Flexible risk aware sequential decision making

Nadjet Bourdache

Flexible risk aware sequential decision making

Nadjet Bourdache[1] [https://orcid.org/0009−0000−9005−7673]

Université de Caen Normandie, ENSICAEN, CNRS, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France nadjet.bourdache@unicaen.fr

Abstract. In this work, we study risk aware sequential decision making in a Markov Decision Process (MDP). Unlike many works in the literature, where MDPs are solved by optimizing expected rewards (ER), and thus assuming neutrality w.r.t. risk, we use a more sophisticated operator: the Weighted Ordered Weighted Average (WOWA), a parameterized operator that allows to model a wide range of behaviors, from extreme risk seeking to extreme risk aversion (as well as compromises between both behaviors). This operator has thus a high descriptive capacity, but is rather difficult to optimize in an MDP because of its non-linearity that makes standard solving algorithms sub-optimal. In this paper, we introduce and justify a ranking algorithm that allows to determine an optimal (or nearly optimal) policy for a wide range of attitudes w.r.t. risk (averse, seeking, neutral, intermediate) using WOWA. Empirical results are given to illustrate the relevance and the efficiency of the approach.

Keywords: Sequential decision making · Markov decision processes · decision theory under risk · preference modeling.

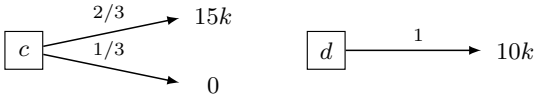
1 Introduction

Designing a decision support system or an automated decision making system for sequential decision making under risk is a widely studied task [1, 2, 21]. On the one hand, the need for such systems is considerably important as sequential decision making under risk has applications in several fields (business, finance, health, navigation, ...). On the other hand, dealing with risk in decision making is a hard task, since it requires to both know and consider all the probable consequences of a decision (or action) at short, medium and/or long term.

An agent faces risk when the outcomes of her decisions (or actions) are uncertain. Uncertainty can be found in different forms, we consider here situations where several outcomes can result from an action, and where the agents knows these outcomes, as well as their probabilities, but cannot predict with certainty which one will occur. For example, playing the roulette game enables to win 35 times the bet played with a probability 1/37, and to lose the bet otherwise. In such decision processes, the uncertainty makes the decision complex, and when decisions are taken sequentially, uncertainty can greatly increase, as the succession of stochastic events can greatly increase the number of possible consequences, making the decision much harder.

Markov Decision Processes (MDPs) [19] provide natural tools to formalize sequential decision making under risk processes. Thus, solving the decision problem consists in computing an optimal policy for the corresponding MDP. In the literature, the most commonly used operator to solve an MDP is the expected reward (ER) criterion, suggesting the (very strong) assumption that any agent, human or artificial, is perfectly risk-neutral. Yet, this assumption is often unrealistic. Let us consider a simple example :

Example 1. An agent has to chose between two actions: action c that gives $15k$\$ with probability $2/3$ and 0 otherwise; and action d that gives $10k$\$ for sure.



In this example, both actions have the same ER value, yet, it seems obvious that any rational agent would not choose indifferently one of the two actions. The first option offers the opportunity of a greater gain in the best case, but it is riskier as it involves having nothing at all in the worst case. It is clear that, in such situations, optimality is subjective, and we can easily admit that neutrality does not coincide with any agent behavior. Some, being strongly risk-averse, would always prefer the certainty over the risk (choosing action d in the example), while others, being strongly risk-seeking, would always be attracted by the possibility of a bigger payoff (choosing action c in the example). And finally, some decision makers would have intermediate behaviors being risk-averse in some situations and risk-seeking in others. Using ER fails to express such basic behaviors, simply because it does not allow to measure risk, even in the simple case where a random variable X is opposed to the certain gain $ER(X)$. Whereas in real-life situations, agents are often faced with more complex situations where they have to chose between actions representing different levels of risk.

As we will see in section 2, several works in the literature proposed to replace ER by other operators to model non-neutral attitudes w.r.t. risk. Most of them focused on risk-averse decision making and by proposing pessimistic operators that evaluate the worst possible outcome(s): the min-max (optimizing the worst possible outcome), the min-max regret (minimizing the worst regret over a choice), the expected shortfall (optimizing the average reward of a given proportion of the worst possible outcomes). However, it is important to note that risk-averse decision making is not only about avoiding risk or danger, it is also about the compromise between the risk and the potential benefit from this risk, which can lead to intermediate behaviors that are not model by the cited operators. For example, many operators (including many of those discussed later) fail to express the widely observed Allais paradox [3] that states that the independence axiom (which is often considered as verified) is not always relevant. We illustrate this paradox through the following example:

Example 2. Before the agent of example 1 makes her choice, we flip a coin. If the coin lands on head (with probability α), the agent chooses her action and gets her reward as described in example 1. If it lands on tail (with probability $1-\alpha$), she gets nothing. When $\alpha = 0.9$, this game is equivalent to a choice between actions c' and d' pictured below: The independence axiom states that

action c' is preferred to d' ($c' \succ d'$) *if and only if* $c \succ d$ (whatever the value of α). Nevertheless, the experimental study of Allais showed that the majority of surveyed agents violates this axiom.

The aim of this work is to propose a generic approach to solve sequential decision making problems under risk, considering as many behaviors as possible w.r.t. risk, including the one highlighted by the Allais paradox. For this, we will focus on the Weighted Ordered Weighted Average operator (WOWA for short). An operator that offers a rich descriptive power thanks to its parameter which, when well chosen, allows to model a wide range of attitudes w.r.t. risk. We will see that the main drawback of WOWA is its non-linearity and its lack of time-consistency, which prevents the use of standard algorithms like dynamic and linear programming. We propose in this paper a ranking algorithm that allows to determine a WOWA-optimal or nearly WOWA-optimal policy in an MDP.

2 Related works

There are many works in the literature addressing risk-aware optimization in MDPs. The most popular operator used to substitute ER is the expected utility (EU) criterion [25], that consists in replacing the reward values in ER by their subjective utility values. Thus, in order to determine an optimal policy following a risk-averse (resp. risk-seeking) attitude, EU is optimized using a concave (resp. convex) utility function [15, 23]. However, this operator has some descriptive limits, including the inability to explain the Allais paradox described above. Other criticisms of this operator can be found in [7, 12, 16].

Prospect theory [16], that generalizes EU, has been proposed to overcome the limits of ER and EU. The main idea is to use a transform function on probability values in order to express the subjective way agents perceive them. Nevertheless, this model has not been widely used in the literature because it can promote dominated solutions.

Another popular risk measure is the Conditional Value at Risk (CVaR), a.k.a. expected shortfall. It consists in optimizing the total mean reward in the worst α-fraction of runs, α being a parameter to be priorly fixed. This operator has been widely used to solve MDPs [5,8,9,22], but it has two major drawbacks: first,

it only allows to model more or less pessimistic attitudes w.r.t. risk, and second, it can promote dominated solutions since the evaluation is focused on the worst possible consequences. Authors of [20] indeed show that there could exist several policies that optimize the CVaR value, and thus, an algorithm optimizing CVaR, without considering all the consequences of the policies, would indifferently return one of them, whether dominated or not. Authors of [20] introduced in their paper a lexicographic method that returns the best ER policy among the CVAR-optimal ones, thus responding to the criticism of the dominated solutions. But the lack of flexibility in terms of descriptive capacity remains.

To sum up, many papers studied the question of modeling an agent's attitude w.r.t. risk, but these papers have two main limits that we want to overcome in this paper. First, they are focused on the representation of risk-averse attitudes while other types of behaviors can be observed (risk-seeking and intermediate behaviors). Secondly, they suffer from a lack of flexibility regarding the modeling of different levels of attitudes. An agent can in fact be more or less risk-averse than another, or even have a behavior mixing the two attitudes (as in the Allais paradox). There are therefore an infinite number of possible behaviors that are not modeled by many operators used in the literature.

In this paper, we will focus on the Weighted Ordered Weighted Average operator (WOWA for short) [26]. A parameterized operator that offers a rich descriptive power. It allows to model different levels of risk-aversion, risk-seeking, and intermediate behaviors. However, optimizing WOWA in an MDP is a challenging issue because of its non-linearity and time inconsistency. Thus, the use of standard algorithms as dynamic and linear programming is not possible (or at least not obvious). Note that there exists an LP formulation for optimizing WOWA [18], but this solution does not apply here. The linear program indeed only apply to cases were the number of scenarios/consequences is fixed and priorly known. In an MDP, as we will see in the next section, policies have a variable (see example 1) and unpredictable number of outcomes. We propose in this paper a ranking algorithm that allows to determine a WOWA-optimal or nearly WOWA-optimal policy. The idea of a ranking algorithm is not new. Outside the MDP framework, the authors of [6] proposed a ranking algorithm to determine a robust solution for the assignment and shortest path problems. They assume that the result of a decision is certain but the satisfaction of the agent is not. This uncertainty is related to the existence of a fixed and known number of scenarios. The adaptation of their method to MDPs is not easy as the context is different. The number of consequences is neither known or fixed in our case. In addition, their approach focused on the case of risk-averse behaviors, while we extend the method to deal with a larger range of behaviors, we will see next that this generalization implies additional difficulties.

To conclude this section, it is important to note that WOWA has already been used in the MDP framework but for multi-criteria decision making [17]. The main difference with our work lies in the policy evaluation, which implies both a descriptive difference and different algorithmic issues. They associate a reward vector to each pair (state, action) to describe the immediate reward of

the pair according to every considered criterion. A policy is then associated to a vector giving, for every criteria, the expected discounted reward of the policy. Thus assuming neutrality to obtain every criterion value. WOWA is only used to evaluate the compromise between criteria values. While in our paper, we consider only one reward function, and evaluate a policy with a lottery that summarizes all its possible consequences (thus not assuming neutrality). The lottery is then evaluated using WOWA. Thus, the difference between both works is similar to the difference with [6] except that one considers different criteria while the other considers different scenarios.

3 Backround and notations

We define a Risk aware MDP (R-MDP) by a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, H, f \rangle$ where $\mathcal{S}$ is a finite set of fully observable states; $\mathcal{A}$ is a finite set of actions; $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function where $\mathcal{P}(s, a, s')$ gives the probability of reaching state s' after performing action a in state s; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^+$ is the reward function where $\mathcal{R}(s, a, s')$ is the reward value obtained when performing action a in state s and reaching state s'; H is the maximum time step; and finally f is a function that model the attitude of the agent w.r.t. risk. Depending on the considered problem, an R-MDP can be defined with two additional elements: an initial state $s_0 \in \mathcal{S}$ and a set of final states $\mathcal{S}_f \subset \mathcal{S}$ if applicable.

A solution of an MDP is a policy $\pi : H \times \mathcal{S} \to \mathcal{A}$ that gives the action to perform in state $s \in \mathcal{S}$ at time step $h < H$. Due to the stochasticity of the process, the result of an action is uncertain, and thus, a policy π induces a lottery $\langle r_1^\pi : p_1^\pi, \ldots, r_{|\mathcal{T}_\pi|}^\pi : p_{|\mathcal{T}_\pi|}^\pi \rangle$ where $\mathcal{T}_\pi = \{t_1^\pi, \ldots, t_{|\mathcal{T}_\pi|}^\pi\}$ is the set of trajectories induced by π, each trajectory t_i^π occurring with probability p_i^π (defined by $\mathcal{P}$) and leading to a cumulative reward r_i^π (defined by $\mathcal{R}$). Note that we can obtain the lottery using a dynamic programming method similar to value iteration [4].

In order to evaluate policies, function f will generally be defined on its induced lottery. Thus, solving an R-MDP means determining the best policy according to f. We call such a policy an f-optimal policy. In the literature, the most commonly used operator to evaluate policies is the Expected Reward (ER):

$$\mathrm{ER}(\pi) = \sum_{i=1}^{|\mathcal{T}_\pi|} p_i^\pi r_i^\pi \tag{1}$$

As ER is linear, equation (1) can be formulated dynamically thanks to Bellman's equations. Thus, the value of each state $s \in \mathcal{S}$ at each time step $h \in \{0, \ldots, H-1\}$ is given by:

$$V_h^\pi(s) \quad = \quad \sum_{s' \in \mathcal{S}} \mathcal{P}(s, \pi(h, s), s') \quad \times \quad \left[\mathcal{R}(s, \pi(h, s), s') + V_{h+1}^\pi(s')\right] \tag{2}$$

with $V_H^\pi(s') = 0$. Using equation (2), an ER-optimal policy can be obtained with linear programming [11] or using the well known value or policy iteration [4, 14].

As illustrated in the introduction, ER's has descriptive limits that represent a serious drawback when it comes to model realistic behaviors in risky situations. Thus, we will not use it to model the agent preferences w.r.t. risk, but it is still useful as we will see in the remaining of the paper. The following section gives a definition of the operator we use to better model risk-aware agents as well as the motivation behind this choice.

4 WOWA model for risk-aware optimization

The Weighted Ordered Weighted Average (WOWA) model is also known as the Yaari's model [24,26] because it has been introduced and justified by Yaari in [26] in the context of decision making under risk. It is a parameterized model that generalizes ER and offers a much richer descriptive power. It is define as follows.

Definition 1 *For any policy π and transform probability function $\varphi : [0,1] \to [0,1]$ that is continuous, increasing and such that $\varphi(0) = 0$ and $\varphi(1) = 1$, we have:*

$$\mathrm{W}_\varphi(\pi) = \sum_{i=1}^{|\mathcal{T}_\pi|} \left[(r^\pi_{\sigma(i)} - r^\pi_{\sigma(i-1)}) \right] \varphi \left(\sum_{k=i}^{|\mathcal{T}_\pi|} p^\pi_{\sigma(k)} \right) \tag{3}$$

where $r^\pi_{\sigma(0)} = 0$, and σ is a permutation of $\{1, \dots, |\mathcal{T}_\pi|\}$ that reorders the elements of r^π in the increasing order of reward values, i.e., $r^\pi_{\sigma(1)} \leq \dots \leq r^\pi_{\sigma(|\mathcal{T}_\pi|)}$.

Example 3. Let us consider a policy π that induces the lottery : $\langle 0 : \frac{1}{3}, 10 : \frac{1}{2}, 15 : \frac{1}{6} \rangle$, and a function $\varphi(p) = p^2, \forall p \in [0,1]$. We have:

$$\mathrm{W}(\pi) = 0 + (10-0)\varphi(\frac{1}{2}+\frac{1}{6}) + (15-10)\varphi(\frac{1}{6}) \approx 4.58$$

The parameter function φ allows to model a subjective perception of the probability values. Note that, as long as φ is increasing on $[0,1]$, the preferences induced by W_φ are monotonic with respect to the first order stochastic dominance (FSD). FSD expresses the rational behavior of preferring ℓ to ℓ' as long as for all $x \in \mathbb{R}$, the probability of getting a reward higher than x with ℓ is greater than with ℓ'. In addition to this objective and rational preference, φ allows to control the type of decision behavior we want to model, depending on its specific shape: a convex function (resp. concave) allows to model risk-aversion (resp. risk-seeking) [13, 26]; and a linear function allows to model neutrality as $\mathrm{W}_\varphi(\pi) = ER(\pi)$ when $\varphi(p) = p, \forall p \in [0,1]$. We can also use S-shaped or inverse S-shaped functions in order to model more sophisticated behaviors. In particular, the function proposed by *Kahneman and Tversky* [16], defined by $\varphi(p) = \exp^{-\sqrt{-ln(p)}}, \forall p \in [0,1]$, allows to express the behavior highlighted by the Allais paradox. This is illustrated by the following examples.

Example 1 (continued) *Table 1 gives WOWA values of actions c and d for three different φ functions: $p^2, \sqrt{p}$ and $\exp^{-\sqrt{-ln(p)}}$ (noted kt[1]) for short:*

[1] after Kahneman and Tversky

action	W_{p^2}	$W_{\sqrt{p}}$	W_{kt}
c	6666.6667	**12247.4487**	7935.0431
d	**10000**	10000	**10000**

Table 1. Wowa values for actions c and d.

Here, unlike ER, WOWA gives different values for c and d. We can see that the convex function $\varphi(p) = p^2$ (as well as the Kahneman and Tversky function) favors d which is less risky, while the concave function $\varphi(p) = \sqrt{p}$ gives the opposite preference order.

The following example shows that the precise definition of φ allows to model different level of a certain type of preferences, as well as the behavior highlighted by the Allais paradox.

Example 4. Let us consider a simple sequential decision problem modeled by the decision tree pictured in figure 1, where circle nodes (s_0 to s_6) are states and rectangle nodes (a to d) are actions. For any transition (s, a, s'), the probability $\mathcal{P}(s, a, s')$ is given above the edge (a, s'), and the reward $\mathcal{R}(s, a, s')$ is given on the right of node state s'.

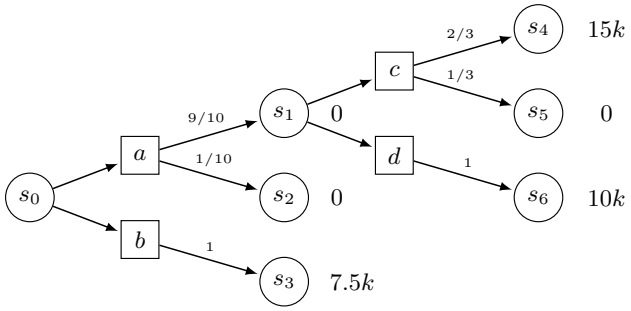


Fig. 1. Decision tree of example 4.

There are three possible policies: choosing a in s_0 and c in s_1 noted π_{ac}; choosing a in s_0 and d in s_1 noted π_{ad}; choosing b in s_0 noted π_b. The lotteries induced by these policies and their evaluations with ER and WOWA with 4 different φ functions are given in table 2.

The table illustrates several aspect of the descriptive power of WOWA: 1. it allows to discriminate between policies having equal ER values (π_{ac} and π_{ad} for example), but also to favor a policy that is strictly dominated w.r.t. ER (π_b and π_{ac} for example); 2. different shapes for φ can give different preference orders. We can also see that two functions modeling the same kind of behavior can also give different preference orders as they represent different levels of the behavior. For example, the two convex functions $\varphi(p) = p^2$ and $\varphi(p) = p^5$ do not give the

Policy	Induced lotteries	ER	W_{p^2}	W_{p^5}	$W_{\sqrt{x}}$	W_{kt}
π_{ac}	$\langle 15k : \frac{3}{5}, 0 : \frac{2}{5} \rangle$	**9000**	5400	1166.4	**11618.95**	7339.93
π_{ad}	$\langle 10k : \frac{9}{10}, 0 : \frac{1}{10} \rangle$	**9000**	**8100**	5904.9	9486.83	7228.21
π_b	$\langle 7.5k : 1 \rangle$	7500	7500	**7500**	7500	**7500**

Table 2. Lotteries of example 1 and their WOWA values.

same preference order for π_{ad} and π_b; 3. WOWA explains the Allais paradox with the Kahneman and Tversky function. We can indeed see that π_{ac} is preferred to π_{ad} while table 1 shows that c is preferred to d (note that the lotteries induced by π_{ac} and π_{ad} are the same as those induced by c' and d' in example 2); 4. the preference orders in the previous point also show that WOWA is not time consistent. Note that this is a consequence of the non-linearity of the operator, which comes from the sorting operation in WOWA.

The absence of time consistency has a dual effect. On the one hand it reflects a very realistic behavior, expressing (among other things) the Allais paradox. On the other hand, it prevents the use of practical tools as linear programming and value/policy iteration that are based on Bellman's equations. In the remaining of the paper, we introduce a ranking algorithm allowing to determine a WOWA-optimal or nearly WOWA-optimal policy.

5 Computing a WOWA-optimal policy

The main idea of the algorithm is to enumerate policies using a linear operator, that is easier to optimize, until a satisfactory policy has been found. The algorithm has the following steps:

1. determine a close linear bound for WOWA noted B (see subsection 5.1),
2. enumerate policies by decreasing order of B values (see subsection 5.2),
3. stop enumeration when we can prove that a (nearly) WOWA-optimal policy has been enumerated (see subsection 5.3).

The steps are detailed in the remaining of the section.

5.1 Bounding WOWA

The next proposition allows to define a bounding linear (on $\mathrm{ER}(\pi)$) function $B : \Pi \to \mathbb{R}^+$, where Π is the set of all possible policies.

Proposition 1 *Let $\pi \in \Pi$ be a feasible policy, $\mathcal{T}_\pi$ be the set of induced trajectories, and $\langle r_1^\pi : p_1^\pi, \dots, r_{|\mathcal{T}_\pi|}^\pi : p_{|\mathcal{T}_\pi|}^\pi \rangle$ be the associated lottery. For any linear function $g : [0,1] \to \mathbb{R}^+$ of the form $g(p) = ap + b$ where $a, b \in \mathbb{R}^+$ chosen such that $\varphi(p) \leq g(p), \forall p \in [0,1]$, we have:*

$$W_\varphi(\pi) \leq B(\pi) = aER(\pi) + b \max_{i \in \{1,\dots,|\mathcal{T}_\pi|\}} r_i^\pi \tag{4}$$

Proof. As $\varphi(p) \leq g(p) = ap + b, \forall p \in [0,1]$, and using equation (3) we have:

$$\begin{aligned}
\mathrm{W}_{\varphi}(\pi) &\leq \sum_{i=1}^{|\mathcal{T}_\pi|} \left[r^\pi_{(i)} - r^\pi_{(i-1)} \right] \left[a \left(\sum_{k=i}^{|\mathcal{T}_\pi|} p^\pi_{(k)} \right) + b \right] \\
&\leq a \sum_{i=1}^{|\mathcal{T}_\pi|} r^\pi_{(i)} \left[\sum_{k=i}^{|\mathcal{T}_\pi|} p^\pi_{(k)} - \sum_{k=i+1}^{|\mathcal{T}_\pi|} p^\pi_{(k)} \right] + b\ r^\pi_{(|\mathcal{T}_\pi|)} \leq a \sum_{i=1}^{|\mathcal{T}_\pi|} r^\pi_{(i)} p^\pi_{(i)} + b\ r^\pi_{(|\mathcal{T}_\pi|)} \\
&\leq a \sum_{i=1}^{|\mathcal{T}_\pi|} r^\pi_i p^\pi_i + b\ r^\pi_{(|\mathcal{T}_\pi|)} \leq aER(\pi) + b \max_{i \in \{1,\ldots,|\mathcal{T}_\pi|\}} r^\pi_t
\end{aligned}$$

Thus, to define B we only need to find values for a and b such that $\varphi(p) \leq ap + b$ in $[0,1]$. Note that there is an infinite number of values verifying this inequality. In order to obtain the most efficient algorithm possible, we will take values that makes g as close as possible to φ in $[0,1]$. Thus, for convex functions, we define g by $g(p) = p$, for concave and (inverse) S-shaped functions we will take a tangent line having the minimum distance to φ[2].

5.2 Enumerating policies

The idea of the enumeration method is simple, it consists in exploring and partitioning Π in a specific way to enumerate policies by decreasing order of B values, and this without missing any policy. For this, we will need to: 1. determine a procedure to efficiently partition Π and to explore its subsets. 2. find an optimal policy in a specific subset of Π.

Partitioning Π The procedure described below is very similar to the one proposed in [10]. The idea is simple: we first optimize B for the initial MDP (see the MILP formulation paragraph below), and we obtain a policy π^*_1. Using π^*_1 we partition Π as follows. Let us note E the support of π^*_1[3] and $\{e_1, ..., e_{|E|}\}$ the different elements of E. We define the set $\mathcal{X}^i, \forall i \in \{0, \ldots, |E|\}$, as

$$\{\pi \in \Pi | \pi(e_i) \neq \pi^*_1(e_i) \wedge \pi(e_k) = \pi^*_1(e_k), \forall k < i\}$$

It is easy to see that $\{\pi^*_1\} \cup \mathcal{X}^0 \cup \cdots \cup \mathcal{X}^{|E|}$ is a partition of Π. Thus, the second best policy of Π (according to B), noted π^*_2, will be the B-optimal policy of one of the sets $\mathcal{X}^i, i \in \{0, \ldots, |E|\}$. Thus, to find it we optimize B in every subset and take the best one. Let us note j the index of the set $\mathcal{X}^j$ containing π^*_2. The next step is to partition $\mathcal{X}^j$ using π^*_2 as we did for Π using π^*_1. The total partition of Π is then obtained by replacing $\mathcal{X}^j$ (in the initial partition) by the union of $\{\pi^*_2\}$ and the obtained partition of $\mathcal{X}^j$. These operations are repeated as many times as necessary to find a satisfying policy (see subsection 5.3).

[2] The distance of a function f to a function g on $[0,1]$ is defined by $\int_0^1 |f(x) - g(x)| dx$. Note that in our experiments, we used a sum to approximate this integral.

[3] The set of couples $(h, s) \in \{1, \ldots, H\} \times \mathcal{S}$ such that s is reachable at time h when we apply π^*_1.

Optimizing B To optimize the B value in each subset of the partition, we propose in the following a MILP formulation that is composed of the main MILP (given below), which solves the initial problem (considering Π as the set of possible policies), and additional constraints to determine the optimal policy for a subset $\mathcal{X}^i$ of Π. The main MILP is:

$$\max \sum_{h=0}^{H-1} \sum_{(s,a,s') \in \mathcal{T}} \mathcal{R}(s,a,s') \left(x^h_{sa} \mathcal{P}(s,a,s') + y^h_{sas'} \right)$$

$$\sum_a x^h_{sa} - \sum_{s'} \sum_a \mathcal{P}(s',a,s) x^{h-1}_{s'a} \leq \mathbb{1}_{[s=s_0,h=0]} \quad \forall s \in \mathcal{S}, h < H \tag{5}$$

$$y^h_{sas'} \leq x^h_{sa} + 1 - \epsilon \quad \forall s \in \mathcal{S}, a \in \mathcal{A}, h < H \tag{6}$$

$$\sum_{s'a} y^h_{s'as} = \sum_{s'a} y^{h+1}_{sas'} \quad \forall s \in \mathcal{S}, h < H \tag{7}$$

$$\sum_{sas'} y^h_{sas'} = 1 \quad \forall h < H \tag{8}$$

$$x^h_{sa} \geq 0 \quad \forall s \in \mathcal{S} \setminus \mathcal{S}_f, a \in \mathcal{A}, h < H - 1 \tag{9}$$

$$y^h_{sas'} \in \{0,1\} \quad \forall s, s' \in \mathcal{S} \setminus \mathcal{S}_f, a \in \mathcal{A}, h < H - 1 \tag{10}$$

Two type of variables are used:
Continuous variables x^h_{sa} that are the standard decision variables of the LP formulation of an MDP [11]. Their values can be interpreted as follows:

$$x^h_{sa} = \begin{cases} p(s,h \mid s_0) & \text{if } a \text{ is performed in state } s \text{ at time } h \\ 0 & \text{otherwise} \end{cases}$$

where $p(s,h \mid s_0)$ is the probability of reaching s at time h when s_0 is the initial state. Note that these variables will be implicitly constrained to be less than 1 (by definition of the MILP).
Binary variables $y^h_{sas'}$ that indicates whether the transition (s,a,s') is in the maximum reward trajectory induced by the policy (the max term in proposition 1) or not.

The objective function of the MILP expresses the linear function B defined in (4). In this expression, $\mathcal{T}$ is the set of all probable transitions, i.e., all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ such that $\mathcal{P}(s,a,s') > 0$. The first constraint is the standard constraint of the MDP's LP formulation [11]. The second constraint expresses the fact that a transition (s,a,s') (at time h) can be in the maximum reward trajectory only if a is performed in s at time h in the policy. Here we consider that a probability is significant only if it is greater than a given ϵ. The third constraint expresses the fact that a the trajectory gets out of a state $s \in \mathcal{S}$ at time $h+1$ if and only

if it enters in at time h. The forth constraint expresses the fact that only one transition is crossed at time h.

An optimal solution to the defined MILP gives a B-optimal solution for the initial MDP. In order to determine the optimal policy for a subset $\mathcal{X}^i$ of Π, we need to define some additional constraints:

- for any tuple (h, s, a) for which there is a constraint $\pi(h, s) \neq a$:

$$x^h_{sa} = 0$$

- for any tuple (h, s, a) for which there is a constraint $\pi(h, s) = a$, we add the constraint $\pi(h, s) \neq a', \forall a' \neq a$ to prevent the choice of another action[4].

5.3 Stopping condition

Proposition 2 gives a condition that, when fulfilled at a given step of the ranking algorithm, guaranties that a WOWA-optimal policy has been found.

Proposition 2 *Let Π be the set of all possible policies, and let $\Pi^k = (\pi^1, \ldots, \pi^k)$ be the list of the k best elements of Π according to the linear bound B. We have:*

$$\max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi) > B(\pi^k) \Rightarrow \max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi) = \max_{\pi \in \Pi} \mathrm{W}_\varphi(\pi)$$

Proof. We will prove that, at enumeration k, if there exists a policy $\pi' \in \arg\max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi)$ satisfying $\mathrm{W}_\varphi(\pi') > B(\pi^k)$, then any policy in $\Pi \setminus \Pi^k$ cannot be WOWA-optimal: a policy $\pi'' \in \Pi \setminus \Pi^k$ is such that $B(\pi'') \leq B(\pi^k)$ (otherwise it would have been enumerated before π^k). From proposition 1 we have $\mathrm{W}_\varphi(\pi'') \leq B(\pi'')$, we then deduce that $\mathrm{W}_\varphi(\pi'') \leq \mathrm{W}_\varphi(\pi') = \max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi)$, which concludes the proof.

The enumeration can then be stopped as soon as this condition is fulfilled. Note that this algorithm is anytime, we can, at any step k, stop the enumeration and return the best found policy (a policy in $\arg\max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi)$) with $B(\pi^k) - \max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi)$ as a bound on the distance to the optimal value.

Finally, we could relax the stopping condition and stop enumerations as soon as $\max_{\pi \in \Pi^k} \mathrm{W}_\varphi(\pi) > B(\pi^k) - \delta$ where δ is a prefixed threshold in order to save some computation time and determine a δ-optimal policy.

Proposition 3 *The ranking algorithm ends and returns an (δ-)optimal solution.*

The optimality results from proposition 2, while the termination results from the definition of the algorithm: either the stopping condition is verified and the algorithm stops, or the condition is never fulfilled and all the policies of Π are enumerated exactly once. Since Π is a finite set, the algorithm necessarily stops.

[4] Note that we could instead define a constraint $x^h_{sa} \geq \epsilon$ but our experiments showed that this option is slower.

6 Experimental results

We have implemented[5] and tested[6] the ranking algorithm introduced in the previous section, and we give in this section a part of the obtained results.

We tested the algorithm on 100 randomly generated[7] MDPs with 10 states, 3 actions, and $H = 5$. In order to simulate the attitude of an agent w.r.t. risk, we used a WOWA operator with multiple parameter functions φ: $p^5, p^{0.25}$ and $\exp -\sqrt{-ln(p)}$. The histograms on Figure 2 give the rank (in the enumeration) of the WOWA-optimal policy. These histograms focus on the first 1000 enumerations, but we can see that this was not much constraining as, most of the time, the optimal policy is found before 1000 enumerations. Finally, the figure in the bottom right of Figure 2 shows the evolution of ER (or more accurately $aER + b$), WOWA, and CVaR values throughout the running of the ranking algorithm on a specific instance[8].

We can see in the histograms that, for any considered φ function, the WOWA-optimal policy is generally different from the ER-optimal policy. Note that there is a proportion of instances for which the first enumerated policy is optimal. This is due to the fact that random generation often produces uninteresting instances w.r.t. decision under risk, as it is the case when, for example, there exists a policy that (strongly) dominates all the other ones, regardless of the considered type of preferences. Besides these instances, the rank of the optimal policy is relatively well distributed between rank 2 and rank 1000. Thus, the WOWA-optimal policy is often far from the ER-optimal one in term of ranks. Considering the figure on the bottom right, it clearly shows that WOWA can discriminate between solutions considered equivalent either by ER or by the CVaR operator. It also shows that the stopping condition can take time to be verified (more than 1000 enumerations here), but the approach has the advantage of allowing the determination of a policy in a more accurate (and may be less hazardous) way than by simply optimizing ER and/or CVaR with value iteration or lexicographic algorithms such as the one in [20].

7 Conclusion

We have introduced an algorithm to solve risk-aware MDPs. This work differentiates from previous works in the MDP literature in proposing a more flexible risk measure that has never been used (as far as we know) in MDPs as a measure of risk. In comparison with the decision making under risk literature, this work provides an extension to sequential decision problems and to a more general case in term of preference modeling. Similar works indeed focus on the case of risk-averse agents.

[5] The implementation was performed in Python, and the linear programs were solved using the `gurobi` Python library.
[6] Tests were performed on an Intel(R) Core(TM) i7-1165G7 CPU with 15 GB of RAM.
[7] using the `mdptoolbox` library: `https://pymdptoolbox.readthedocs.io`
[8] A betting game instance, a definition can be found in [20]

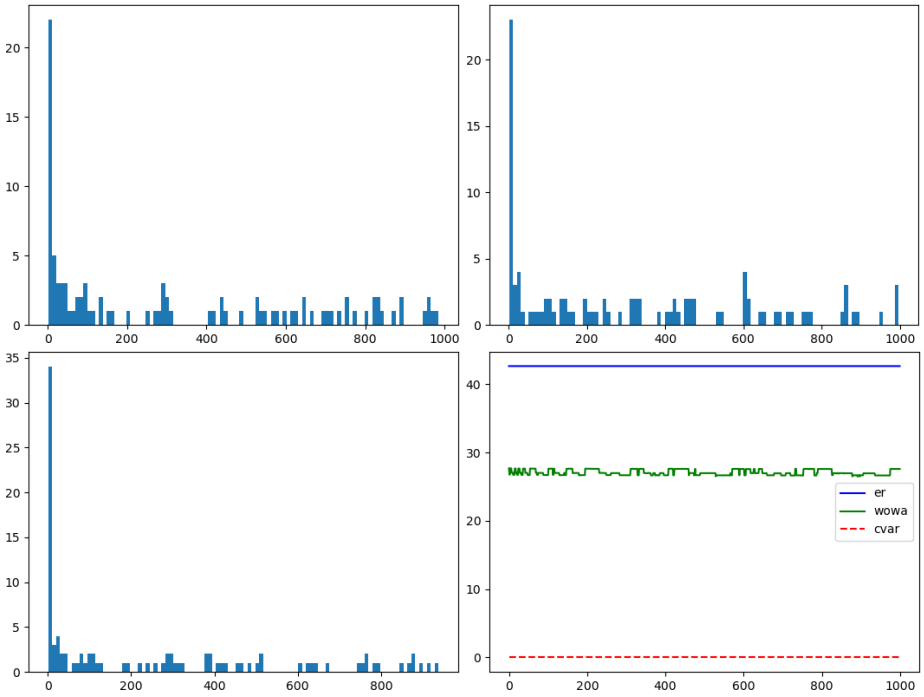


Fig. 2. Rank of the optimal policies for $\varphi(p) = p^5$ (top left), $\varphi(p) = p^{0.25}$ (top right), $\varphi(p) = e^{-\sqrt{-ln(p)}}$ (bottom left) and an execution trace of the algorithm (bottom right).

In this work, we make the assumption that the attitude of the agent w.r.t. risk is precisely known, and that the parameter function φ can be fixed accordingly. However, this assumption is restrictive and may be unrealistic in some situations. Thus, our next step is to extend this approach to the case of imprecisely known preferences w.r.t. risk. This is a challenging task because if φ is not fixed, its bound cannot be defined with the same precision, and consequently, it is harder to determine an efficient bound on WOWA.

References

1. Ahiska, S.S., Appaji, S.R., King, R.E., Warsing Jr, D.P.: A markov decision process-based policy characterization approach for a stochastic inventory control problem with unreliable sourcing. International Journal of Production Economics **144**(2), 485–496 (2013)
2. Alexander, G.J., Baptista, A.M.: A comparison of var and cvar constraints on portfolio selection with the mean-variance model. Management science **50**(9), 1261–1273 (2004)

3. Allais, M.: Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. Econometrica: journal of the Econometric Society pp. 503–546 (1953)
4. Bellman, R.: A markovian decision process. Journal of Mathematics and Mechanics **6**(5), 679–684 (1957)
5. Borkar, V., Jain, R.: Risk-constrained markov decision processes pp. 2664–2669 (2010)
6. Bourdache, N., Perny, P.: Anytime algorithms for adaptive robust optimization with owa and wowa pp. 93–107 (2017)
7. Chateauneuf, A., Cohen, M., Meilijson, I.: Four notions of mean-preserving increase in risk, risk attitudes and applications to the rank-dependent expected utility model. Journal of Mathematical Economics **40**(5), 547–571 (2004)
8. Chow, Y., Ghavamzadeh, M.: Algorithms for cvar optimization in mdps. Advances in neural information processing systems **27** (2014)
9. Chow, Y., Tamar, A., Mannor, S., Pavone, M.: Risk-sensitive and robust decision-making: a cvar optimization approach. Advances in neural information processing systems **28** (2015)
10. Dai, P., Goldsmith, J.: Finding best k policies pp. 144–155 (2009)
11. Denardo, E.V.: On linear programming in a markov decision problem. Management Science **16**(5), 281–288 (1970)
12. Gonzales, C., Perny, P.: Decision under uncertainty. A Guided Tour of Artificial Intelligence Research: Volume I: Knowledge Representation, Reasoning and Learning pp. 549–586 (2020)
13. Hong, C.S., Karni, E., Safra, Z.: Risk aversion in the theory of expected utility with rank dependent probabilities. Journal of Economic theory **42**(2), 370–381 (1987)
14. Howard, R.A.: Dynamic programming and markov processes. (1960)
15. Howard, R.A., Matheson, J.E.: Risk-sensitive markov decision processes. Management science **18**(7), 356–369 (1972)
16. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica **47**(2), 263–291 (1979)
17. Ogryczak, W., Perny, P., Weng, P.: A compromise programming approach to multiobjective markov decision processes. International Journal of Information Technology & Decision Making **12**(05), 1021–1053 (2013)
18. Ogryczak, W., Śliwiński, T.: On efficient wowa optimization for decision support under risk. International Journal of Approximate Reasoning **50**(6), 915–928 (2009)
19. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Inc., USA, 1st edn. (1994)
20. Rigter, M., Duckworth, P., Lacerda, B., Hawes, N.: Planning for risk-aversion and expected value in mdps **32**, 307–315 (2022)
21. Rigter, M., Lacerda, B., Hawes, N.: Risk-averse bayes-adaptive reinforcement learning. Advances in Neural Information Processing Systems **34**, 1142–1154 (2021)
22. Rockafellar, R.T., Uryasev, S., et al.: Optimization of conditional value-at-risk. Journal of risk **2**, 21–42 (2000)
23. Rothschild, M., Stiglitz, J.E.: Increasing risk: I. A definition. Elsevier (1978)
24. Torra, V.: The weighted owa operator. International journal of intelligent systems **12**(2), 153–166 (1997)
25. Von Neumann, J., Morgenstern, O.: Theory of games and economic behavior, 2nd rev (1947)
26. Yaari, M.E.: The dual theory of choice under risk. Econometrica: Journal of the Econometric Society pp. 95–115 (1987)