Ordinal Decision Models for Markov Decision Processes

Paul Weng[1]

Abstract. Setting the values of rewards in Markov decision processes (MDP) may be a difficult task. In this paper, we consider two ordinal decision models for MDPs where only an order is known over rewards. The first one, which has been proposed recently in MDPs [23], defines preferences with respect to a reference point. The second model, which can been viewed as the dual approach of the first one, is based on quantiles. Based on the first decision model, we give a new interpretation of rewards in standard MDPs, which sheds some interesting light on the preference system used in standard MDPs. The second model based on quantile optimization is a new approach in MDPs with ordinal rewards. Although quantile-based optimality is state-dependent, we prove that an optimal stationary deterministic policy exists for a given initial state. Finally, we propose solution methods based on linear programming for optimizing quantiles.

1 INTRODUCTION

Planning under uncertainty is an important task in Artificial Intelligence (AI) [19]. Such problems can be modeled as Markov decision processes (MDP) [16]. In standard MDPs, uncertainty is described by probabilities and preferences are represented by numeric rewards. MDPs have proved to be very powerful to solve many different planning problems. However, in some real-life problems, setting the numeric parameters (probabilities and rewards) may be a difficult task. As optimal solutions could be impacted by even slight variations of the parameter values, it should not be considered lightly.

This observation has motivated much work on uncertain parameter MDPs and on robust MDPs. In [12], parameters are represented by intervals and an interval version of value iteration is proposed. In [1, 15, 5], the case where only probabilities are not accurately known is considered and solving methods are proposed for searching for robust solutions, i.e. optimizing the worst case. More generally, a unifying extension of MDPs allowing for different kinds of uncertainty have been proposed by [21]. Recently, the dual case where only rewards are partially known has been studied. The approach developed in [17, 24] is based on the minimax criterion. In that setting, finding an optimal stationary deterministic policy is NP-hard, which makes this approach difficult to put into practice for large size problems for the moment. Assuming that rewards can only be ranked, [23] proposed a decision model based on reference points.

Following [23], we also assume that uncertainty is probabilistic and rewards are ordinal. While probabilities can be estimated (experiments, observations...), setting numeric rewards may be difficult. This is especially the case when rewards do not represent some physical measure (e.g., money, length, weight, duration...) to be optimized. The problem of assessing rewards, called preference or utility elicitation in decision theory, is known to be a troublesome task [3]. Besides, the assumption that only ordinal information is known about rewards follows a long tradition in AI [4, 11, 10, 9, 2] where qualitative preference models and representations have been investigated. Our work could allow the integration of those qualitative preference representations (e.g. CP-nets) with MDPs.

In the framework of MDPs with ordinal rewards, we consider two preference systems that could be seen as dual one to the other: the first based on reference points and the second based on quantiles. Thanks to the first, we give a new interpretation of the preference system of standard MDPs. The second based on quantiles is a new preference model in MDPs with ordinal rewards. It enjoys some nice properties. For instance, no assumption is made about the commensurability between preferences and uncertainty. Here, it allows for completely ordinal rewards with probabilistic uncertainty. The preferences it defines are more robust than those in standard MDPs, in the sense that slight variations away from the quantile does not impact optimal solutions. As a consequence, quantiles do not depend on extreme values. [8] also proposed to optimize quantiles in MDPs. However, our works differ in so far as quantiles in [8] are defined over distributions of cumulated rewards while we define quantiles over distributions of ordinal rewards. Thus, contrarily to their approach, we do not need to assume the quantitative nature of rewards.

The paper is structured as follows. In Section 2, we recall some needed definitions and give some motivation to our work. In Section 3, we present MDPs with ordinal rewards and how one can interpret the value of a policy in a state as a distribution over ordinal rewards. In Section 4, we present two decision models that can be exploited in this ordinal setting. Finally, we present in Section 5 a solving method for quantile maximization.

2 BACKGROUND

We now recall the model of *Markov decision processes* (MDP). It is defined as a quadruplet (S, A, T, r) [16] where S is a finite set of states, A is a finite set of actions, $T\colon S\times A\times S \to [0,1]$ is a transition function and $r\colon S \times A \to X \subset \mathbb{R}$ is a reward function. The *transition function* gives the probability that a future state occurs after performing an action in a state, i.e., for all s, a, $\sum_{s'\in S} T(s,a,s') = 1$. The *reward function* gives the immediate reward received 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. However, 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 $\mathcal{P}(A)$.

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 to be *deterministic* when it only contains deterministic decision rules and *randomized* otherwise. At the infinite horizon, a policy is simply

[1] LIP6, UPMC, France, email: paul.weng@lip6.fr

an infinite sequence of decision rules. A policy is said *stationary* if at each decision step, the same decision rule is applied.

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. A *history* γ starting from state $s_0 \in S$ corresponds to a sequence $\gamma = (s_0, a_1, s_1, a_2, s_2, \ldots)$ where $\forall i \in \mathbb{N}, (a_i, s_i) \in A \times S$. The value of history γ can be defined by:

$$r^{\beta}(\gamma) = \sum_{i=0}^{\infty} \beta^i r(s_i, a_{i+1}) \quad (1)$$

where $\beta \in [0, 1[$ is a discount factor. 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 $T(s, \delta(s), .)$ in initial state s. By induction, a policy π in a given initial state s can be associated to a probability distribution over histories. Hence, a policy also induces a probability distribution over the values of histories. Consequently, it is possible to define the expected cumulated reward that a policy can yield from an initial state. The function $v^{\pi} : S \to \mathbb{R}$, which associates to each state s the expected reward that can be obtained by the policy π is called the *value function* of π:

$$v^{\pi}(s) = E_s^{\pi}(r^*(\Gamma)) \quad (2)$$

where E_s^{π} is the expectation with respect to the probability distribution induced by the application of π from state s and Γ is a random variable over histories. Then, a policy π is *preferred* to a policy π':

- from an initial state s if $\pi \succsim_s \pi' \Leftrightarrow v^{\pi}(s) \geq v^{\pi'}(s)$
- from any initial state if $\pi \succsim \pi' \Leftrightarrow \forall s \in S, \pi \succsim_s \pi'$

In the standard framework, the preference relation defined with Eq. 2 guarantees that an optimal stationary deterministic policy exists. In this paper, for a preference relation $\succsim$, we will denote $\succ$ (resp. $\sim$) its asymmetric (resp. symmetric) part.

Solving Method. There are three main approaches for solving MDPs at the infinite horizon. Two are based on dynamic programming: value iteration and policy iteration. The third is based on linear programming. We recall the last approach as it is needed for the exposition of our solving methods. The linear program $(\mathcal{P})$ for solving MDPs can be written as follows:

$$\begin{aligned} \min \quad & \sum_{s \in S} \mu(s) v(s) \\ \text{s.t.} \quad & v(s) - \beta \sum_{s' \in S} T(s, a, s') v(s') \geq r(s, a) \quad \forall s, \forall a \end{aligned}$$

where weights μ can be interpreted as the probability of starting in a given state. The dual $(\mathcal{D})$ of this program has a nice property as it separates preferences, which are expressed in the objective function, and the dynamics of the system, which is expressed in the constraints:

$$\left.\begin{aligned} \max \quad & \sum_{s \in S} \sum_{a \in A} r(s, a) x_{sa} \\ \text{s.t.} \quad & \sum_{a \in A} x_{sa} - \beta \sum_{s' \in S} \sum_{a \in A} x_{s'a} T(s', a, s) = \mu(s) \quad \forall s \\ & x_{sa} \geq 0 \quad \forall s, \forall a \end{aligned}\right\} (\mathcal{C})$$

To interpret the x_{sa}'s, we recall two propositions relating feasible solutions of $(\mathcal{D})$ to stationary randomized policies in the MDP [16].

Proposition 1. *For a policy π, if x^{π} is defined by $\forall s, \forall a, x_{sa}^{\pi} = \sum_{t=0}^{\infty} \beta^t p_t^{\pi}(s, a)$ where $p_t^{\pi}(s, a)$ is the probability of reaching state s and choosing a at step t, then x^{π} is a feasible solution of $(\mathcal{D})$.*

Proposition 2. *If x_{sa} are a solution of $\mathcal{D}$, then the stationary randomized policy δ^{∞}, defined by $\forall s, \forall a, \delta(s, a) = x_{sa} / \sum_{a' \in A} x_{sa'}$ defines $x_{sa}^{\delta^{\infty}}$ (as in Proposition 1), that are equal to x_{sa}.*

Hence, the set of randomized policies is completely characterized by constraints $(\mathcal{C})$. Besides, the basic solutions of $(\mathcal{D})$ correspond to deterministic policies. Randomized policies are in the convex hull of those basic solutions. Note that in an MDP, any feasible value function can be obtained with a stationary randomized policy.

Motivations of this work. We now present some useful results concerning rewards and the preference relation they induce over policies. A strictly increasing affine transformation of the reward function does not change the preferences over policies.

Lemma 1. *For $\lambda > 0$ and $c \in \mathbb{R}$, the preferences over policies defined by Eq. 2 in (S, A, T, r) and $(S, A, T, \lambda r + c)$ are identical.*

Proof. Assume first $c = 0$. By Eq. 1 and 2, the value function of a policy π in (S, A, T, br) is equal to bv^{π} where v^{π} is the value function of π in (S, A, T, r). Then, the result obviously holds in this case. Consider now $c \neq 0$. We can assume $\lambda = 1$. When comparing two policies in a state, the histories in the expectations are all of the same length (see Lemma 2). Therefore, adding c to the rewards would affect all the value functions by the same constant. □

As a side note, the result does not hold anymore for an MDP containing an absorbing state where no action is taken.

[23] showed that in an MDP where there is only one non null positive (resp. negative) reward, any positive (resp. negative) would do. Using our previous lemma, we can prove a slightly stronger result.

Corollary 1. *If a reward function can take $n(\geq 2)$ values $r_1 > \cdots > r_n$, r_1 and r_n can be arbitrarily set to 1 and 0.*

The case where $n = 2$ implies there is no need to elicitate rewards as any (order preserving) reward values would yield the same preference relation over policies. However, in a problem where one needs more than three different rewards, their values must be set carefully as they may have an undesirable impact over the preferences over policies. In such cases, when reward values are not precisely known, it is questionable to use arbitrary values and apply directly the theory developed for standard MDPs. We propose instead to start with ordinal rewards – this information is generally available – and build a preference system more suitable to this qualitative setting.

3 MDP WITH ORDINAL REWARDS

Let $E = \{r^1 < r^2 < \ldots < r^n\}$ be a qualitative completely ordered set. An *MDP with ordinal rewards* (ORMDP) is an MDP (S, A, T, r) where the reward function $r\colon S \times A \to E$ is defined to take its values on the scale $(E, <)$. As in standard MDPs, a history $(s_0, a_1, s_1, a_2, s_2, \ldots)$ in an ORMDP is associated to a sequence of rewards $(r(s_0, a_1), r(s_1, a_2), \ldots)$. Due to the ordinal nature of the rewards, one cannot add them. A natural way to value a history is then to count the number of rewards obtained in a history. For the finite horizon case, a history $\gamma = (s_0, a_1, s_1, a_2, s_2, \ldots, a_h, s_h)$ can be valued by:

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

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

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

where $\chi_{r(s_i,a_{i+1})=r^k} = 1$ if $r(s_i, a_{i+1}) = r^k$ and 0 otherwise.

More generally, one can introduce a discount factor $\beta \in [0, 1[$ with a similar semantic as in standard MDPs, meaning that one reward r obtained h steps from now is worth β^{h-1} of r now. In this case, the value of a history γ can be defined by:

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

where for $k = 1, \ldots, n, N_k^\beta(\gamma)$ 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 is defined by:

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

N^β can be extended to the infinite horizon as factor β guarantees the convergence of the sums.

Once the values of histories are defined, it is then natural to value policies in a state as the expectation of those values. Like in standard MDPs, we define the value function $\hat{v}^\pi : S \to \mathbb{R}_+^n$ of π:

$$\hat{v}^\pi(s) = (\hat{v}_1^\pi(s), \ldots, \hat{v}_n^\pi) = (E_s^\pi(N_1^\beta(\Gamma)), \ldots, E_s^\pi(N_n^\beta(\Gamma))) \quad (3)$$

where E_s^π is the expectation with respect to the probability distribution induced by the application of π from state s and Γ is a random variable over histories. Then, $\hat{v}_i^\pi(s)$ is the expected number of reward r^i's obtained when applying policy π from initial state s.

As one would like to compare policies via their vectorial value functions, we need to define a preference relation over vectors. By abuse of notation, $\succsim$ will also denote this preference relation as there is no risk of confusion. In order to make explicit the assumptions that are made during the construction of the preference system in an ORMDP, we introduce two requirements that the preference relation over policies and the preference relation over vectors have to satisfy.

H1. For any two policies π, π', any s, $\pi \succsim_s \pi' \Leftrightarrow \hat{v}^\pi(s) \succsim \hat{v}^{\pi'}(s)$
H2. For any two vectors $v, v' \in \mathbb{R}_+^n$, any λ, $v \succsim v' \Leftrightarrow \lambda v \succsim \lambda v'$

Assumption H1 states that the preferential information of a policy in a state is completely given by the expectation of the values of its histories, which themselves are defined by counting the ordinal rewards. H2 implies that preferences over vectors should be homothetic. They both seem natural and we will assume both of them in the remainder of the paper. As a side note, they are both naturally assumed in standard MDPs. Indeed, if E were a numerical scale, we have:

Proposition 3. *For any policy π, $v^\pi(s) = \sum_{i=1}^n \hat{v}_i^\pi(s) r^i$.*

In a state s, the vectorial value function $\hat{v}^\pi(s)$ can be viewed as describing the composition of a population of ordinal rewards where each reward r^i appears $\hat{v}_i^\pi(s)$ times. Based on this observation, one could import some tools used in descriptive statistics [14]. For a vector N representing the composition of a population of rewards, the *distribution* f^N associated to N is defined by $f^N = (f_1^N, f_2^N, \ldots, f_n^N)$, i.e., $f_i^N = N_i / \sum_{i=1}^n N_i$ is the frequency or the proportion of r^i in the population described by N. The *cumulative distribution* of N is defined by $F^N = (F_1^N, \ldots, F_n^N)$ where $F_i^N = \sum_{j=1}^i f_j^N$ being the frequency of elements lower or equal to r^i. The *decumulative distribution* of N is defined by $G^N = (G_1^N, \ldots, G_n^N)$ where $G_i^N = \sum_{j=i}^n f_j^N$ being the frequency of elements greater or equal to r^i. In the remainder of the paper, we identify vector N and the population it describes. Besides, for convenience, we introduce dummy components $F_0^N = G_0^N$.

The following lemma shows that the sums of vectors considered in ORMDPs are all equal for a fixed horizon and discount factor β.

Lemma 2. *For any history γ, any policy π, for any state $s \in S$:*

$$\sum_{i=1}^n N_i^\beta(\gamma) = \begin{cases} \frac{1-\beta^h}{1-\beta} & \text{for any finite horizon } h \\ \frac{1}{1-\beta} & \text{for the infinite horizon} \end{cases}$$

$$\sum_{i=1}^n \hat{v}_i^\pi(s) = \begin{cases} \frac{1-\beta^h}{1-\beta} & \text{for any finite horizon } h \\ \frac{1}{1-\beta} & \text{for the infinite horizon} \end{cases}$$

Proof. At each step along a history, a reward is obtained. For the finite case, $\sum_{i=1}^n N_i^\beta(\gamma) = 1 + \beta + \ldots + \beta^{h-1} = \frac{1-\beta^h}{1-\beta}$. The infinite case is obtained by taking the limit. For value functions, the expectation is computed over histories having the same length. □

Under assumptions H1 and H2, Lemma 2 implies that comparing policies in a state is equivalent to comparing the associated distributions as the sums of the vector components are equal. Therefore, defining how to compare policies in a state boils down to defining how to compare the associated distributions.

When comparing distributions, one requires the following natural dominance property : $\forall N, N' \in \mathbb{R}_+^n$,

D. For all $i = 1, \ldots, n, G_i^N \geq G_i^{N'} \Rightarrow N \succsim N'$

Property D states that if for any reward r^i, the proportion of rewards equal or better than r^i is greater in N than in N' then N should be preferred. Note that when considering probability distributions, it is simply the *first-order stochastic dominance*. This property alone does not yield an exploitable preference system as it defines a partial order. In the case of MDPs, one would obtain too many non-dominated policies. In order to define a more suitable preference system, it is then natural to consider preference relations that refines requirement D. We present in the next section two ordinal decision models. In that section, for the sake of simplicity, preferences are written with distributions instead of vectors as it is equivalent under H1 and H2.

4 ORDINAL DECISION MODELS

4.1 Reference Point-Based Preferences

In order to compare vectors, one can introduce a reference point denoted $\tilde{N} \in \mathbb{R}_+^n$ and compare two vectors $N, N' \in \mathbb{R}_+^n$ by:

$$f^N \succsim f^{N'} \Leftrightarrow \phi_{\tilde{N}}(N) \geq \phi_{\tilde{N}}(N') \text{ with } \phi_{\tilde{N}}(N) = \sum_{i=1}^n f_i^N \sum_{j=1}^i f_j^{\tilde{N}}$$

$\phi_{\tilde{N}}(N)$ can be interpreted as the proportion of times a drawing from population N yields a better result than an independent drawing in $\tilde{N}$. In a probabilistic language, an informal intuitive interpretation would be: vector N is preferred to N' if the probability of N getting a better value than $\tilde{N}$ is greater than that with N'. Such a criterion has been proposed by [6] for decision making under risk. They showed that it is formally identical to expected utility [3]. It has also

been studied by [23] in ORMDPs. As underlined by [23], formally such an approach boils down to standard MDPs as choosing a reference point $\tilde{N}$ induces a numeric reward ($r^i = F_i^{\tilde{N}}$). However, such a preference system is indeed ordinal in so far as one does not set arbitrarily numeric values for rewards, but the numeric values comes from a carefully chosen reference point. In situations of partial and imperfect knowledge, the choice of a reference point allows for a certain justification and a natural interpretation, which is not the case for a direct arbitrary choice of numeric values for rewards. For instance, the reference point could be chosen as one of the ordinal reward or the vectorial value induced by a history or a policy [23].

In the reference point-based preference system, the interpretation is simple using the language of probability distributions: one wants to maximize the probability of getting better rewards than a reference point. With this semantic, we can propose for two simple cases a nice interpretation of the preferences induced by numeric rewards in standard MDPs. When the reward function in a standard MDP can only take two different values, the reference-point preference tells us that the implicit reference point used in that case is a Bernoulli distribution (by Corollary 1) and the optimal policies are those that maximize the probability of doing better than a Bernoulli random variable over the rewards.

For the class of standard MDPs whose reward function can take many different values but is regularly-spaced, i.e. if $r^1 < \ldots < r^n$ are the different possible reward values, then there exists a constant C such that $\forall i, r^{i+1} = r_i + C$, the reference point-based preference tells us that one implicitly chooses a uniform probability distribution over rewards as a reference point. Therefore, the interpretation is that in such MDPs, optimal policies are those that maximize the probability of doing better than a uniform random variable. Thus, this preference system gives a new and interesting understanding of the preferences used in standard MDPs. However, the limit of this approach is that it moves the difficulty of setting the numeric reward values to that of picking a reference point. As the choice of a reference point may not be obvious at all in some problems, we propose in the next section a new approach in ORMDPs.

4.2 Quantile Optimization

In descriptive statistics [14], three values can be considered to characterize the central tendency of a population N: the *mean*, the *median* and the *mode*. The mean, which is defined only if the elements of the population are numbers, is the average of those elements. In our setting, the mean can be written as $\sum_{i=1}^{n} f_i^N r^i$. One can then recognize the approach taken in standard MDPs. The median of N, which is a special case of quantiles, is the value $\bar{m}(N)$ such that half of population N is lower than $\bar{m}(N)$ and the other half is greater than $\bar{m}(N)$. We present below the quantile-based approach for decision-making. Finally, the mode of N is the (possibly not unique) element that occurs the most frequently in N. Unfortunately, it cannot be used as a rational decision criterion because it does not satisfy requirement D.

Example 1. *Assume $E = \{r^1 < r^2 < r^3\}$. Let N, N' be two vectors with the following associated distributions $f^N = (0.4, 0.6, 0)$ and $f^{N'} = (0.39, 0.3, 0.31)$. Then, the mode of N is r^2 and that of N' is r^1. The mode as a decision criterion indicates that N should be strictly preferred. However, $\forall i = 1, \ldots, n, G_i^{N'} \geq G_i^N$.*

Intuitively, the τ^{th} *quantile* of population N for $\tau \in [0, 1]$ is the value r such that τ percent of N is equal or lower than r and $1 - \tau$ percent of N is equal or greater than r. Formally, in our framework, it is defined as follows. First, we define the *lower* τ^{th} *quantile* of N for $\tau \in]0, 1]$ by:

$$Q_-^\tau(N) = r^i \text{ such that } F_{i-1}^N < \tau \text{ and } F_i^N \geq \tau$$

Then, we define the *upper* τ^{th} *quantile* of N for $\tau \in [0, 1[$ by:

$$Q_+^\tau(N) = r^i \text{ such that } G_i^N \geq 1 - \tau \text{ and } G_{i+1}^N < 1 - \tau$$

When only one of $Q_-^\tau(N)$ or $Q_+^\tau(N)$ is defined (i.e., $\tau = 0$ or $\tau = 1$), then we define the τ^{th} quantile $Q^\tau(N)$ as that value. When both are defined, by construction, we have $Q_+^\tau(N) \geq Q_-^\tau$. Note that they are generally equal and $Q^\tau(N)$ is defined as equal to them. For instance, this is the case in continuous settings ($E = \mathbb{R}$) for continuous distributions. However, in our discrete setting, although rare, it could happen that those values differ, as shown by:

Example 2. *Let $E = \{r^1 < r^2 \ldots < r^6\}$. Let N define distribution $f^N = (0, 0.1, 0.4, 0, 0.3, 0.2)$. Then, $Q^0(N) = r^2$ gives the minimum of N, $Q^1(N) = r^6$ gives the maximum of N, $Q^{0.75}(N) = r^5$ gives the third quartile of N. Here, the median needs to be properly defined as the upper median $Q_+^{0.5}(N) = r^5$ is strictly greater than the lower median $Q_-^{0.5}(N) = r^3$.*

In such cases, we explain how Q^τ can be defined. When computing the lower and the upper quantiles of a vector N, one gets a couple $(Q_+^\tau(N), Q_-^\tau(N)) \in E_2$ where $E_2 = \{(r_+, r_-) \in E \times E : r_+ \geq r_-\}$. A partially order $\geq$ is defined over E_2: $(r_+, r_-) \geq (r'_+, r'_-)$ iff $r_+ \geq r'_+$ and $r_- \geq r'_-$. Poset $(E_2, \geq)$ is in fact a lattice. Assume we have an aggregation function $\phi : E_2 \to E$ satisfying the following two properties: For all $(r_+, r_-), (r'_+, r'_-) \in E_2$,

P1. $r_+ \geq \phi(r_+, r_-) \geq r_-$
P2. $\phi(r_+, r_-) \geq \phi(r'_+, r'_-)$ iff $r_+ \geq r'_+$ and $r_- \geq r'_-$

We define the τ^{th} *quantile* by $Q^\tau(N) = \phi(Q_+^\tau(N), Q_-^\tau(N))$. P1 and P2 are two natural properties one wants an aggregation function to satisfy. P1 guarantees $\phi(r, r) = r$ and P2 is a monotony property. In our ordinal setting, function ϕ needs to be provided by the decision-maker. It could model in a decision-theoretic sense, her optimistic or pessimistic attitude by taking $\phi = \max$, $\phi = \min$ or some other values in between. As a side note, ϕ could be defined to take its values in a finer scale than E allowing for a finer preference representation.

Quantiles generalizes different known criteria. The $\min$ of the rewards in N is given by $Q^0(N)$. Symmetrically, the $\max$ of the rewards in N is given by $Q^1(N)$. The median of N is given by $Q^{0.5}(N)$ (with $\phi(r_+, r_-) = (r_+ + r_-)/2$ when $E \subset \mathbb{R}$). Besides, the well-known risk measure, *Value-at-Risk* [20], used notably in finance is a quantile.

Interestingly, in decision theory, quantiles as a criterion have been studied and axiomatized recently [7, 18] in quantitative settings. They mainly satisfy two properties: first-order stochastic dominance (property D in our setting) and ordinal invariance (i.e., preferences only depends on the order over rewards). This makes quantiles particularly suitable in our setting. In this paper, we extend their use to sequential decision-making under uncertainty. The value of a policy π in a state s can be defined as a quantile of $\hat{v}^\pi(s)$ viewed as a population of ordinal rewards. Then, policies can be compared in a state s via their associated quantiles:

$$\pi \succsim_s \pi' \Leftrightarrow Q^\tau(\hat{v}^\pi(s)) \geq Q^\tau(\hat{v}^\pi(s'))$$

In such a preference system, the decision-maker needs to specify the value τ. A natural value for τ would be 0.5, which implies that policies are compared via the median. This draws a nice parallel with

standard MDPs where the mean is used. Naturally, other values for τ would be possible, depending on the preferences of the decision-maker. For instance, setting $\tau = 0.05$ amounts to finding policies that maximizes the lowest reward achieved 95% of the times. This would lead to a less extreme approach than that in robustness which focuses on the worst case.

Note that optimizing a quantile can be viewed as the dual approach to the reference point-based approach. Indeed, in the latter, one sets a reference point and maximizes the probability of beating this reference point. In the former, one sets a probability τ and maximizes the value v that ensures getting at least v.

Preference relations induced by quantiles satisfy property D. This is known for lower and upper quantiles, this is still true in our setting for any ϕ.

Proposition 4. $\forall i = 1, \ldots, n, G_i^N \geq G_i^{N'} \Rightarrow Q^\tau(N) \geq Q^\tau(N')$

Proof. Let i be the greatest index such that $G_i^N \geq 1 - \tau$. By assumption, $1 - \tau > G_{i+1}^N \geq G_{i+1}^{N'}$. Then, $Q_+^\tau(N) \geq Q_+^\tau(N')$.

Besides, note that $G_i^N \geq G_i^{N'}$ implies $F_i^N \leq F_i^{N'}$ for all i. Let j be the lowest index such that $F_j^{N'} \geq \tau$. Then, $\tau > F_{j-1}^{N'} \geq F_{j-1}^N$ and $Q_-^\tau(N') \leq Q_-^\tau(N)$. Finally, by P2, $Q^\tau(N) \geq Q^\tau(N')$ □

As property D is satisfied, quantiles can be considered a good candidate for comparing policies in a state. However, to be able to exploit this decision criterion in ORMDPs at the infinite horizon, there should exist an optimal (w.r.t. quantiles) stationary deterministic policy. Interestingly, this is the case as shown in Theorem 1. In order to prove this result, we first state a lemma showing an interesting property of the quantiles of the linear convex combination of vectors.

Lemma 3. *For any $\tau \in [0, 1]$, for any $\lambda \in [0, 1]$, we have:*

$$Q^\tau(N) \vee Q^\tau(N') \geq Q^\tau(\lambda N + (1-\lambda)N') \geq Q^\tau(N) \wedge Q^\tau(N')$$

where $\vee$ and $\wedge$ are respectively max *and* min *over E.*

Proof. We prove for $\tau \in]0, 1[$. For $\tau = 0$ or $\tau = 1$, the proof is similar. Assume $Q^\tau(N) \geq Q^\tau(N')$ (the other case is symmetric) and let $Q_-^\tau(N) = r^i$, $Q_+^\tau(N) = r^j$, $Q_-^\tau(N) = r^{i'}$ and $Q^\tau(N') = r^{j'}$ with $i \leq j$ and $i' \leq j'$. Denote $N'' = \lambda N + (1-\lambda)N'$ for any $\lambda \in]0, 1[$ (for $\lambda = 0$ or $\lambda = 1$, the result is obvious). Let $Q_-^\tau(N) = r^{i''}$ and $Q_+^\tau(N) = r^{j''}$ with $i'' \leq j''$. Then, by definition of $Q^\tau(N'')$ and $Q^\tau(N'')$, we have $\max(i, i') \geq i'' \geq \min(i, i')$ and $\max(j, j') \geq j'' \geq \min(j, j')$. By assumption, we cannot have $i \leq i'$ and $j \leq j'$ with one or both of the inequalities strict. We consider the other cases. If $i \geq j'$, then $i \geq i'' \geq i'$ and $j \geq j'' \geq j'$ and the inequalities of the lemma are true by monotony of ϕ.

Note that $Q_-^\tau(N) = r^i \neq Q_+^\tau(N) = r^j$ means that $F_i^N = \tau$, $G_j^N = 1-\tau$ and $f_k^N = 0, \forall k = j+1, \ldots, i-1$. If $i \geq i'$ and $j \geq j'$, $Q_+^\tau(N'') = Q_+^\tau(N')$ and $Q_-^\tau(N'') = Q_-^\tau(N)$. By monotony of ϕ, $Q^\tau(N) \geq Q^\tau(N'') \geq Q^\tau(N')$. If $i < i'$ and $j > j'$, $Q^\tau(N'') = Q^\tau(N')$. If $i > i'$ and $j < j'$, $Q^\tau(N'') = Q^\tau(N)$. In both case, the result holds. □

We can now state the following theorem that shows quantiles can be used in ORMDPs at the infinite horizon.

Theorem 1. *For an initial state s, there exists a stationary deterministic policy π such that:*

$$Q^\tau(\hat{v}^\pi(s)) = \max_{\pi'} Q^\tau(\hat{v}^{\pi'}(s))$$

Proof. We give an outline of the proof. First, we only need to consider stationary policies because for any policy π, there is a stationary policy π' such that $\hat{v}^\pi = \hat{v}^{\pi'}$ (The proof is similar to that for standard MDPs). By Proposition 1, one can identify a stationary policy π to the vector $x^\pi \in \mathbb{R}^{S \times A}$. Then, comparing stationary policies is equivalent to comparing those vectors. It is well-known that the space of the vectors representing stationary policies is a convex set. By abuse of notation, $\succsim_s$ also denotes the preference relation over those vectors. For such a vector x, we denote π_x its associated stationary policy (given by Proposition 2) and $Q_x^\tau = Q^\tau(\hat{v}^{\pi_x}(s))$.

We prove that $\succsim_s$ is concave, i.e., for any vectors x, y, z and any $\lambda \in [0, 1]$, $z \succsim_s x$ and $z \succsim_s y \Rightarrow z \succsim_s \lambda x + (1-\lambda)y$. Assume $z \succsim_s x$ and $z \succsim_s y$, i.e., $Q_z^\tau \geq Q_x^\tau$ and $Q_z^\tau \geq Q_y^\tau$. By Lemma 3, we have $Q_x^\tau \vee Q_y^\tau \geq Q_{\lambda x+(1-\lambda)y}^\tau$, which implies $\succsim_s$ is concave.

Now, consider a stationary randomized policy π. Its associated vector x^π can be expressed as the linear convex combination of vectors $x_1, \ldots x_k$ representing stationary deterministic policies. As $\succsim_s$ is concave, π is dominated with respect to the τ^{th} quantile by a stationary deterministic policy. □

The previous theorem justifies the use of quantiles in ORMDPs. However, one needs to be careful when using this criterion. While in standard MDPs, there is an optimal policy in every state, the quantile-based optimality is state-dependent. We illustrate this point on a small example.

Example 3. *Assume that $E = \{r^1 < r^2 < r^3\}$. Consider three vectors N, N' and N'' with their associated distributions $f = (0.48, 0, 0.52)$, $f' = (0.38, 0.62, 0)$ and $f'' = (0.6, 0.4, 0)$. We have $\bar{m}(N) = r^3 > \bar{m}(N') = r^2$. Take $\lambda = 0.5$. Then, the associated distribution of $\lambda N + (1-\lambda)N''$ is $(0.54, 0.2, 0.26)$ with a median of r^1 and that of $\lambda N' + (1-\lambda)N''$ is $(0.49, 0.51, 0)$ with a median of r^2. Thus, we have an inversion of preferences: $N \succ N'$ and $\lambda N' + (1-\lambda)N'' \succ \lambda N + (1-\lambda)N''$.*

In an ORMDP, assume that there are two policies π and π' whose value functions yield those vectors N and N' in a state s, i.e., $\pi \succ_s \pi'$. Now, from a state s_0, there is an action a that leads to state s with probability λ and to another state s' with probability $1 - \lambda$. In state s', there is a policy whose value function yields N''. Then, by choosing action a, π' would be preferred to π viewed from s_0.

5 SOLVING METHODS

We now present how optimal (w.r.t. the τ^{th} quantile) stationary deterministic policies can be computed. In this section, the initial state is assumed to be s_0, i.e., $\mu(s_0) = 1$ and $\mu(s) = 0$ for $s \neq s_0$.

First, note that Eq. 3 can be computed as the value function of a policy in a *vector-reward MDP* (VMDP) [22], i.e., an MDP $(S, A, T, \hat{r})$ where $\hat{r}(s, a)$ is a vector in $\mathbb{R}^n$. For our purpose, $\hat{r}(s, a)$ is defined from $r(s, a)$ of the ORMDP as the vector whose i-th component is equal to 1 if $r(s, a) = r^i$ and null on the other components. It is then obvious that summing vectorial rewards along a history in this VMDP amounts to computing N^β.

A first method relies on the following linear program $(\mathcal{D}_-^j)$ for a given $j = 1, \ldots, n$:

$$\left.\begin{array}{ll} \min & \displaystyle\sum_{i=1}^{j} \sum_{s \in S} \sum_{a \in A} \hat{r}_i(s, a) x_{sa} \\ \text{s.t.} & \displaystyle\sum_{a \in A} x_{sa} - \beta \sum_{s' \in S} \sum_{a \in A} x_{s'a} T(s', a, s) = \mu(s) \quad \forall s \\ & x_{sa} \geq 0 \quad \forall s, \forall a \end{array}\right\} (\mathcal{C})$$

$(\mathcal{D}^j_-)$ amounts to optimizing the cumulative distribution. It has the same number of variables and constraints as program $(\mathcal{D})$. Its solution yields a deterministic policy that minimizes the number of rewards equal or worse than r^j. Then, one can solve sequentially $(\mathcal{D}^1_-)$, $(\mathcal{D}^2_-)$, $\dots(\mathcal{D}^k_-)$ until finding the first k such that the value v of objective function is greater or equal to $\tau/(1-\beta)$. Recall that the sum of the components of $v^\pi(s_0)$ is equal to $1/(1-\beta)$. This procedure optimizes the lower τ^{th} quantile, and thus the case $\phi = \min$. In the rare cases where the value of the objective function is exactly equal to $\tau/(1-\beta)$, one needs to check whether the upper quantile is strictly greater than the lower one. In such a case, one needs to use the information about ϕ and solve $(\mathcal{D}^j_-)$ for $j = k+1, \dots$ for finding the optimal solution.

Instead of optimizing the cumulative distribution, one could also symmetrically optimize the decumulative distribution with a linear program $(\mathcal{D}^j_+)$ obtained by replacing the objective of $(\mathcal{D}^j_-)$ by $\max \sum_{i=j}^n \sum_{s\in S} \sum_{a\in A} \hat{r}_i(s,a) x_{sa}$. But, a more direct approach exists exploiting the following fact. In quantitative settings, it is well-known that the τ^{th} quantile of a vector N can be found by minimizing an absolute loss function [13]:

$$Q^\tau(N) = \operatorname*{argmin}_{r\in E} \sum_{i=1}^n \rho_\tau(r^i - r) N_i \tag{4}$$

where $\rho_\tau(x) = (\tau-1)x$ if $x \leq 0$ and $\rho_\tau(x) = \tau x$ otherwise. Function ρ_τ is somewhat an absolute value with different slopes for negative and positive values. In ordinal settings, one can set arbitrary values for the ordinal rewards as long as the order is respected. Then, we can assume $E \subset \mathbb{R}$ and solve the following linear program:

$$\begin{array}{ll} \max & r \\ \text{s.t.} & N_i = \sum_{s\in S} \sum_{a\in A} \hat{r}_i(s,a) x_{sa} \quad \forall i \\ & r^i - r = r^i_+ - r^i_- \quad \forall i \\ & \sum_{i=1}^n (\tau r^i_+ + (1-\tau) r^i_-) N_i \leq \\ & (1-\tau) \sum_{i=1}^{j-1} (r^j - r^i) N_i + \tau \sum_{i=j+1}^n (r^i - r^j) N_i \quad \forall j \\ & r \geq 0 \quad r^i_+ \geq 0 \quad r^i_- \geq 0 \quad \forall i \\ & \sum_{a\in A} x_{sa} - \beta \sum_{s'\in S} \sum_{a\in A} x_{s'a} T(s',a,s) = \mu(s) \quad \forall s \\ & x_{sa} \geq 0 \quad \forall s, \forall a \end{array} \Bigg\} (\mathcal{C})$$

Variables N_i in the first set of constraints are introduced for convenience. The second set of constraints expresses the absolute value in ρ_τ. They could be slightly simplified for $i = 1$ and $i = n$ as the sign is known. The third set of constraints states that r should be solution of Eq. 4. Without counting the N_i's, this program has $2n+1$ extra variables and $2n$ extra constraints compared to $(\mathcal{D})$.

This program optimizes the upper τ^{th} quantile and thus solves the case $\phi = \max$. It also provides an optimal policy when the lower and upper quantiles are equal, i.e. $\sum_{i=k}^n \sum_{s\in S} \sum_{a\in A} \hat{r}_i(s,a) x_{sa} > (1-\tau)/(1-\beta)$ for $r = r^k$. Otherwise, one again needs to use the information of ϕ and optimizes $\mathcal{D}^j_+$ for $j = k-1, \dots$ for finding the optimal solution. As a side note, one may combine in a lexicographic fashion quantile optimization and reference point-based preference. Indeed, after finding the τ^{th} optimal quantile, one can maximize the proportion of rewards better than that quantile.

6 CONCLUSION

Although of great practical interest, the case where preferences are qualitative and uncertainty is probabilistic has been rarely investigated in the literature. In this paper, we considered ordinal decision models in problems of planning under probabilistic uncertainty modeled as MDPs with ordinal rewards (ORMDP). For this model, we considered two preference systems dual one to the other: reference point-based preferences and quantile-based preferences. Based on the first one, already proposed by [23] in ORMDPs, we gave a new interpretation of rewards in standard MDPs. We studied the second in the framework of ORMDPs and proved that for a fixed initial state, there is a stationary deterministic policy optimal with respect to quantile optimization. However, some caution is needed as contrary to the preference system used in standard MDPs, quantile-based optimality is state-dependent. Finally, we proposed solving methods based on linear programming.

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

REFERENCES

[1] J.A. Bagnell, A.Y. Ng, and J.G. Schneider, 'Solving uncertain Markov decision processes', Technical report, CMU, (2001).

[2] C. Boutilier, R. Brafman, C. Domshlak, H. Hoos, and D. Poole, 'CP-nets: A tool for representing and reasoning with conditional ceteris paribus preference statements', *J. Artif. Intell. Research*, **21**, (2004).

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

[4] R.I. Brafman and M. Tennenholtz, 'On the axiomatization of qualitative decision criteria', in *AAAI*, volume 14, pp. 76–81, (1997).

[5] O. Buffet and D. Aberdeen, 'Robust planning with (L)RTDP', in *IJCAI*, pp. 1214–1219, (2005).

[6] E. Castagnoli and M. Li Calzi, 'Expected utility without utility', *Theory and Decision*, **41**, 281–301, (1996).

[7] C.P. Chambers, 'Quantiles and medians'. mimeo, Caltech, 2005.

[8] E. Delage and S. Mannor, 'Percentile optimization in uncertain Markov decision processes with application to efficient exploration', in *ICML*, pp. 225–232, (2007).

[9] D. Dubois, H. Fargier, and P. Perny, 'Qualitative decision theory with preference relations and comparative uncertainty: An axiomatic approach', *Artificial Intelligence*, **148**, 219–260, (2003).

[10] D. Dubois, L. Godo, H. Prade, and A. Zapico, 'Making decision in a qualitative setting: from decision under uncertainty to case-based decision', in *KR*, volume 6, pp. 594–607, (1998).

[11] D. Dubois, H. Prade, and R. Sabbadin, 'Qualitative decision theory with Sugeno integrals', in *UAI*, volume 14, pp. 121–128, (1998).

[12] R. Givan, S. Leach, and T. Dean, 'Bounded-parameter Markov decision process', *Artif. Intell.*, **122**(1-2), 71–109, (2000).

[13] R. Koenker, *Quantile Regression*, Cambridge university press, 2005.

[14] P.S. Mann, *Introductory Statistics*, Wiley, 2006.

[15] A. Nilim and L. El Ghaoui, 'Robustness in Markov decision problems with uncertain transition matrices', in *NIPS*, (2003).

[16] M.L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*, Wiley, 1994.

[17] K. Regan and C. Boutilier, 'Regret-based reward elicitation for Markov decision processes', in *UAI*, pp. 444–451, (2009).

[18] M.J. Rostek, 'Quantile maximization in decision theory', *Review of Economic Studies*, **77**(1), 339371, (2010).

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

[20] B. Schachter, 'An irreverent guide to Value-at-Risk', *Financial Engineering News*, **1**, (1997).

[21] F.W. Trevizan, F.G. Cozman, and L.N. de Barros, 'Planning under risk and Knightian uncertainty', in *IJCAI*, pp. 2023–2028, (2007).

[22] B. Viswanathan, V.V. Aggarwal, and K.P.K. Nair, 'Multiple criteria Markov decision processes', *TIMS Studies in the Management Sciences*, **6**, 263–272, (1977).

[23] P. Weng, 'Markov decision processes with ordinal rewards: Reference point-based preferences', in *ICAPS*, volume 21, pp. 282–289, (2011).

[24] H. Xu and S. Mannor, 'Parametric regret in uncertain Markov decision processes', in *IEEE Conference on Decision and Control*, (2009).