POMDP solving: what rewards do you really expect at execution?

Caroline Ponzoni Carvalho CHANEL [a,b] Jean-Loup FARGES [a] and Florent TEICHTEIL-KÖNIGSBUCH [a] and Guillaume INFANTES [a]
[a] *ONERA - Office National d'Etudes et de Recherches Aérospatiales, Toulouse, France. Email: name.surname@onera.fr*
[b] *ISAE - Institut Supérieur de l'Aéronautique et de l'Espace*

Abstract. Partially Observable Markov Decision Processes have gained an increasing interest in many research communities, due to sensible improvements of their optimization algorithms and of computers capabilities. Yet, most research focus on optimizing either average accumulated rewards (AI planning) or direct entropy (active perception), whereas none of them matches the rewards actually gathered at execution. Indeed, the first optimization criterion linearly averages over all belief states, so that it does not gain best information from different observations, while the second one totally discards rewards. Thus, motivated by simple demonstrative examples, we study an additive combination of these two criteria to get the best of reward gathering and information acquisition at execution. We then compare our criterion with classical ones, and highlight the need to consider new hybrid non-linear criteria, on a realistic multi-target recognition and tracking mission.

Keywords. POMDP, active perception, optimization criterion.

Introduction

Many real-world AI applications require to plan actions with incomplete information on the world's state. For instance, a robot has to find its way to a goal but without perfect sensing of its current localization in the map. As another example, the controller of a camera must plan the best optical tasks and physical orientations to precisely identify an object as fast as possible. If action effects and observations are probabilistic, Partially Observable Markov Decision Processes (POMDPs) are an expressive but long-neglected — due to prohibitive complexity — model for sequential decision-making with incomplete information [5]. Yet, new recent strides in POMDP solving algorithms [8,12,13] have revived an intensive research on algorithms and applications of POMDPs.

A POMDP is a tuple $\langle S, A, \Omega, T, O, R, b_0 \rangle$ where S is a set of states, A is a set of actions, Ω is a set of observations, $T : S \times A \times S \to [0; 1]$ is a transition function such that $T(s_t, a, s_{t+1}) = P(s_{t+1} \mid a, s_t)$, $O : \Omega \times S \to [0; 1]$ is an observation function such that $O(o_t, s_t) = P(o_t | s_t)$, $R : S \times A \times S \to \mathbb{R}$ is a reward function associated with transitions, and b_0 is a probability distribution over initial states. We note

B the set of probability distributions over the states, named *belief state space*. At each time step t, the agent updates its *belief state* defined as an element $b_t \in B$.

The aim of POMDP solving is to construct a policy function $\pi : B \to A$ such that it maximizes some criterion generally based on rewards or belief states. In robotics, where symbolic rewarded goals must be achieved, it is usually accepted to optimize the long-term average discounted accumulated rewards from any initial belief state [2,11]: $V^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b\right]$. Following from optimality theorems, the optimal value function is piece-wise linear, what offers a relatively simple mathematical framework for reasoning, on which most, if not all, algorithms are based. However, as highlighted and explained in this paper, the linearization of belief states' average value comes back to flatten observations and finally to loose distinctive information about them. Therefore, the optimized policy does not lead the agent to acquire sufficient information about the environment before acting to gather rewards: as discussed in this paper, such a strategy unfortunately results in less reward gathering at execution than expected if the initial belief state is very far from actual state.

This confusing but crucial point deserves more explanations for better understanding of what is at stake in this paper. At first, it may seem strange that the strategy which maximizes accumulated rewards is not optimal at actual execution: in what sense are the optimized criterion and the rewards gathered at execution different? In fact, the average accumulated rewards criterion is defined over belief states (because the agent applies a strategy based only on its belief), whereas the rewards gathered at execution are accumulated on the basis of the actual successive states, hidden from the agent. With total observability (MDP case), such issue does not arise since actual states are observed, so the criterion is averaged over actual probabilistic execution paths. But in the POMDP case, the criterion is *averaged over probabilistic believed paths*, which are generally different from the *actual* execution paths. Strangely enough, this bias between optimized criterion and actual rewards gathered at execution has not been much studied: to our knowledge, most robotics research on POMDPs has considered more and more efficient methods to optimize this average accumulated rewards criterion, despite the lack of explicit separation between possible observations during optimization.

In spite of better explaining the idea raised here, we seek to show the existence of a Δ, more or equal à zero, defined by 1, that expresses the difference between the criterion optimized by the classical POMDP framework and the rewards cumulated at policy execution.

$$\Delta = \left| E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(b_t)) | b_0\right] - E\left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) | b_0\right] \right| \tag{1}$$

b_t represents de belief state, i.e. the probability distibution over states at an instant t (at each time step b_t updated with the Bayes' rule after each action done and observation perceived). And s_t represents de hidden state of the system, and depends only on the dynamic of the system. The difference is more or equal to zero every time step. Equal to zero for anytime step in which the agents' belief is a Dirac's delta over a state of the system ($b_t = \delta_{s_t}$), and more than zero otherwise

($b_t \neq \delta_{s_t}$). Formally, $r(b_t, \pi(b_t))$ is defined by: $r(b_t, \pi(b_t)) = \sum_s r(s, \pi(b_t)) b_t(s)$, and we introduce it in the Eq. (1):

$$\Delta = \left| E\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(b_t)) | b_0 \right] - \sum_{t=0}^{\infty} \gamma^t \left(\sum_s r(s, \pi(b_t)) b_t(s) \right) | b_0 \right] \right| \tag{2}$$

Using the norm and the expected value properties, we get:

$$\begin{aligned} \Delta &= \left| E\left[\sum_{t=0}^{\infty} \gamma^t \left(r(s_t, \pi(b_t)) - \left(\sum_s r(s, \pi(b_t)) b_t(s) \right) \right) | b_0 \right] \right| \\ &\leq \left| E\left[r(s_0, \pi(b_0)) - \left(\sum_s r(s, \pi(b_0)) b_0(s) \right) | b_0 \right] \right| + \cdots + \\ &\quad \gamma^t \left| E\left[r(s_t, \pi(b_t)) - \left(\sum_s r(s, \pi(b_t)) b_t(s) \right) | b_0 \right] \right|, \; t \to \infty \end{aligned} \tag{3}$$

$\sum_s r(s, \pi(b_t)) b_t(s)$ clearly average rewards $r(s, \pi(b_t))$ over states. More precisely, for a given state s_n and a given time step t, if $b_t = \delta_{s_n}$, the reward will be $\sum_s r(s, \pi(b_t)) b_t(s) = r(s_n, \pi(b_t))$, on the contrary for a $b_t \neq \delta_{s_n}$ the reward will be different of $r(s_n, \pi(b_t))$. Denoting:

$$\triangle R_t(s_t, b_t) = E\left[r(s_t, \pi(b_t)) - \left(\sum_s r(s, \pi(b_t)) b_t(s) \right) | b_0 \right]$$

and re-writing Eq. (3), we obtain:

$$\Delta \leq \triangle R_0(s_0, b_0) + \gamma \triangle R_1(s_1, b_1) + \cdots + \gamma^t \triangle R_t(s_t, b_t), \text{ with } t \to \infty \tag{4}$$

If for a given time step $t = k$ we have $b_k = \delta_{s_k}$, it is easy to see that $\triangle R_k(s_k, b_k) = 0$. This allows us to infer that if $b_0 \neq \delta_{s_0}$, i.e, the probability distribution over states is not a Dirac's delta over the initial hidden state s_0, the difference Δ is more than zero already in $t = 0$. And successively, for all time steps where $b_t \neq \delta_{s_t}$.

On the other hand, researchs on active sensing aim at maximizing knowledge of the environment [3,4,7]; thus minimizing *Shannon's entropy* criterion, which assesses the accumulated quantity of information in the initial belief state b_0: $H(b_0) = \sum_{t=0}^{+\infty} \gamma^t \sum_{s \in S} b_t(s) \log(b_t(s))$. Contrary to the previous criterion, this criterion is non-linear over belief states so it makes a clear distinction between observations to promote one that update the belief state in the right direction. But this criterion does not take into account rewards, so it is not appropriate for goal reaching problems.

Thus, considering approaches from both research communities, it is natural to search for new non-linear reward-based optimization criteria by aggregating the average accumulated rewards criterion and the entropy one into a single mixed criterion. This way, optimized strategies would consist in alternating information acquisition and state-modification actions to maximize reward gathering at execu-

tion, provided both criteria are appropiately balanced. Formally, noting $J_\lambda(V,H)$ a mixed criterion depending on some $\lambda \in \Lambda$ parameter, the *general problem we address* is formalized as follows:

$$\max_{\lambda\in\Lambda} E\left[\sum_{t=0}^{+\infty}\gamma^t r_t \mid s_0,\pi_\lambda\right] \text{ such that } \pi_\lambda = \underset{\pi\in A^S}{\operatorname{argmax}}\, J_\lambda(V(b_0),H(b_0))$$

In other words, what is the value λ balancing $V(b_0)$ and $H(b_0)$ that maximizes the average accumulated rewards gathered at execution, starting from an initial state s_0 unknown to the agent, when applying the policy that maximizes the mixed criterion based on the agent's initial belief state? Solutions to this problem depend on the class of functions to which J_λ belongs. Yet, even for simple classes like $\{J_\lambda : J_\lambda(V,H) = (1-\lambda)V + \lambda H, 0 \leqslant \lambda \leqslant 1\}$, we could not find algebraic general solutions. Some authors studied applications of such criterion for some fixed λ with 1-step optimization of the entropy [1]. Others formalized active sensing problems as POMDP optimization based on the previous class, but without solving them nor studying the impact of λ on rewards gathered at execution [6].

A recent work [10] considers the problem of dynamical sensor selection in camera networks based on user-defined objectives, such as maximizing coverage or improved localization uncertainty. The criterion optimized is the POMDP classical one, but the key of this work relies in the model of the reward function. For exemple, for improving localization uncertainty, the authors use the determinant of the variance matrix as additional information in the reward function. This variance matrix is obtained for each sensor and possible location of the target. In this way, the reward function continues to be linear, and the classical criterion is applied.

Therefore, in the next section, we highlight the importance of mixed non-linear criteria as introduced above on a simple but illustrative example. We show the impact of different values of λ on rewards gathered at execution, depending on the initial belief state. Then, in the next section, we formally define an additive criterion that may be of interest for better optimization of POMDP robotics problems. Finally, before concluding the paper, we point out the relevance of considering non-linear mixed criteria on a realistic multi-target recognition and tracking robotics mission, which we solved with a state-of-the-art POMDP planner modified for our new criterion.

1. Illustrative Examples

This section intends to study the difference of behavior obtained at execution by modifying the classical POMDP's criterion on a given problem. The objective is to show that the change of criterion induces agent caution in relation to its belief state, reducing the chances of mistake at policy execution.

Let us define a problem with four states $\{s_0, s_1, s_2, s_3\}$ and two observations $\{o_1, o_2\}$. Initially, the agent can be in s_0 or s_2, so that $b_0(s_0) = 1 - b_0(s_2)$, and o_1 (resp. o_2) corresponds to observe if it is in s_0 (resp. s_2). It can perform three actions: a_0 is a perception action that costs c and does not change the state,

while a_1 and a_2 deterministically lead to absorbing states as shown in Figure 1. Depending on the actual state s_0 or s_2, actions a_1 and a_2 give opposite rewards (either R or $-R$), meaning that a_1 should be chosen if actual state is s_0, a_2 if it is s_2. Note that $R > c > 0$.

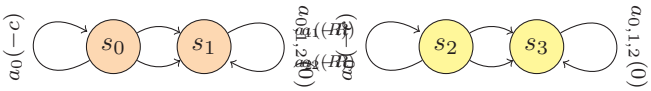

Figure 1. Transitions of the POMDP (rewards between parenthesis)

Intuitively, there are two "good" strategies here, depending on the initial belief state:
- try to avoid the observation cost and directly choose a_1 or a_2;
- first observe with action a_0 then act with action a_1 or a_2.

The observation matrix is defined as: $p(o|s') = \begin{bmatrix} 1 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 1 & 0.5 \end{bmatrix}$ We can compute the Q-values over $b(s)$, which are the best values of each action if the optimal policy is applied next. Note that, in this simple example, the optimal policy is obvious after the first action, starting either in s_0 or s_2. Q-values depend on $b_0(s_0)$ and $b_0(s_2)$:

$$Q^\pi(b, a_0) = (R - c)(b_0(s_0) + b_0(s_2))$$
$$Q^\pi(b, a_1) = R(b_0(s_0) - b_0(s_2))$$
$$Q^\pi(b, a_2) = R(b_0(s_2) - b_0(s_0))$$

Q-values over $b_0(s_0)$ are shown in Figure 2-left along with the value function, which is the best Q-value (for $R = 1$ and $c = 0.5$). We see that the optimal policy depends on the initial belief state, as expected. Also, is represented the actual value gathered by the agent according to its initial belief when the initial state of the system is s_0.

1.1. Criterion Modification

Now, we add Shannon's entropy of $b(s)$ to the criterion at every time step, i.e. we add the expected entropy $H^\pi(b)$ denoted by : $H^\pi = \sum_{t=0}^{N} H(b_t)$. And so as, the new criterion becomes: $J^\pi(b, \lambda) = (1 - \lambda)V^\pi(b) + \lambda H^\pi(b)$ The value of the belief state entropy $H(b)$ almost does not change when the first strategy is executed. In other hand, when the second strategy is chosen, the entropy lowers to zero at the second step. After taking action a_1 or a_2, the entropy value decreases, but less than after action a_0 which brings the entropy to zero. So, the mixed criterion of the first strategy is more penalized than the one of the second strategy, because it takes into account the total value of entropies (at $t = 0$ and $t = 1$).

$$Q^\pi(b, a_0) = (1 - \lambda)(R - c)(b_0(s_0) + b_0(s_2)) + \lambda H(b_0)$$
$$Q^\pi(b, a_1) = (1 - \lambda)R(b_0(s_0) - b_0(s_2)) + \lambda(H(b_0) + H(b_1))$$
$$Q^\pi(b, a_2) = (1 - \lambda)R(b_0(s_2) - b_0(s_0)) + \lambda(H(b_0) + H(b_1)))$$

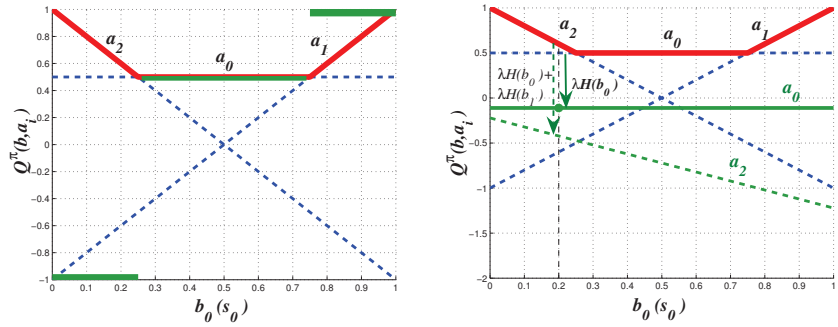

Figure 2. Left: Q-values (in blue), value function (in red), also actual value gathered by the agent (in green) when system's initial state is s_0, all over $b_0(s_0)$. Right: Q-values for $b_0(s_0) = 0.2$ before (blue) and after (green) criterion modification.

In order to illustrate the change in the criterion, we have computed the α-vectors for a given belief state. In the Figure 2-right the Q-values for the classical criterion and the Q-values for the modified criterion are presented for a $b_0(s_0) = 0.2$. It can be verified that the addition of the weighted entropy changes the gradient of the α-vectors. The new criterion penalizes much more the first strategy than the second one. In other words, when the weighted entropy is taken into account for this belief, the new criterion reflects in the Q-value for this belief the uncertainty, and brings on the α-vector for a_0 as dominant.

To show the change in the criterion and as a consequence, the change in the shape of the value function, we have computed the best mixed criterion in function of $b_0(s_0)$ for different values of λ. Figure 3-left shows that the mixed criterion's shape changes a lot while varying the value of λ from zero to one: the higher the λ value is, the more the first strategy gets penalized. It also shows that the criterion is no longer linear.

Some assumptions can be overcome with this change. Figure 3-right presents the actual rewards gathered by the agent when it acts based on $b(s)$, without knowing it is actually in state s_0 at the beginning. Note the differences with

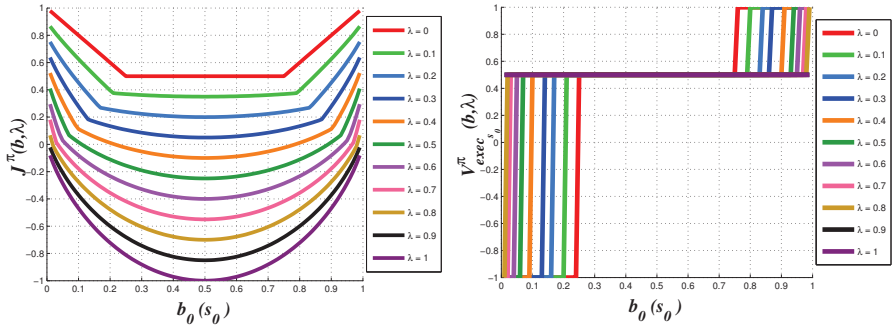

Figure 3. Left: best mixed-criterion based on the agent's initial belief for different λ values (λ increases from top curves to bottom ones); Right: rewards gathered at execution for different λ values depending on the agent's initial belief when the initial state of the system is s_0– which determines its policy (λ increases from inside to outside).

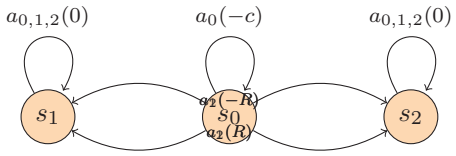

Figure 4. Counter example of entropy addition in criterion.

the average rewards which the agent believes to gather on Figure 2-left. The closer λ is to one, the more the agent prefers observing first, so that it is less penalized if its initial belief is wrong (0.5 instead of -1). But the rewards gathered if it is right decrease also (0.5 instead of 1). So, we would like to establish some degree of confidence over $b(s_0)$ and then figure out the appropriate λ for the problem. For this simple example, we can calculate a function $\lambda = f(b_{s_0})$, with $b_{s_0} = b_0(s_0)$, using the point where $Q^\pi(b, a_0) = Q^\pi(b, a_2)$ and taking advantage that $H(b_1) = H(b_0)$ when action a_2 is done.

$$\lambda(b_{s_0}) = \frac{2Rb_{s_0} - c}{2Rb_{s_0} - c + b_{s_0}\ln(b_{s_0}) + (1 - b_{s_0}))\ln(1 - b_{s_0}))}$$

This kind of modification of criterion is necessary when the agent's initial belief (or prior) $b_0(s)$ does not correspond to real frequencies of the initial states. In real situations, this kind of mistake often happens: the $b_0(s)$ used for the policy calculation may not be the best approximation of the reality.

A counter example is detailed in Figure 4. We see that there is no gain in adding the belief state entropy value at every time-step, because there is no ambiguity in the initial state this case. The value function only depends on the arrival state and it will be equally penalized by the belief state entropy for each action.

$$Q^\pi(b, a_0) = R|b_0(s_1) - b_0(s_2)| - cb_0(s_0)$$
$$Q^\pi(b, a_1) = Q^\pi(b_0, a_2) \;=\; R|b_0(s_1) - b_0(s_2)|$$

In the following, a mixed non-linear criterion for POMDPs is presented and the modification we made to the state-of-the-art algorithm Symbolic-PERSEUS [9] to optimize it. The next sections present some results obtained by modeling and solving a simple realistic problem which confirms the intuitions raised in this example.

2. Hybrid Optimization Criterion for POMDPs

The criterion proposed in this section models the expected cumulative discounted reward, attributed to the chosen actions, added to the expected cumulative discounted entropy of the belief (computed over the successive stochastic belief states), both in infinite-horizon. This two values are themselves weighted by a constant λ:

$$J^{\pi}(b) = (1-\lambda)V^{\pi}(b) + \lambda H^{\pi}(b), \text{ with} \tag{5}$$
$$V^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(b_t, \pi(b_t)) \mid b_0 = b\right]$$
$$H^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t H(b_t) \mid b_0 = b\right]$$

Theorem : Bellman's equation of the additive criterion. *The optimal value function of the additive criterion is the limit of the vector sequence defined by:*

$$J_{n+1}(b) = \max_{a \in A}\left\{(1-\lambda)r(b,a) + \lambda H(b) + \gamma \sum_{o \in \Omega} p(o|b,a) J_n(b_a^o(s'))\right\} \tag{6}$$

Proof. The equation 5 can be rewritten as:

$$J^{\pi}(b) = E_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(\beta r(b_t, \pi(b_t)) + \rho H(b_t)\right) \mid b_0 = b\right] \tag{7}$$

This shows that this criterion corresponds to the classic γ-discounted criterion which is the current reward added to the current entropy of the belief. It is therefore a maximization problem over γ-discounted artificial rewards equals to the real rewards plus the actual belief's entropy. □

This new Bellman's equation permits *via* dynamic-programming the computation of a policy that weights the immediate reward by the immediate entropy of the belief.

Heuristic. Symbolic-PERSEUS uses a heuristic function to determine the set of reachable belief states. It initializes the belief states search with $V^{\pi}_{degrad} = \max_{s,a} r(s,a)$, calculated from a depleted model, e.g. with an admissible heuristic whose value must be smaller than the optimal value. The definition of the new criterion therefore requires to change this heuristic as well, in order to take into account a minimal value for H in the initialization of the reachable beliefs search calculation. The heuristic is now defined by J_0 shown below.

Theorem: Heuristic for the additive criterion. *An admissible heuristic to the additive criterion is given by:*

$$J_0 = \frac{(1-\lambda)V^{\pi}_{degrad} - \lambda \log_{10}(n)}{1-\gamma} \tag{8}$$

Proof. The minimal value to $H(b)$ constrained by $\sum_{i=1}^{n} b(s_i) = 1$ is given by the Lagrangian optimization:

$$H(b)_{min} = n\frac{1}{n}\log_{10}\left(\frac{1}{n}\right) = -\log_{10}(n) \tag{9}$$

$$\text{So: } J^{\pi} \geq \frac{(1-\lambda)V^{\pi}_{degrad} - \lambda \log_{10}(n)}{1-\gamma} = J_0 \tag{10}$$

□

Discussion. The criterion presented in this section is no more piecewise linear, but algorithms such as PBVI [8], PERSEUS [11] and Symbolic-PERSEUS [9], which approach the criterion by stochastic generation of local belief states, can approximate this nonlinear criterion, given that every function can be approximate by a piecewise linear function.

3. Robotics Example

The studied model deals with an autonomous helicopter that tries to identify and track two targets. These targets are of different types, A or B. Objective of the helicopter is to land onto the target of type A, without initially knowing types of targets . This scenario combines both mission and perception objectives. Thus, this is relevant for this work because the reward optimization (implicitly) implies reducing the belief's entropy: actually, it is necessary to reduce uncertainty over the nature of targets in order to achieve the mission. Initially, the autonomous helicopter has an *a priori* knowledge about the targets. It needs to track and identify each target by its actions in order to accomplish its goal. In the simulations studied in this work, the agent was given an initial belief state weighted and not uniforme over all possible combinations of targets types. The initial belief's values with respect to the targets types are shown in table 1.

Table 1. Initial belief about the targets.

Targets	A	B
target 1	0.2	0.8
target 2	0.8	0.2

The motion space of the helicopter is modeled by a $3 \times 3 \times 3$ grid, and the one of the targets by a $3 \times 3 \times 1$ grid: the targets moves only on the ground (see Figure 5). The helicopter can do 7 actions: forward in x, y and z (go up), backward in x, y and z (go down), and land. It cannot realizes more than one action at any time-step. Motions of helicopter can fail with a 10% probability, except the land action which always succeeds. Target positions are completely observable to the helicopter agent. Nevertheless, the targets change position in x and/or y, 1 time for 20 helicopter motions, but the latter cannot predict the evolution. Helicopter is allowed to land only if it is directly above a target ($z = 2$, as ground is $z = 1$). Once the helicopter has landed on the target (A) or (B), neither the helicopter nor the targets can move, and helicopter is not allowed to take-off. The observation model of the type of the targets depends on the euclidean distance from helicopter to targets as show below.

$$p(o' = A \mid s' = A) = \frac{1}{2}\left(e^{\frac{-d}{D}} + 1\right) \text{ and } p(o' = A \mid s' = B) = \frac{1}{2}\left(1 - e^{\frac{-d}{D}}\right)$$

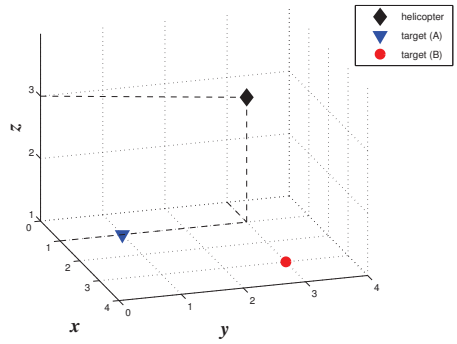

Figure 5. Initial position of helicopter, target 1 (A) and target 2 (B).

where d is the euclidean distance between helicopter and target, and D a factor of adjustment of the exponential. This observation function allows to model the gain of information when the helicopter comes near to the target : the closer the helicopter is from the observed target, the higher the probability to observe the actual nature of the target is. The helicopter completely observes others state variables, as the positions of targets. Note that, this model does not aim at testing the effectiveness of an algorithm in terms of time of calculation nor memory used, but it is meant to illustrate different optimization criteria for the same problem.

3.1. Experimental Protocol

The optimization criterion for the POMDP is a function of the agent's belief state, so it is optimized in terms of agent's pessimistic and subjective belief, which have not access to the actual state of world.

On the other hand, the criterion proposed is based on rewards which *really* are accumulated at policy execution, weighting only the uncertainty over effects of actions. Thus we want to measure the criterion optimality from an external viewpoint i.e. from an observer outside the system, who knows perfectly the state of environment at any moment. In this paper, this omniscient observer based on a policy simulator.

For each optimized policy, we have performed 500 simulations for a 50 horizon time, i.e. for 50 successive actions executed. For $\gamma = 0.9$, this horizon is considered large enough to obtain a good approximation of the criterion for an infinite horizon. The objective value function, which depends only on the current state of the environment is compute by means of Eq. (11).

$$V^{\pi}(s_t) = \frac{1}{500} \sum_{500\, simulations} \left[\sum_{k=0}^{t} \gamma^k r^{\pi}(s_k) \mid s_t \right] \tag{11}$$

In this paper the optimized policies with different optimized criteria are compared on the basis of the same objective value, which, whatever the optimization criterion is, will always be the rewards actually collected by the agent at policy execution. To study convergence speed of the entropy of the agent's belief, the the

current entropy of the belief entropy was calculated (statistically averaged over the runs), as shown in Eq. (12).

$$H^{\pi}(b_t) = \frac{1}{500} \sum_{500\, simulations} \left[\sum_{k=0}^{t} \gamma^k H^{\pi}(b_k) \mid b_t \right] \tag{12}$$

Note that this measure is subjective, specific to the agent, unlike the previous measure which is objective and specific to the simulator.

3.2. Simulation and Results

Policies have been computed for different λ values : $0, 0.5$ and 1. The first value is the γ-weighted classic criterion. Note that this criterion tries to optimize only the expected cumulative discounted reward assigned to the actions and task completion. The second λ value seeks to give the same importance to the accomplishment of the mission and information acquisition. And the third one optimizes only the information gain, i.e. the reduction of the entropy of the belief.

On Figure 6-left, the average of the value functions $V^{\pi}(s_t)$, Eq. (11), for the 3 values of λ are presented. In the first case ($\lambda = 0$), the value function $V^{\pi}(s_t)$ starts with negatives value; this is because the landing actions on the correct target (achieved after more than 10 simulation steps) weight less than those on the wrong one (done after 3 or 4 steps simulation) in the calculation of $V^{\pi}(s_t)$ due to γ-weighting. The authors think that the landing actions on the wrong target are probably due to the small size of the grid preventing the autonomous helicopter to acquire more information before the landing happens. The helicopter, which starts with a belief state weighted towards the type of the targets, leads to land as fast as possible over a target of type (B) after only 4 or 5 steps because, at this point, it still believes that this one can be a correct target of type (A). On the other side, in simulations in which the autonomous helicopter has acquired more information from its environment, the autonomous helicopter has inverted its belief state (not presented in this section due to space) and lands on the correct target of type (A). Hence the observed reversal of the value function, which shows that the agent reacts well even to the worst case.

For the second λ value, the criterion value starts also with negative values, corresponding to the same reasons raised above. The main difference reposes on the total value gathered when $t \to 50$. The value for the $\lambda = 0.5$ is more important than the value for $\lambda = 0$. Here the landing actions are made earlier than with the classical criterion, and so they weight more in the calculation of the $V^{\pi}(s_t)$ because of the γ-weighting. The paper contribution is here illustrated, since the agent now explicitly tries to acquire more information from its environment, allowing it to reverse earlier its belief and finally to land more frequently (402 times *versus* 390) on the correct target of type (A). A problem with the classical criterion is actually its linearity in $b(s)$: it considers that a belief state with 60% of chance of giving a reward actually brings 60% of this reward in. The non-linearity of this new criterion allows to better evaluate the value of $b(s)$, giving less weight to smaller uncertainties. Thus we conclude that taking into account actual belief state entropy at every time-step in policy computation forces the autonomous

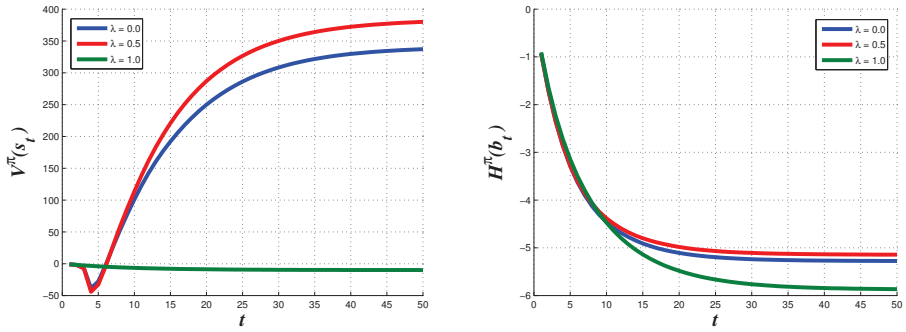

Figure 6. Averaged value function of current state $V^{\pi}(s_t)$ (left); And cumulated weighted entropy of belief state $H^{\pi}(b_t)$ (right).

agent to better sense its environment and then, carry out its mission and receive more rewards at policy execution, which is the overall objective for the designers of an autonomous system.

With the third value of λ , the criterion optimizes only the information gain. The Figure 6-left shows that the average of $V^{\pi}(s_t)$ remains close to zero: the helicopter does not try to land, it only gathers information. This is because the reward related with the task completion (landing on the target of type (A)) is no longer taken into account.

The Figure 6-right shows comparison of the subjective criterion $H^{\pi}(b_t)$, Eq. (12), for the 3 values of λ. Note that in the second case, average of the sum of entropy is bigger than with the classical criterion. Actually, the helicopter needs to observe its environment before task completion, which makes him reduce uncertainty over its belief state faster. This Figure also shows that the optimization of $J^{\pi}(b)$ allows to reduce the uncertainty faster than optimizing only $H^{\pi}(b)$, because it is implicitly necessary to reduce uncertainty in order to land on the correct target. Furthermore, this work shows that the $H^{\pi}(b)$ optimization does not necessarily optimizes its rate of growth at strategy execution. The mission objective, integrated to the additive criterion, gives a faster lowering of uncertainty by optimizing only $H^{\pi}(b)$.

To conclude, the additive criterion is as good as the purely entropic criterion in terms of information gathering, and it converges to a bigger value than the two others criteria (purely entropic or classic γ-weighted). This is due to a second fundamental fact: the additive criterion, by forcing the optimized policy to choose actions to gain rewards over actions to gain information from time to time, actually accumulates more rewards than running the classical criterion.

4. Conclusions and Future Works

In this paper have been presented a new mixed optimization criterion for POMDPs, which aggregate the cumulative information gain (perception) and the cumulative rewards gain (mission), weighted and averaged over an infinite horizon. Optimality of Bellman's equation has been underlined and proved for this

new criterion. Furthermore new admissible heuristic have been proposed for this criterion in order to use with algorithms such as Symbolic-PERSEUS.

We have experimentally demonstrated that this criterion allows the autonomous agent to gather information about this environment and to estimate faster its real state, compared to classic criteria (which take into account only either information gain, or rewards). Furthermore, for the additive criterion, the agent accumulates more rewards at policy execution than when it executes a policy obtained with a classical criterion (which does not explicitly take into account the information gain). In some way, explicit consideration of the entropy of the belief state in addition to rewards makes the autonomous agent acquire more information in order to weight his subjective view of the rewards, that it believes receiving, but may not due to its imperfect knowledge of the environment.

In the future, the influence of the λ coefficient in the additive criterion will be studied in more details. The authors believe that there is an optimal λ coefficient depending on the problem model, allowing to maximize the objective rewards actually accumulated at policy execution. The authors may also propose an optimization algorithm that optimizes policy and λ coefficient at the same time with respect to the modeled problem.

References

[1] W. Burgard, Dieter Fox, and Sebastian Thrun. Active mobile robot localization. In *Proceedings of IJCAI-97*. Morgan Kaufmann, 1997.

[2] A.R. Cassandra, L.P. Kaelbling, and J.A. Kurien. Acting under uncertainty: Discrete Bayesian models for mobile-robot navigation. In *In Proceedings of IEEE/RSJ*, 1996.

[3] F. Deinzer, J. Denzler, and H. Niemann. Viewpoint selection-planning optimal sequences of views for object recognition. *Lecture notes in computer science*, pages 65–73, 2003.

[4] R. Eidenberger, T. Grundmann, W. Feiten, and RD Zoellner. Fast parametric viewpoint estimation for active object detection. In *Proceeding of the IEEE International Conference on Multisensor of Fusion and Integration for Intelligent Systems (MFI 2008), Seoul, Korea*, 2008.

[5] L.P. Kaelbling, M.L. Littman, and A.R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[6] L. Mihaylova, T. Lefebvre, H. Bruyninckx, K. Gadeyne, and J. De Schutter. Active sensing for robotics – a survey. In *5th Intl Conf. On Numerical Methods and Applications*, pages 316–324, 2002.

[7] Lucas Paletta and Axel Pinz. Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, 31:71–86, 2000.

[8] J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. of IJCAI*, 2003.

[9] P. Poupart. *Exploiting structure to efficiently solve large scale partially observable Markov decision processes*. PhD thesis, University of Toronto, 2005.

[10] M.T.J. Spaan and P.U. Lima. A decision-theoretic approach to dynamic sensor selection in camera networks. In *Int. Conf. on Automated Planning and Scheduling*, pages 279–304, 2009.

[11] M.T.J. Spaan and N. Vlassis. A point-based POMDP algorithm for robot planning. In *IEEE International Conference on Robotics and Automation*, volume 3, pages 2399–2404. IEEE; 1999, 2004.

[12] M.T.J. Spaan and N. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *JAIR*, 24:195–220, 2005.

[13] M. Sridharan, J. Wyatt, and R. Dearden. HiPPo: Hierarchical POMDPs for Planning Information Processing and Sensing Actions on a Robot. In *Proc. of ICAPS*, 2008.