitions

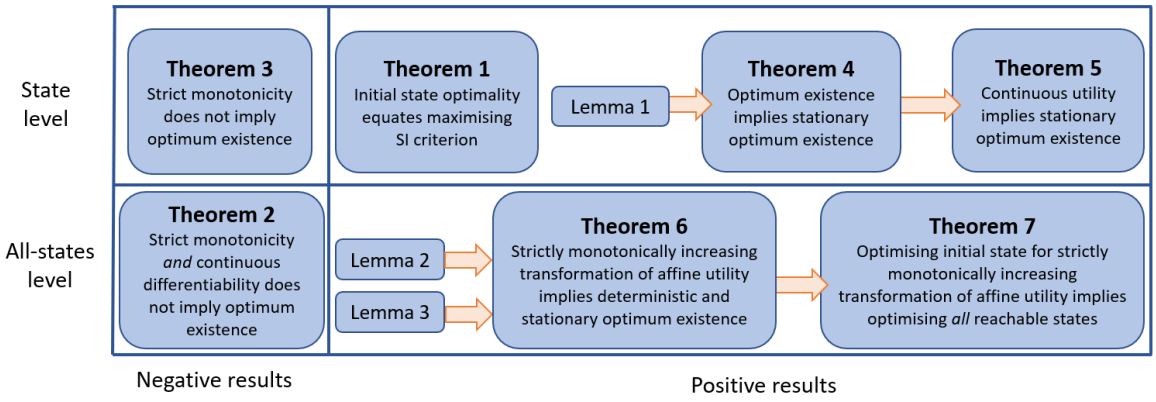

Figure 1: All theorems and lemmas presented in Section 3, which relate to utility-optimal policies under the SER criterion. Small rectangles indicate lemmas, and large rectangles indicate theorems. Arrows indicate if a given lemma or theorem requires a previous result for its proof. Formal results are classified in two axes: either state-level results (related with Definition 18), or all-state-level results (related with Definition 19), and either negative (located in Section 3.1) or positive results (located elsewhere in Section 3).

that take into account each state of an MOMDP individually. An in-depth analysis on why the state-dependent criterion is also necessary is offered at later Section 3.5, after all necessary theorems to back up this claim have been presented.

Moreover, in single-objective MDPs, by virtue of Banach's fixed-point theorem, a deterministic and stationary optimal policy is guaranteed to exist for any finite MDP (and in particular for every state) (Feinberg, 2011). These guarantees weaken substantially in the multi-objective setting. In particular, there exist finite MOMDPs for which no optimal policy exists at any state.

These more fragile formal properties further motivate the study of optimal policy existence in MOMDPs at two distinct levels: the *state* level (i.e., focusing on a single state) and the *all-states* level (i.e., considering the entire state space).

We proceed by formalising *utility-optimality* at the state level. We say that a policy π is optimal at a state s for some utility function u if and only if it yields a higher scalarised expected discounted return than all remaining policies at this state s. Formally:

Definition 18 (SER utility-optimal policy at a state) *Let $\vec{\mathcal{M}}$ be an MOMDP with the state set $\mathcal{S}$, the policy set $\Pi(\vec{\mathcal{M}})$, and a utility function u. We say that a policy $\pi_* \in \Pi(\vec{\mathcal{M}})$ is* optimal with respect to utility function u at state $s \in \mathcal{S}$ *under the SER criterion if and only if:*

$$(u \circ \vec{V}^{\pi_*})(s) \geq (u \circ \vec{V}^{\pi})(s), \tag{22}$$

for every policy $\pi \in \Pi(\vec{\mathcal{M}})$. We say that π_* is $\langle u, s \rangle_{SER}$-optimal for short.

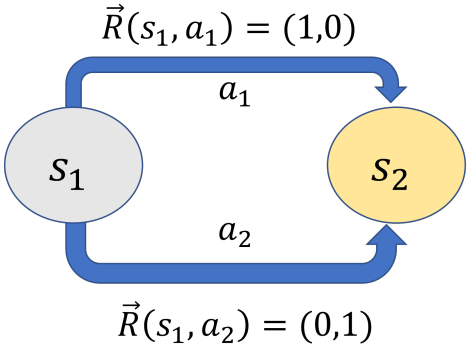

Figure 2: MOMDP of Example 1, Example 2, Theorem 3, Example 5, and Example 6. Circles indicate states and arrows indicate transitions. The yellow circle is a terminal state.

By abuse of notation, in this section, we refer to any utility-optimal policy at state s as $\langle u, s\rangle$-optimal, removing the SER subscript.

Example 1 *Consider an MOMDP $\vec{\mathcal{M}}$ with two states: an initial state s_1 and a terminal state s_2. An agent can perform two actions (a_1, a_2) in this environment, with rewards $\vec{R}(s_1, a_1) = (1, 0)$, $\vec{R}(s_1, a_2) = (0, 1)$ respectively. Figure 2 illustrates such MOMDP.*

Consider the utility function $u(x, y) = x + sin(y)$. Any stochastic and stationary policy π will obtain scalarised value $\sum_a u(\pi(s, a) \cdot \vec{R}(s, a))$. The formula for non-stationary policies is analogous. Importantly, the scalarised value of any policy resides within $u(k, 1 - k)$ for some $k \in [0, 1]$.

Hence, the deterministic and stationary policy $\pi_*(s_1) = a_1$ that obtains vectorial value $\vec{V}^{\pi_*}(s_1) = (1, 0)$ is clearly $\langle u, s_1\rangle$-optimal since $sin(k) < k$ for any $k \in [0, 1]$. To illustrate, the $\langle u, s_1\rangle$-optimal policy $\pi_1(s_1) = a_1$ achieves scalarised value $u(1, 0) = 1 + sin(0) = 1$, while the policy $\pi_2(s_1) = a_2$ achieves $u(0, 1) = 0 + sin(1) \approx 0.84$.

Notice that there is a clear connection between the prevailing state-independent (SI) criterion in MORL, and $\langle u, s\rangle$-optimal policies. Trivially, for environments with a single initial state s_0, maximising a policy according to a utility function u following the SI criterion amounts to computing a $\langle u, s_0\rangle$-optimal policy. In the general case, with multiple initial states, we can interpret the policy maximising the SI criterion as an $\langle u, s'_0\rangle$-optimal policy of an almost identical MOMDP with a unique initial state s'_0 that subsumes all initial states of the original MOMDP. Theorem 1 provides the formal details:

Theorem 1 *Let $\vec{\mathcal{M}}$ be a finite MOMDP with state set $\mathcal{S}$ and let u be any utility function of $\vec{\mathcal{M}}$. Let $\mu : \mathcal{S} \to [0, 1]$ be the probability distribution of each state $s \in \mathcal{S}$ being the initial state, and let $\gamma \in [0, 1]$ be the discount factor.*

Then, consider an alternative MOMDP $\vec{\mathcal{M}}'$ identical to $\vec{\mathcal{M}}$ except that:

- *Its state space contains one more state $\mathcal{S}' = \mathcal{S} \cup \{s_0\}$ such that s_0 is the only initial state of $\vec{\mathcal{M}}'$.*

- *Its reward set $\mathcal{R}'$ multiplies any reward of the original reward set $\mathcal{R}$ by $\frac{1}{\gamma}$.*
- *A single action a can be applied to s_0. The reward obtained by applying it at s_0 is null $\vec{R}'(s_0, a) = \vec{0}$.*
- *The transition dynamics of state s_0 are defined as follows: $\mathbb{P}(s \mid s_0, a) = \mu(s)$.*

Then, for any policy $\pi' \in \Pi(\vec{\mathcal{M}}')$, consider its associate $\pi \in \Pi(\vec{\mathcal{M}})$ such that $\pi = \pi'|_{\mathcal{S}}$ is the restricted version of π' in $\mathcal{S}$:

- *π maximises u under the SI criterion in $\vec{\mathcal{M}}$* if and only if *π' is $\langle u, s_0 \rangle$-optimal in $\vec{\mathcal{M}}'$.*

Proof It suffices to prove that for any pair of policies π, π' such that $\pi = \pi'|_{\mathcal{S}}$, the values $\vec{V}^{\pi}_{SI}$ and $\vec{V}^{\pi'}(s_0)$ are identical:

$$\begin{aligned}
\vec{V}^{\pi'}(s_0) =& \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k \vec{R}'(S_{t+k}, A_{t+k}) \mid S_t = s_0, \pi'] && (23)\\
=& \mathbb{E}[\vec{0} + \sum_{k=1}^{\infty} \gamma^k R'(S_{t+k}, A_{t+k}) \mid S_t = s_0, \pi'] && (24)\\
=& \mathbb{E}[\sum_{k=1}^{\infty} \gamma^k R'(S_{t+k}, A_{t+k}) \mid S_t = S_0, \pi] && (25)\\
=& \mathbb{E}[\sum_{k=1}^{\infty} \gamma^k \cdot \frac{1}{\gamma} R(S_{t+k}, A_{t+k}) \mid S_t = S_0, \pi] && (26)\\
=& \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R(S_{t+k}, A_{t+k}) \mid S_t = S_0, \pi] && (27)\\
=& \vec{V}^{\pi}_{SI}, && (28)
\end{aligned}$$

where recall that $\mathcal{S}_0$ is the random variable of possible initial states of $\vec{\mathcal{M}}$, determined by μ.

Thus, we proved that:

$$u(\vec{V}^{\pi}_{SI}) = \max_{\rho \in \Pi(\vec{\mathcal{M}})} u(\vec{V}^{\rho}_{SI}) \quad \iff \quad \pi' \text{ is } \langle u, s_0 \rangle\text{-optimal in } \vec{\mathcal{M}}'. \tag{29}$$

■

Importantly, Theorem 1 proves that any formal result of $\langle u, s \rangle$-optimal policies will apply to policies maximising the SI criterion and vice-versa. In fact, this relationship also holds under the ESR criterion, but we will need to wait until Section 4.1 to prove it. Henceforth, by abuse of notation but without loss of generality, we will refer to policies that maximise the SI criterion as $\langle u, \mathcal{S}_0 \rangle$-policies.

In direct analogy with single-objective reinforcement learning, we define a policy in an MOMDP to be utility-optimal at the *all-states* level (or simply utility-optimal) if it is utility-optimal in every state of the MOMDP simultaneously. Formally:

Definition 19 (SER utility-optimal policy) *Let $\vec{\mathcal{M}}$ be an MOMDP with the state set $\mathcal{S}$, the policy set $\Pi(\vec{\mathcal{M}})$, and a utility function u. We say that a policy $\pi_* \in \Pi(\vec{\mathcal{M}})$ is* optimal with respect to utility function *u under the SER criterion if and only if:*

$$(u \circ \vec{V}^{\pi_*})(s) \geq (u \circ \vec{V}^{\pi})(s), \tag{30}$$

for every policy $\pi \in \Pi(\vec{\mathcal{M}})$, and every state $s \in \mathcal{S}$. We say that π_* is u_{SER}-optimal for short.

By abuse of notation, in this section, we refer to any utility-optimal policy as u-optimal, removing the SER subscript.

Example 2 *In the MOMDP in Example 1 (illustrated in Figure 2), policy $\pi(s_1) = a_1$ is u-optimal since there are only two states, and the second one is terminal.*

3.1 Negative Results

It is well-known that, for any linear utility function, a deterministic and stationary u-optimal policy is guaranteed to exist (see Section 2.3), as in the single-objective case. However, this guarantee does not extend to arbitrary utility functions. In fact, it is well-known that for some utility functions, the only $\langle u, s, \rangle$-optimal policies are stochastic and non-stationary (e.g. (Perny and Weng, 2010; Siddique et al., 2020)). In this section, we go a step further and prove that there are MOMDPs and utility functions for which there is no policy maximising them, neither at the state level nor at the all-states level. Next, we characterise several *insufficient* conditions for utility functions that are not enough to guarantee the existence of (non-stationary) deterministic or (non-stationary) stochastic $\langle u, s \rangle$-optimal or u-optimal policies in finite MOMDPs. These conditions are:

1. **For u-optimal policies (all-states level):** being simultaneously *strictly monotonically increasing* and *continuously differentiable*. Theorem 2 shows a counter-example.

2. **For $\langle u, s \rangle$-optimal policies (state level):** being *strictly monotonically increasing*. Theorem 3 shows a counter-example.

Notice that strictly monotonically increasing functions are a family of utility functions specially studied in MORL (Roijers and Whiteson, 2017; Hayes et al., 2022)). Before proving both theorems, as a preliminary, Example 3 below illustrates why restricting to deterministic policies is non-viable in MOMDPs, a well-known fact in the MORL literature (Roijers and Whiteson, 2017). In more detail, Example 3 illustrates how some utility functions are not guaranteed to have associated (non-stationary) deterministic u-optimal policies. To the best of our knowledge, this fact was first noticed in Example 4 of Perny and Weng (2010) but only for stationary policies. Example 3 considers the *Chebyshev* function (also called *Tchebycheff*), a well-known utility function in MORL (Perny and Weng, 2010; Roijers et al., 2013; Roijers and Whiteson, 2017; Hayes et al., 2022). The Chebyshev function assigns higher scalar values to inputs x that are closer to a given *reference* value $\vec{r}$. Additionally, this function is also monotonically increasing (Roijers and Whiteson, 2017).

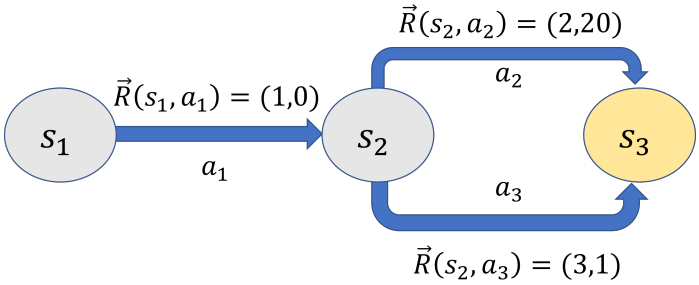

Figure 3: MOMDP of Example 3 and Theorem 2. Circles indicate states, and arrows indicate transitions. The yellow circle is a terminal state.

Example 3 *Let $\vec{r} \in \mathbb{R}^n$ be a reference vector and $\vec{w} \in \mathbb{R}^n$ a non-negative weight vector (i.e., $w_i \geq 0$ for all i). The Chebyshev scalarisation function $\psi_{\vec{r},\vec{w}} : \mathbb{R}^n \to \mathbb{R}$ is formally defined as (Steuer and Choo, 1983; Perny and Weng, 2010):*

$$\psi_{\vec{r},\vec{w}}(x) \doteq -\max_i w_i \, |r_i - x_i|. \tag{31}$$

Consider a 2-objective deterministic MDP $\vec{\mathcal{M}}$ with states s_1, s_2, and terminal state s_3.

- *In s_1, there is a single action a_1 with reward $\vec{R}(s_1, a_1) = (1, 0)$, which transitions to s_2.*
- *In s_2, there are two actions: a_2 with reward $\vec{R}(s_2, a_2) = (2, 20)$ and a_3 with reward $\vec{R}(s_2, a_3) = (3, 1)$. Both actions lead to the terminal state s_3.*

Figure 3 illustrates the whole MOMDP. There are only two deterministic policies in $\vec{\mathcal{M}}$. The first policy is $\pi_1(s_1) = a_1, \pi_1(s_2) = a_2$. This policy obtains values $\vec{V}^{\pi_1}(s_1) = (3, 20), \vec{V}^{\pi_1}(s_2) = (2, 20)$. The second policy is $\pi_2(s_1) = a_1, \pi_2(s_2) = a_3$. This policy obtains values $\vec{V}^{\pi_2}(s_1) = (4, 1), \vec{V}^{\pi_2}(s_2) = (3, 1)$. We define a Chebyshev function with reference point $\vec{r} = (3.5, 20)$, associated weights $\vec{w} = (1, 1/19)$. For such a configuration, the scalarised values per state of the two deterministic policies would be:

- *For policy π_1, $\psi(\vec{V}^{\pi_1}(s_1)) = -0.5$, $\psi(\vec{V}^{\pi_1}(s_2)) = -1.5$.*
- *For policy π_2, $\psi(\vec{V}^{\pi_2}(s_1)) = \psi(\vec{V}^{\pi_2}(s_2)) = -1$.*

Clearly, π_1 is the only deterministic $\langle \psi, s_1 \rangle$-optimal policy, while π_2 is the only deterministic $\langle \psi, s_2 \rangle$-optimal policy. Thus, no deterministic ψ-optimal policy exists.

Next Theorem 2 goes beyond Example 2 and provides a utility function that is both *strictly* monotonically increasing *and continuously differentiable* for which there cannot be any u-optimal policy:

Theorem 2 *Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ be a function under the SER criterion that is both strictly monotonically increasing and continuously differentiable in the set $\vec{\mathcal{X}} \subseteq \mathbb{R}^n$. Then, there are finite MOMDPs $\vec{\mathcal{M}}$ for which u is an utility function but no policy $\pi \in \Pi(\vec{\mathcal{M}})$ is u-optimal.*

Proof Consider $\vec{\mathcal{X}} = \mathbb{R}^+ \times \mathbb{R}$, and the function $u(x, y) = \sqrt{x^2 + 1} + \frac{y}{20}$, which is both strictly monotonically increasing and continuously differentiable for any $(x, y) \in \vec{\mathcal{X}}$. Let $\vec{\mathcal{M}}$ be the same 2-objective deterministic MDP from Example 3, which is graphically depicted in Figure 3. Clearly u is an utility function of $\vec{\mathcal{M}}$, satisfying Definition 9.

This MOMDP $\vec{\mathcal{M}}$ has the same two deterministic policies from Example 3. The first policy is $\pi_1(s_1) = a_1, \pi_1(s_2) = a_2$, which obtains scalarised values $u(\vec{V}^{\pi_1}(s_1)) \approx 4.16, u(\vec{V}^{\pi_1}(s_2)) \approx 3.24$. The second policy is $\pi_2(s_1) = a_1, \pi_2(s_2) = a_3$, which obtains scalarised values $u(\vec{V}^{\pi_2}(s_1)) \approx 4.17, u(\vec{V}^{\pi_2}(s_2)) \approx 3.21$.

It is easy to check that π_1 is a $\langle u, s_2 \rangle$-optimal policy, while π_2 is a $\langle u, s_1 \rangle$-optimal policy. Any stochastic policy π can be expressed as a convex combination of two deterministic policies, namely $\pi = p\,\pi_1 + (1 - p)\,\pi_2$, with $p \in [0, 1]$. Under this representation we can compute the value $\vec{V}^{\pi}(s)$ and scalarised value $u(\vec{V}^{\pi}(s))$ of any policy π for both states:

$$\vec{V}^{\pi}(s) = (\alpha_s - p,\, 19p + 1) \implies u(\vec{V}^{\pi}(s)) = \sqrt{(\alpha_s - p)^2 + 1} + \frac{19p + 1}{20}, \tag{32}$$

where $\alpha_{s_1} = 4$ and $\alpha_{s_2} = 3$. Notice that the scalarised value $u(\vec{V}^{\pi}(s))$ of any policy π is completely determined by parameters $\langle p, s \rangle$. Hence, let us define a function $f(p, s) = u(\vec{V}^{\pi}(s))$. We can compute its derivative with respect to p as:

$$f'(p, s) = \frac{p - \alpha_s}{\sqrt{p^2 - 2\alpha_s p + \alpha_s^2 + 1}} + \frac{19}{20}, \tag{33}$$

The derivative $f'(p, s_1)$ has a root at $r_1 \approx 0.958$, while $f'(p, s_2)$ has a root at $r_2 \approx -0.042$. In both cases, these roots correspond to global minima. Consequently, over the interval $[0, 1]$, the maximum of $f'(p, s_1)$ is attained at $p = 0$, and the maximum of $f'(p, s_2)$ is attained at $p = 1$.

Therefore, π_1 is the unique $\langle u, s_2 \rangle$-optimal policy, whereas π_2 is the unique $\langle u, s_1 \rangle$-optimal policy. It follows that no u-optimal policy exists. ■

Next, we focus on state-level optimality. Here, we prove that there are finite MOMDPs whit utility functions u such that, despite being strictly monotonically increasing, not even stochastic non-stationary $\langle u, s \rangle$-optimal policy exist for a single state. Formally:

Theorem 3 *Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ be a function under the SER criterion that is strictly monotonically increasing in the set $\vec{\mathcal{X}} \subseteq \mathbb{R}^n$. Then, there are finite MOMDPs $\vec{\mathcal{M}}$ for which u is an utility function but no policy $\pi \in \Pi(\vec{\mathcal{M}})$ is $\langle u, s \rangle$-optimal policy for any state s of the $\vec{\mathcal{M}}$.*

Proof Consider the set $\vec{\mathcal{X}} = \mathbb{R}^2$ and the utility function $u : \vec{\mathcal{X}} \to \mathbb{R}$ defined by

$$u(x, y) = \frac{x}{2} + y + \mathbf{1}_{\{x > 0\}},$$

where $\mathbf{1}_{\{x>0\}}$ is equal to 1 if $x > 0$ and 0 otherwise. It is clear that this function u is strictly monotonically increasing over $\mathbb{R}^2$. Moreover, we can quickly notice that this policy prioritises the second objective y over the first objective x, but requires that x is not completely neglected.

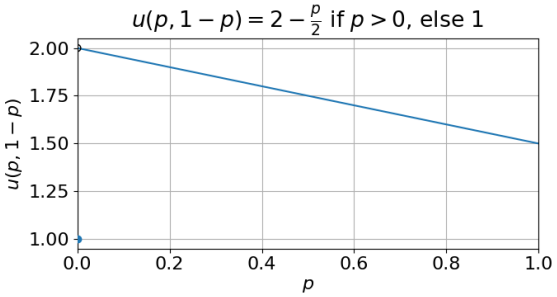

Figure 4: Visual representation of all possible scalarised values of all policies of the MDP of Theorem 3. For every policy π of this MDP, there is some $p \in [0, 1]$ such that $u(p, 1-p)$ is exactly the scalarised value of π. The x-axis ranges from 0 to 1, while the y-axis shows that no policy attains a scalarised value of exactly 2.

Let $\vec{\mathcal{M}}$ be the two-objective deterministic MDP introduced in Example 1, consisting of two states s_1 and s_2, where s_1 is the initial state and s_2 is the terminal state. At state s_1, there are two available actions with associated rewards $\vec{R}(s_1, a_1) = (1, 0)$ and $\vec{R}(s_1, a_2) = (0, 1)$.

Any policy in $\vec{\mathcal{M}}$ can be written as $\pi(s_1, a_1) = p$ and $\pi(s_1, a_2) = 1-p$, for some $p \in [0, 1]$. The corresponding value vector at state s_1 is then

$$\vec{V}^{\pi}(s_1) = (p,\, 1-p).$$

For any policy with $p > 0$ (i.e., with at least some probability of selecting action a_1 with positive reward in x), we have

$$u(p, 1-p) = \frac{p}{2} + (1-p) + 1 = 2 - \frac{p}{2}.$$

Therefore:

$$\sup_{p>0} u(p, 1-p) = \lim_{p \to 0^+} \left(2 - \frac{p}{2}\right) = 2.$$

However, the value 2 is not attained by any policy because for $p = 0$, the utility is $u(0, 1) = 1$. Despite that fact, we can get arbitrarily close to the value of 2 without never reaching it: for every policy π with $\pi(s_1, a_1) = p > 0$, there exists $\varepsilon > 0$ such that $0 < p - \varepsilon < p$, and the policy π' satisfying $\pi'(s_1, a_1) = p - \varepsilon$ obtains

$$u(\vec{V}^{\pi'}(s_1)) = 2 - \frac{p-\varepsilon}{2} > 2 - \frac{p}{2} = u(\vec{V}^{\pi}(s_1)).$$

Hence, no $\langle u, s_1 \rangle$-optimal policy exists in this MOMDP. Finally, as a visual illustration of our proof, Figure 4 illustrates the scalarised values of all possible policies in this MOMDP. ■

The conclusions of Theorem 3 might seem especially counter-intuitive. According to Lu et al. (2023), the Pareto front is convex (recall from Def. 13: the set of $\langle u, \mathcal{S}_0 \rangle$-optimal policies for any given strictly monotonically increasing function), and thus all its policies are guaranteed to be easily obtained by solving a series of single-objective MDPs. The subtlety to understand this seeming paradox is that *if an utility function has an $\langle u, \mathcal{S}_0 \rangle$-optimal policy,* it belongs to the Pareto front. However, there is no guarantee that such utility function can be maximised in the first place, as Theorem 3 just proved.

Moreover, the result of Theorem 3 is particularly noteworthy, as much of the MORL literature focuses on computing $\langle u, S_0 \rangle$-optimal policies for strictly monotonically increasing utility functions (Roijers et al., 2013; Hayes et al., 2022) (i.e., following the state-independent criterion, which considers only the initial states S_0). However, as established in Theorem 3, the existence of such optimal policies is not guaranteed.

These negative results naturally lead to the question of identifying which families of utility functions ensure the existence of at least one global stationary utility-optimal policy, or at least one stationary utility-optimal policy per state. The interest in stationary policies is customary in (single-objective and multi-objective) reinforcement learning Roijers and Whiteson (2017). Formally, in the following subsections, we address the two problems below, focusing on the SER criterion:

Problem 1 *For which families of utility functions is the existence of a stationary $\langle u, s \rangle$-optimal policy for every state s in any finite MOMDP guaranteed?*

Problem 2 *For which families of utility functions is the existence of a stationary u-optimal policy in any finite MOMDP guaranteed?*

Following Sections 3.2 and 3.3, then provide *sufficient conditions* for the existence of utility-optimal policies at the state level and at the all-states level, respectively.

3.2 Existence of a Utility-Optimal Policy in a State

Next, we tackle Problem 1 by presenting a sufficient condition for utility functions u that ensures the existence of stationary (possibly stochastic) $\langle u, s \rangle$-optimal policies for each state s of any finite MOMDP. We remark that those stationary policies need to be $\langle u, s \rangle$-optimal among the set of all possible policies Π, including non-stationary stochastic ones.

This sufficient condition to prove the existence of stationary $\langle u, s \rangle$-optimal policies is that the utility function u must be continuous. To prove it, we require first the indispensable Theorem 4 that constrains the search space for the utility-optimal policy to that of stationary policies.

Lemma 1 *Let $\vec{\mathcal{M}}$ be any finite MOMDP. Then, for every policy π, there exists a stationary policy π' such that, for each state s of $\vec{\mathcal{M}}$, it obtains the same expected returns as π: $\vec{V}^{\pi}(s) = \vec{V}^{\pi'}(s)$.*

Proof It follows directly by generalisation from single-objective MDPs, where for any policy π there exists a stationary policy π' such that, for every state s, both policies yield the same value (i.e., $V^{\pi}(s) = V^{\pi'}(s)$). Theorem 5.5.1 of (Puterman, 1998) provides a complete proof for the single-objective case. ■

Theorem 4 *Let $\vec{\mathcal{M}}$ be any finite MOMDP, let $\mathcal{S}$ be its state set, let $\Pi(\vec{\mathcal{M}})$ be its policy set, and let $\Pi(\vec{\mathcal{M}}^s)$ be its set of stationary policies. Consider any utility function u of $\vec{\mathcal{M}}$ under the SER criterion. If an $\langle u, s\rangle$-optimal policy $\pi_* \in \Pi(\vec{\mathcal{M}})$ exists for a state $s \in \mathcal{S}$, there exists at least one stationary policy $\pi'_* \in \Pi(\vec{\mathcal{M}}^s)$ such that π'_* is also $\langle u, s\rangle$-optimal.*

Proof Let π_* be such that $u(\vec{V}^{\pi_*})(s) \geq u(\vec{V}^{\pi})(s)$. Then, by Lemma 1, there exists another stationary policy π'_* such that $\vec{V}^{\pi_*}(s) = \vec{V}^{\pi'_*}(s)$. Thus, $u(\vec{V}^{\pi'_*}(s)) = u(\vec{V}^{\pi_*}(s)) \geq u(\vec{V}^{\pi}(s))$, and so π'_* is also $\langle u, s\rangle$-optimal. ■

Lemma 1 and Theorem 4 allow us to finally prove why the family of continuous utility functions satisfies Problem 1:

Theorem 5 *Let $\vec{\mathcal{M}}$ be a finite MOMDP and let $\mathcal{S}$ be its set of states. Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ be a utility function of $\vec{\mathcal{M}}$ that is continuous in $\vec{\mathcal{X}}$. Then, for each state $s \in \mathcal{S}$, there exists at least one stationary $\langle u, s\rangle$-optimal policy.*

Proof Our proof consists of proving that the restricted dominion $\vec{\mathcal{D}} \subseteq \vec{\mathcal{X}}$ of any utility function u over all possible value vectors of a finite MOMDP $\vec{\mathcal{M}}$ is necessarily a closed and bounded set. Thereafter, we can apply the Extreme Value Theorem (Theorem 4.16 in Rudin (1976), Theorem 8 of Section 14 in Stewart (2011)), which states that any continuous function $f : \vec{\mathcal{D}} \to \mathbb{R}$ whose dominion is a closed and bounded set $\vec{\mathcal{D}} \subset \mathbb{R}^n$ (i.e., a compact subset of $\mathbb{R}^n$) is guaranteed to have absolute maxima and minima in $\vec{\mathcal{D}}$.

Given an MOMDP $\vec{\mathcal{M}}$, consider the polytope (i.e., an n-dimensional bounded polyhedron) induced at a state s by a convex coverage set CCS of $\vec{\mathcal{M}}$ (recall Equation 15). This polytope has a finite number of vertices, each corresponding to a deterministic stationary policy (Roijers and Whiteson, 2017).

By definition, the CCS contains all attainable value vectors at state s by stationary policies. Moreover, we know that the CCS also contains all attainable value vectors at state s by non-stationary policies by virtue of Theorem 4. Furthermore, any such value vector can be expressed as a convex combination of the value vectors associated with the deterministic stationary policies in the CCS (Roijers and Whiteson, 2017).

Let us define $\vec{\mathcal{D}} \subseteq \mathbb{R}^n$ to the polytope of all value vectors contained within the CCS. It is clear that such set $\vec{\mathcal{D}} \subseteq \vec{\mathcal{X}}$ contains the dominion of all possible inputs of utility function u in a given MOMDP. Moreover, it is also clear that $\vec{\mathcal{D}}$ is both bounded and closed by CCS.

Therefore, since u is a continuous function defined over a set $\vec{\mathcal{D}}$ that is closed and bounded, by the extreme value theorem, there exists some element $\vec{V}_*(s) \in \vec{\mathcal{D}}$ for which $u(\vec{V}_*(s))$ is the absolute maximum of u.

By definition, this policy π is a stationary $\langle u, s\rangle$-optimal policy. ■

It is important to note that Theorem 5 establishes only sufficient conditions for utility-optimal policy existence. Consequently, $\langle u, s\rangle$-optimal policies may still exist for certain discontinuous utility functions. Indeed, the proof of the upcoming Theorem 12 provides an example of a discontinuous utility function for which $\langle u, s\rangle$-optimal policies exist for every state s.

An interesting consequence of Theorem 5 is that, for any initial state s_0 of any MOMDP, the undominated set (recall Definition 11) is guaranteed to contain at least one stationary $\langle u, s_0 \rangle$-optimal policy for every continuous utility function u. For continuous *and strictly monotonically increasing* utility functions, such policy will also belong to the Pareto front (Definition 13). Finally, all linear utility functions are in particular continuous and it is already known that their utility-optimal policies belong to the Convex hull (Definition 14).

3.3 Existence of a Utility-Optimal Policy

Next, we focus on tackling Problem 2. We want to find stationary policies that are u-optimal for every state in an MOMDP. Again, we remark that we aim to find stationary (possibly stochastic) policies that are u-optimal among the set of all possible policies Π, including non-stationary stochastic ones. Problem 2 is substantially more demanding than Problem 1, which only required optimality at a single state. In this broader setting, neither continuous differentiability nor strict monotonicity of the utility function is sufficient, as shown in Theorem 2.

It is well known that a u-optimal policy always exists when u is linear. This naturally raises the question of whether other families of utility functions also guarantee the existence of such policies. Theorem 6 addresses this by identifying a broader family: utility functions obtained by composing an affine transformation with a strictly monotonically increasing function. We first present several auxiliary results that will be instrumental in proving this theorem.

Lemma 2 *For every finite single-objective MDP $\mathcal{M}$, any utility function u that is strictly monotonically increasing preserves the ordering between policies based on their values. That is, for every two value functions V_1 and V_2, for every state s:*

$$(u \circ V_1)(s) > (u \circ V_2)(s) \iff V_1(s) > V_2(s), \tag{34}$$

$$(u \circ V_1)(s) = (u \circ V_2)(s) \iff V_1(s) = V_2(s). \tag{35}$$

In particular, any optimal policy is also u-optimal in $\mathcal{M}$ and vice-versa.

Proof Direct from the definition of a strictly monotonic function. ■

Next Lemma 3 proves another auxiliary result: that a deterministic and stationary u-optimal policy always exists when u is affine.

Lemma 3 *For every finite MOMDP $\vec{\mathcal{M}}$, for any affine utility function u, a deterministic and stationary u-optimal policy exists in $\vec{\mathcal{M}}$.*

Proof We offer a constructive proof: we find this deterministic and stationary u-optimal policy. Consider any affine utility function $u(\vec{x}) = \vec{w} \cdot \vec{x} + \vec{b}$ with $\vec{w}, \vec{b} \in \mathbb{R}^n$. Next, given an MOMDP $\vec{\mathcal{M}}$ with reward set $\vec{\mathcal{R}}$, consider an associated single-objective MDP $\mathcal{M}$ with identical states, actions, and transition function, but with a reward set $\mathcal{R}$ defined as:

$$\vec{r} \in \vec{\mathcal{R}} \iff \vec{w} \cdot \vec{r} \in \mathcal{R}. \tag{36}$$

It is clear that there exists at least one deterministic and stationary optimal policy π_* in $\mathcal{M}$. This policy π_* (or any other optimal policy of $\mathcal{M}$) satisfies for every state s:

$$V^{\pi_*}(s)_{\mathcal{M}} = \max_{\pi} \mathbb{E}[\sum_{t=0}^{\infty} \gamma^k \vec{w} \cdot \vec{R}(S_{t+k}, A_{t+k}) \mid S_t = s, \pi] \tag{37}$$

$$= \max_{\pi} \vec{w} \cdot \mathbb{E}[\sum_{t=0}^{\infty} \gamma^k \vec{R}(S_{t+k}, A_{t+k}) \mid S_t = s, \pi] \tag{38}$$

$$= \max_{\pi} \vec{w} \cdot \vec{V}(s)_{\vec{\mathcal{M}}} \tag{39}$$

$$= \max_{\pi} \vec{w} \cdot \vec{V}(s)_{\vec{\mathcal{M}}} + \vec{b} \tag{40}$$

$$= \max_{\pi} u(\vec{V}(s)_{\vec{\mathcal{M}})}, \tag{41}$$

where the notation $\vec{V}^{\pi}_{\vec{\mathcal{M}}}$ indicates the (MO)MDP where π is being evaluated. Thus, by definition, π_* is also u-optimal in the MOMDP $\vec{\mathcal{M}}$. ■

Building on these two lemmas, we can establish a family of non-linear utility functions for which a stationary u-optimal policy exists. Moreover, for this family of utility functions, there is always at least one deterministic and stationary utility-optimal policy. Specifically, this class consists of utility functions obtained by composing an affine transformation with a strictly monotonically increasing function.

Theorem 6 *Let $\vec{\mathcal{M}}$ be a finite multi-objective MDP, and let u be a utility function that can be decomposed as $u(x) = h(g(x))$, where g is an affine utility function of $\vec{\mathcal{M}}$ and h is a strictly monotonically increasing function over the image of g. Then, there exists at least one deterministic and stationary u-optimal policy.*

Proof First, we show that any utility function u that can be written as the composition of a linear function and a strictly monotonically increasing function is guaranteed to have u-optimal policies:

(i) By Lemma 2, applying a strictly monotonically increasing utility function to a single-objective MDP does not change the set of optimal policies.

(ii) Any linear utility function l can be used to convert a multi-objective MDP $\vec{\mathcal{M}}$ into a standard single-objective MDP $\mathcal{M}$. The optimal policies of $\mathcal{M}$ are exactly the l-optimal policies of $\vec{\mathcal{M}}$.

(iii) Combining these observations, consider a utility function u' such that $u'(x) = h(l(x))$, where h is strictly monotonically increasing and l is linear. Then u preserves the same set of deterministic and stationary optimal policies as l. In particular, there exists at least one u'-optimal policy that is both deterministic and stationary.

Next, we prove that two families of utility functions share the same utility-optimal policies (i.e., two utility functions u, u' for which every u-optimal policy is u'-optimal and

vice-versa) : those formed by composing a strictly monotonically increasing function with either: (1) an affine or (2) a linear function.

(iv) The proof of Lemma 3 shows that any affine utility function $g(x)$ shares the same utility-optimal policies as a linear utility function $l(x) = g(x) - g(0)$. Specifically, in any MOMDP, a policy is g-optimal if and only if it is l-optimal. This follows from the fact that any affine function can be written as $g(x) = A(x) + b$, where $A(x)$ is linear and $b = g(0)$ is a constant. Thus, $l(x) = g(x) - g(0) = A(x) + b - g(0) = A(x)$ is always a linear function.

(v) Now consider a function $u(x) = h(g(x))$, where h is strictly monotonically increasing and g is affine. Define a second function $u'(x) = h(g(x) - g(0))$. Since $g(x) - g(0)$ is linear, u' is the composition of a strictly monotonically increasing function and a linear function. Therefore, by part (iii), u' admits deterministic and stationary optimal policies.

(vi) From part (iv), the functions $g(x)$ and $l(x) = g(x) - g(0)$ share the same utility-optimal policies. Because strictly monotonically increasing functions preserve order, it follows that $u(x) = h(g(x))$ and $u'(x) = h(g(x) - g(0))$ also share the same utility-optimal policies.

Finally, since there is at least one u'-optimal policy that is both deterministic and stationary, and u and u' share the same optimal policies, we conclude that there exists at least one deterministic and stationary u-optimal policy. ■

An interesting consequence of Theorem 6 is that, for any MOMDP, the Convex Hull (recall Definition 14) is guaranteed to contain at least one deterministic and stationary u-optimal policy for every utility function u that is a strictly monotonically increasing transformation of an affine function.

Notice that linear utility functions are a particular case covered by Theorem 6. Such functions are among the most commonly studied in MORL (Roijers and Whiteson, 2017; Hayes et al., 2022). To conclude this section, we present an example of a utility function that is non-linear, not absolutely continuous, and non-differentiable at some points, and nevertheless a u-optimal policy exists (by virtue of Theorem 6).

Example 4 *Let $\vec{\mathcal{M}}$ be any 2-objective MDP. Let us define the utility function u such that:*

$$u(x, y) = e^{2x-y+5} + C(\frac{1}{1 + e^{y-2x-5}}), \tag{42}$$

where $C : [0, 1] \to [0, 1]$ is the Cantor function.

This utility function is decomposable as $u(x, y) = h(g(x, y))$ with the affine function $g(x, y) = 2x - y + 5$ being affine, and $h(x) = e^x + C(\frac{1}{1+e^{-x}})$ is a strictly monotonically increasing function. By Theorem 6, an u-optimal policy exists. Interestingly, u is in general not absolutely continuous nor differentiable at all points due to the Cantor function component.

Notice that Theorem 6 only provides sufficient conditions that a utility function u must satisfy to guarantee the existence of u-optimal policies. In fact, Example 2 shows a non-affine utility function for which a u-optimal policy exists in a particular MOMDP.

3.4 An Algorithmic Perspective

We present one last theorem in this sections that connects our previous formal results with practical MORL. As we have repeatedly mentioned, the MORL literature has focused on developing algorithms that find the policies that maximise utility functions under the SI criterion (i.e., computing $\langle u, s\rangle$-policies, as proved by Theorem 1). Hence, a natural question is whether such a computed $\langle u, s\rangle$-optimal policy is *also* an u-optimal policy for the whole MOMDP. Next, Theorem 7 states that the answer is *yes* (with a very minor *but*) if the utility function belongs to the family of functions described in Theorem 6, independently of whether such policy is stochastic, deterministic, non-stationary or stationary. Formally:

Theorem 7 *Let $\vec{\mathcal{M}}$ be a finite multi-objective MDP, and let u be a utility function of $\vec{\mathcal{M}}$ that can be decomposed as $u(x) = h(g(x))$, where g is an affine function and h is a strictly monotonically increasing function over the image of g. Let $\mathcal{S}_0$ be the set of initial states. Then, for any policy $\pi \in \Pi(\vec{\mathcal{M}})$:*

1. ***(Weak version):** π maximises g according to the SI criterion if and only if π is $\langle g, s\rangle$-optimal $\forall s \in \mathcal{S}_R$,*
2. ***(Strong version):** π maximises u according to the SI criterion if and only if π is $\langle u, s\rangle$-optimal $\forall s \in \mathcal{S}_R$,*

where $\mathcal{S}_R \subseteq \mathcal{S}$ is the subset of reachable *states such that for each state $s \in \mathcal{S}_R$ there exists at least one trajectory of state-actions pairs starting at some initial state $s_0 \in \mathcal{S}_0$ with a non-zero probability of reaching s.*

Proof We prove the weak version of this theorem by induction. Without loss of generality, we assume that $g(x)$ is a linear function $g(x) = \vec{w} \cdot x$. Also, by abuse of notation, we refer to a policy maximising $g(x) = \vec{w} \cdot x$ at the state level as $\langle \vec{w}, s\rangle$-optimal, and at the all-states level as $\vec{w}$-optimal.

First of all, we notice that, if a policy π_0 is optimal for some weight vector $\vec{w}$ under the SI criterion (i.e., $\langle \vec{w}, S_0\rangle$-optimal), this implies that it necessarily is $\langle \vec{w}, s_0\rangle$-optimal for all initial states $s_0 \in \mathcal{S}_0$. This implication occurs because achieving maximum value $\vec{w} \cdot \vec{V}_{SI}$ with respect to the expectation over initial states means that such a policy has to reach the absolute maximum value in at least one state. However, we know that there exists at least one policy π_* that is both deterministic and stationary and reaches the maximum value for every state (in particular, for every initial state) simultaneously. Thus, either the policy π_0 is optimal at every initial state, or the policy π_* achieves a higher scalarised value according to the SI criterion.

Now consider any of the initial states s_0 of the MOMDP, and let $\mathcal{S}_1^\pi$ be the set of states with a non-zero probability of being transitioned from s_0 when following the policy π_0. Similarly, consider $\mathcal{S}_n^\pi$ to be the set of states with a non-zero probability of being

transitioned from n steps from s_0 when following the policy π_0. By induction, we consider that up to n, the policy π_0 is $\langle \vec{w}, s \rangle$-optimal for any state in $\mathcal{S}_0^\pi, \dots, S_{n-1}^\pi$.

Now, assume by reductio ad absurdum that in any of the states $s_n \in \mathcal{S}_n^\pi$, the policy π_0 was not $\langle \vec{w}, s_n \rangle$-optimal. Due to the Bellman equation (Sutton and Barto, 1998), this would imply that π_0 would obtain a sub-optimal value for some state $s_{n_1} \in \mathcal{S}_{n-1}^\pi$, contradicting our induction hypothesis.

It is clear that, for any state s with a non-zero probability of being visited when following policy π_0, the policy π_0 will also be $\langle \vec{w}, s \rangle$-optimal if π_0 was $\vec{w}$-optimal under the SI criterion.

Finally, the proof of the strong version of this theorem is direct from the weak version due to Theorem 6. ■

Notice that, in practice, any (multi-objective) reinforcement learning algorithm only deals with the set of reachable states $\mathcal{S}_R$ of an MOMDP. Thus, in all cases, either the set $\mathcal{S}_R$ is equivalent to the whole state set $\mathcal{S}_R = \mathcal{S}$, or can be safely considered as so. Under those circumstances, for any utility function u satisfying the properties of Theorem 7, the existing algorithms that compute $\langle u, S_0 \rangle$-optimal policies are actually computing u-optimal policies. In particular, Theorem 7 provides a formal justification for the use of the SI criterion in the plethora of MORL work dealing with linear utility functions.

Nevertheless, we remark that, as stated even in the major surveys of multi-objective reinforcement learning such as (Roijers and Whiteson, 2017; Hayes et al., 2022), there is much more than linear utility functions (or combinations of linear and SMI functions) in MORL. Hence, our formal analysis in Section 3 motivates the need for developing new algorithms for computing u-optimal policies for other families of utility functions. We expand on this point in Section 7.

3.5 The Need for the State-Dependent Criterion

To conclude this section, we provide a deeper discussion on a topic briefly commented on in previous sections: why do we need the state-dependent (SD) criterion? Which can also be rephrased as: under which conditions would we need dedicated algorithms for computing u-optimal policies?

First of all, there are two main reasons why it would be tempting to focus exclusively on the state-independent (SI) criterion:

1. **Simplicity:** it is much easier to develop and test algorithms that aim at computing $\langle u, \mathcal{S}_0 \rangle$-optimal policies for the random variable of initial states $\mathcal{S}_0$ of the MOMDP, than to compute policies that are $\langle u, s \rangle$-optimal for *all* states s of the MOMDP. Recall from Theorem 1 that any $\langle u, \mathcal{S}_0 \rangle$-optimal policy is also optimal according the SI criterion.

2. **Apparent unnecessariness:** The state-dependent criterion is seemingly not necessary in single-objective reinforcement learning. In any single-objective MDP, a policy that is optimal at the initial state is also optimal for every other state reachable from the initial one. A proof of a more general statement is provided in Theorem 7, in which we even generalised this claim for multi-objective MDPs and the family of strictly monotonically increasing transformations of affine utility functions.

Unfortunately, the second reason does not hold for arbitrary utility functions in MOMDPs. We offered a counter-example in Theorem 2. There, we showed that there exist utility function for which in some MOMDPs there does not exist any u-optimal policy, despite an $\langle u, s\rangle$-optimal policy existing for every state s of the environment. Moreover, the utility function of Theorem 2 satisfied several desired properties such as strict monotonicty and continuous differentiability. Thus, Theorem 2 showed us a problem endemic to MOMDPs: some $\langle u, \mathcal{S}_0\rangle$-optimal policies are so-called **contradictory** policies. That is, they are policies that at the initial state they seem to optimise the utility function u, but as the agent keeps following the policy, it may start performing suboptimal actions according to u.

Contradictory policies, a phenomenon exclusive to MOMDPs, lead to three major algorithmic problems when exclusively applying SI-criterion techniques. We enumerate here these three issues:

1. **Initial-state bias:** Imagine two policies, the first one π_1 is $\langle u, s_0\rangle$-optimal only for the initial state of the environment, and suboptimal for all remaining ones. The second one, π_2, is suboptimal only for the initial state s_0 of the environment, but $\langle u, s\rangle$-optimal for all remaining ones. Policy π_2 might seem at least a reasonable alternative to take into account. Nevertheless, all current SI-criterion-based algorithms will prefer π_1 over π_2.

2. **Initial-state noise:** Imagine two policies, the first one π_1 is $\langle u, s_0\rangle$-optimal only for the initial state of the environment, and suboptimal for all remaining ones. The second one π_2 is u-optimal. Clearly, policy π_2 is the best choice of the two. However, all current SI-criterion-based algorithms will be unable to differentiate between π_1 and π_2.

3. **Problematic-utility hiding**: For some utility functions, a u-optimal policy might not exist, as Theorem 2 proves. If no u-optimal policy exists, it is reasonable to think that the utility designer might want to modify its utility function to a more appropriate one. However, it may occur that an $\langle u, s\rangle$-optimal policy does exist for every single state s of the given MOMDP, as the conditions required by Theorem 5 are much weaker than its all-state-level counterpart, Theorem 6. Under such circumstances, all current SI-criterion-based algorithms will compute some $\langle u, \mathcal{S}_0\rangle$-optimal policy, completely oblivious of any issue with the considered utility function.

The only way to tackle the three issues of contradictory policies is to also focus on the SD criterion, and develop algorithms that compute u-optimal policies, or at least test if they are $\langle u, s\rangle$-optimal for all states of the MOMDP.

4. Utility-Optimal Policies under the ESR Criterion

After studying utility functions following the SER criterion in the previous section, here we will examine them under the second-most widespread MORL criterion for defining utility functions: the Expected Scalarised Returns (ESR) criterion. We recall that the ESR criterion has received a lot less attention in the MORL literature (Hayes et al., 2022). For the remainder of this section, unless explicitly stated otherwise, we assume that all utility

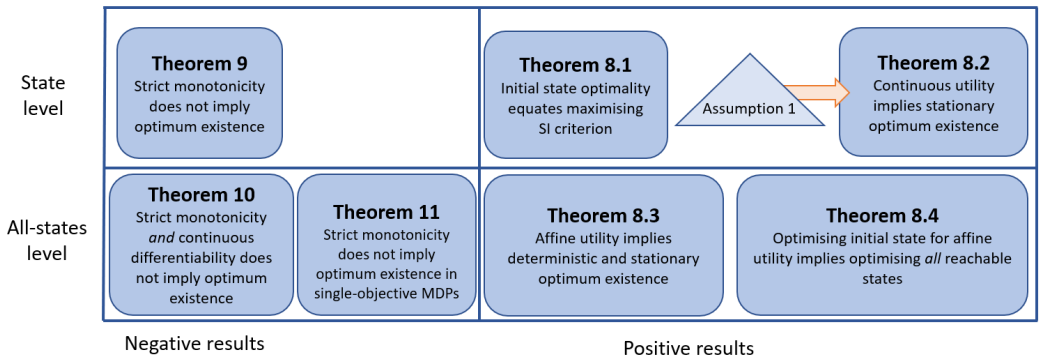

Figure 5: All theorems and assumptions presented in Section 4, which relate to utility-optimal policies under the ESR criterion. Triangles indicate assumptions, and large rectangles indicate theorems. Arrows indicate if a given theorem requires an assumption for its proof. Formal results are classified in two axes: either state-level results (related to Definition 18), or all-state-level results (related to Definition 19), and either positive (located in Section 4.1) or negative results (located in Section 4.2).

functions follow the ESR criterion (Def. 10). A graphical summary of the formal results introduced in this section is presented in Figure 5.

As in the previous section, we start by introducing the notion of *utility-optimality* at the state, adapting it to the particularities of the ESR criterion. We say that a policy π is optimal in state s, with respect to a utility function u, when it achieves a higher expected scalarised discounted return from s than any alternative policy. Formally:

Definition 20 (ESR utility-optimal policy at a state) *Let $\vec{\mathcal{M}}$ be an MOMDP with state set $\mathcal{S}$, and policy set $\Pi(\vec{\mathcal{M}})$. Let u be a utility function of $\vec{\mathcal{M}}$. A policy $\pi_* \in \Pi(\vec{\mathcal{M}})$ is* optimal at a state $s \in \mathcal{S}$ with respect to u under the ESR criterion *if and only if:*

$$\mathbb{E}[u(\vec{G}_t)|S_t = s, \pi_*] \geq \mathbb{E}[u(\vec{G}_t)|S_t = s, \pi], \tag{43}$$

for every policy $\pi \in \Pi(\vec{\mathcal{M}})$. We say that π_* is $\langle u, s \rangle_{ESR}$-optimal for short.

By abuse of notation, in this Section, we refer to any utility-optimal policy at state s as $\langle u, s \rangle$-optimal, removing the ESR subscript.

Example 5 *Consider the same MOMDP $\vec{\mathcal{M}}$ from Example 1. It has two states: an initial state s_1 and a terminal state s_2. An agent can perform two actions (a_1, a_2) in this environment, with rewards $\vec{R}(s_1, a_1) = (1, 0)$, $\vec{R}(s_1, a_2) = (0, 1)$ respectively.*

Consider the utility function $u(x, y) = x + sin(y)$. Any stochastic and stationary policy π will obtain scalarised value $\sum_a \pi(s, a) \cdot u(\vec{R}(s, a))$. The formula for non-stationary policies is analogous. Importantly, the scalarised value of any policy resides within $k + (1 - k) \cdot sin(1)$ for some $k \in [0, 1]$.

The deterministic policy $\pi(s_1) = a_1$ *that obtains vectorial value* $\vec{V}^{\pi}(s_1) = (1, 0)$ *is clearly* $\langle u, s_1 \rangle$*-optimal since* $k \cdot sin(1) < k$ *for any* $k \in [0, 1]$.

Next, following the same structure of the previous section, we also define *utility-optimal* policies at the *all-states* level as policies that are utility-optimal in every state in the MOMDP. Formally:

Definition 21 (ESR utility-optimal policy) *Let* $\vec{\mathcal{M}}$ *be an MOMDP with state set* $\mathcal{S}$*, and policy set* $\Pi(\vec{\mathcal{M}})$*. Let* u *be a utility function of* $\vec{\mathcal{M}}$*. A policy* $\pi_* \in \Pi(\vec{\mathcal{M}})$ *is* optimal with respect to u under the ESR criterion *if and only if*

$$\mathbb{E}[u(\vec{G}_t)|S_t = s, \pi_*] \geq \mathbb{E}[u(\vec{G}_t)|S_t = s, \pi], \tag{44}$$

for every policy $\pi \in \Pi(\vec{\mathcal{M}})$*, and every state* $s \in \mathcal{S}$*. We say that* π_* *is* u_{ESR}*-optimal for short.*

By abuse of notation, in this Section, we refer to any utility-optimal policy as u-optimal, removing the ESR subscript.

Example 6 *In the MOMDP in Example 5 above, policy* $\pi(s_1) = a_1$ *is u-optimal since there are only two states, and the second one is terminal.*

The remainder of this section, analogously to the previous Section 3, is devoted to providing formal results for utility functions under the ESR condition. Next Section 4.1 provides all *positive* results (such as sufficiency conditions), while following Section 4.2 provides all *negative* results (insufficiency conditions). You will notice that both subsections are more succinct than their SER-criterion counterparts of Section 3, to avoid unnecessary repetitiveness.

4.1 Positive Results

In this section, for the same reasons stated in Section 3, we aim at addressing Problem 1 and Problem 2 for utility functions under the ESR criterion. As mentioned in the previous section, our aim is to find stationary (possibly stochastic) policies that are either $\langle u, s \rangle$-optimal or u-optimal among the set of all possible policies Π, including non-stationary stochastic ones.

To address both problems, it would be very practical to know which formal results of Section 3 generalise for the ESR criterion as well. The answer is that most of them do, but not all of them. Next, Theorem 8 contains all such formal results for the ESR criterion.

Only one formal result of the next Theorem 8 also requires a new assumption: that any discounted return can be obtained by stationary policies[5] Formally:

Assumption 1 *Let* $\mathcal{M}$ *be a finite single-objective Markov decision process. Let* G_π *be the random variable of possible discounted returns that can be achieved by any policy* π *(stationary or not) of* $\mathcal{M}$*. Let* $G \in \mathbb{R}^n$ *be any value belonging to the support* $supp(G_\pi)$ *of*

5. This assumption is analogous to Theorem 5.5.1 of (Puterman, 1998), which states that any value V can be obtained by stationary policies.

G_π. Then, there exists at least one stationary (possibly stochastic) policy $\pi' \in \Pi^S(\mathcal{M})$ for which one of its possible discounted returns is G:

$$\forall \pi : G \in supp(G_\pi) \implies \exists \pi' \in \Pi^S(\mathcal{M}) : G \in supp(G_{\pi'}). \tag{45}$$

We conjecture that Assumption 1 holds for every finite MDP, but it is beyond the scope of this paper to prove such a theoretical (single-objective) result. Such a result would prove that the sets of possible discounted returns for stationary and non-stationary policies are identical. In this work, we limit ourselves to studying the consequences that such an Assumption 1 would have if it were true.

Finally, we provide the *omnibus* theorem, which summarises all positive results that generalise from the SER criterion, which are:

1. **Regarding the state-independent criterion:** Maximising an utility function u under the SI criterion in some MOMDP is equivalent to being $\langle u, s \rangle$-optimal for the unique initial state of an associated MOMDP.

2. **State-level optimality:** If Assumption 1 is satisfied, for any continuous utility function u, for any state of every finite MOMDP, a stationary $\langle u, s \rangle$-optimal policy exists.

3. **All-states-level optimality:** For every MOMDP, for any affine utility function u, a deterministic and stationary u-optimal policy exists.

4. **Regarding algorithms:** For any affine utility function u, a policy that maximises u under the SI-criterion is also $\langle u, s \rangle$-optimal for all *reachable* states s from any initial state of the MOMDP (i.e., the policy is virtually u-optimal).

Next Theorem 8 formalises each of these four statements and proves all of them.:

Theorem 8 (ESR Omnibus) *Under the ESR criterion:*

1. ***Regarding the state-independent criterion:*** *Theorem 1 holds.*

2. ***State-level optimality:*** *Theorem 5 holds if Assumption 1 is satisfied.*

3. ***All-states-level optimality:*** *Lemma 3 holds.*

4. ***Regarding algorithms:*** *The weak version of Theorem 7 holds.*

Proof We prove each statement separately.

1. **Regarding the state-independent criterion:** It suffices to prove that for any pair of policies π, π' such that $\pi = \pi'|_{\mathcal{S}}$, the scalarised values $\mathbb{E}[u(\vec{G}^\pi)]_{SI}$ and $\mathbb{E}[u(\vec{G}_t) \mid$

$S_t = s_0, \pi']$ are identical:

$$\mathbb{E}[u(\vec{G}_t) \mid S_t = s_0, \pi'] = \mathbb{E}[u(\sum_{k=0}^{\infty} \gamma^k \vec{R}'(S_{t+k}, A_{t+k})) \mid S_t = s_0, \pi'] \tag{46}$$

$$= \mathbb{E}[u(\sum_{k=1}^{\infty} \gamma^k R'(S_{t+k}, A_{t+k})) \mid S_t = s_0, \pi'] \tag{47}$$

$$= \mathbb{E}[u(\sum_{k=1}^{\infty} \gamma^k R'(S_{t+k}, A_{t+k})) \mid S_t = S_0, \pi] \tag{48}$$

$$= \mathbb{E}[u(\sum_{k=0}^{\infty} \gamma^k R(S_{t+k}, A_{t+k})) \mid S_t = S_0, \pi] \tag{49}$$

$$= \mathbb{E}[u(\vec{G}^{\pi})]_{SI}. \tag{50}$$

Thus, we proved that:

$$\mathbb{E}[u(\vec{G}^{\pi})]_{SI} = \max_{\rho \in \Pi(\vec{\mathcal{M}})} \mathbb{E}[u(\vec{G}^{\rho})]_{SI} \quad \iff \quad \pi' \text{ is } \langle u, s_0 \rangle\text{-optimal in } \vec{\mathcal{M}}'. \tag{51}$$

2. **State-level optimality:**

   - For any MOMDP, its vectorial reward set is bounded and closed (i.e., compact). Now, every possible return of the MOMDP is a convergence sequence of elements from $\vec{\mathcal{R}}$.
   - Consider the space of possible discounted returns $\mathcal{G}$ of $\vec{\mathcal{M}}$. This space is clearly bounded by $\frac{1}{1-\gamma}\|R_{max}\|$. We want to prove that $\mathcal{G}$ is also closed to conclude that it is compact.
   - To prove that $\mathcal{G}$ is closed, we will check that every convergent series of elements $G_n \in \mathcal{G}$ converges to a point in $\mathcal{G}$. For that, we notice that any element $G \in \mathcal{G}$ has at least one stationary policy π associated with it, such that π achieves the discounted return G, by Assumption 1. The set of stationary policies $\Pi^s(\vec{\mathcal{S}})$ is compact. Thus, any associated convergent sequence of policies will also converge to another policy. There might be several associated policy sequences, but all of them converge within $\Pi^s(\vec{\mathcal{S}})$. Moreover, all of them will have an associated discounted return. Thus, $\mathcal{G}$ is closed.
   - Since $\mathcal{G}$ is bounded and closed, it is compact. Now, since u is continuous, the image $u(\mathcal{G})$ is also compact. Thus, $\mathbb{E}[u(\mathcal{G}) \mid s, \pi]$, the image of a linear function, is also compact.
   - Finally, any compact set in the metric space has a maximum element. The maximum element of $\mathbb{E}[u(\mathcal{G}) \mid s, \pi \in \Pi^s(\vec{\mathcal{S}})]$ is the $\langle u, s \rangle$-optimal policy. Thus, there exists a stationary $\langle u, s \rangle$-optimal policy for every state s of the MOMDP.

3. **Remaining statements:** Affine utility functions behave identically under the ESR and SER criteria. Hence, the proof of Lemma 3 and Theorem 7 also apply here.

■

Theorem 8 addresses both Problem 1 and Problem 2 for utility functions under the ESR. Importantly, it provides weaker results for both problems:

- First, at the state level, we now require Assumption 1 to prove that continuous utility functions have associated $\langle u, s\rangle$-optimal policies at a given state s.
- Second, at the all-states level, we could not generalise Theorem 6 nor the complete Theorem 7 to the ESR criterion. Thus, we can only prove that affine utility functions are guaranteed to have associated deterministic and stationary u-optimal policies.

If Assumption 1 holds, Theorem 8 provides an important connection between $\langle u, s\rangle$-optimal policies and the distributed undominated set (recall Definition 4). Under Assumption 1, for any initial state s_0 of any MOMDP, the distributed undominated set is guaranteed to contain at least one stationary $\langle u, s_0\rangle$-optimal policy for every continuous and strictly monotonically increasing utility function u. Moreover, all linear utility functions are in particular continuous and it is already known that their utility-optimal policies belong to the convex hull (Definition 14).

Moreover, another interesting consequence of Theorem 8 is that, for any MOMDP, the convex hull is guaranteed to contain at least one deterministic and stationary u-optimal policy for every affine utility function u.

In the following subsection, we will delve into the reasons why the results obtained by Theorem 8 are weaker for the ESR criterion. Following the methodology of the previous section, we will provide several examples of *insufficient conditions* for utility functions under the ESR criterion.

4.2 Negative Results

Next, we characterise several *insufficient* conditions that are not enough to guarantee the existence of (non-stationary) deterministic or (non-stationary) stochastic $\langle u, s\rangle$-optimal or u-optimal policies in finite MOMDPs under the ESR criterion. In short, (i) any insufficient condition for utility functions under the SER criterion is also insufficient under the ESR criterion, and (ii) some sufficient conditions for utility functions under the SER criterion now are insufficient under the ESR criterion:

1. **For $\langle u, s\rangle$-optimal policies (state level):** being *strictly monotonically increasing*. Theorem 9 shows a counterexample.
2. **For u-optimal policies (all-states level):**
   - being simultaneously *strictly monotonically increasing* and *continuously differentiable*. Theorem 10 shows a counterexample.
   - being *strictly monotonically increasing*, even in *single-objective* MDPs. Theorem 11 shows a counterexample. This insufficient condition is exclusive to the ESR criterion.

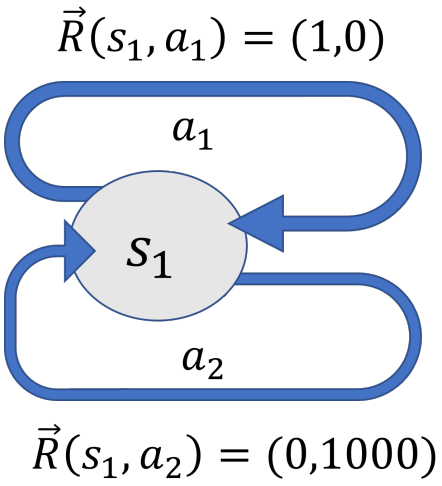

Figure 6: MOMDP of Theorem 9. Circles indicate states and arrows indicate transitions. The grey circle is a non-terminal state, in which all transitions return to it indefinitely.

These conditions prove that, when computing a utility-optimal policy under the ESR criterion, we need to be careful about understanding if such a policy will exist in the first place. The remainder of this section is devoted to providing three counterexamples to prove each of the three stated insufficiency theorems.

Our first counterexample is a finite MOMDP and a monotonically increasing utility function u for which no $\langle u, s\rangle$-optimal policy exists for any state s, even if u belongs to the family of monotonically increasing functions.

Theorem 9 *Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ be a function under the ESR criterion that is strictly monotonically increasing in the set $\vec{\mathcal{X}} \subseteq \mathbb{R}^n$. Then, there are finite MOMDPs $\vec{\mathcal{M}}$ for which u is an utility function but no policy $\pi \in \Pi(\vec{\mathcal{M}})$ is $\langle u, s\rangle$-optimal for any state s of $\vec{\mathcal{M}}$.*

Proof Consider the set $\vec{\mathcal{X}} = \mathbb{R}^+ \times \mathbb{R}^+$, and the utility function $u : \vec{\mathcal{X}} \to \mathbb{R}$ such that

$$u(x, y) = \begin{cases} x + y, & \text{if } x > 0 \text{ and } y > 0, \\ -\frac{1}{x+y}, & \text{otherwise.} \end{cases}$$

Notice that the utility function u is strictly monotonically increasing for all possible vectors of X.

Consider an MOMDP with a unique state s_1 such that all actions upon s_1 transition to itself indefinitely, represented in Figure 6. There are two possible actions in s_1 with associated rewards $\vec{R}(s_1, a_1) = (1, 0)$ and $\vec{R}(s_1, a_2) = (0, 1000)$.

It is clear that all stationary policies (both deterministic and stochastic) in this environment yield a negative scalarised value and can be easily discarded as suboptimal.

Intuitively, we observe that the optimal policy for u is a non-stationary policy that almost always performs action a_2, but at some later point, it performs action a_1. The later

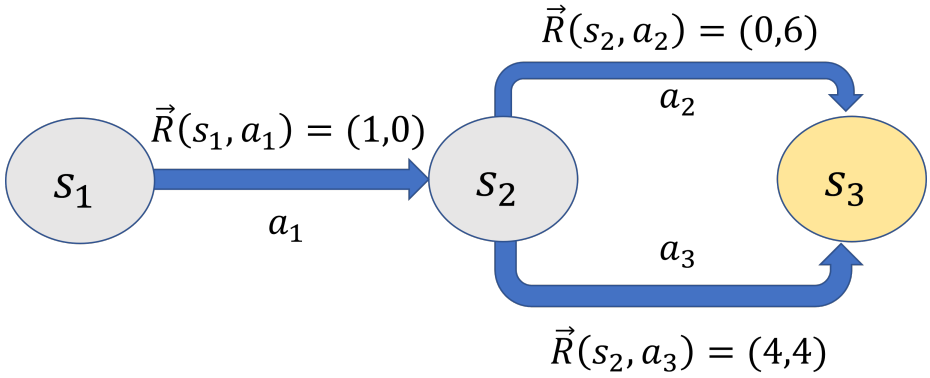

Figure 7: MOMDP of Theorem 10. Circles indicate states and arrows indicate transitions. The yellow circle is a terminal state.

it performs a_1, the more scalarised value it obtains. Let us consider the discount factor $\gamma = 0.1$ for the computations:

- If it performs action a_1 first and then only performs action a_2, it obtains a return of $(1,0) + 0.1 \cdot (0,1000) + 0.01 \cdot (0,1000) + 0.001 \cdot (0,1000) + \cdots = (1, 111.\overline{1})$, which obtains a scalarised value of $u(1, 111.\overline{1}) = 112.\overline{1}$.
- If it performs action a_1 second and the rest of the time performs action a_2, it obtains a return of $(0,1000) + 0.1 \cdot (1,0) + 0.01 \cdot (0,1000) + 0.001 \cdot (0,1000) + \cdots = (1, 111.\overline{1})$, which obtains a scalarised value of $u(0.1, 1011.\overline{1}) = 1011.2\overline{1}$.
- If it performs action a_1 third and the rest of the time performs action a_2, it obtains a return of $(0,1000) + 0.1 \cdot (0,1000) + 0.01 \cdot (1,0) + 0.001 \cdot (0,1000) + \cdots = (1, 111.\overline{1})$, which obtains a scalarised value of $u(0.01, 1101.\overline{1}) = 1101.12\overline{1}$.

Hence, there is no $\langle u, s_1 \rangle$-optimal policy in $\vec{\mathcal{M}}$ for utility function u. ■

The following Theorem 10 shows that being finite is also insufficient for an MOMDP to guarantee the existence of stochastic u-optimal policies. Moreover, the utility function used in the proof of Theorem 10, the *squared norm* n is both *strictly* monotonically increasing *and continuously differentiable*, and is also a utility function commonly used in the MORL literature following the ESR criterion (e.g., (Hayes et al., 2021), or (Roijers et al., 2018), which use a more complex version of this utility function):

Theorem 10 *Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ be a function under the ESR criterion that is both strictly monotonically increasing and continuously differentiable in the set $\vec{\mathcal{X}} \subseteq \mathbb{R}^n$. Then, there are finite MOMDPs $\vec{\mathcal{M}}$ for which u is an utility function but no policy $\pi \in \Pi(\vec{\mathcal{M}})$ is u-optimal.*

Proof Consider the set $\vec{\mathcal{X}} = \mathbb{R}^+ \times \mathbb{R}^+$ and the utility function $n(x, y) = x^2 + y^2$, which is strictly monotonically increasing for any $(x, y) \in \mathbb{R}^+ \times \mathbb{R}^+$, and continuously differentiable in $\mathbb{R}^2$.

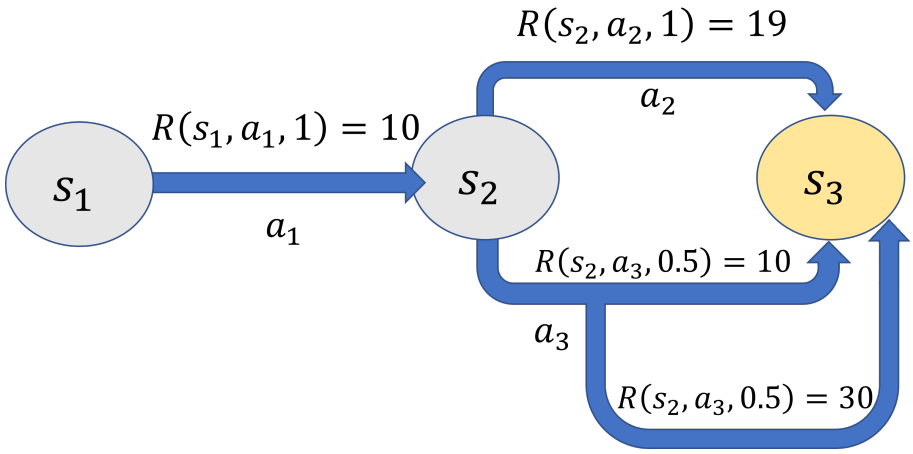

Figure 8: MDP of Theorem 11. Circles indicate states, and arrows indicate transitions. Notice how action a_3 has two possible transitions, each with probability 0.5. The yellow circle is a terminal state.

Let $\vec{\mathcal{M}}$ be a 2-objective deterministic MDP with three states s_1, s_2, and s_3 such that s_3 is the terminal state. Regarding actions, there is one possible action in s_1, which has associated rewards $\vec{R}(s_1, a_1) = (1, 0)$. Action a_1 transitions the state to s_2. Then, in state s_2, there are two possible actions with associated rewards $\vec{R}(s_2, a_2) = (0, 6)$ and $\vec{R}(s_2, a_3) = (4, 4)$. All actions in s_2 transition to the terminal state s_3. The whole MOMDP is represented in Figure 7.

This MOMDP has two possible deterministic policies. The first policy is $\pi_1(s_1) = a_1, \pi_1(s_2) = a_2$. This policy obtains values $\vec{V}^{\pi_1}(s_1) = (1, 6), \vec{V}^{\pi_1}(s_2) = (0, 6)$. The second policy is $\pi_2(s_1) = a_1, \pi_2(s_2) = a_3$. This policy obtains values $\vec{V}^{\pi_2}(s_1) = (5, 4), \vec{V}^{\pi_2}(s_2) = (4, 4)$.

According to utility function u, policy π_1 obtains scalarised value $1^2 + 6^2 = 37$ at state s_1, and $0^2 + 6^2 = 36$ at state s_2. Then, policy π_2 obtains scalarised value $5^2 + 4^2 = 41$ at state s_1, and $4^2 + 4^2 = 32$ at state s_2.

Then, any stochastic policy π will have scalarised value $37k + (1 - k) \cdot 41$ at state s_1 and scalarised value $36k + (1 - k) \cdot 32$ at state s_2, both for some $k \in [0, 1]$.

In other words, π_1 is the only $\langle u, s_2 \rangle$-optimal policy, while π_2 is the only $\langle u, s_1 \rangle$-optimal policy. Thus, no u-optimal policy exists. ■

To finish this section, we recall that, under the SER criterion, we were able to prove Theorem 6. This theorem states that there is always a utility-optimal policy for utility functions that result from composing an affine function with a strictly monotonically increasing function. However, such a condition is not enough under the ESR criterion, as the following counterexample proves.

Theorem 11 *Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ be a function under the ESR criterion that is strictly monotonically increasing in the set $\vec{\mathcal{X}} \subseteq \mathbb{R}^n$. Then, there are finite single-objective MDPs $\mathcal{M}$ for which u is an utility function but no policy $\pi \in \Pi(\vec{\mathcal{M}})$ is u-optimal.*

Proof Let $\mathcal{M}$ be a MDP in Figure 8, with three states s_1, s_2, and s_3 such that s_3 is the terminal state. Regarding actions, there is one possible action in s_1, which has associated rewards $R(s_1, a_1) = 10$. Action a_1 transitions the state to s_2. Then, in state s_2, there are two possible actions with associated rewards $R(s_2, a_2, 0.5) = 30$, $R(s_2, a_2, 0.5) = 10$, and $R(s_2, a_3, 1) = 19$. All actions in s_2 transition to the terminal state s_3.

This environment has two possible deterministic policies. The first policy is $\pi_1(s_1) = a_1, \pi_1(s_2) = a_2$. This policy obtains values $V^{\pi_1}(s_1) = 30, V^{\pi_1}(s_2) = 20$. The second policy is $\pi_2(s_1) = a_1, \pi_2(s_2) = a_3$. This policy obtains values $V^{\pi_2}(s_1) = 29, V^{\pi_2}(s_2) = 19$. Consider the following piece-wise utility function $u : \mathbb{R} \to \mathbb{R}$, which is clearly strictly monotonically increasing over $\mathbb{R}$:

$$\begin{cases} x & x \leq 10 \\ 200 + 19 \cdot (x - 20) & 10 \leq x \leq 20 \\ 10 \cdot x & \text{otherwise.} \end{cases}$$

According to the utility function u, policy π_1 obtains scalarised value 300 at state s_1, and 155 at state s_2. Then, policy π_2 obtains scalarised value 290 at state s_1, and 181 at state s_2.

Then, any stochastic policy π will have scalarised value $300k + (1 - k) \cdot 290$ at state s_1 and scalarised value $155k + (1 - k) \cdot 181$ at state s_2, both for some $k \in [0, 1]$.

In other words, π_1 is the only $\langle u, s_1 \rangle$-optimal policy, while π_2 is the only $\langle u, s_2 \rangle$-optimal policy. Thus, no u-optimal policy exists. ■

Theorem 11 provides a clear counterexample to any attempt to prove Theorem 6 under the ESR criterion. In turn, this formal result clearly hampers the search for sufficient conditions for stationary optimality under the ESR criterion. Hence, we anticipate that algorithms for computing solutions under the ESR criterion will need to focus on computing non-stationary policies.

5. Preference Relations in MOMDPs

In the past two sections, we identified families of utility functions that are guaranteed to admit utility-optimal policies under both the ESR and SER criteria. A more fundamental issue, however, remains: which user preferences can actually be represented by utility functions in a given MOMDP?

To address this question, we must formalise preference relations and their maximal elements in MOMDPs. Preference relations (or binary relations) specify, for any two elements, which is preferred (Maschler et al., 2013; Steele and Stefánsson, 2015). Although much of the MORL literature treats preference relations and utility functions as interchangeable (Hayes et al., 2022), it is conceptually important to distinguish them.

We begin by defining preference relations for MOMDPs, following (Steele and Stefánsson, 2015). To provide a definition that is valid for both ESR and SER criteria, we define preference relations as orderings over random vectors of (discounted) returns. Formally:

Definition 22 (Preference relation) *We say that any binary relation $\succeq$ over at least one pair of n-dimensional real-based random vectors $\vec{Z}_1, \vec{Z}_2$ is a* preference relation. *In particular, we say that:*

- $\vec{Z}_1$ *is* weakly preferred *to $\vec{Z}_2$ if and only if $\vec{Z}_1 \succeq \vec{Z}_2$.*
- $\vec{Z}_1 \in \mathbb{R}^n$ *is* strictly preferred *to $\vec{Z}_2$ if and only if $\vec{Z}_1 \succeq \vec{Z}_2$ and not $\vec{Z}_2 \succeq \vec{Z}_1$. In short, we denote $\vec{Z}_1 \succ \vec{Z}_2$.*
- $\vec{Z}_1 \in \mathbb{R}^n$ *is* indifferent *to $\vec{Z}_2$ if and only if $\vec{Z}_1 \succeq \vec{Z}_2$ and $\vec{Z}_2 \succeq \vec{Z}_1$. In short, we denote $\vec{Z}_1 \approx \vec{Z}_2$.*

Notice that constant vectors are a degenerate case of random vectors. Thus any preference relation *exclusively* defined over pairs of constant vectors $\vec{v} \in \mathbb{R}^n$ would also satisfy Def. 22. Such a degenerate case is especially important when considering utility functions under the SER criterion. Meanwhile, for utility functions under the ESR criterion, we will need to consider the general case of Def. 22.

Moreover, notice that Def. 22 does not make any assumption about the properties of the preference relation. The preference relation of the MOMDP need not be a pre-order, partial order, or total order (Harzheim, 2005). This generality is intentional, as MORL applications consider a wide variety of utility functions, and restricting the definition would be unnecessarily limiting.

That said, if the preference relation is assumed to be at least a quasi-order (i.e., reflexive and transitive), then its maximal elements can be defined (Harzheim, 2005). In our case, these maximal elements correspond to random vectors of discounted returns that satisfy as much as possible the given preference relation. Formally:

Definition 23 (Maximal element) *Let $\succeq$ be a preference relation that is at least reflexive and transitive (a quasi-order) over the set of random vectors $\vec{\mathcal{Z}}$. Then, the random vector $\vec{Z}_*$ is a* maximal element in $\vec{\mathcal{Z}}$ *if and only if for every other possible random vector $\vec{Z}$ of $\vec{\mathcal{Z}}$:*

$$\vec{Z} \succeq \vec{Z}_* \implies \vec{Z}_* \succeq \vec{Z}. \tag{52}$$

In any (single-objective or multi-objective) reinforcement learning algorithm, its solution concept is either a policy or multiple policies. Thus, it would be very convenient to adapt Def. 23 to policies of MOMDPs. Hence, we formalise the concept of *maximal policies*: policies whose associated discounted returns or value vector is a maximal element of some preference relation. Following our methodology with utility functions, we will define preference-maximality twice: first for a single state (state level), and second for every state (all-states level). Formally:

Definition 24 (Maximal policy) *Let $\vec{\mathcal{M}}$ be an n-objective MDP with state set $\mathcal{S}$, let $\Pi(\vec{\mathcal{M}})$ be the policy set, let $\mathcal{S}$ be the state set of $\vec{\mathcal{M}}$. Then, let $\vec{\mathcal{G}}_s$ be the sets of all possible*

return random vectors of $\vec{\mathcal{M}}$ at each state $s \in \mathcal{S}_R$, and $\vec{\mathcal{V}}_s$ be the sets of all possible value random vectors of $\vec{\mathcal{M}}$ at each state $s \in \mathcal{S}$. Finally, let $\succeq$ be a preference relations, that is a quasi-order, over an associated set of random vectors $\vec{\mathcal{Z}}$. Then, when focusing a single state s, we say that a policy $\pi_* \in \Pi(\vec{\mathcal{M}})$ is a:

- ***Maximal policy at state** $s \in \mathcal{S}$ **for** $\succeq$ **under the ESR criterion** if and only if its associated random vector of discounted returns $\vec{G}^{\pi}(s)$ at state s is a maximal element of $\succeq$ over $\vec{\mathcal{Z}} \cap \vec{\mathcal{G}}_s$. In short, we say that π_* is $\langle \succeq, s \rangle_{ESR}$-maximal.*
- ***Maximal policy at state** $s \in \mathcal{S}$ **for** $\succeq$ **under the SER criterion** if and only if its associated value vector $\vec{V}^{\pi}(s)$ at state s is a maximal element of $\succeq$ over $\vec{\mathcal{Z}} \cap \vec{\mathcal{V}}_s$. In short, we say that π_* is $\langle \succeq, s \rangle_{SER}$-maximal.*

And, when considering all states, we also say that:

- ***Maximal policy for** $\succeq$ **under the ESR criterion** if and only if its associated random vector of discounted returns $\vec{G}^{\pi}(s)$ for every state s is a maximal element of $\succeq$ over $\vec{\mathcal{Z}} \cap \vec{\mathcal{G}}_s$. In short, we say that π_* is $\succeq_{ESR}$-maximal.*
- ***Maximal policy for** $\succeq$ **under the SER criterion** if and only if its associated value vector $\vec{V}^{\pi}(s)$ for every state s is a maximal element of $\succeq$ over $\vec{\mathcal{Z}} \cap \vec{\mathcal{V}}_s$. In short, we say that π_* is $\succeq_{SER}$-maximal.*

Notice that Def. 24 of preference-maximal policies is a more general solution concept than utility-optimal policies. As we have previously mentioned, for every utility function, we can always consider its implicit preference relation.

Preference relations (alongside the search for their maximal policies) are ubiquitous in both single-objective and multi-objective MDPs, as the following three examples will show. The first example is presented in the trivial case of a preference relation defined exclusively over constants, and for single-objective MDPs.

Example 7 *In finite single-objective MDPs, for every state s, the optimal policy π_* is a maximal policy for the preference relation $\succeq$ defined as $V(s) \succeq_0 V'(s) \iff V(s) \geq V'(s)$.*

Our second example also considers the degenerate case of a preference relation defined exclusively over constants, but now over an MOMDP. This preference relation, the lexicographic order, has been extensively studied in MORL (Gábor et al., 1998; Vamplew et al., 2018):

Example 8 *For any finite n-objective MDP, we consider the preference relation $\succeq_l$ over all possible value vectors $\vec{v} \in \mathbb{R}^n$ established by the lexicographic order $l = \{1, \ldots, n\}$ such that:*

$$v \succeq_l v' \iff \forall i \in \{1, \ldots, k-1\} : v_i = v'_{\sigma(i)} \qquad \text{and} \quad v_{\sigma(k)} \geq v'_{\sigma(k)}.$$

Since the lexicographic preference relation is a total order, it is always guaranteed for any MOMDP that there is at least one policy whose value is a maximal element of $\succeq_l$ for every state s of $\mathcal{M}$ simultaneously.

Our third example presents a well-known preference relation defined over return random vectors (Hayes et al., 2021; Röpke et al., 2023).

Example 9 *In finite MOMDPs, let us consider the preference relation $\succeq$ defined over return random vectors at some state s as*

$$\vec{G}_1 \succeq \vec{G}_2 \iff F_{\vec{G}_1}(\vec{x}) \leq F_{\vec{G}_2}(\vec{x}) \quad \forall \vec{x} \in \mathbb{R}^n,$$

that is, if and only if $\vec{G}_1$ first-order stochastically dominates $\vec{G}_2$ (recall Def. 4

Then, for every state s, the discounted returns random vector $\vec{G}_*$ is a maximal element at state s if and only if it, for every other returns random vector $\vec{G}$ at s satisfies:

$$\vec{G} \succeq \vec{G}_* \implies F_{\vec{G}}(\vec{x}) = F_{\vec{G}_*}(\vec{x}) \quad \forall \vec{x} \in \mathbb{R}^n \implies \vec{G} = \vec{G}_*. \tag{53}$$

5.1 Preference Relations and Utility Functions

Now that these examples have established the ubiquity of preference relations in MOMDPs, we proceed to study their connection to utility functions. In line with decision theory, humans possess preferences that can *sometimes* be represented by utility functions, but not all preference relations can be corresponded with utility functions (Steele and Stefánsson, 2015; Peterson, 2017; Hansson and Hendricks, 2018; Maschler et al., 2013). When such a function does exist, it is referred to as a *representative* function of the preference relation $\succeq$. Formally:

Definition 25 (Representative function) *Let $\succeq$ be a preference relation over a set of random vectors $\vec{\mathcal{Z}}$. Then, we say that a function $u : \vec{\mathcal{X}} \to \mathbb{R}^n$ with $\vec{\mathcal{Z}} \subseteq \vec{\mathcal{X}}$ is* representative *of the preference relation $\succeq$ if and only if, for every pair of random vectors $\vec{Z}_1$, $\vec{Z}_2 \in \vec{\mathcal{Z}}$:*

$$\vec{Z}_1 \succeq \vec{Z}_2 \iff \mathbb{E}[u(\vec{Z}_1)] \geq \mathbb{E}[u(\vec{Z}_2)]. \tag{54}$$

As usual, we remark that for the degenerate case in which our preference relation only compares constant vectors $\vec{V}_1, \vec{V}_2 \in \mathbb{R}^n$, then Eq. 54 gets simplified into:

$$\vec{V}_1 \succeq \vec{V}_2 \iff u(\vec{V}_1) \geq u(\vec{V}_2). \tag{55}$$

Notice that Def. 25 is typically highly restrictive, as it requires a utility function to completely mimic a preference relation without fail for a given set of random vectors. Thus, some (but not all) preference relations have representative utility functions capable of representing them for any possible random vector. Nevertheless, any utility function u represents some preference relation $\succeq_u$ that exactly follows Equation 54.

A prominent example of a preference relation that cannot be represented with a utility function is the *lexicographic order* (Debreu, 1997; Barbera et al., 1998), a fruitful and extensively explored preference relation in MORL (Gábor et al., 1998; Vamplew et al., 2018; Li and Czarnecki, 2019; Vamplew et al., 2021; Skalse et al., 2022; Skalse and Abate, 2023). To illustrate this limitation, we provide in Example 10 an MOMDP in which we attempt unsuccessfully to build a representative linear function for a lexicographic order. For the general impossibility proof covering arbitrary functions, see Corollary 2 of (Skalse and Abate, 2023).

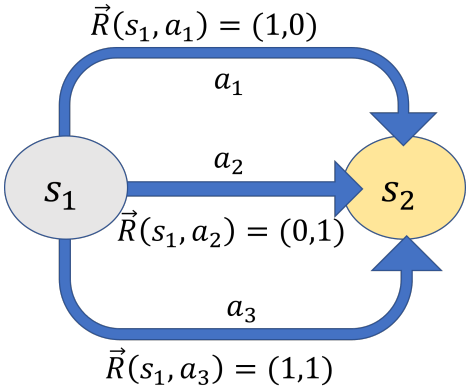

Figure 9: MOMDP of Example 10, Example 11, and Example 12. Circles indicate states, and arrows indicate transitions. The yellow circle is a terminal state.

Example 10 *Consider an MOMDP $\vec{\mathcal{M}}$ with two states: an initial state s_1 and a terminal state s_2. An agent can perform three actions in this environment (a_1, a_2, a_3), with rewards $\vec{R}(s_1, a_1) = (1, 0)$, $\vec{R}(s_1, a_2) = (0, 1)$, and $\vec{R}(s_1, a_3) = (1, 1)$, respectively. This MOMDP is represented in Figure 9.*

Consider now the lexicographic order $\succeq$ over constant vectors such that objective 1 is always preferred to objective 2. Hence, $\vec{R}(s_1, a_3) \succ \vec{R}(s_1, a_1) \succ \vec{R}(s_1, a_2)$. Any linear utility function here will be of the form $u_\alpha(x, y) = \alpha \cdot x + (1-\alpha) \cdot y$. For u_α to represent $\succ$, it must satisfy $u_\alpha(1, 1) > u_\alpha(1, 0)$ and $u_\alpha(1, 0) > u_\alpha(0, 1)$, for example, $u_{0.9}(x, y) = 0.9x + 0.1y$.

Notice that we could create such a utility function because there was a finite number of policies that u needed to order. However, policies can be stochastic, and thus, the utility function u needs to order infinite policies.

We can have for instance any policy π such that $\pi(s_1, a_1) = p, \pi(s_1, a_2) = 1 - p$, with $1 \geq p \geq 0$, which has associated value $\vec{V}^\pi(s_1) = (p, 1 - p)$. Hence, the utility function must also satisfy $u(1, 0) > u(0.999, 0.001)$, which would require an increased $\alpha' > 0.9$. And it needs to be absolutely precise: $u(p + \epsilon, 1 - p - \epsilon) > u(p, 1 - p)$ for every $\epsilon > 0$ arbitrarily small.

Thus, in practice, no $\alpha < 1$ would allow us to represent the lexicographic order. However, $u_1(x, y) = x$ is not a satisfactory option, as it even fails to represent $(1, 1) \succ (1, 0)$. Thus, it is impossible to represent the lexicographic order as a linear utility function.

Although lexicographic orders cannot be represented by utility functions in MOMDPs, they have maximal elements, associated with the maximal policies that we are most interested in, as illustrated in Example 8. This fact begs for the identification of utility functions that at least share the same maximal policies as a given preference relation. Such functions would allow us to recover the maximal policies using standard MORL algorithms. We call this kind of utility function *quasi-representative*. Formally:

Definition 26 (Quasi-representative utility function) *Let $\vec{\mathcal{M}}$ be an n-objective MDP with state set $\mathcal{S}$, let $\Pi(\vec{\mathcal{M}})$ be the policy set, let $\succeq$ be a preference relation that is a quasi-order, over an associated set of random vectors $\vec{\mathcal{Z}}$. Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ with $\vec{\mathcal{Z}} \subseteq \vec{\mathcal{X}}$. Then, we*

say that u is a quasi-representative utility function at state $s \in \mathcal{S}$ for $\succeq$ *if and only if every* $\langle \succeq, s \rangle$*-maximal policy* $\pi_* \in \Pi(\vec{\mathcal{M}})$ *is* $\langle u, s \rangle$*-optimal and vice-versa.*

Notice that Def. 26 is valid for both the ESR or the SER criteria. For a given utility function, if the quasi-representativity is satisfied for all utility-optimal policies according the SER criterion, we will say that the utility function is quasi-representative under the SER criterion, and likewise for the ESR criterion.

Notice also that if a utility function u is quasi-representative of some preference relation $\succeq$ for all states of a given MOMDP, then every $\succeq$-maximal policy $\pi_* \in \Pi(\vec{\mathcal{M}})$ is u-optimal and vice-versa.

Example 11 *Consider the MOMDP $\vec{\mathcal{M}}$ in Example 10. The utility function $u(x, y) = 10x + y$ is quasi-representative of the lexicographic order over value vectors because $u(1, 1) > u(1, 0)$ and $u(1, 1) > u(0, 1)$.*

Furthermore, quasi-representative utility functions naturally induce an equivalence relation among utility functions. Accordingly, with a slight abuse of notation, we will say that two utility functions are *quasi-representative* for a given MOMDP if and only if they share the same set of utility-optimal policies at every state s.

Example 12 *Consider the same MOMDP $\vec{\mathcal{M}}$ and the lexicographic order $\succeq$ from Example 10. For example, utility functions $u(x, y) = 10x + y$ and $u'(x, y) = 15x + y + 30$ share the same utility-optimal policy (which is $\pi(s_1) = a_3$), and hence are quasi-representative for $\vec{\mathcal{M}}$ and $\succeq$ under SER criteria.*

Quasi-representative utility functions allow us to focus on our last research question: in a given MOMDP, which families of preference relations can be maximised by utility functions? This question is of great importance, as the majority of MORL algorithms rely on utility functions. Formally, in the following subsection, we address the problem below, focusing on both the SER and ESR criterion:

Problem 3 *For which families of preference relations is the existence of a quasi-representative utility function u policy for every state s in any finite MOMDP guaranteed?*

5.2 Formal Results

The following theorem presents one family of preference relations that satisfy Problem 3: those for which there exists at least one maximal policy at every state (but not necessarily the same for every state). For these preference relations, quasi-representative utility functions (under both the ESR and SER criteria) always exist in any finite MOMDP. Formally:

Theorem 12 (Quasi-representative function existence) *Let $\vec{\mathcal{M}}$ be any finite MOMDP, let $\mathcal{S}_R \subseteq \mathcal{S}$ be a subset of all possible states of $\vec{\mathcal{M}}$.*

Then, let $\succeq$ be a preference relation over a set of random vectors $\vec{\mathcal{Z}}$. Assume that $\succeq$ is :

1. ***Complete:*** *for every two return random vectors $\vec{Z}_1, \vec{Z}_2 \in \vec{\mathcal{Z}}$, either $\vec{Z}_1 \succeq \vec{Z}_2$ or $\vec{Z}_2 \succeq \vec{Z}_1$ or both.*

2. ***Transitive:*** *for every three return random vectors $\vec{Z}_1, \vec{Z}_2, \vec{Z}_3 \in \vec{\mathcal{Z}}$), if $\vec{Z}_1 \succeq \vec{Z}_2$, and $\vec{Z}_2 \succeq \vec{Z}_3$, then $\vec{Z}_1 \succeq \vec{Z}_3$.*

3. ***State-maximisable:*** *for every state $s \in \mathcal{S}_R$, at least one $\langle \succeq, s \rangle$-maximal policy exists, for either the ESR criterion or the SER criterion.*

Then, if $\succeq$ is state-maximisable under the ESR criterion, there exists at least one quasi-representative utility function for $\succeq$ at every state $s \in \mathcal{S}_R$ under the ESR criterion.

Otherwise, if $\succeq$ is state-maximisable under the SER criterion, there exists at least one quasi-representative utility function for $\succeq$ at every state $s \in \mathcal{S}_R$ under the SER criterion.

Proof We offer a constructive proof of such a utility function focusing exclusively on return random vectors $G \in \vec{\mathcal{G}}$ (i.e., for the ESR criterion). The proof for constant value vectors (for the SER criterion) is direct from it:

1. For every state $s \in \mathcal{S}_R$, consider its set of maximal elements $\vec{\mathcal{G}}_*(s)$ according to $\succeq$, which is non-empty for every state due to Condition (3).

2. Since the preference relation is complete, for every state s all its maximal elements can safely share the same scalarised value (i.e., $\mathbb{E}[u(\vec{G}_*(s))] = \mathbb{E}[u(\vec{G}'_*(s))]$ for every two $\vec{G}_*, \vec{G}'_* \in \vec{\mathcal{G}}_*(s)$).

3. Hence, without loss of generality, we consider that there is a single maximal element per state, without repeated maximal elements between states. That is, $|\vec{\mathcal{G}}_*(s)| = 1$ for every state s. We denote $\vec{G}_*(s)$ as the maximal element of state s.

4. Then, the number of maximal elements is the same as the number of considered states $|\mathcal{S}_R|$, and we can order them according to $\succeq$ (we can order them because $\succeq$ is total and transitive). In other words, we always know if $\vec{G}_*(s_i) \succeq \vec{G}_*(s_j)$ or $\vec{G}_*(s_j) \succeq \vec{G}_*(s_i)$ or $\vec{G}_*(s_i) \approx \vec{G}_*(s_j)$ for any two states s_i, s_j.

5. Then, we set a number between 1 and $|\mathcal{S}_R|$ for each maximal element $\vec{G}_*(s_i)$, ordered by $\succeq$. Without loss of generality, we assume that $\succeq$ orders last the maximal element of state s_1, then s_2, and so on until the maximal element of state $s_{|\mathcal{S}_R|}$ is the most preferred one. Then, we set $\mathbb{E}[u(\vec{G}_*(s_1))] = 1, \mathbb{E}[u(\vec{G}_*(s_2))] = 2, \ldots, \mathbb{E}[u(\vec{V}_*(s_{|\mathcal{S}_R|}))] = |\mathcal{S}_R|$.

6. For every other random vector $\vec{G} \in \cup_s \vec{\mathcal{G}}(s)$ such that $\vec{G} \neq \vec{G}_*(s)$ for any state s, we set $\mathbb{E}[u(\vec{G})] = 0$.

Now, by construction, for every state s we have that

$$\arg\max_{\vec{\mathcal{G}}(s)} \mathbb{E}[u(\vec{G})] = \vec{G}_*(s) \in \vec{\mathcal{G}}_*(s). \tag{56}$$

In other words, for every state $s \in \mathcal{S}_R$, every maximal element $\vec{G}_*(s)$ of $\succeq$ at that state will maximise u at s. Hence, the policy π_* associated with $\vec{G}_*(s)$ will be both $\langle \succeq, s \rangle$-optimal and $\langle u, s \rangle$-optimal. Thus, by Def. 26, the utility function u is a quasi-representative utility function of $\succeq$ at state s under the ESR criterion. ■

Completeness and transitivity are standard assumptions for preference relations in decision and game theory (Steele and Stefánsson, 2015; Maschler et al., 2013). The third condition arises specifically in the MOMDP setting, where the policy space may be infinite. Indeed, as illustrated in Theorem 3, even in a finite MOMDP there may be no maximal policy for any state of the MOMDP.

5.3 An Algorithmic Perspective

To conclude this section, we remark that MORL algorithms should especially focus on preference relations that are guaranteed to have associated quasi-representative utility functions. Quasi-representative utility functions enable us to easily evaluate and compare policies quantitatively, aiding agents' learning. Theorem 12 describes sufficient conditions for a preference relation to have an associated quasi-representative utility function, as we just proved.

The family of preference relations that satisfy the conditions of Theorem 12 is broad, encompassing, for example, the family of lexicographic orders discussed earlier. In addition, any continuous utility function induces a total order that meets all the requirements of the theorem.

Nevertheless, we understand that, in practice, it may be difficult to compute utility functions that exactly express a given preference relation. For instance, in single-objective MDPs, so-called optimal algorithms like value iteration always terminate at a policy that is nearly optimal up to a very small threshold $\epsilon > 0$. Considering the practical constraints of reinforcement learning algorithms, we introduce next a relaxed version of quasi-representative utility functions, which only requires its utility-optimal elements to be nearly optimal up to some threshold $\epsilon > 0$. Formally:

Definition 27 (near-representative utility function) *Let $\vec{\mathcal{M}}$ be an MOMDP, let $\epsilon \geq 0$ be a small positive number, and let $n : \mathbb{R}^n \mapsto \mathbb{R}^+$ be a norm function. Let $\Pi(\vec{\mathcal{M}})$ be the policy set, let $\succeq$ be a preference relation that is a quasi-order, over an associated set of random vectors $\vec{\mathcal{Z}}$, and for which at least one $\langle \succeq, s \rangle$-optimal policy π_* exists for a given state s of $\vec{\mathcal{M}}$. Let $u : \vec{\mathcal{X}} \to \mathbb{R}$ with $\vec{\mathcal{Z}} \subseteq \vec{\mathcal{X}}$ such that for every policy $\pi \in \Pi(\vec{\mathcal{M}})$:*

$$\pi \text{ is } \langle u, s \rangle\text{-optimal} \implies n(\vec{V}^{\pi}(s) - \vec{V}^{\pi_*}(s)) < \epsilon. \tag{57}$$

Then, we say that u is near-representative *of $\succeq$ in $\vec{\mathcal{M}}$ for threshold ϵ and norm n in $\vec{\mathcal{Z}}$ for the state s.*

Clearly, any quasi-representative utility function is, in particular, a near-representative utility function for threshold $\epsilon = 0$. Hence, Theorem 12 guarantees their existence for any complete preference relation.

Recent examples of near-representative utility functions include the work of Macau et al. (2025), where their goal (applying our terminology) is to apply deep reinforcement learning to compute near-representative linear utility functions of some given lexicographic order in MOMDPs with large state and action spaces, where exact methods would be infeasible.

6. Related Work

The majority of the literature on MOMDPs focuses on developing novel solution concepts and on algorithms to compute them (e.g., Vamplew et al., 2009; Van Moffaert and Nowé, 2014; Roijers et al., 2015; Skalse et al., 2022; Röpke et al., 2023). For example, Van Moffaert and Nowé (2014) developed an algorithm for learning the Pareto front of any given MOMDP. Hence, they assumed that this Pareto front would always contain the solution policy (i.e., a u-optimal policy). However, as we proved in Theorem 2, this assumption does not hold for arbitrary MOMDPs.

Here, alternatively, we focus on describing some families of utility functions for which these solutions exist, a primarily overlooked formal problem despite its relevant implications. Next, we compare the similarities and differences in the state of the art of formal and algorithmic MORL literature with our work across several key areas.

6.1 Utility Functions in MOMDPs

We start by comparing our work with existing formal results on utility functions in MOMDPs. First of all, in some cases, this paper includes formal results, for which we provided a step-by-step formal proof for completeness, even though the proof could be easily extracted. We are especially referring to Lemma 1 (which we comment on in more detail later) and Lemma 3, which proves that stationary and deterministic u-optimal policies exist for any affine utility function. While its proof is a natural consequence of the fact that such policies also exist for linear utility functions (a fact that is well-known in the literature, see for instance Weng, 2011), we deemed that it was important to at least offer a proof with the minimal steps to make a connection between the two facts (and hence why it is merely a Lemma and not a Theorem).

Next, we wanted to point out some connections of Assumption 1 with existing reinforcement learning literature. Assumption 1 states that, for any finite (multi-objective) Markov decision process, all possible discounted returns can be achieved even if we restrict ourselves to stationary policies. This unproven assumption resembles the previously mentioned Puterman's Theorem 5.5.1 in Puterman (1998). Recall that Puterman's theorem states that all possible values V (i.e., the *expected* discounted returns) can be achieved even if we restrict ourselves to stationary policies. However, this theorem does not specify whether this result can also be applied to all the individual returns aggregated to compute the policy value V. As mentioned above, while we conjecture that Assumption 1 holds for any MOMDP, it was beyond the scope of this paper to provide a proof.

Then, we wanted to focus on the topic of finding properties of utility functions that are insufficient to ensure that at least one utility-optimal policy exists, either at the state level or the all-states level (i.e., our so-called *negative* results of Sections 3.1 and 4.2). Indeed, previous works in the literature noticed that for some specific utility functions u there are no stationary $\langle u, s\rangle$-optimal policies (Weng, 2011) nor stationary u-optimal policies (Siddique et al., 2020) under the SER criterion. Our work here presents examples of utility functions for which neither the non-stationary u-optimal nor the $\langle u, s\rangle$-optimal policy exists, for both the ESR and SER criteria. Importantly, we focused on finding negative results for utility functions that satisfy *attractive* properties, such as strict monotonicity or continuous differentiability. By identifying these *insufficient* properties, in this work, we are building

towards a future theory of all necessary and sufficient conditions for the existence of utility-optimal policies.

Finally, before our paper, the work of Rodriguez-Soto et al. (2024) was also particularly concerned with the formal analysis of utility functions. Unlike our work, Rodriguez-Soto et al. focused exclusively on utility functions under the SER criterion. Moreover, even for the SER criterion, they considered a definition of utility function more restrictive than what we apply in Def. 9 (among other things, requiring the utility function to be defined on every point of $\mathbb{R}^n$). Hence, in this work, we generalise all formal results of Rodriguez-Soto et al. (2024) for both the ESR and SER criteria. Furthermore, we also studied several problems that Rodriguez-Soto et al. avoided. Specifically, we formally connect the large body of algorithmic MORL work under the state-independent (SI) criterion with the concepts of $\langle u, s\rangle$-policies and u-optimal policies: Theorem 1 provide a unifying bridge between the SI criterion and $\langle u, s\rangle$-optimal policies, and Theorem 7 provides likewise for u-optimal policies.

6.2 Preference Relations in MOMDPs

Next, with respect to the analysis of preference relations in MOMDPs, to the best of our knowledge, the only contributions in the MORL field beyond our own are Weng (2011); Lu et al. (2023); Skalse and Abate (2023); Subramani et al. (2023) and Rodriguez-Soto et al. (2024). Weng (2011) studied preference relations in *ordinal* MDPs, a specific family of MOMDPs in which each reward function can take only two values (either 0 or some $r > 0$). Instead, we offer a formalisation of preference relations that is valid for any MOMDP.

Next, Lu et al. (2023) focused on the Pareto-dominance preference relation. We recall that the Pareto-dominance preference relation and the strictly monotonically increasing utility functions are strongly related: the only possible $\langle u, s\rangle$-optimal policies for such utility functions are necessarily Pareto-dominant. Lu et al. proved that the Pareto front (Definition 13) is convex, leading to the result that it is sufficient to solve a series of single-objective MDPs to compute the $\langle u, s\rangle$-optimal policy for any strictly monotonically increasing function u *if such a policy exists in the first place.* Thereafter, the final $\langle u, s\rangle$-optimal policy will be guaranteed to be stationary and, in some cases, also deterministic. This formal result is consistent with our Theorem 4, which proves that, under the SER criterion, *if an $\langle u, s\rangle$-optimal policy exists for an arbitrary utility function u (not necessarily monotonous), then an alternative* stationary *$\langle u, s\rangle$-optimal policy is also guaranteed to exist.*

The theoretical results of Skalse and Abate (2023) complement ours by showing that, for every so-called *objective* (i.e., a preorder over policies), a u-optimal policy exists if and only if the objective admits a linear utility representation. This finding is consistent with our result in Theorem 6. However, they do not address whether broader families of utility functions may also ensure the existence of a u-optimal policy, which is precisely what we establish in Theorem 6. Furthermore, our notion of preference enables the ordering of policies at the levels of individual states and all states, and adapts between the SER and ESR criteria. Hence, we offer a finer level of granularity than their notion of objective. This distinction is very relevant, as we have repeatedly mentioned that these different criteria have different limitations, as illustrated in Theorems 3, 2, 9, and 10.

Next, Subramani et al. (2023) follow on the work of Skalse and Abate (2023), but they address a separate research problem than this paper. They focus on comparing the

MORL framework's expressivity with other reinforcement learning frameworks. They aim to understand the objectives (defined identically to (Skalse and Abate, 2023)) that each framework can represent. However, like Skalse and Abate (2023), their objective definition cannot provide different orderings for the state level, or the all-state level, while our formalisation of preference relations allows such granularity. Interestingly, Subramani et al. (2023) provide results for both the ESR and SER criteria, which are referred to as *inner MORL* (for ESR) and *outer MORL* (for SER), respectively, treating them as two separate reinforcement learning frameworks. One of their most striking results is that, if we restrict ourselves to stationary policies, any utility function u under the ESR criterion provokes the same ordering between policies than another utility function u' under the SER criterion (the opposite claim is not true). Subramani et al. (2023) leave as a conjecture that the same relationship may hold for non-stationary policies. Interestingly, none of our formal results contradict this claim, but a definitive proof is outside of the scope of this work.

Then, Rodriguez-Soto et al. (2024) conducted a study of preference relations over value vectors in MOMDPs, thereby establishing a family of preference relations for which we can construct a quasi-representative utility function under the SER criterion. Here, we have extended the results in Rodriguez-Soto et al. (2024) by defining preference relations over random vectors in an MOMDP, including constant-value vectors as a special case. This modification allowed us to generalise their results, and in particular prove in Theorem 12 that there is a family of preference relations for which we can build a quasi-representative utility function under both the SER and ESR criteria.

Finally, we notice that the MOMDP community has tackled the study of specific families of preference relations, especially lexicographic orders (Wray et al., 2015; Vamplew et al., 2018; Li and Czarnecki, 2019; Vamplew et al., 2021; Skalse et al., 2022; Rodriguez-Soto et al., 2026) or Pareto-dominance relations (Röpke et al., 2023; Van Moffaert and Nowé, 2014; Reymond et al., 2022; Vamplew et al., 2008). Both families of works typically focus on developing algorithms to obtain policies that optimise a specific preference relation (either the ESR or the SER criteria). Our work here formalised both problems as particular instances of the problem of maximising a given preference relation. Thus, we aim to bring to both communities' attention that there might be more in common than they believe.

6.3 Preference Relations in Constrained MDPs

Continuing with the topic of preferences, closely related to our work, Miura (2023) provides a formalisation of preferences and their main characteristics in constrained MDPs (Altman, 1999). In that setting, preferences are defined as sets of *acceptable policies*, and the objective is to determine, for a given environment, whether one can specify constraints and reward functions in a constrained MDP (CMDP) such that the acceptable policies become optimal. Although CMDPs and MOMDPs share several similarities, they ultimately belong to different research areas. In a Constrained MDP, there are also multiple reward functions, like in MOMDPs. However, there is a major difference in how these reward functions are handled between the two frameworks. In an MOMDP, the classical approach is to consider a utility function for aggregating all of them (or even trying to find a set of possible solutions, considering that the utility function might be unknown). In a Constrained MDP, all reward functions except one are associated with a constraint. The goal is to maximise the

remaining reward function while ensuring that the remaining reward functions satisfy the constraints (which are always known). Thus, it is important to consider that formal results in the framework of MOMDPs are not always applicable to CMPDs or vice versa.

Take, for instance, our Lemma 1, which proves that restricting to stationary policies is enough when searching for $\langle u, s\rangle$-optimal policies under the SER criterion. This Lemma is strongly related to Theorem 3.1 in Altman (1999), which states that the space of stationary policies is complete in any CMDP. In fact, Altman's and our formal results are both a direct consequence of Theorem 5.5.1 in Puterman (1998). Indeed, other related work in the field of MOMDPs has directly applied Altman's theorem to prove their formal results, such Lemma 3.1 in Siddique et al. (2020). However, we preferred to explicitly provide a formal proof for MOMDPs analogous to Altman's Theorem 3.1 first, even if it was a relatively minor result.

7. Conclusions and Future Work Directions

Multi-objective Markov decision processes (MOMDPs) are widely regarded as the most suitable framework for sequential multi-objective decision-making problems. In an MOMDP, an agent must balance competing objectives through a utility function that reflects the user's preferences. Nevertheless, current MORL applications largely overlooks two fundamental theoretical questions concerning utility functions: (1) under which conditions does a utility function guarantee the existence of an optimal policy? and (2) which preference relations admit a representation in terms of a utility function? These questions are particularly relevant to the two principal solution criteria for MOMDPs: the Expected Scalarised Returns (ESR) and the Scalarised Expected Returns (SER) criteria.

In this work, we advance the theoretical foundations of MOMDPs by formalising both problems for the ESR and SER criteria and providing a systematic analysis of each. Specifically, we begin by introducing a formal notion of *utility optimality* in MOMDPs, and subsequently derive both sufficient and insufficient conditions for the existence of such policies in any finite MOMDP. Importantly, we have analysed utility functions at two levels: the *state* level (requiring utility-optimality at a single state, such as the initial one), and the *all-states* level (requiring utility-optimality at all states simultaneously). This separation into two levels has allowed us to identify the level at which current MORL algorithms operate (mostly at the state level, also known as the *state-independent* criterion), and hence the formal results on which these algorithms can rely.

Turning to preference relations, we formalised them within MOMDPs. We provided a definition of *preference maximality* in MOMDPs. Then we established minimal conditions under which we can represent a preference relation with some utility function, namely the so-called *quasi-representative* utility function that transforms preference-maximal policies into utility-optimal policies and vice-versa.

For both concepts (utility functions and preference relations), we have focused on connecting our formal results with algorithmic applications. For utility functions, we have proven that current *state-independent* (SI) algorithms always compute $\langle u, s\rangle$-optimal policies, and also compute complete u-optimal policies for an identified family of utility functions. Then, for preference relations, we have introduced the concept of *near-representativity* (a more relaxed notion than quasi-representativity) and shown that any preference relation

that admits a quasi-representative utility functions also admits a near-representative one. Moreover, all these algorithmic connections apply to both the ESR and SER criteria.

Finally, we anticipate that these formal contributions will stimulate further research in both the theoretical and applied domains of MORL. Indeed, our results carry direct practical implications: the MORL community needs to decide at which level they want to work (either the state level, the all-states level or both). As we have observed, an algorithm that computes $\langle u, s\rangle$-optimal policies (i.e., at the state level) may not necessarily learn u-optimal policies. In fact, it may occur that such an algorithm will learn contradictory policies at the all-state level, even if they are perfectly fine at the state level. In more detail, by contradictory policies we refer to policies that follow trajectories that seem optimal from an initial state, but then become suboptimal as the trajectory progresses (e.g., the policy π_2 in the proof of Theorem 10).

We envision multiple future research directions. The rest of this paper will provide an overview of the topics we consider the most significant and urgent challenges in theoretical research on multi-objective reinforcement learning.

7.1 Existence of Utility-Optimal Policies

This work's theoretical results have focused on proving the sufficient and insufficient conditions for the existence of utility-optimal policies (for both the ESR and the SER criteria).

A natural next question is to determine which conditions are necessary for the existence of utility-optimal policies. A first step towards answering this question would be to find a family of utility functions for which u-optimal policies exist that is larger than the one found by Theorem 6. While Theorem 6 already covers all linear utility functions, non-linear utilities are also of significant importance in MORL. Probably the most famous non-linear utility is the Chebyshev function (see (Perny and Weng, 2010; Roijers and Whiteson, 2017)).

Moreover, while our research has focused on proving the existence of utility-optimal policies in general finite MOMDPs, it is also of great interest to study whether specific properties of MOMDPs guarantee the existence of even more utility-optimal policies.

7.2 Novel Metrics and Algorithms

From an algorithmic perspective, we anticipate the development of methods that leverage our theoretical results to better calculate utility-optimal policies. This exploitation will probably require two steps. The first will be the development of metrics to determine whether a policy is truly utility-optimal for all states (and again, these metrics will need to exist for the ESR and SER criteria). Once these metrics are developed, they will need to be incorporated into the structure of novel MORL algorithms that aim to compute utility-optimal policies.

This algorithmic research should first be developed in controlled environments wherein we can apply tabular reinforcement learning (as opposed to deep reinforcement learning (Sutton and Barto, 1998)). This restriction stems from the fact that tabular reinforcement learning algorithms generally provide convergence guarantees in single-objective reinforcement learning, which we are especially interested in maintaining in the multi-objective case.

7.3 More General Models

In this work, we have focused on finite multi-objective Markov decision processes. The finite MOMDP model is the most well-known and applied model for multi-objective sequential decision-making problems (Roijers et al., 2013; Roijers and Whiteson, 2017; Rădulescu et al., 2019; Hayes et al., 2022). Nevertheless, more general formal multi-objective models exist that could be of great use if the same theoretical research studied here were replicated for them. Since these models generalise the finite MOMDP model, all their theoretical results would benefit the theoretical research in finite MOMDPs. Some examples include MOMDPs with a countable state set or a Borel state set.

7.4 Closing Remarks

This paper aims to make clear the importance of theoretical research in multi-objective reinforcement learning. Such research has direct implications for the future development of real-world applications. To this end, we have provided a study of several key theoretical areas in MORL that were still largely unexplored and showed their importance. We aspire for this paper to inspire and contribute to the future development and expansion of the field of theoretical multi-objective reinforcement learning.

Acknowledgments

This work has been supported by the EU-funded VALAWAI (# 101070930), and VAAV (# BOOS_01_10) projects and the Spanish-funded EVASAI (# PID2024-158227NB-C31), and EMOROBCARE (# IASOMM24002) projects.

The author declares no conflicting interest relevant to this work.

References

E. Altman. *Constrained Markov decision processes*. Routledge, 1999.

S. Barbera, P. J. Hammond, and C. Seidl, editors. *Handbook of Utility Theory: Volume 1: Principles*. Kluwer Academic Publishers, 1998.

M. G. Bellemare, W. Dabney, and R. Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 449–458. JMLR.org, 2017.

D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability*. Athena Scientific, 2008.

A. Charpentier, R. Élie, and C. Remlinger. Reinforcement learning in economics and finance. *Comput. Econ.*, 62(1):425–462, 2021. ISSN 0927-7099. doi: 10.1007/s10614-021-10119-4. URL https://doi.org/10.1007/s10614-021-10119-4.

A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109: 101964, 2020. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2020.101964. URL https://www.sciencedirect.com/science/article/pii/S093336572031229X.

G. Debreu. On the preferences characterization of additively separable utility. In A. Tangian and J. Gruber, editors, *Constructing Scalar-Valued Objective Functions*, pages 25–38, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. ISBN 978-3-642-48773-6.

B. B. Elallid, N. Benamar, A. S. Hafid, T. Rachidi, and N. Mrani. A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving. *Journal of King Saud University - Computer and Information Sciences*, 34(9): 7366–7390, 2022. ISSN 1319-1578. doi: https://doi.org/10.1016/j.jksuci.2022.03.013. URL https://www.sciencedirect.com/science/article/pii/S1319157822000970.

E. A. Feinberg. *Total Expected Discounted Reward MDPS: Existence of Optimal Policies*. John Wiley and Sons, Ltd, 2011. ISBN 9780470400531. doi: https://doi.org/10.1002/9780470400531.eorms0906.

Z. Gábor, Z. Kalmár, and C. Szepesvári. Multi-criteria reinforcement learning. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 197–205, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 1558605568.

S. O. Hansson and V. Hendricks. *Introduction to Formal Philosophy*. Springer, Berlin, 2018.

E. Harzheim. *Ordered sets*, volume 7. Springer Science & Business Media, 2005.

C. F. Hayes, T. Verstraeten, D. M. Roijers, E. Howley, and P. Mannion. Expected scalarised returns dominance: A new solution concept for multi-objective decision making. *Neural Computing and Applications*, abs/2106.01048, 2021. URL https://api.semanticscholar.org/CorpusID:235293838.

C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *Autonomous Agents and Multi-Agent Systems*, 36, 2022.

L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *J. Artif. Int. Res.*, 4(1):237–285, May 1996. ISSN 1076-9757.

C. Li and K. Czarnecki. Urban driving with multi-objective deep reinforcement learning. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 359–367, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Y. Li. Deep reinforcement learning: An overview, 2017. URL http://arxiv.org/abs/1701.07274. cite arxiv:1701.07274.

H. Lu, D. Herman, and Y. Yu. Multi-objective reinforcement learning: Convexity, stationarity and pareto optimality. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=TjEzIsyEsQ6.

A. M. Macau, M. R. Soto, E. Marchesini, M. S. Fibla, M. López-Sánchez, J. A. Rodríguez-Aguilar, and A. Farinelli. An approximate embedding for designing ethical reinforcement learning environments. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*, 2025.

M. Maschler, E. Solan, and S. Zamir. *Game Theory, 2nd Edition.* Cambridge University Press, 2013.

J. McColl. *Multivariate Probability.* Wiley, 2004. ISBN 978-0-470-68926-4.

S. Miura. On the expressivity of multidimensional Markov reward. In *Proceedings of the 2022 Conference on Reinforcement Learning and Decision Making.*, 07 2023.

P. Perny and P. Weng. On finding compromise solutions in multiobjective Markov decision processes. *Frontiers in Artificial Intelligence and Applications*, 215:969–970, 01 2010. doi: 10.3233/978-1-60750-606-5-969.

M. Peterson. *An introduction to decision theory.* Cambridge University Press, 2017.

M. L. Puterman. Markov decision processes: Discrete stochastic dynamic programming. *Wiley*, 1998.

M. Reymond, E. Bargiacchi, and A. Nowé. Pareto conditioned networks. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1110–1118, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

M. Rodriguez-Soto, M. Lopez-Sanchez, and J. A. Rodríguez-Aguilar. An analytical study of utility functions in multi-objective reinforcement learning. In *Proceedings of the Thirty-Eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

M. Rodriguez-Soto, R. Rădulescu, F. Bistaffa, O. Ricart, A. Mayoral-Macau, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, and A. Nowé. Multi-objective reinforcement learning for provably incentivising alignment with value systems. *Artificial Intelligence*, 351: 104460, 2026. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2025.104460. URL `https://www.sciencedirect.com/science/article/pii/S0004370225001791`.

D. Roijers and S. Whiteson. *Multi-Objective Decision Making.* Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool, California, USA, 2017. doi:10.2200/S00765ED1V01Y201704AIM034.

D. Roijers, D. Steckelmacher, and A. Nowe. Multi-objective reinforcement learning for the expected utility of the return. In *Adaptive and Learning Agents Workshop (AAMAS 2018)*, 07 2018.

D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley. A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.*, 48(1):67–113, Oct. 2013. ISSN 1076-9757.

D. M. Roijers, S. Whiteson, and F. A. Oliehoek. Computing convex coverage sets for faster multi-objective coordination. *J. Artif. Intell. Res.*, 52:399–443, 2015. URL `https://api.semanticscholar.org/CorpusID:1821001`.

W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3 edition, 1976. ISBN 9780070542358.

W. Röpke, C. Hayes, P. Mannion, E. Howley, A. Nowe, and D. Roijers. Distributional multi-objective decision making. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 05 2023. doi: 10.48550/arXiv.2305.05560.

R. Rădulescu, P. Mannion, D. M. Roijers, and A. Nowé. Multi-objective multi-agent decision making: a utility-based analysis and survey. *Autonomous Agents and Multi-Agent Systems*, 34:1–52, 2019.

R. Rădulescu, P. Mannion, Y. Zhang, D. M. Roijers, and A. Nowé. A utility-based analysis of equilibria in multi-objective normal-form games. *The Knowledge Engineering Review*, 35, 2020. doi: 10.1017/S0269888920000351.

U. Siddique, P. Weng, and M. Zimmer. Learning fair policies in multiobjective (deep) reinforcement learning with average and discounted rewards. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

J. Skalse and A. Abate. On the limitations of Markovian rewards to express multi-objective, risk-sensitive, and modal tasks. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org, 2023.

J. Skalse, L. Hammond, C. Griffin, and A. Abate. Lexicographic multi-objective reinforcement learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 3405–3411, 07 2022. doi: 10.24963/ijcai.2022/473.

K. Steele and H. O. Stefánsson. *Decision theory*. Stanford University, 2015. ISBN 1096-5054.

R. Steuer and E. Choo. An interactive weighted tchebycheff procedure for multiple objective programming. *Mathematical Programming*, 26:326–344, 10 1983. doi: 10.1007/BF02591870.

J. Stewart. *Calculus: Early Transcendentals*. Brooks/Cole, Cengage Learning, 7 edition, 2011. ISBN 9781111426682.

R. Subramani, M. Williams, M. Heitmann, H. Holm, C. Griffin, and J. Skalse. On the expressivity of objective-specification formalisms in reinforcement learning. In *Eleventh International Conference on Learning Representation.*, 2023.

R. S. Sutton and A. G. Barto. *Reinforcement learning - an introduction*. Adaptive computation and machine learning. MIT Press, 1998. ISBN 0262193981.

P. Vamplew, J. Yearwood, R. Dazeley, and A. Berry. On the limitations of scalarisation for multi-objective reinforcement learning of pareto fronts. In *Proceedings of the 21st Australasian Joint Conference on Artificial Intelligence: Advances in Artificial Intelligence*, pages 372–378, 12 2008. ISBN 978-3-540-89377-6. doi: 10.1007/978-3-540-89378-3_37.

P. Vamplew, R. Dazeley, E. Barker, and A. Kelarev. Constructing stochastic mixture policies for episodic multiobjective reinforcement learning tasks. In A. Nicholson and X. Li, editors, *AI 2009: Advances in Artificial Intelligence*, pages 340–349, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-642-10439-8.

P. Vamplew, R. Dazeley, C. Foale, S. Firmin, and J. Mummery. Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology*, 20, 03 2018. doi: 10.1007/s10676-017-9440-6.

P. Vamplew, C. Foale, R. Dazeley, and A. Bignold. Potential-based multiobjective reinforcement learning approaches to low-impact agents for ai safety. *Engineering Applications of Artificial Intelligence*, 100, 04 2021. doi: 10.1016/j.engappai.2021.104186.

P. Vamplew, B. J. Smith, J. Källström, G. Ramos, R. Rădulescu, D. M. Roijers, C. F. Hayes, F. Heintz, P. Mannion, P. J. K. Libin, R. Dazeley, and C. Foale. Scalar reward is not enough: a response to silver, singh, precup and sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2), 2022. ISSN 1387-2532. doi: 10.1007/s10458-022-09575-5. URL `https://doi.org/10.1007/s10458-022-09575-5`.

P. Vamplew, C. Foale, C. F. Hayes, P. Mannion, E. Howley, R. Dazeley, S. Johnson, J. Källström, G. Ramos, R. Radulescu, W. Röpke, and D. M. Roijers. Utility-based reinforcement learning: Unifying single-objective and multi-objective reinforcement learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, page 2717–2721, Richland, SC, 2024. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9798400704864.

K. Van Moffaert and A. Nowé. Multi-objective reinforcement learning using sets of pareto dominating policies. *J. Mach. Learn. Res.*, 15(1):3483–3512, Jan. 2014. ISSN 1532-4435.

K. Van Moffaert, T. Brys, A. Chandra, L. Esterle, P. Lewis, and A. Nowe. A novel adaptive weight selection algorithm for multi-objective multi-agent reinforcement learning. In *Proceedings of the 2014 International Joint Conference on Neural Networks*, 07 2014. doi: 10.1109/IJCNN.2014.6889637.

P. Weng. Markov decision processes with ordinal rewards: reference point-based preferences. In *Proceedings of the Twenty-First International Conference on International Conference on Automated Planning and Scheduling*, ICAPS'11, page 282–289. AAAI Press, 2011.

K. Wray, S. Zilberstein, and A.-i. Mouaddib. Multi-objective mdps with conditional lexicographic reward preferences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29, 01 2015. doi: 10.1609/aaai.v29i1.9647.

W. Zhao, J. P. Queralta, and T. Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 737–744, 2020. doi: 10.1109/SSCI47803.2020.9308468.