Markov Decision Processes with Ordinal Rewards: Reference Point-Based Preferences

Paul Weng
LIP6, UPMC
4 place Jussieu
75005 Paris, France

Abstract

In a standard Markov decision process (MDP), rewards are assumed to be precisely known and of quantitative nature. This can be a too strong hypothesis in some situations. When rewards can really be modeled numerically, specifying the reward function is often difficult as it is a cognitively-demanding and/or time-consuming task. Besides, rewards can sometimes be of qualitative nature as when they represent qualitative risk levels for instance. In those cases, it is problematic to use directly standard MDPs and we propose instead to resort to MDPs with ordinal rewards. Only a total order over rewards is assumed to be known. In this setting, we explain how an alternative way to define expressive and interpretable preferences using reference points can be exploited.

1 Introduction

The model of Markov decision processes (MDP) is a general model for solving sequential decision-making problems under uncertainty (Puterman 1994; Russell and Norvig 2003). Its exploitation in practice can sometimes be difficult as it requires a precise knowledge of its parameters (transition probabilities and rewards). In many real situations, they are only known imprecisely because it can be difficult or costly, even impossible to determine them precisely. As the solution of an MDP can naturally be very sensitive to those parameters, it is then often very delicate to set those values.

This observation has motivated some recent work on finding robust policies in sequential decision-making under uncertainty problems (Givan, Leach, and Dean 2000; Bagnell, Ng, and Schneider 2001; Nilim and El Ghaoui 2003; Trevizan, Cozman, and de Barros 2007; Jeantet and Spanjaard 2009; Regan and Boutilier 2009). In those works, probabilities alone or probabilities with rewards are assumed to be known imperfectly. The parameters can then be represented for instance by intervals instead of precise values. Generally, this line of research mainly focused on ignorance or partial knowledge of probabilities. An exception to that is the recent work of (Regan and Boutilier 2009) where the authors propose a method for eliciting numeric rewards with the minmax regret criterion. However, the proposed method has a high computation cost.

Furthermore, in some situations, rewards are really of qualitative nature as it is the case when they represent qualitative risk levels for instance. Then their representation by additive quantitative values is inappropriate. Possibilistic MDPs (Dubois et al. 1996; Sabbadin 1999) allow for taking into account the situations where rewards are qualitative. However, the uncertainty representation then has to be possibilistic, which can be unsuitable when the dynamic of the system is stochastic. To the best of our knowledge, there is no extension of MDPs allowing for both a probabilistic representation of uncertainty and qualitative rewards.

In this work, we are interested in the case where rewards are qualitative or ill-known while probabilities can be precisely determined. We think that this case can be common in practice. Indeed, if there generally exists some means to evaluate the transition probabilities of a system (experiments and statistical estimation for instance), preferences entailed by reward values are more difficult to define. On which grounds shall we choose to set a reward to a certain value rather than another? The works in preference elicitation in decision theory (Keeney and Raiffa 1976) suggest that this is a difficult task. That is why we assume in our framework that only an ordering of the different rewards is known.

In this paper, we propose a variant of MDPs that only relies on ordinal rewards (OMDPs) and we show how to build a preference system in that setting by introducing a reference point. By doing so, we reveal also the assumptions that are implicitly made in standard MDPs. Although the proposed reference point-based preference model shares some similarity with that of standard MDPs, their semantics is quite different. In our ordinal setting, this point is important as the introduced reference point allows to give a natural interpretation to the constructed preferences over policies. An optimal policy in our context is a policy that maximizes the proportion of times it has better rewards than the reference point. We give some ideas about how to pick a reference point. Then we show that an OMDP can be interpreted as a particular MDP with numerical vectorial rewards (VMDP). We explain how to solve the problem by recasting the original problem into a standard MDP. Finally, we illustrate our proposition with some simple examples.

2 Background

Markov Decision Processes

The model of *Markov decision processes* (MDP) is generally defined as a quadruplet (S, A, P, R) (Puterman 1994): S a finite set of states, A a finite set of actions, $P\colon S \times A \to \mathcal{P}(S)$ a transition function where $\mathcal{P}(S)$ is the set of probability distributions over S and $R\colon S \times A \to X \subset \mathrm{I\!R}$ a reward function. The *transition function* provides the probability that a future state occurs after performing an action in a state. As usually, we write $P(s, a, s')$ for $P(s, a)(s')$. The *reward function* gives the immediate reward that the agent receives after executing an action in a state. The set X – a finite set as S and A are finite – represents the set of all possible rewards.

A *decision rule* δ indicates which action to choose in each state for a given step. It can be *deterministic* : $\delta : S \to A$ is then a function from the set of states S into the set of actions A. But it can also be *randomized* : $\delta : S \to \mathcal{P}(A)$ is then a function from the set of states S into the set of probability distributions over actions A. The choice of the action to be executed in a state is therefore chosen randomly according to the probability distribution associated to that state. Notice that a deterministic decision rule is simply a randomized rule whose probability distributions are degenerate (only one action having a probability of 1 in each state). For conciseness' sake, we write $R^\delta(s) = R(s, \delta(s))$ and $P^\delta(s, s') = P(s, \delta(s), s')$ for any δ and all $s, s' \in S$. A *policy* π at an horizon h is a sequence of h decision rules, denoted $\pi = (\delta_1, \ldots, \delta_h)$ where each δ_i is a decision rule. It is said *deterministic* when it only contains deterministic decision rules and *randomized* otherwise. The set of deterministic policies (resp. randomized) at horizon h is denoted Π_h^D (resp. Π_h^M). At the infinite horizon, a policy is simply an infinite sequence of decision rules. The set of those deterministic policies (resp. randomized) is denoted Π_∞^D (resp. Π_∞^M). A policy is said *stationary* if at each decision step, the same decision rule is applied.

A *preference relation* $\succsim$, which is simply a binary relation, is assumed to be defined over policies. A policy π is said *preferred to* or *$\succsim$-dominates* (or simply *dominates*) another policy π' when $\pi \succsim \pi'$. We write $\pi \succ \pi'$ (resp. $\pi \sim \pi'$) when $\pi \succsim \pi'$ and not $\pi' \succsim \pi$ (resp. $\pi \succsim \pi'$ and $\pi' \succsim \pi$). A policy π is *preferred* or *non $\succsim$-dominated* (or simply *non-dominated*) when there is no policy π' such that $\pi' \succ \pi$. It is *optimal* when it is preferred to any other policy. Remark that an optimal policy does not necessarily exist because as the preference relation can be partial in the general case, some policies can be incomparable. Solving an MDP amounts to determining a preferred policy for a certain preference system. We now recall how those preferences are defined in the standard framework.

Histories, which can be finite or infinite correspond to the following sequences, starting from state $s_0 \in S$: $(s_0, a_1, s_1, a_2, s_2, \ldots)$ where $\forall i \in \mathrm{I\!N}, (a_i, s_i) \in A \times S$. The value of a history $\gamma = (s_0, a_1, s_1, a_2, s_2, \ldots, a_h, s_h)$ can be defined in several ways. One can simply sum all the rewards obtained along a history:

$$R^\Sigma(\gamma) = \sum_{i=0}^{h-1} r(s_i, a_{i+1})$$

One can also use a discounted sum of the rewards:

$$R^\beta(\gamma) = \sum_{i=0}^{h-1} \beta^i r(s_i, a_{i+1})$$

where $\beta \in [0, 1[$ is a discount factor. Finally, it is possible to consider the average of the rewards:

$$R^\mu(\gamma) = \frac{1}{h} \sum_{i=0}^{h-1} r(s_i, a_{i+1})$$

Those values can be extended when the history is infinite. For the total sum R^Σ, there could be convergence problems. The discounted sum is well defined thanks to the discount factor. For the reward average, the value of a history is defined as a limit, if there exists:

$$R^\mu(\gamma) = \lim_{h\to\infty} \frac{1}{h} \sum_{i=0}^{h-1} r(s_i, a_{i+1})$$

A decision rule δ from an initial state s induces a probability distribution over histories (of length 1). As a value can be associated to every history, δ also induces a probability distribution over the set X of possible rewards. This probability distribution is equal to $P(s, \delta(s))$. By induction, a policy π in a given initial state s can be associated to a probability distribution over histories. Thus a policy also induces a probability distribution over the values of histories. Consequently, it is possible to define the expected reward that a policy can yield from an initial state.

The function $v^\pi : S \to \mathrm{I\!R}$, which associates to each state s the expected reward that can be obtained from the policy π is called the *value function* of π :

$$v^\pi(s) = E_s^\pi(R^*(\Gamma))$$

where E_s^π is the expectation with respect to the probability distribution induced by the application of π from state s, $* \in \{\Sigma, \beta, \mu\}$ and Γ is a random variable over histories. Here, a policy π is *preferred* to another policy π' if:

$$\pi \succsim \pi' \Leftrightarrow \forall s \in S, v^\pi(s) \geq v^{\pi'}(s)$$

In the classical framework, the preference relation defined in such a way guarantees that an optimal stationary deterministic policy exists.

Depending on the use of the total sum, the discounted sum or the average, the value function is said to rely on the *expected total criterion*, *discounted criterion* or *average criterion*. Those three criteria are based on expectation, the difference residing in how histories are valued. In fact, those criteria are instances of *expected utility* (Bouyssou et al. 2000) where the history values play the role of utilities. Thus, they are expected utilities for which one assume that utilities are additively decomposable. To remain simple, in the following, we only consider the discounted criterion to avoid any problem at the infinite horizon. Our approach could naturally be extended to the other two criteria.

Motivations of this work

The classical approach is not questionable when all the parameters of the model are numeric and precisely known.

However, when only an ordering over rewards is known, it can be problematic to set arbitrarily their values as the preferred policies can be very sensitive to them. The problem of determining the values of rewards, called *elicitation of preferences* is a thorny problem studied in decision theory (Bouyssou et al. 2000).

The problem that we consider in this paper only appears when preferences have a certain degree of complexity, that is to say when one needs at least two distinct non null values for defining the reward function. When only one non null value is sufficient, as it is the case for instance in problems where there is only one set of goal states (all identical) and where all the other states are considered equivalent and of null value, the choice of the two reward values is not important (as long as the order is respected) because it does not influence the optimal policy (or policies if there are more than one).

Proposition 1. *Let $r, r' \in \mathbb{R}$ such that r and r' are both positive or both negative. Let an MDP (S, A, P, R) whose reward function can only take two different values 0 and r. Define R' the function from R by substituting r' for r. Then, the MDPs (S, A, P, R) and (S, A, P, R') have the same optimal policy (or policies if there are more than one).*

Proof. The two MDPs have the same policies. It is sufficient to show that the preference direction between two policies remains the same whether R or R' is used. Notice that it is possible to transform one into the other by multiplying by a positive constant r'/r or r/r', which are positive as r and r' have the same sign. As the discounted criterion is linear, such a transformation preserves the inequalities and thus the direction of preferences between two policies. □

In the case where the reward function needs at least two distinct non null values, the choice of those values can have an important impact on the optimal policies as we illustrate it on a simple example.

Example 2. *Consider the following MDP where $S = \{1, 2\}$ and $A = \{a, b\}$. The discount factor is set at $\beta = 0.5$. The transition function is defined as follows:*

$$P(1, a, 1) = 1 \quad P(1, b, 1) = 0.5 \quad P(2, a, 1) = 1$$

To simplify, we assume that action b is not possible in state 2. In this case, there are only two deterministic stationary policies depending on the choice of the action in the first state. Assume that we only know $R(1, b) > R(1, a) > R(2, a)$. $R(2, a)$ represents no reward, $R(1, a)$ a small reward and $R(1, b)$ a big reward. If the reward function is arbitrarily defined as follows:

$$R(1, b) = 2 \quad R(1, a) = 1 \quad R(2, a) = 0$$

then we can easily check that the best policy is the one consisting in choosing action b in state 1 The value function obtained in that state equals to $\frac{16}{5} = 3.2$ against 2 for the other policy.

Now, if the reward function were defined as follows:

$$R(1, b) = 10 \quad R(1, a) = 9 \quad R(2, a) = 0$$

the best policy would have been the one choosing action a. The value function in state 1 would be equal to 18 against 16 for the other policy.

Although the two functions respect the order imposed on rewards, we observe an inversion of preferences. Thus the choice of the scale of valuation for rewards can have an important impact on the optimal policy. This can be problematic in some situations.

3 Ordinal Reward MDP

As the preferences over policies can be sensitive to the choice of the values of rewards, we propose in this paper not to introduce arbitrarily this information when it is unknown. In the situations where only an ordinal information about rewards is available, a semi-qualitative model of MDPs, called *Ordinal Reward MDP* (OMDP), can be exploited. The reward function $R : S \times A \to E$ is then defined over a qualitative scale $(E, >)$ totally ordered, the number of steps of this scale being the number of different values that is needed to model the preferences of the considered problem. The scale is necessarily finite as the sets S and A are assumed to be finite. Let $n \in \mathbb{N}$ be the number of steps of the scale $E = \{r_1 > r_2 \ldots > r_n\}$. We assume that there exists $k_0 \in \{1, \ldots, n\}$ such that reward r_{k_0} is considered a neutral reward, i.e. receiving r_{k_0} is considered neither good nor bad.

Preferences over histories

In the same way as in classical MDPs, before defining preferences over policies, we need first to define a preference relation over histories. In OMDPs, a history $(s_0, a_1, s_1, a_2, s_2, \ldots)$ is valued by a sequence of ordinal rewards $(R(s_0, a_1), R(s_1, a_2), \ldots)$. Exploiting directly those sequences of rewards to compare histories is infeasible when the horizon is high or infinite. We propose to summarize the preferential information contained in a sequence by counting the number of occurrences of each different ordinal rewards in that sequence. We will state explictly the assumption we make after defining how the rewards are counted.

When the horizon h is finite, the preferential information on a history $\gamma = (s_0, a_1, s_1, \ldots, a_h, s_h)$ is defined by:

$$N^{\Sigma}(\gamma) = (N_1^{\Sigma}(\gamma), \ldots, N_n^{\Sigma}(\gamma))$$

where for any $k = 1, \ldots, n, N_k^{\Sigma}(\gamma)$ is the number of occurrences of reward r_k in the sequence of ordinal rewards associated to γ. The $N_k^{\Sigma}(\gamma)$'s can be defined by:

$$N_k^{\Sigma}(\gamma) = \sum_{i=0}^{h-1} \chi_{R(s_i, a_{i+1}) = r_k}$$

where $\chi_{R(s_i, a_{i+1}) = r_k}$ is the indicator function[1].

Alternatively, one can introduce a discount factor $\beta \in [0, 1[$ with the same semantic as in classical MDPs, meaning that a reward r obtained h steps from now is worth $\beta^{h-1} \times r$ now. In this case, the preferential information on a history γ is defined by:

$$N^{\beta}(\gamma) = (N_1^{\beta}(\gamma), \ldots, N_n^{\beta}(\gamma))$$

[1] $\chi_{R(s_i, a_{i+1}) = r_k} = 1$ if $R(s_i, a_{i+1}) = r_k$ and 0 otherwise.

where for any $k = 1, \ldots, n, N_k^\beta(\gamma)$ is the discounted number of occurrences of reward r_k in the sequence of ordinal rewards associated to γ. The $N_k^\beta(\gamma)$'s can be defined by:

$$N_k^\beta(\gamma) = \sum_{i=0}^{h-1} \beta^i \chi_{R(s_i, a_{i+1}) = r_k}$$

Finally, one can also consider the average number of occurrences:

$$N^\mu(\gamma) = (N_1^\mu(\gamma), \ldots, N_n^\mu(\gamma))$$

where for any $k = 1, \ldots, n, N_k^\mu$ is the average number of occurrences of reward r_k in the sequence of ordinal rewards associated to γ. The $N_k^\mu(\gamma)$'s can be defined by:

$$N_k^\mu(\gamma) = \frac{1}{h} \sum_{i=0}^{h-1} \chi_{R(s_i, a_{i+1}) = r_k}$$

In the same way as in the classical case, those values could be extended to the infinite horizon when they are well defined.

For the sake of simplicity, we use the same notation for the preference relations over histories, vectors and policies as they are of the same nature. The context will tell if they are defined over histories, over vectors or over policies. We can then state the following assumptions, for $*$ in $\{\Sigma, \beta, \mu\}$:

H*. For any two histories γ, γ', we would have:

$$\gamma \succsim \gamma' \Leftrightarrow N^*(\gamma) \succsim N^*(\gamma')$$

Depending on how we choose to evaluate a history in OMDPs, we will assume H$^\Sigma$, H$^\beta$ or H$^\mu$. As a side note, these natural assumptions are also made in classical MDPs. Indeed, if E were a numerical scale, we have:

Proposition 1. *$R^*(\gamma) = \sum_{k=1}^n N_k^* r_k$ where $* \in \{\Sigma, \beta, \mu\}$ and γ any history.*

Preferences over vectors

Under one of the assumptions H*, comparing histories amounts to comparing vectors. Therefore, we now need to define how we compare those occurrence vectors. Let us review some possible preference relations over vectors in $\mathbb{R}^n$.

Pareto Dominance. In our context, we cannot compare vectors directly with Pareto Dominance[2]. Indeed, as an order is defined over the different ordinal rewards, we have for example $e_i \succsim e_j$ for any $i < j$, where e_i (resp. e_j) is the vector null everywhere except on component i (resp. j) where it is equal to 1.

Pareto Dominance over Cumulative Vectors. A natural dominance relation $\succsim_D$ can be defined between any two vectors $N, N' \in \mathbb{R}^n$:

$$N \succsim_D N' \Leftrightarrow \forall i = 1, \ldots, n, \sum_{k=1}^i N_k \geq \sum_{k=1}^i N'_k \qquad (1)$$

This relation has a natural interpretation. It states that for any reward r_i, the number of rewards better than r_i is higher in N than in N'. This dominance is the first-order stochastic dominance (Shaked and Shanthikumar 1994) expressed in our settting. It can also be viewed as Pareto dominance over transformed vectors $L(N) = \left(N_1, N_1 + N_2, \ldots, \sum_{k=1}^n N_k\right)$.

Unfortunately, this dominance is generally not very discriminating as it is a partial order due to the condition "for all i". As equation 1 is a rational condition to impose on a preference relation, it is natural to want to refine it with a more discriminating preference relation.

Lexicographic Orders. Then, a simple and natural idea is to rely on the order on scale E to compare the valuation vectors with the lexicographic order as follows, for any $x, y \in \mathbb{R}^n$:

$$x \succsim_L y \Leftrightarrow \exists i = 1, \ldots, n, \begin{cases} \forall j < i, x_j = y_j \\ \text{et } x_i > y_i \end{cases} \qquad (2)$$

Its interpretation is simple in our framework. In the comparison of two vectors, we want the first component of the vector to be the highest possible as it corresponds to the best reward. In case of equality, we would look at the second component and so on.

However, the drawback with lexicographic order is that compensation is forbidden between different rewards. Thus, for instance, e_1 would be preferred to $100 \times e_2$, which can be questionable. Besides, we would prefer to have a higher degree of expressiveness in the definition of the preferences.

Proposed Criterion. In order to define a reasonable preference relation $\succsim$ over vectors to use in our setting, we will use results from Measurement Theory (Krantz et al. 1971). To that end, we list some natural properties, called axioms, that we want $\succsim$ to satisfy. Those axioms will define completely which criterion to use to represent $\succsim$.

We first assume that the preference relation $\succsim$ over those vectors satisfies:

A1. $\succsim$ is a complete preorder[3].

Axiom A1 implies that any two vectors can be compared and that the relation is transitive, which is a very natural property to impose on preference relation $\succsim$. Then, we assume that, $\forall N, N' \in \mathbb{R}^n$:

A2. $N \succsim N' \Leftrightarrow \forall i = 1, \ldots, n, N + e_i \succsim N' + e_i$

Axiom A2 entails that the common part of any two vectors does not influence the direction of the preference relation. We assume also an Archimedian property, $\forall N$, N', M, $M' \in \mathbb{R}^n$:

A3. $N \succ N' \Rightarrow \exists n \in \mathbb{N}, nN + M \succsim nN' + M'$

A3 can be interpreted as follows: the preference "difference" between N and N', which is positive as $N \succ N'$, can be

[2] A vector x Pareto-dominates a vector y if and only if for all i, $x^{(i)} \geq y^{(i)}$ and there exists j, $x^{(j)} > y^{(j)}$

[3] $\succsim$ over X is a complete preorder iff:

- (complete) $\forall x, y \in X, x \succsim y$ or $y \succsim x$
- (reflexive) $\forall x \in X, x \succsim x$
- (transitive) $\forall x, y, z \in X, x \succsim y$ and $y \succsim z \Rightarrow x \succsim z$

made great enough to compensate the preference "difference" of M and M' or in English, any positive preference "difference" can be made as great as we want. Those axioms are all satisfied in the classical setting for R^Σ, R^β and R^μ. Using those three axioms, we have the following representation theorem:

Theorem 2. *The two following propositions are equivalent:*
(i) $\succsim$ satisfies Axioms A1, A2 and A3.
(ii) there exists a function $u : E \to \mathbb{R}$ such that $\forall N, N' \in \mathbb{R}^n$:

$$N \succsim N' \Leftrightarrow \sum_{k=1}^{n} N_k u(r_k) \geq \sum_{k=1}^{n} N'_k u(r_k)$$

Proof. (i) ⇒ (ii): By Axioms A1, A2 and A3, we easily show that the structure $(\mathbb{R}^n, \succsim, +)$ is a closed extensive structure[4] where $+$ is the componentwise addition. Then, by Theorem 1 in Section 3 of (Krantz et al. 1971) that states that a closed extensive structure can be "measured", there exists a function $\phi : \mathbb{R}^n \to \mathbb{R}$ such that $\forall N, N' \in \mathbb{R}^n$:

$$N \succsim N' \Leftrightarrow \phi(N) \geq \phi(N') \tag{3}$$
$$\phi(N + N') = \phi(N) + \phi(N') \tag{4}$$

For any $N \in \mathbb{R}^n$, we have $N = \sum_{k=1}^{n} N_k e_k$ and thus, by Eq. 4:

$$\phi(N) = \sum_{k=1}^{n} N_k \phi(e_k) \tag{5}$$

We can finally define a function $u : E \to \mathbb{R}$ with $u(r_k) = \phi(e_k)$ with $k = 1, \ldots, n$.

(ii) ⇒ (i): It is easy to check that the proposed criterion satisfies the three axioms. □

The previous theorem shows what are the assumptions that are implicitly made when we use a real reward function in standard MDPs. Remark that u is a function that values each ordinal reward and therefore allows a numeric reward function to be defined. As Axioms A1, A2 and A3 seem to be natural, we want to assume them as well in our ordinal setting. However, we will interpret the reward function differently.

We add two other simple axioms:

A4. $e_1 \succsim e_2 \succsim \ldots \succsim e_n$

Axiom A4 simply expresses the preference order over the ordinal rewards. And, finally, we assume:

A5. $N \sim N + e_{k_0}$

Axiom A5 states explicitly that r_{k_0} is a null reward.

To ease the exposition, we assume at first that all rewards are positive feedbacks ($k_0 = n$). For a fixed positive vector $\tilde{N} \in \mathbb{R}^n_+$, called a *reference point*, we denote $\phi_{\tilde{N}} : \mathbb{R}^n \to \mathbb{R}$ a function defined by:

$$\phi_{\tilde{N}}(N) = \sum_{k=1}^{n-1} N_k \sum_{j=k}^{n-1} \tilde{N}_j$$

By applying the previous theorem, we have the following representation theorem with criterion $\phi_{\tilde{N}}$:

Corollary 3. *The two following propositions are equivalent:*
(i) $\succsim$ satisfies Axioms A1 to A5.
(ii) there exists a reference point $\tilde{N} \in \mathbb{R}^n_+$ such that $\forall N, N' \in \mathbb{R}^n$:

$$N \succsim N' \Leftrightarrow \phi_{\tilde{N}}(N) \geq \phi_{\tilde{N}}(N')$$

Proof. (i) ⇒ (ii): We give the proof for the general case when $k_0 = 1, \ldots, n$ and $\phi_{\tilde{N}}$ is defined by Eq. 9. By Axioms A1, A2 and A3, we know that there exists a function $\phi : \mathbb{R}^n \to \mathbb{R}$ such that vectors can be compared with:

$$\phi(N) = \sum_{k=1}^{n} N_k \phi(e_k) \tag{6}$$

Let us define $\tilde{N}$ by:

$$\begin{aligned} \tilde{N}_k &= \phi(e_k) && \text{if } k = k_0 \\ &= \phi(e_k) - \phi(e_{k+1}) && \text{if } k < k_0 \\ &= \phi(e_{k-1}) - \phi(e_k) && \text{if } k > k_0 \end{aligned} \tag{7}$$

We have $\tilde{N}_{k_0} = \phi(e_{k_0}) = 0$ as by Axiom A5, $\phi(e_{k_0}) = \phi(e_{k_0}) + \phi(e_{k_0})$. By Axiom A4, we have $\phi(e_i) \geq \phi(e_j)$ if $i < j$. Thus, $\forall k = 1, \ldots, n$, $\tilde{N}_k \geq 0$. Finally, we can check that it is equivalent to compare vectors with ϕ or $\phi_{\tilde{N}}$.

(ii) ⇒ (i): It is easy to check that the proposed criterion satisfies the five axioms. □

Interpretation. Assume that a reference point $\tilde{N} \in \mathbb{R}^n_+$ has been chosen. We will explain later how it could be chosen. Any vector of $\mathbb{R}^n_+$ can be interpreted as the composition of a bag (or multiset) of rewards. $\phi_{\tilde{N}}(N)$ can then be interpreted as the cardinal of the Cartesian product of the bags defined by N and $\tilde{N}$ with the constraint that the reward in N is better than that of $\tilde{N}$. In English, $\phi_{\tilde{N}}(N)$ represents the number of times N yields better rewards than $\tilde{N}$.

Remark that for a given initial state s and a given horizon h, in MDPs, we only compare histories of equal length (which depends on h) and starting from s. Thus, for $* \in \{\Sigma, \beta, \mu\}$, when we compare $N = N^*(\gamma)$ and $N' = N^*(\gamma')$ obtained from two histories γ, γ', as their lengths are identical, we have $\sum_{k=1}^{n} N_k = \sum_{k=1}^{n} N'_k$. Quantities $N / \sum_{k=1}^{n} N_k$, where the division is componentwise, represent the proportion of each reward in the associated history. Let us define $\phi'_{\tilde{N}} : \mathbb{R}^n \to \mathbb{R}$ by:

$$\phi'_{\tilde{N}}(N) = \frac{\phi_{\tilde{N}}(N)}{\sum_{k=1}^{n} N_k \sum_{k=1}^{n} \tilde{N}_k} \tag{8}$$

Comparing vectors with $\phi_{\tilde{N}}$ or $\phi'_{\tilde{N}}$ yields the same results as long as we compare vectors whose sums are equal, which is the case in MDPs.

[4] $(A, \succsim, \circ)$ is a closed extensive structure iff $\forall a, b, c, d \in A$:

- $\succsim$ is a complete preorder
- $a \circ (b \circ c) \sim (a \circ b) \circ c$
- $a \succsim b \Leftrightarrow a \circ c \succsim b \circ c \Leftrightarrow c \circ a \succsim c \circ b$
- $a \succ b \Rightarrow \exists n > 0, na \circ c \succsim nb \circ d$

As all rewards are considered positive feedbacks, $\phi'_{\tilde{N}}(N)$ could be interpreted as the proportion of times that N has better rewards than the reference point $\tilde{N}$. In the probabilistic language, this relation could be interpreted as the probability that the random variable with probability distribution $N/\sum_{k=1}^{n} N_k$ yields a better reward than the random variable with distribution $\tilde{N}/\sum_{k=1}^{n}\tilde{N}_k$. In other words, a history γ is valued by the probability that a random drawing of a reward in the sequence of rewards associated to γ yields a better reward than an independent random drawing of a reward with respect to a probability distribution defined by the reference point.

Discussion. As the value $\phi_{\tilde{N}}(N)$ measures the extent to which N is better than $\tilde{N}$, another natural idea for defining a preference relation over vectors would be to use the following relation, $\forall N, N' \in \mathrm{I\!R}^n$:

$$N \succsim_C N' \Leftrightarrow \phi_{N'}(N) \geq \phi_N(N')$$

This preference relation could be interpreted in a simple manner: N is preferred to N' if and only if the extent to which N is better than N' is greater than that of the opposite outcome.

Unfortunately, although this preference relation is total, there can be cycles. We adapt an example from (Perny and Pomerol 1999) in our context:

Example 4. *Assume that $E = \{r_1 > r_2 > r_3 > r_4 > r_5\}$. Consider three vectors N, N', N'' over E defined by:*

	r_1	r_2	r_3	r_4	r_5
N	0	51	0	0	49
N'	0	0	100	0	0
N''	49	0	0	51	0

It is easy to check that $N \succ N' \succ N'' \succ N$.

Therefore, this relation cannot be used to define a rational preference relation.

Preferences over policies

The preference system in an OMDP is completely specified when a reference point has been chosen. Indeed, the extension of the previous preference relation over histories to a preference relation over policies is quite straightforward. Similarly to classical MDPs, we can value a policy π in a state by taking the expectation over the values of all histories generated by π. Here, a policy would be valued in a state by the vector of expected numbers of occurrences of each reward. Then, a policy π is preferred to a policy π' for a given reference point $\tilde{N}$ if and only if for all state $s \in S$:

$$\phi_{\tilde{N}}(E_s^{\pi}(N^*(\Gamma))) \geq \phi_{\tilde{N}}(E_s^{\pi'}(N^*(\Gamma)))$$

where E_s^{π} is the expectation with respect to the probability distribution induced by the application of π from the initial state s, $* \in \{\Sigma, \beta, \mu\}$ and Γ is a random variable over histories. Since, for comparing two policies in a state at a certain horizon, using $\phi_{\tilde{N}}$ or $\phi'_{\tilde{N}}$ is completely equivalent, the proposed preference system values a policy in a state by the expected proportion of times it yields better rewards than a reference point. Thus, an optimal policy in a state is a policy that maximizes that proportion.

Choosing a Reference Point

We now discuss how a reference point can be chosen for a given problem.

A first and natural idea would be to choose a step of scale E as a reference point, i.e. $\tilde{N} = e_k$ for a certain k. Such a choice could be interpreted as the desire of maximizing the number of rewards better than r_k, generated by the optimal policy. In some situations, it may suitable. But, the drawback of this approach is that it may lead to not very discriminating preferences as it amounts to defining a reward function that can only take one non-null value.

A second idea would be to use a probability distribution over E (or a vector of proportions of each reward) as a reference point. Using the probabilistic interpretation of the proposed preference relation, optimizing with such a reference point amounts to finding a policy π that maximizes the probability of drawing a reward in a history generated by π better than one drawn with respect to the probability distribution of the reference point.

A third idea would be to choose a history as a reference point. Indeed, depending on the problem that we want to solve, there may be a natural history that could be used as a reference point. In Section 5 , we illustrate this case with a simple problem. Extending this idea, one could also want to use a set of histories and combine their associated vectors with a mean for instance.

A final idea would be to choose a policy as a reference point. In certain situations, it could be easy to pick some policy as a reference point. For instance, for a problem where there is already a policy that is implemented and used, one could pick that policy in order to find a policy that would be even better. To that effect, the associated VMDP (see next section) could be utilized to find the expected vector of occurrences of the reference policy. We illustrate this case in Section 5.

Remark that for the last two approaches, the initial state has to be known as the reward function defined by the reference point depends on it.

Extension to the General Case

Negative Feedbacks We assume now that $k_0 = 1$, i.e. all rewards are negative feedbacks. For a reference point $\tilde{N} \in \mathrm{I\!R}_+^n$, we define $\phi_{\tilde{N}} : \mathrm{I\!R}^n \to \mathrm{I\!R}$ by:

$$\phi_{\tilde{N}}(N) = -\sum_{k=2}^{n} N_k \sum_{j=2}^{k} \tilde{N}_j$$

Corollary 3 applies here as well and shows that if Axiom A1 to A5 are assumed then one should use $\phi_{\tilde{N}}$ with a chosen reference point $\tilde{N}$ to compare vectors.

When comparing histories (or policies) from a certain initial state and a given horizon, using $\phi_{\tilde{N}}$ or $\phi'_{\tilde{N}}$ defined by:

$$\phi'_{\tilde{N}}(N) = 1 + \frac{\phi_{\tilde{N}}(N)}{\sum_{k=1}^{n} N_k \sum_{k=1}^{n} \tilde{N}_k}$$

gives the same results. Then, the interpretation is similar to the positive case: $\phi'_{\tilde{N}}$ gives the proportion of times that N has strictly better rewards than $\tilde{N}$.

Positive and Negative Feedbacks Finally, when $1 < k_0 < n$, both positive and negative feedbacks are allowed in the OMDP. For a reference point $\tilde{N} \in \mathbb{R}^n$, we denote $\phi_{\tilde{N}} : \mathbb{R}^n \to \mathbb{R}$ a function defined by:

$$\phi_{\tilde{N}}(N) = \sum_{k=1}^{k_0-1} N_k \sum_{j=k}^{k_0-1} \tilde{N}_j - \sum_{k=k_0+1}^{n} N_k \sum_{j=k_0+1}^{k} \tilde{N}_j \quad (9)$$

Again, by Corollary 3, Axiom A1 to A5 imply that one should use $\phi_{\tilde{N}}$ with a chosen reference point $\tilde{N}$ to compare vectors.

As for histories (or policies) starting from a certain initial state and a given horizon, using $\phi_{\tilde{N}}$ or $\phi'_{\tilde{N}}$ (of the previous subsection) is equivalent. In a probabilistic language, $\phi'_{\tilde{N}}(N)$ is the sum of the probability that N yields a better rewards than $\tilde{N}$ while $\tilde{N}$ yields a positive reward and the probability that N yields a strictly better reward than $\tilde{N}$ while $\tilde{N}$ yields a negative reward.

4 Solving OMDPs

We now show how those ideas can be used to solve an OMDP. To simplify the exposition, we will assume from now on that rewards are counted with a discount factor, i.e. $* = \beta$.

Vectorial Reward MDP

From the previous section, it is easy to see that an OMDP can be viewed as a particular MDP with vectorial rewards (VMDP). The reward function $\hat{r} : S \times A \to \mathbb{R}^n$ of this VMDP is defined from R by $\forall s \in S, \forall a \in A, \hat{r}(s,a) = e_i$ if $R(s,a) = r_i$. This VMDP is a standard multicriteria MDP (Viswanathan, Aggarwal, and Nair 1977) in which a preference order over criteria is defined.

In the same fashion as in standard multicriteria MDPs, in this VMDP, we can recursively define the value function $\hat{v}_h^\pi$ of a policy π at a finite horizon h by $\forall s \in S, \forall t > 0$:

$$\hat{v}_0^\pi(s) = (0, \dots, 0) \in \mathbb{R}^n \quad (10)$$

$$\hat{v}_t^\pi(s) = \hat{r}^{\delta_t}(s) + \beta \sum_{s' \in S} P^{\delta_t}(s,s') \hat{v}_{t-1}^\pi(s') \quad (11)$$

where $\pi = (\delta_h, \dots, \delta_1)$ and the sums and products over vectors are componentwise. Moreover, those value functions are also well defined at the infinite horizon thanks to the discount factor $\beta \in [0, 1[$. They are simply denoted $\hat{v}^\pi$. Finally, one can notice that:

Proposition 1.

$$\forall t > 0, \forall s \in S, \hat{v}_t^\pi(s) = E_s^\pi(N^*(\Gamma))$$

and

$$\forall s \in S, v^\pi(s) = E_s^\pi(N^*(\Gamma))$$

This VMDP is needed if we choose a policy as a reference point. The previous equations compute the vector $\tilde{N}$ of expected numbers of occurrences of each reward for a policy.

As a side note, remark that if finally rewards are numeric and known, it is possible to relate value functions in a standard MDP and a VMDP.

Proposition 2. *We have:*

$$\forall t > 0, \forall s \in S, v_t^\pi(s) = \sum_{i=1}^{n} (\hat{v}_t^\pi(s))_i r_i$$

and

$$\forall s \in S, v^\pi(s) = \sum_{i=1}^{n} (\hat{v}^\pi(s))_i r_i$$

where $(\hat{v}^\pi(s))_i$ is i-th component of $\hat{v}^\pi(s)$.

Building a Reward Function

As seen previously, picking a reference point implicitly defines a reward function. In the proof of Corollary 3, we showed how one can define a reference point from numeric reward values. Here, we show how to define reward values and then the reward function from a reference point.

From a reference point defined by a vector $\tilde{N} \in \mathbb{R}^n$, one can define reward values $u^{\tilde{N}} : E \to \mathbb{R}$ by, $\forall k = 1, \dots, n$:

$$\begin{aligned} u^{\tilde{N}}(r_k) &= 0 && \text{if } k = k_0 \\ &= \textstyle\sum_{j=k}^{k_0-1} \tilde{N}_j && \text{if } k < k_0 \\ &= -\textstyle\sum_{j=k_0+1}^{k} \tilde{N}_j && \text{if } k > k_0 \end{aligned} \quad (12)$$

Finally, a reward function $R^{\tilde{N}}$ can be defined as follows, $\forall s \in S, a \in A$:

$$R^{\tilde{N}}(s,a) = u^{\tilde{N}}(R(s,a))$$

Once a reward function has been defined from a chosen reference point, any standard method for finding an optimal policy can be exploited as searching for optimal policies in the OMDP (S, A, P, R) with a reference point is equivalent to solving the standard MDP $(S, A, P, R^{\tilde{N}})$ by Proposition 2.

5 Examples

In this section, we illustrate some use cases of our approach. In the first example, rewards are qualitative and the reference point is chosen as a history. In the second example, rewards are numeric but difficult to quantify and the reference point is a policy. In the last example, our approach is used to obtain a policy that could be considered better than an optimal policy (in a standard MDP) when the policy has to be applied only once or a few times.

Navigation in a Hostile Environment

Let us first consider the problem of navigation of an autonomous robot in a hostile environment. This problem can be modeled by an OMDP where the states are the different possible positions of the robot, the actions are the directions where it can move to, the transitions are probabilistic as the ground can be slippery, the robot does not control perfectly its motors... The rewards represent qualitative risk levels (safe, mildly dangerous, dangerous, very dangerous). For example, we know that some zones are more dangerous than others from past observations. The robot is in an initial state and wants to reach a goal state. And we would like the robot to avoid dangerous zones as much as possible.

In this problem, it could be easy to find acceptable histories, which are simply paths in this context. In those paths, we could count the numbers of occurrences of the different risk levels to define a reference point.

Production Planning in a Strategy Game

Now, let us assume that we want to build an AI for a non-player character (NPC) in a strategy video game, like Stratagus (Ponsen et al. 2005). In such games, the NPC has to decide in which order to build its units and/or its buildings in order to develop quickly. This problem can be modeled as an OMDP. A state would then be a description of the game board. An action would be the unit and/or the building to build. The transition function would describe the dynamics of the game. In this problem, it is particularly difficult to value numerically the feedback for each action. Using ordinal rewards could ease this task. In addition, in order to define completely the preference system, one can then use as a reference point histories generated by or more generally the policy of a good human player, which can be deduced from the recorded past games.

Doing Better than an Optimal Policy

For this last example, we want to show that our approach can even be exploited when the values of rewards are known. Let us assume that the planning problem that we want to solve has already been modeled by an MDP. In that MDP, an optimal policy π^* can be determined. However, π^* is considered better than other policies only when it is applied a high number of times as it is optimal for the expectation criterion. When we know in advance that we are going to run a policy only once or a few times, a much better approach can be adopted. Using π^* as a reference point in our preference system, we can find a better policy π^{**}. In the probabilistic language, π^{**} would be the policy whose probability of getting rewards better than those of π^* is the greatest. Our preference system may permit to obtain reliable policies.

6 Conclusion

We have extended the model of Markov decision processes (MDP) for taking into account ordinal rewards. We have shown how to define in such a setting expressive and interpretable preferences. This model is useful in the situations where rewards are difficult to assess but also to those where the nature of rewards are really qualitative. Even when rewards can be defined easily and naturally, our approach can be useful for finding good policies when we know in advance that a policy is going to be applied only once or a few times, as illustrated in the last example.

Acknowledgments. Funded by the French National Research Agency under grant ANR-10-BLAN-0215.

References

Bagnell, J.; Ng, A.; and Schneider, J. 2001. Solving uncertain Markov decision processes. Technical report, CMU.

Bouyssou, D.; Marchant, T.; Perny, P.; Pirlot, M.; Tsoukiàs, A.; and Vincke, P. 2000. *Evaluation and decision models: a critical perspective*. Kluwer.

Dubois, D.; Fargier, H.; Lang, J.; Prade, H.; and Sabbadin, R. 1996. Qualitative decision theory and multistage decision making: A possibilistic approach. In *Proc. of the European Workshop on Fuzzy Decision Analysis for Management, Planning and Optimization (EFDAN'96)*.

Givan, R.; Leach, S.; and Dean, T. 2000. Bounded-parameter Markov decision process. *Artif. Intell.* 122(1-2):71–109.

Jeantet, G., and Spanjaard, O. 2009. Optimizing the hurwicz criterion in decision trees with imprecise probabilities. In *1st Int. Conf. on Algorithmic Decision Theory*, Lecture Notes in Artificial Intelligence, 340–352.

Keeney, R., and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley.

Krantz, D.; Luce, R.; Suppes, P.; and Tversky, A. 1971. *Foundations of measurement*, volume Additive and Polynomial Representations. Academic Press.

Nilim, A., and El Ghaoui, L. 2003. Robustness in Markov decision problems with uncertain transition matrices. In *NIPS*.

Perny, P., and Pomerol, J. 1999. Use of artificial intelligence in MCDM. In Gal, T.; Stewart, T.; and Hanne, T., eds., *Multicriteria Decision Making Advances in MCDM Models, Algorithms Theory, and Applications*. Kluwer Academic. 15:1–15:43.

Ponsen, M.; Lee-Urban, S.; Muñoz-Avila, H.; Aha, D.; and Molineaux, M. 2005. Stratagus: An open-source game engine for research in real-time strategy games. In *Workshop IJCAI*.

Puterman, M. 1994. *Markov decision processes: discrete stochastic dynamic programming*. Wiley.

Regan, K., and Boutilier, C. 2009. Regret-based reward elicitation for Markov decision processes. In *UAI*.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence: A Modern Approach*. Prentice-Hall, 2nd edition.

Sabbadin, R. 1999. A possibilistic model for qualitative sequential decision problems under uncertainty in partially observable environments. In *UAI*, volume 15, 567–574.

Shaked, M., and Shanthikumar, J. 1994. *Stochastic Orders and Their Applications (Probability and Mathematical Statistics)*. Academic press.

Trevizan, F.; Cozman, F.; and de Barros, L. 2007. Planning under risk and knightian uncertainty. In *IJCAI*, 2023–2028.

Viswanathan, B.; Aggarwal, V.; and Nair, K. 1977. Multiple criteria Markov decision processes. *TIMS Studies in the Management Sciences* 6:263–272.