Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Optimizing Quantiles in Preference-Based Markov Decision Processes

Hugo Gilbert
Sorbonne Universités,
UPMC Univ Paris 06, CNRS,
LIP6 UMR 7606, Paris, France
hugo.gilbert@lip6.fr

Paul Weng, Yan Xu
SYSU-CMU Joint Institute of Engineering, Guangzhou, China
School of Electronics and Information Technology, Guangzhou, China
SYSU-CMU Shunde International Joint Research Institute, Shunde, China
{paweng,xuyan}@cmu.edu

Abstract

In the Markov decision process model, policies are usually evaluated by expected cumulative rewards. As this decision criterion is not always suitable, we propose in this paper an algorithm for computing a policy optimal for the quantile criterion. Both finite and infinite horizons are considered. Finally we experimentally evaluate our approach on random MDPs and on a data center control problem.

1 Introduction

Sequential decision-making in uncertain environments is an important task in artificial intelligence. Such problems can be modeled as Markov Decision Processes (MDPs). In an MDP, an agent chooses at every time step actions to perform according to the current state of the world in order to optimize a criterion in the long run. In standard MDPs, uncertainty is described by probabilities over the possible action outcomes, preferences are represented by numeric rewards and the expectation of future cumulated rewards is used as the decision criterion. And yet, for numerous applications, the expectation of cumulated rewards may not be the most appropriate criterion. For instance, in one-shot decision-making problems an alternative and well motivated objective for the agent is to insure a certain level of satisfaction with high probability.

In this paper we focus on the decision criterion that consists in maximizing a quantile. Intuitively, the τth quantile of a population is the value x such that $100 \cdot \tau$ percent of the population is equal or lower than x and $100 \cdot (1 - \tau)$ percent of the population is equal or greater than x. Optimizing a quantile criterion offers nice properties: i) no assumption is made about the commensurability between preferences and uncertainty, ii) preferences over actions or trajectories can be expressed on a purely ordinal scale, iii) preferences induced over policies are more robust than with the standard criterion of maximizing the expectation of cumulated rewards.

As a result, maximizing a quantile is used in many applications. For instance, the *Value-at-Risk* criterion (Jorion 2006) widely used in finance is in fact a quantile. Moreover, in the Web industry (Wolski and Brevik 2014; DeCandia et al. 2007), decisions about performance or Quality-Of-Service are often made based on quantiles. For instance, Amazon reports (DeCandia et al. 2007) that they optimize the 99.9% quantile for their cloud services. More generally, in the service industry, because of skewed distributions (Benoit and Van den Poel 2009), one generally does not want that customers are satisfied on average, but rather that most customers (*e.g.*, 99% of them) to be as satisfied as possible.

Our contribution: We show that optimizing the quantile criterion amounts to solving a sequence of MDP problems using an Expected Utility criterion with a target utility function. We provide a binary search algorithm using functional backward induction (Liu and Koenig 2006) as a subroutine for computing an optimal policy. Moreover, we investigate some properties of the optimal policies in the finite and infinite cases. Finally, we provide the results of experiments testing our algorithm in a variety of settings.

The paper is organized as follows. Section 2 introduces the necessary background to present our approach and state formally our problem. Section 3 presents the details of our solving algorithm for the finite horizon case. Section 4 provides some theoretical results in the infinite horizon case. In Section 5, we experimentally evaluate our proposition. Section 6 discusses the related work and Section 7 concludes.

2 Background

In this section, we provide the background information necessary for the sequel.

2.1 Markov Decision Process

Markov Decision Processes (MDPs) offer a general and powerful formalism to model and solve sequential decision-making problems (Puterman 1994). An MDP is formally defined as a tuple $\mathcal{M}_T = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, s_0)$ where T is a time horizon, $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is a transition function with $\mathcal{P}(s, a, s')$ being the probability of reaching state s' when action a is performed in state s, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a bounded reward function and $s_0 \in \mathcal{S}$ is a particular state called initial state.

In a nutshell, at each time step t, the agent knows her current state s_t. According to this state, she decides to perform an action a_t. This action results in a new state $s_{t+1} \in \mathcal{S}$ according to probability distribution $\mathcal{P}(s_t, a_t, .)$, and a reward signal $r(s_t, a_t)$ which penalizes or reinforces the choice of this action. At time step $t = 0$, the agent is in the initial state s_0. We will call t-history h_t a succession of t state-action pairs starting from state s_0 (*e.g.,* $h_t = (s_0, a_0, s_1, \ldots, s_{t-1}, a_{t-1}, s_t)$). We call episode a T-history and denote $\mathcal{E}$ the set of episodes.

The goal of the agent is to determine a policy, *i.e.,* a procedure to select an action in a state, that is optimal for a given criterion. More formally, a *policy* π at an horizon T is a sequence of T decision rules $(\delta_1, \ldots, \delta_T)$. *Decision rules* are functions which prescribe the actions that the agent should perform. They are *Markovian* if they only depend on the current state. Moreover, a decision rule is either *deterministic* if it always selects the same action in a given state or *randomized* if it can prescribe a probability distribution over possible actions. A policy can be *Markovian*, *deterministic* or *randomized* according to the type of its decision rules. Lastly, a policy is *stationary* if it uses the same decision rule at every time step, *i.e.,* $\pi = (\delta, \delta, \ldots)$.

Different criteria can be defined in order to compare policies. One standard criterion is *expected cumulated reward*, for which it is known that an optimal deterministic Markovian policy exists at any horizon T. This criterion is defined as follows. First, the value of a history $h_t = (s_0, a_0, s_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ is described as the sum of rewards obtained along it, *i.e.,* $r(h_t) = \sum_{i=0}^{t-1} r(s_i, a_i)$. Then, the value of a policy $\pi = (\delta_1, \ldots, \delta_T)$ in a state s is set to be the expected value of the histories that can be generated by π from s. This value, given by the *value function* $v_1^\pi : \mathcal{S} \to \mathbb{R}$ can be computed iteratively as follows:

$$
\begin{aligned}
&v_{T+1}^\pi(s) = 0 \\
&\quad v_t^\pi(s) = r(s, \delta_t(s)) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s, \delta_t(s), s') v_{t+1}^\pi(s') \qquad (1)
\end{aligned}
$$

The value $v_t^\pi(s)$ is the expectation of cumulated rewards obtained by the agent if she performs action $\delta_t(s)$ in state s at time step t and continues to follow policy π thereafter. The higher the values of $v_t^\pi(s)$ are, the better. Therefore, value functions induce a preference relation $\succsim_\pi$ over policies in the following way:

$$\pi \succsim_\pi \pi' \Leftrightarrow \forall s \in \mathcal{S}, \forall t = 1, \ldots, T, v_t^\pi(s) \geq v_t^{\pi'}(s)$$

A solution to an MDP is a policy, called *optimal policy*, that ranks the highest with respect to $\succsim_\pi$. Such a policy can be found by solving the *Bellman equations*.

$$
\begin{aligned}
&v_{T+1}^*(s) = 0 \\
&\quad v_t^*(s) = \max_{a \in \mathcal{A}} r(s, a) + \sum_{s' \in \mathcal{S}} \mathcal{P}(s, a, s') v_{t+1}^*(s')
\end{aligned}
$$

As can be seen, the preference relation $\succsim_\pi$ over policies is directly induced by the reward function r.

The decision criterion, based on the expectation of cumulated rewards, may not always be suitable. Firstly, unfortunately, in many cases, the reward function r is not known. One can therefore try to uncover the reward function by interacting with an *expert* of the domain considered (Regan and Boutilier 2009; Weng and Zanuttini 2013). However, even for an expert user, the elicitation of the reward function can be burdensome. Indeed, this process can be cognitively very complex as it requires to balance several criteria in a complex manner and as it can imply a large number of parameters. In this paper, we address this problem by only assuming that we have a strict weak ordering on episodes.

Secondly, for numerous applications, the expectation of cumulated reward, as used in Equation 1, may not be the most appropriate criterion (even when a numeric reward function is defined). For instance, in the Web industry, most decisions about performance are based on the minimal quality of 99% of the possible outcomes. Therefore, in this article we aim at using a quantile (defined in Section 2.3) as a decision criterion to solve an MDP.

2.2 Preferences over Histories

For generality's sake, contrary to standard MDPs, we define in this work the reward function to take values in a set $\mathcal{R}$. Moreover, we assume that the values of histories take values in a set $\mathcal{W}$, called the wealth level space, and that the value of a history $h_t = (s_0, a_0, s_1, \ldots, s_t)$ is defined by:

$$w(h_0) = w_0 \qquad w(h_t) = w(h_{t-1}) \circ r(s_{t-1}, a_{t-1})$$

where $h_{t-1} = (s_0, a_0, s_1, \ldots, s_{t-1})$, $\circ$ is a binary operation from $\mathcal{W} \times \mathcal{R}$ to $\mathcal{W}$ and $w_0 \in \mathcal{W}$ is the left identity element of $\circ$. Let $\mathcal{W}_T \subset \mathcal{W}$ be the set of wealth levels of T-histories. We make three assumptions about $\mathcal{W}_T$:

- It is ordered by a total order $\preceq_\mathcal{W}$, which defines how T-histories are compared,
- It admits a lowest element, denoted $w_{\min}$ and a greatest element, denoted $w_{\max}$ for order $\preceq_\mathcal{W}$.
- A distance consistent with $\preceq_\mathcal{W}$ is defined over $\mathcal{W}_T$. It is denoted $d(w, w')$ for any pair $(w, w') \in \mathcal{W}_T \times \mathcal{W}_T$.

Note that when a distance is defined, for any pair (w, w'), its set of mid-elements is also defined $mid(w, w') = \arg\inf\{\max(d(w, w''), d(w', w'')) \mid w'' \in \mathcal{W}_T\}$.

In a numerical context, the possible wealth levels of a state are the possible sums (resp. γ-discounted sums) of rewards that can be obtained during an episode. We have $w_{\max} = R_{\max} T$ (resp. $w_{\max} = R_{\max} \frac{(1-\gamma)^T}{1-\gamma}$) with $R_{\max}$ being the highest reward and $mid(w, w') = \{(w + w')/2\}$. In the most general case, the possible wealth levels of a state are the possible histories (or more precisely their equivalent classes) that can be obtained during an episode. Here, if the equivalence classes are known and denoted by $w_1 \prec_\mathcal{W} w_2 \prec_\mathcal{W} \ldots \prec_\mathcal{W} w_m$ and if $d(w_i, w_j) = |j - i|$, then $w_{\min} = w_1$, $w_{\max} = w_m$ and $mid(w_i, w_j) = \{w_{\lfloor (i+j)/2 \rfloor}, w_{\lceil (i+j)/2 \rceil}\}$ (where $\lfloor x \rfloor$is the greatest integer smaller than x and $\lceil x \rceil$ is the smallest integer greater than x).

The goal of the agent is then to make sure that most of the time, it will generate episodes that have the highest possible wealth levels. This can be implemented by optimizing a quantile criterion as explained in the next subsection.

2.3 Quantile Criterion

Intuitively, the τ-quantile of a population of ordered elements, for $\tau \in [0, 1]$, is the value q such that $100 \cdot \tau\%$ of the population is equal or lower than q and $100 \cdot (1 - \tau)\%$ of the population is equal or greater than q. The 0.5-quantile, also known as the median, can be seen as the ordinal counterpart of the mean. More generally, quantiles define decision criteria that have the nice property of not requiring numeric valuations, but only an order. They have been axiomatically studied as decision criteria by Rostek (2010).

We now give a formal definition of quantiles. For this purpose we define the probability distribution p^π over wealth levels induced by a policy π, *i.e.,* $p^\pi(w)$ is the probability of getting a wealth level $w \in \mathcal{W}_T$ when applying policy π from the initial state. The *cumulative distribution* induced by p^π is then defined as F^π where $F^\pi(w) = \sum_{w' \preceq_\mathcal{W} w} p^\pi(w')$ is the probability of getting a wealth level not preferred to w when applying policy π. Similarly, the *decumulative distribution* induced by p^π is defined as $G^\pi(w) = \sum_{w \preceq_\mathcal{W} w'} p^\pi(w')$ is the probability of getting a wealth level "not lower" than w.

These two notions of cumulative and decumulative enable us to define two kinds of criteria. First, given a policy π, we define the lower τ-quantile for $\tau \in (0, 1]$ as:

$$\underline{q}^\pi_\tau = \min\{w \in \mathcal{W}_T \mid F^\pi(w) \geq \tau\} \tag{2}$$

where the $\min$ operator is with respect to $\prec_\mathcal{W}$.

Then, given a policy π, we define the upper τ-quantile for $\tau \in [0, 1)$ as:

$$\overline{q}^\pi_\tau = \max\{w \in \mathcal{W}_T \mid G^\pi(w) \geq 1 - \tau\} \tag{3}$$

where the $\max$ operator is with respect to $\prec_\mathcal{W}$.

If $\tau = 0$ or $\tau = 1$ only one of $\underline{q}^\pi_\tau$ or $\overline{q}^\pi_\tau$ is defined and we define the τ-quantile q^π_τ as that value. When both are defined, by construction, we have $\underline{q}^\pi_\tau \preceq_\mathcal{W} \overline{q}^\pi_\tau$. If those two values are equal, q^π_τ is defined as equal to them. For instance, this is always the case in continuous settings for continuous distributions. However, in our discrete setting, it could happen that those values differ, as shown by Example 1.

Example 1. *Consider an MDP where $\mathcal{W}_T = \{w_1 \prec_\mathcal{W} w_2 \prec_\mathcal{W} w_3\}$. Now assume a policy π attains each wealth level with probabilities 0.5, 0.2 and 0.3 respectively. Then it is easy to see that $\underline{q}^\pi_{0.5} = w_1$ whereas $\overline{q}^\pi_{0.5} = w_2$.*

When the lower and upper quantiles differ, one may define the quantile as a function of the lower and upper quantiles (Weng 2012). For simplicity, we show in this paper how to optimize (approximately) the lower and the upper quantiles.

Definition 1. *A policy π^* is optimal for the lower (resp. upper) τ-quantile criterion if:*

$$\underline{q}^{\pi^*}_\tau = \max_\pi \underline{q}^\pi_\tau \quad (\textit{resp. } \overline{q}^{\pi^*}_\tau = \max_\pi \overline{q}^\pi_\tau) \tag{4}$$

where the $\max$ *operator is with respect to $\prec_\mathcal{W}$ and taken over all policies π at horizon T.*

Even in a numerical context where a numerical reward function is given and the quality of an episode is defined as the cumulative of rewards received along the episode, this criterion is difficult to optimize, notably due to the two following related points:

- It is *non-linear* meaning for instance that the τ-quantile $q^{\tilde{\pi}}_\tau$ of the mixed policy $\tilde{\pi}$ that generates an episode using policy π with probability p and π' with probability $1 - p$ is not given by $pq^\pi_\tau + (1 - p)q^{\pi'}_\tau$.
- It is *non-dynamically consistent* meaning that at time step t, an optimal policy computed in s_0 with horizon T might not prescribe in state s_t to follow a policy optimal in s_t for horizon $T - t$.

Three solutions are then possible (McClennen 1990): 1) adopting a *consequentialist* approach, *i.e.,* at each time step t we follow an optimal policy for the problem with horizon $T - t$ and initial state s_t even if the resulting policy is not optimal at horizon T; 2) adopting a *resolute choice* approach, *i.e.,* at time step $t = 0$ we apply an optimal policy for the problem with horizon T and initial state s_0 and do not deviate from it; 3) adopting a *sophisticated resolute choice* approach (Jaffray 1998; Fargier, Jeantet, and Spanjaard 2011), *i.e.,* we apply a policy π (chosen at the beginning) that trades off between how much π is optimal for all horizons $T, T - 1, \ldots, 1$.

With non-dynamically consistent preferences, it is debatable to adopt a consequentialist approach, as the sequence of decisions may lead to dominated results. In this paper, we adopt a resolute choice point of view. We leave the third approach for future work.

As optimizing exactly a (lower or upper) quantile is hard, we aim at finding an approximate solution. Let $\underline{q}^*_\tau$ and $\overline{q}^*_\tau$ be equal to the optimal lower and upper quantile respectively.

Definition 2. *Let $\varepsilon > 0$. A policy π^*_ε is said to be ε-optimal for the lower (resp. upper) τ-quantile criterion if $d(\underline{q}^{\pi^*_\varepsilon}_\tau, \underline{q}^*_\tau) \leq \varepsilon$ (resp. $d(\overline{q}^{\pi^*_\varepsilon}_\tau, \overline{q}^*_\tau) \leq \varepsilon$).*

3 Solving Algorithm

In this section, we present a technique for computing an ε-optimal policy for the quantile criterion. Our approach amounts to solving a sequence of MDPs optimizing EU with target utility functions (see Section 3.2).

3.1 Binary Search

In order to justify our algorithm, we introduce two lemmas that characterize the optimal lower and upper quantiles[1]:

Lemma 1. *The optimal lower τ-quantile $\underline{q}^*$ satisfies:*

$$\underline{q}^* = \min\{w : F^*(w) \geq \tau\} \tag{5}$$

$$F^*(w) = \min_\pi F^\pi(w) \quad \forall w \in \mathcal{W} \tag{6}$$

Note the last two equations can be equivalently rewritten:

$$\underline{q}^* = \min\{w : G^*_\prec(w) \leq 1 - \tau\} \tag{7}$$

$$G^*_\prec(w) = \max_\pi G^\pi_\prec(w) \quad \forall w \in \mathcal{W} \tag{8}$$

where $G^\pi_\prec(w) = 1 - F^\pi(w) = \sum_{w \prec_\mathcal{W} w'} p^\pi(w')$.

Lemma 2. *The optimal upper τ-quantile $\overline{q}^*$ satisfies:*

$$\overline{q}^* = \max\{w : G^*(w) \geq 1 - \tau\} \tag{9}$$

$$G^*(w) = \max_\pi G^\pi(w) \quad \forall w \in \mathcal{W} \tag{10}$$

[1]For lack of space, all proofs are in the supplementary material which can be found at hugogilbert.pythonanywhere.com

Algorithm 1: Binary Search for the Lower Quantile (resp. Upper Quantile)

Data: MDP $\mathcal{M}, \tau, \varepsilon$
Result: an ε-optimal policy π
$\overline{w} \leftarrow w_{\max}$; $\underline{w} \leftarrow w_{\min}$; $w \leftarrow mid(\underline{w}, \overline{w})$
while $d(\overline{w}, \underline{w}) > \varepsilon$ **do**
  $(\pi, p) = solve(\mathcal{M}, w)$;
  if $p > 1 - \tau$ *(resp.* $p \geq 1 - \tau$*)* **then**
    $\underline{w} \leftarrow w$; $w \leftarrow \max(mid(\underline{w}, \overline{w}))$; $\pi^* \leftarrow \pi$;
  else
    $\overline{w} \leftarrow w$; $w \leftarrow \min(mid(\underline{w}, \overline{w}))$;
return π^*

Given Lemmas 1 and 2 the problem now reduces to finding the right value of $w \in \mathcal{W}$ that solves the problems defined by Equation 7 or 9. Our solving method is based on binary search (see Algorithm 1) and on the function $solve(\mathcal{M}, w)$ that returns a pair (π, p), the solution of the problems defined by Equation 8 or 10 for a fixed w, *i.e.,* the $\max$ is equal to p and attained at π. Note that while for the upper quantile criterion, $solve(\mathcal{M}, \overline{q}_\tau^{\pi^*})$ returns an optimal policy, for the lower quantile, $solve(\mathcal{M}, \underline{q}_\tau^{\pi^*})$ may not if $\underline{q}_\tau^{\pi^*} \succ_\mathcal{W} \min(\mathcal{W}_T)$. However, $solve(\mathcal{M}, prec(\underline{q}_\tau^{\pi^*}))$ returns an optimal policy where $prec(w)$ is the most preferred element such that $prec(w) \prec_\mathcal{W} w$ (see supplementary material).

In the next subsection, we show how function $solve$ can be computed for the lower and upper quantile.

Note that when $\mathcal{W}_T$ is defined on the real line, Algorithm 1 needs only $\lceil \log_2 d(w_{\max}, w_{\min})/\varepsilon \rceil$ iterations to terminate by using $[w_{\min}, w_{\max}]$ as $\mathcal{W}_T$. In the case where $\mathcal{W}_T$ is finite, binary search can of course determine the optimal policy with $\varepsilon = 1$ and needs $\lceil \log_2(|\mathcal{W}_T|) \rceil$ iterations.

The next proposition asserts that Algorithm 1 is correct:

Proposition 1. *Algorithm 1 returns an ε-optimal policy for the lower (or upper) quantile criterion.*

3.2 Dynamic Programming

For $\lhd \in \{\prec_\mathcal{W}, \preceq_\mathcal{W}\}$, we denote by $U_w^\lhd : \mathcal{W} \to \mathbb{R}$ the function, called *target utility function*, defined as follows:

$$U_w^\lhd(x) = 1 \text{ if } w \lhd x \text{ and } 0 \text{ else.} \quad (11)$$

When optimizing the lower (resp. upper) quantile, function $solve(\mathcal{M}, w)$ can be computed by solving MDP $\mathcal{M}$ using EU as a decision criterion with $U_w^{\prec_\mathcal{W}}$ (resp. $U_w^{\preceq_\mathcal{W}}$) as a utility function. Indeed, we have:

$$\mathbb{E}_\pi[U_w^\lhd\big(w(H_T)\big)] = \mathbb{P}[w \lhd w(H_T) \mid \pi]$$

where H_T is a random variable representing a T-history and $\mathbb{P}[w \lhd w(H_T) \mid \pi]$ denotes the probability that π generates a history whose wealth is strictly better (resp. at least better) than w when $\lhd = \prec_\mathcal{W}$ (resp. $\lhd = \preceq_\mathcal{W}$).

Following (Liu and Koenig 2006), this problem can be solved with a functional backward induction (Algorithm 2). For each state s, it maintains a function $V_t(s, .)$ which associates to each possible wealth level w the expected utility obtained by applying an optimal policy in state s for the remaining $T - t$ time steps with w as initial wealth level. At each time step $(t = T, \dots, 1)$ this function is updated similarly as in backward induction except that operations are not applied to scalars but to functions. The $\max$ and $\times$ operations are extended over functions as pointwise operations. As utility functions defined by Equation 11 are piecewise-linear, $V_t(s, .)$ is also piecewise-linear because all the operations in Line 5 of Algorithm 2 preserve this property.

Algorithm 2: FunctionalBackwardInduction

Data: MDP $\mathcal{M}$, wealth w
Result: an optimal policy π
1 **for** *all* $s \in S$ **do**
2   $V_{T+1}(s, .) \leftarrow U_w^\lhd(.)$
3 **for** $t = T$ *to* 1 **do**
4   **for** *all* $s \in S$ **do**
5     $V_t(s, \cdot) \leftarrow \max_a \sum_{s' \in S} \mathcal{P}(s, a, s') V_{t+1}(s', \cdot \circ r(s, a))$
6 **return** $(\pi_{V_1}, V_1(s_0, w_0))$ \\\\ π_{V_1} = policy corresponding to V_1

Policies returned by Algorithm 2 have a special structure. They are deterministic and wealth-Markovian:

Definition 3. *A policy is said to be* wealth-Markovian *if its decision rules are functions of both the current state and the current wealth level.*

Besides, this is also the case for policies optimal with respect to the quantile criterion.

Proposition 2. *Optimal policies for the lower or upper quantile at horizon T can be found as deterministic wealth-Markovian policies.*

4 Infinite Horizon

We present in this section some results regarding the infinite horizon case. Similarly to the finite horizon setting, the situation for the quantile criterion is not as simple as for the standard case. Indeed, in the infinite horizon case, it may happen that there is no *stationary* deterministic Markovian policy that is optimal (w.r.t. the quantile criterion) among all policies, contrary to standard MDPs.

Example 2. *Consider an MDP with two states s_1 and s_2 and two actions a_1 and a_2. In s_1, the transition probabilities are $\mathcal{P}(s_1, a_1, s_1) = 0.1$, $\mathcal{P}(s_1, a_1, s_2) = 0.9$ and $\mathcal{P}(s_1, a_2, s_2) = 1$. To make this example shorter, we assume that rewards depend on next states. The rewards are $r(s_1, a_1, s_1) = 1$, $r(s_1, a_1, s_2) = -1$ and $r(s_1, a_2, s_2) = 1$. In s_2, the transition probabilities are $\mathcal{P}(s_2, a_1, s_2) = \mathcal{P}(s_2, a_2, s_2) = 1$. Rewards are null for both actions in s_2. Among all decision rules, there are only two distinct rules: $\delta_1(s_1) = a_1$ and $\delta_2(s_1) = a_2$. To ensure that the values of histories are well-defined, we assume that they are defined as discounted sum of rewards with a discount factor $\gamma = 0.9$. One can then check that the 0.95-quantile of the stationary policy using δ_1 is 0.1, that of the stationary policy using δ_2 is 1. Finally, the 0.95-quantile of the policy applying first*

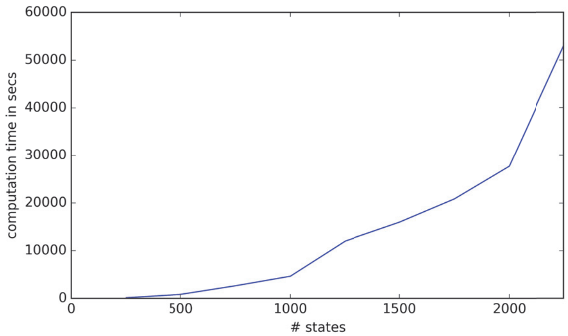

Figure 1: Computation times vs state sizes for Functional Backward Induction.

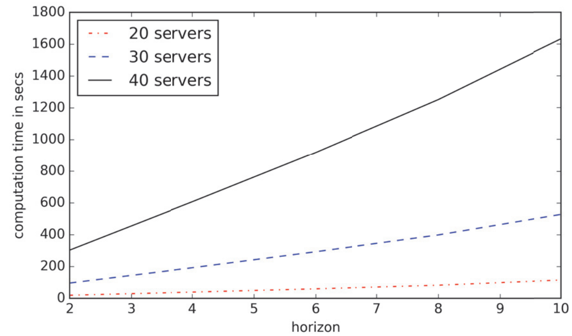

Figure 2: Computation times vs horizon for Functional Backward Induction.

δ_1 *and then* δ_2 *is* 1.9*. Therefore, no stationary deterministic Markovian policy is optimal for the quantile criterion.*

However, considering wealth-Markovian policies, some results can be given when rewards are numeric and wealth levels are undiscounted:

Proposition 3. *Optimal policies for the lower or upper quantile can be found as stationary deterministic wealth-Markovian policies in the two following cases:*

(i) $\forall(s, a) \in \mathcal{S} \times \mathcal{A}, r(s, a) \leq 0$.

(ii) $\forall(s, a) \in \mathcal{S} \times \mathcal{A}, r(s, a) \geq 0$. *Furthermore, we require the existence of a finite upper bound on the optimal lower and upper quantiles.*

Then, a solving algorithm can be obtained from Algorithm 1 by replacing functional backward induction (Alg. 2) by functional value iteration (Liu and Koenig 2006) in the binary search. This amounts to do the for loop over t (line 4) until convergence of V_t, *i.e.*, $\|V_t - V_{t-1}\|_\infty \leq \epsilon'$. Binary search will then return an $(\epsilon+\epsilon')$-optimal for the τ-quantile. However, note that in the first (resp. second) case, a lower (resp. upper) bound on the optimal lower or upper quantile is required to do the binary search.

5 Experimental Results

We experimentally evaluated our approach on a server equipped with four Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz and 64Gb of RAM. The algorithms were implemented in Matlab and ran only on one core. We expect the running times to be improved with a more efficient programming language and by exploiting a multicore architecture.

We designed three sets of experiments. Although our approach could be used in a preference-based setting, we performed the experiments with numerical rewards for simplicity. The first shows the running time of functional backward induction for different varying state sizes on random MDPs. The second set of experiments shows the running time of functional backward induction for different horizons on a data center control problem with various number of servers. Finally, the third compares the cumulative distributions of a policy optimal for the quantile criterion and a policy optimal for the standard criterion on a fixed MDP.

The first set of experiments was conducted on Garnets (McKinnon and Thomas 1995), which designate random MDPs with a constrained branching factor. A Garnet $G(n_S, n_A, b)$ is characterized by n_S a number of states, n_A a number of actions and b the number of successor states for every state and action. For our experiments, $n_S \in \{250, 500, 750, 1000, 1250, 1500, 1750, 2000, 2250\}$ and we set $n_A = 5$ and $b = \lceil \log_2 n_S \rceil$. Rewards are randomly chosen in $[0, 1]$ and the values of histories are simply cumulated rewards. The horizon of the problem was set to 5. The results are presented in Figure 1 where the x-axis represents the state size and the y-axis the computation time. Each point is the average over 10 runs. Naturally, computation times increases with state sizes. In this setting, binary search would call functional backward induction $\lceil \log_2(1/\varepsilon) \rceil = 10$ times if $\varepsilon = 10^{-3}$.

The second set of experiments was performed on a more realistic domain, which is a data center control problem inspired by the model proposed by Yin and Sinopoli (2014). In this problem, one needs to decide every time step how many servers to switch on or off, while maximizing Quality-of-Service and minimizing power consumption. In the model proposed by Yin and Sinopoli, the two objectives are simply combined into one cost, which defines our reward signal. The state is defined as the number of servers that are currently on and the number of jobs that needs to be processed during a time step. The action represents the number of servers that will be on at the next time step. We assume for simplicity that the maximum number of jobs that can arrive at one timestep is three times the total number of servers. For instance, in a problem with $n = 30$ servers, the total number of states is $30 \times 3 \times 30 = 2700$. Besides, the distribution of the next number of jobs is modeled as a Poisson distribution whose parameter can be $\lceil n/2 \rceil$, $\lceil 3n/2 \rceil$ or $\lceil 5n/2 \rceil$ (to model different regimes) depending on the current number of jobs. Figure 2 shows the computation times of functional backward induction for $n \in \{20, 30, 40\}$ and different horizons. We can see that for more structured problems, the computation time is much more reasonable than on random MDPs.

In the last set of experiments, to give an intuition of the kind of policy obtained when optimizing a quantile, we

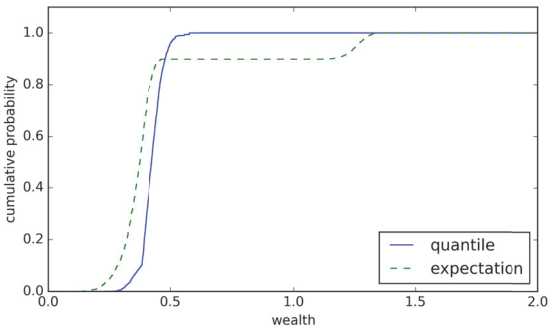

Figure 3: Comparison of cumulative distributions under the quantile criterion and standard criterion

compare the cumulative distribution of a policy optimal for the quantile criterion and that of a policy optimal for the standard criterion. This experiment is performed on an instance of Garnet $G(100, 5, \lceil \log_2 100 \rceil)$ whose rewards are slightly modified to make the distribution of the optimal policy skewed, as it is often the case in some real applications (Benoit and Van den Poel 2009). The horizon is set to 5 and we optimize the 0.1-quantile with $\varepsilon = 0.001$ in binary search. The two cumulative distributions are plotted in Figure 3. We can observe that although the optimal policy for the standard criterion maximizes the expectation, it may be a risky policy to apply as the probability of obtaining a high reward is low. On the contrary, the optimal policy for the τ-quantile criterion will guarantee a reward as high as possible with probability at least $1 - \tau$.

6 Related Work

Much work in the MDP literature (Boussard et al. 2010) considered decision criteria different to the standard ones (i.e., expected discounted sum of rewards, expected total rewards or expected average rewards). For instance, in the operations research community, White (1987) considered different cases where preferences over policies only depend on sums of rewards: Expected Utility (EU), probabilistic constraints and mean-variance formulations. In this context, he showed the sufficiency of working in a state space augmented with the sum of rewards obtained so far. Recently, (Prashanth and Ghavamzadeh 2013) and (Mannor and Tsitsiklis 2011) provided algorithms for this mean-variance formulation. Filar, Kallenberg, and Lee (1989) investigated decision criteria that are variance-penalized versions of the standard ones. They formulated the obtained optimization problem as a non-linear program. Several researchers (White 1993; Bouakiz and Kebir 1995; Yu, Lin, and Yan 1998; Wu and Lin 1999; Ohtsubo and Toyonaga 2002; Hou, Yeoh, and Varakantham 2014; Fan, Kalaba, and Moore II 2005) worked on the problem of optimizing the probability that the total (discounted) reward exceeds a given threshold.

Additionally, in the artificial intelligence community, (Liu and Koenig 2005; 2006; Ermon et al. 2012) also investigated the use of EU as a decision criterion in MDPs. In the continuation of this work, Gilbert et al. (2015) investigated the use of Skew-Symmetric Bilinear (SSB) utility (Fishburn 1981) functions — a generalization of EU with stronger descriptive abilities — as decision criteria in finite-horizon MDPs. Interestingly, SSB also encompasses probabilistic dominance, a decision criterion that can be employed in preference-based sequential decision-making (Busa-Fekete et al. 2014).

Recent work in MDP and reinforcement learning considered conditional Value-at-risk (CVaR), a criterion related to quantile, as a risk measure. Bäuerle and Ott (2011) proved the existence of deterministic wealth-Markovian policies optimal with respect to CVaR. Chow and Ghavamzadeh (2014) proposed gradient-based algorithms for CVaR optimization. In contrast, Borkar and Jain (2014) used CVaR in inequality constraints instead of as objective function.

Closer to our work, several quantile-based decision models have been investigated in different contexts. In uncertain MDPs where the parameters of the transition and reward functions are imprecisely known, Delage and Mannor (2007) presented and investigated a quantile-like criterion to capture the trade-off between optimistic and pessimistic viewpoints on an uncertain MDP. The quantile criterion they use is different to ours as it takes into account the uncertainty present in the parameters of the MDP. Filar, Krass, and Ross (1995) proposed an algorithm for optimizing the quantile criterion when histories are valued by average rewards. In that setting, they showed that an optimal stationary deterministic Markovian policy exists. In MDPs with ordinal rewards (Weng 2011; 2012; Filar 1983), quantile-based decision models were proposed to compute policies that maximize a quantile using linear programming. While quantiles in those works are defined on distributions over ordinal rewards, we defined them as distributions over histories.

More recently, in the machine learning community, quantile-based criteria have been proposed in the multi-armed bandit (MAB) setting, a special case of reinforcement learning. Yu and Nikolova (2013) proposed an algorithm in the pure exploration setting for different risk measures, including Value-at-Risk. Carpentier and Valko (2014) studied the problem of identifying arms with extreme payoffs, a particular case of quantiles. Finally, Szörenyi et al. (2015) investigated MAB problems where a quantile is optimized instead of the mean.

7 Conclusion

In this paper we have developed a framework to solve sequential decision problems in a very general setting according to a quantile criterion. Modeling those problems as MDPs we developed an offline algorithm in order to compute an ϵ-optimal policy and investigated the properties of the optimal policies in the finite and infinite horizon cases. Lastly, we provided experimental results, testing those two algorithms in a variety of settings.

As future work, we plan to investigate how this work can be extended to the case of reinforcement learning, a framework more involved than the one of MDPs where the dynamics of the problems are unknown and must be learned.

References

Bäuerle, N., and Ott, J. 2011. Markov decision processes with average value-at-risk criteria. *Mathematical Methods of Operations Research* 74(3):361–379.

Benoit, D., and Van den Poel, D. 2009. Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services. *Expert Systems with Applications* 36:10475–10484.

Borkar, V., and Jain, R. 2014. Risk-constrained Markov decision processes. *IEEE Trans. on Automatic Control* 59(9):2574–2579.

Bouakiz, M., and Kebir, Y. 1995. Target-level criterion in Markov decision processes. *Journal of Optimization Theory and Applications* 86(1):1–15.

Boussard, M.; Bouzid, M.; Mouaddib, A.-I.; Sabbadin, R.; and Weng, P. 2010. *Markov Decision Processes in Artificial Intelligence*. Wiley. chapter Non-Standard Criteria, 319–359.

Busa-Fekete, R.; Szörenyi, B.; Weng, P.; Cheng, W.; and Hüllermeier, E. 2014. Preference-based Reinforcement Learning: Evolutionary Direct Policy Search using a Preference-based Racing Algorithm. *Machine Learning* 97(3):327–351.

Carpentier, A., and Valko, M. 2014. Extreme bandits. In *NIPS*.

Chow, Y., and Ghavamzadeh, M. 2014. Algorithms for CVaR optimization in MDPs. In *NIPS*.

DeCandia, G.; Hastorun, D.; Jampani, M.; Kakulapati, G.; Lakshman, A.; Pilchin, A.; Sivasubramanian, S.; Vosshall, P.; and Vogels, W. 2007. Dynamo: Amazon's highly available key-value store. *ACM SIGOPS Operating Systems Review* 41(6):205–220.

Delage, E., and Mannor, S. 2007. Percentile optimization in uncertain Markov decision processes with application to efficient exploration. In *ICML*, 225–232.

Ermon, S.; Gomes, C.; Selman, B.; and Vladimirsky, A. 2012. Probabilistic planning with non-linear utility functions and worst-case guarantees. In *AAMAS*, 965–972.

Fan, Y.; Kalaba, R.; and Moore II, J. 2005. Arriving on time. *Journal of Optimization Theory and Applications* 127(3):497–513.

Fargier, H.; Jeantet, G.; and Spanjaard, O. 2011. Resolute choice in sequential decision problems with multiple priors. In *IJCAI*.

Filar, J. A.; Kallenberg, L. C. M.; and Lee, H.-M. 1989. Variance-penalized Markov decision processes. *Mathematics of Operations Research* 14:147–161.

Filar, J.; Krass, D.; and Ross, K. 1995. Percentile performance criteria for limiting average Markov decision processes. *IEEE Trans. on Automatic Control* 40(1):2–10.

Filar, J. A. 1983. Percentiles and Markovian decision processes. *Operations Research Letters* 2(1):13 – 15.

Fishburn, P. 1981. An axiomatic characterization of skew-symmetric bilinear functionals, with applications to utility theory. *Economics Letters* 8(4):311–313.

Gilbert, H.; Spanjaard, O.; Viappiani, P.; and Weng, P. 2015. Solving MDPs with skew symmetric bilinear utility functions. In *IJCAI*, 1989–1995.

Hou, P.; Yeoh, W.; and Varakantham, P. R. 2014. Revisiting risk-sensitive MDPs: New algorithms and results. In *ICAPS*.

Jaffray, J.-Y. 1998. Implementing resolute choice under uncertainty. In *UAI*.

Jorion, P. 2006. *Value-at-Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill.

Liu, Y., and Koenig, S. 2005. Risk-sensitive planning with one-switch utility functions: Value iteration. In *AAAI*, 993–999.

Liu, Y., and Koenig, S. 2006. Functional value iteration for decision-theoretic planning with general utility functions. In *AAAI*, 1186–1193.

Mannor, S., and Tsitsiklis, J. 2011. Mean-variance optimization in Markov decision processes. In *ICML*.

McClennen, E. 1990. *Rationality and dynamic choice: Foundational explorations*. Cambridge university press.

McKinnon, T. A. K., and Thomas, L. 1995. On the generation of Markov decision processes. In *Journal of the Operational Research Society*, 354–361.

Ohtsubo, Y., and Toyonaga, K. 2002. Optimal policy for minimizing risk models in Markov decision processes. *Journal of mathematical analysis and applications* 271:66–81.

Prashanth, L., and Ghavamzadeh, M. 2013. Actor-critic algorithms for risk-sensitive MDPs. In *NIPS*, 252–260.

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*, 444–451. Morgan Kaufmann.

Rostek, M. 2010. Quantile maximization in decision theory. *Review of Economic Studies* 77(1):339–371.

Szörenyi, B.; Busa-Fekete, R.; Weng, P.; and Hüllermeier, E. 2015. Qualitative multi-armed bandits: A quantile-based approach. In *ICML*, 1660–1668.

Weng, P., and Zanuttini, B. 2013. Interactive value iteration for Markov decision processes with unknown rewards. In *IJCAI*, 2415–2421.

Weng, P. 2011. Markov decision processes with ordinal rewards: Reference point-based preferences. In *ICAPS*, volume 21, 282–289.

Weng, P. 2012. Ordinal decision models for Markov decision processes. In *ECAI*, volume 20, 828–833.

White, D. J. 1987. Utility, probabilistic constraints, mean and variance of discounted rewards in Markov decision processes. *OR Spektrum* 9:13–22.

White, D. 1993. Minimising a threshold probability in discounted Markov decision processes. *Journal of mathematical analysis and applications* 173(634–646).

Wolski, R., and Brevik, J. 2014. QPRED: Using quantile predictions to improve power usage for private clouds. Technical report, UCSB.

Wu, C., and Lin, Y. 1999. Minimizing risk models in Markov decision processes with policies depending on target values. *Journal of mathematical analysis and applications* 231:41–67.

Yin, X., and Sinopoli, B. 2014. Adaptive robust optimization for coordinated capacity and load control in data centers. In *International Conference on Decision and Control*.

Yu, J., and Nikolova, E. 2013. Sample complexity of risk-averse bandit-arm selection. In *IJCAI*.

Yu, S. X.; Lin, Y.; and Yan, P. 1998. Optimization models for the first arrival target distribution function in discrete time. *Journal of mathematical analysis and applications* 225:193–223.