Active Diagnosis Through Information-Lookahead Planning

Mauricio Araya, Olivier Buffet, Vincent Thomas

Active Diagnosis Through Information-Lookahead Planning

Mauricio Araya-López[1,2], Olivier Buffet[2,1], Vincent Thomas[1,2]

[1]Université de Lorraine / [2]INRIA
LORIA – Equipe MAIA
Campus Scientifique BP 239, 54506 Vandoeuvre-lès-Nancy CEDEX, France
name.surname@loria.fr

Résumé : We consider challenging active diagnosis problems, that is, when smart exploration is needed to acquire information about a hidden target variable. Classical approaches rely on information-greedy strategies or ad-hoc algorithms for specific classes of problems. We propose to model this problem using the generic ρPOMDP formalism, which leads to an information-lookahead planning strategy, where the objective is to gather information-based reward. We empirically evaluate this approach on the Rock Diagnosis problem, which is a variation of the well-known Rock Sample problem, showing that we obtain better performance results than information-greedy techniques.

1 Introduction

An active diagnosis problem consists in performing the best actions to acquire information about a target object or variable. For example, to diagnose the state of the composite parts of an airplane, testing procedures can be performed on the parts. However, not all the parts can be continually tested all the time. The choice of some tests may depend on the results of previous ones, and there may be dependencies between the composite parts, so a smart active diagnosis system should find a conditional plan to optimally acquire information.

More precisely, in active diagnosis problems there is a target variable with a fixed value to estimate, but the information about this hidden variable can be obtained only through indirect observations by performing different actions. The two key characteristics of these problems are (1) that the target variable does not evolve with time (yet the observation of the target could be influenced by previous decisions), and (2) that the objective is to optimally gather information about the target variable, in contrast to other problems like active classification, where the objective is to provide only a final decision about the class of the hidden target.

Probably, the most natural example is medical diagnosis, where there is a latent disease, and the objective is to know which diseases fit better the symptoms and the outcomes of exams. Active medical diagnosis deals with selecting the best policy of exams or procedures to obtain the best possible diagnosis, where certain exams or procedures could conceal or affect the results of other exams (observations) without altering the disease. An example of the decision-theoretic formulation of this problem can be found in (Pellegrini & Wainer, 2003) where the cost-sensitive medical diagnosis problem is addressed.

Other application domains include fault diagnosis (Zheng *et al.*, 2005), mapping (Saigol *et al.*, 2009), visual search (Vogel & Murphy, 2007), network diagnosis (Ishida, 1997), active feature acquisition (Ji & Carin, 2007), sensor management (Williams, 2007) and informative sensing (Singh *et al.*, 2009).

In this paper we focus on solving an active diagnosis problem using an information-lookahead planning strategy. The objective is to show that information-lookahead methods are needed for some active diagnosis problems that require smart exploration, and to show an empirical application of the theoretical tools introduced in (Araya-López *et al.*, 2010a) for solving POMDP problems with information-based rewards (ρPOMDPs). Specifically, we use as an example the *Rock Diagnosis* problem : a variation of the well-known *Rock Sample* problem where the objective is to gather information about the rocks rather than performing destructive sampling. For a detailed ρPOMDP discussion and experimentation on a more general set of problems than active diagnosis, please refer to (Araya-López, 2013).

In Section 2 we present how previous work addresses the active diagnosis problems using (1) information-greedy approaches, or (2) ad-hoc information-lookahead methods under certain constraints. Then, we quickly introduce the POMDP framework as the standard formalism to address sequential decision making problems under uncertainty. In Section 3, we present how to model the active diagnosis problem using the ρPOMDP formalism without assuming dynamical constraints like previous work, and how to solve this problem approximately using affordable point-based methods. This allows to near-optimally solve problems that need smart exploration as the experiments of Section 4 show. We conclude and propose future work in Section 5.

2 State of the Art

The methods found in the literature for solving active diagnosis problems can be divided in two different groups :

- **Myopic approaches for information gathering**. Most of the methods that explicitly define the performance criteria using an information-theoretic measure falls in the category of myopic approaches, meaning that the decision is based only on a one-step information lookahead (information greedy). An example of this are the results found in (Krause & Guestrin, 2005) and (Williams, 2007) for greedy approaches in sensor management problems. Assuming a submodular information-based measure (Nemhauser *et al.*, 1978), it is shown that there are performance guarantees for information-greedy approaches for sensor management problems. Zheng *et al.* (2005) explicitly address the problem of active diagnosis, proposing a greedy approach based on the entropy in Bayesian networks. A more recent approach of Singh *et al.* (2009) proposes that greedy approaches for entropy-based informative sensing can be improved using short term memory. Unfortunately, the performance guarantees and efficiency of these methods are always restricted to specific classes of problems that have stateless observations ; i.e., that not only the target variable is static, but the whole system state is static.
- **Ad-hoc information-lookahead approaches**. There is a small number of methods that use information-lookahead to solve a proper sequential information-gathering problem. However, these methods rely on ad-hoc simplifications due to the properties of the problem at hand, confining their applicability only to the problems that share the same structure. For example, Saigol *et al.* (2009) propose a method for defining an information-lookahead planning problem using occupancy grids and a deterministic observation model, which can be solved using a POMDP model. Krishnamurthy (2002) has proposed a method for solving the sensor management problem modeled as a POMDP when an norm distance from the simplex corners is used as a reward function, but it also forces the observations to be stateless. This method share some similarities with this paper in its methodology, but the reward functions are only indirectly optimizing information. Another example in this line is (Rezaeian, 2007), that addresses the optimal observability problem using entropy and belief-MDPs for sensor management.

2.1 Partially Observable MDPs

In this paper, we are interested in methods that allow information-lookahead planning in active diagnosis, which requires a sequential decision making modeling. *Markov Decision Processes* (MDPs) (Bellman, 1954; Puterman, 1994) is the best known theoretical framework for sequential decision problems, providing a sequential probabilistic model under the mild Markovian assumption. Formally, an MDP consists in a tuple $\langle \mathcal{S}, \mathcal{A}, T, R, s_0 \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ are respectively the state and action spaces, $T(s, a, s') = Pr(S_{t+1} = s' | S_t = s, A_t = a)$ is a transition function, $r(s, a, s')$ is a scalar reward function, and s_0 is an initial state. A solution of an MDP is a mapping from states to actions called policy, and an optimal policy is one that optimizes the expected (discounted) sum of rewards. In this scenario, optimal policies can be found using techniques such as Value Iteration (Bellman, 1954) or Policy Iteration (Howard, 1960).

The MDP framework can be extended to support partially observable sequential decision problems, by adding an observation space $\mathcal{Z}$, and a stochastic observation function $Pr(Z_t = z | S_t = s, A_{t-1} = a)$. Yet, the initial state s_0 is often unknown, so a belief-state is used to define b_0, an initial prior distribution over the states. A *Partially Observable MDP* (POMDP) (Astrom, 1965; Smallwood & Sondik, 1973) consists in a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, T, R, O, b_0 \rangle$, where $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{Z}$ are the state, action and observation sets respectively, $T(s, a, s') = Pr(S_{t+1} = s' | S_t = s, A_t = a)$ is a transition function, $r(s, a, s')$ is a scalar reward function, $O(s', a, z) = Pr(Z_t = z | S_t = s', A_{t-1} = a)$ is an observation function, and $\mathbf{b}_0$ is an initial belief-state

distribution.

The problem with POMDPs is that the techniques used to solve MDPs cannot be applied directly, because observations are not Markovian. Fortunately, by using the belief-state abstraction and the Bayes rule, a proper Markovian process can be constructed over the beliefs, because the Belief update $\mathbf{b}' = Bayes(\mathbf{b}, z, a)$ is Markovian in the sense that the new belief-state depends only on the previous belief and the current observation and action.

Following the general definition of an MDP, a *belief-state MDP* is defined by a tuple $\langle \Delta, \mathcal{A}, \tau, \rho, \mathbf{b}_0 \rangle$, where $\tau(\mathbf{b}, a, \mathbf{b}') = Pr(B' = \mathbf{b}' | A = a, B = \mathbf{b})$ is the belief-state transition function, $\rho(\mathbf{b}, a)$ is the belief-state reward function, and $\mathbf{b}_0 \in \Delta$ is the initial prior distribution over the states. The $\tau(\mathbf{b}, a, \mathbf{b}')$ function can be written in terms of T and O as follows (Cassandra, 1998)

$$\tau(\mathbf{b}, a, \mathbf{b}') = \sum_{z \in \mathcal{Z}} \mathbb{I}(\mathbf{b}', Bayes(\mathbf{b}, a, z)) \sum_{s' \in \mathcal{S}} \sum_{s \in \mathcal{S}} O(s', a, z) T(s, a, s') b(s), \tag{1}$$

where $\mathbb{I}$ is an indicator function. Similarly, the belief-state reward function at any step can be written as

$$\rho(\mathbf{b}, a) = \sum_{s \in \mathcal{S}} b(s) R(s, a). \tag{2}$$

Unfortunately, even when the state, action and observation spaces are discrete, belief-state MDPs are multi-dimensional continuous-state MDPs, which are in general intractable to solve. Auspiciously, due to regularities of the specific simplex space, belief-propagation and transition function, an ϵ-close solution for belief-MDPs can be found in a finite number of iterations, yet typically with an overwhelming complexity (Lovejoy, 1991; Madani *et al.*, 2003).

2.2 Dynamic Programming for POMDPs

The belief-MDP value function verifies Bellman's optimality equation (Bellman, 1954) (for all $b \in \Delta$), which provides a recursive representation of V_t^* for each time step t in the form :

$$V_t^*(\mathbf{b}) = \max_{a \in \mathcal{A}} \left\{ \rho(\mathbf{b}, a) + \gamma \sum_{b' \in \Delta} \tau(\mathbf{b}, a, \mathbf{b}') V_{t+1}^*(\mathbf{b}') \right\}.$$

By using Equations 1 and 2, the belief-MDP value function can be written in POMDP terms as follows :

$$V_t^*(\mathbf{b}) = \max_{a \in \mathcal{A}} \left\{ \sum_{s \in \mathcal{S}} b(s) \left[R(s, a) + \gamma \sum_{z \in \mathcal{Z}} \sum_{s' \in \mathcal{S}} O(s', a, z) T(s, a, s') V_{t-1}^*(Bayes(\mathbf{b}, a, z)) \right] \right\}. \tag{3}$$

If there is a finite number of possible belief-states (like in the finite-horizon case), a policy can be obtained by computing the value function at each step t for each possible $\mathbf{b} \in \mathcal{B}$, where $\mathcal{B} \subset \Delta$ is a finite set. However, solving POMDPs using this dynamic programming technique is only suitable for small problems like the ones presented in (Astrom, 1965). To scale up, more elaborated methods are required.

2.3 Solving POMDPs

Probably the most remarkable advance for solving POMDPs was presented in (Smallwood & Sondik, 1973), where the authors show that the belief-MDP value function is *Piecewise-Linear and Convex* (PWLC) for a finite horizon. This mathematical property leads to a value function representation using a finite number of hyperplanes, and is the cornerstone of most of the modern POMDP solution algorithms.

In fact, due to the Minkowski-Weyl theorem, we know that the value function at time t ($\neq \infty$) can be represented by a finite set of α-vectors Γ_t in the form

$$V_t^*(\mathbf{b}) = \max_{\alpha \in \Gamma_t} \{ \mathbf{b}^\top \alpha \}. \tag{4}$$

An exact and finite solution (policy) can be obtained by computing each Γ_t iteratively until convergence and labeling the vectors with the action used for each computation, by using a vectorial version of Equation 3. This is a form of *exact value iteration* for POMDPs, which properly solves the problem, but with a

number of vectors in Γ_t that grows exponentially at each time step (Littman, 1996) [1]. Therefore, these exact algorithms do not scale well enough to address real-world problems, being useful only for problems of a very limited size.

The most popular approximation techniques in this field are *point-based* (PB) methods. The idea behind a PB approximation is that the value function for close belief-points is usually represented by the same α-vector. Consequently, if a suitable finite set of belief-points is selected, then the value function can be closely approximated using this finite set. This approximation is obtained by propagating the support α-vectors that are maximal in those points, but neglecting the propagation in the rest of the belief-space. The α-vector of a belief-point in the set works not only as an approximation of the value function for that specific point, but also as an approximation for the nearby beliefs that are not in the set.

The support α-vector of a belief-point $\mathbf{b}$ can be obtained by reinterpreting the Bellman equation as follows

$$\begin{aligned} V_t^*(\mathbf{b}) &= \max_{a\in\mathcal{A}}\left\{\mathbf{b}^\top\mathbf{r}^a + \gamma\sum_{z\in\mathcal{Z}}\max_{\alpha\in\Gamma_{t-1}}\left\{\mathbf{b}^\top P^{a,z\top}\alpha\right\}\right\} \\ &= \max_{a\in\mathcal{A}}\left\{\mathbf{b}^\top\mathbf{r}^a + \gamma\sum_{z\in\mathcal{Z}}\mathbf{b}^\top P^{a,z\top}\operatorname*{argmax}_{\alpha\in\Gamma_{t-1}}\left\{\mathbf{b}^\top P^{a,z\top}\alpha\right\}\right\} \\ &= \max_{\alpha\in\Gamma_t^{\mathbf{b}}}\mathbf{b}^\top\alpha, \text{ where} \\ \Gamma_t^{\mathbf{b}} &= \bigcup_{a\in\mathcal{A}}\left\{\mathbf{r}^a + \gamma\sum_{z\in\mathcal{Z}}P^{a,z\top}\operatorname*{argmax}_{\alpha\in\Gamma_{t-1}}\left\{\mathbf{b}^\top P^{a,z\top}\alpha\right\}\right\}. \end{aligned}$$

Using these equations, the *backup* function can be summarized as

$$backup(\Gamma_{t-1},\mathbf{b}) = \operatorname*{argmax}_{\alpha\in\Gamma_t^{\mathbf{b}}}\mathbf{b}^\top\alpha.$$

The work of Lovejoy (1991) was the first to suggest that PB methods could provide good approximations of the value functions. However, the proposed finite-grid discretization of the belief-space suffers from several drawbacks, such as that most of the belief-points in the set are not even reachable from the initial belief-state. To overcome this problem, Hauskrecht (2000) has proposed to collect reachable points in the belief-space, and use them as support points for the α-vector propagation. From here, several efficient and fast algorithms have been proposed, starting with PBVI (Pineau *et al.*, 2003), PERSEUS (Spaan & Vlassis, 2005) and HSVI2 (Smith & Simmons, 2005), and more recent ones such as FSVI (Shani *et al.*, 2007), SARSOP (Kurniawati *et al.*, 2008) and GapMin (Poupart *et al.*, 2011). For a recent and exhaustive study of these methods, please refer to (Shani *et al.*, 2012).

3 Active Diagnosis using ρPOMDPs

The dynamics and observability of active diagnosis can be properly represented by probability distributions, and therefore POMDPs may seem an adequate model to address this type of sequential problems. Unfortunately, POMDPs actually cannot describe properly these problems, because the strict definition of rewards that depends only on the state and actions does not allow defining objectives depending on the information.

For correctly formalizing problems that require information-based rewards, we proposed the ρPOMDPs framework (Araya-López *et al.*, 2010a), where the definition of POMDPs is extended to support arbitrary belief-rewards : rewards that depend directly on the belief-state and not only on the state of the system.

Formally, a ρPOMDP is a generalization of the POMDP framework where the reward function $\rho(\mathbf{b},a)$ is directly defined in terms of belief-states and actions, and not (only) as the expectation of an external reward $r(s,a,s')$. Then, the ρPOMDP tuple is defined as $\langle\mathcal{S},\mathcal{A},\mathcal{Z},T,O,\rho,\mathbf{b}_0\rangle$.

3.1 Measuring Uncertainty for Diagnosis

As stated before, even though active diagnosis needs sequential decisions, it differs from classical sequential decision making since the objective is to optimize an information measure that is not described in terms

1. This can be improved by adding smart pruning phases like in (Cassandra *et al.*, 1997), but the complexity is still exponential.

of state reward. Therefore, to define these problems the first step is to formally characterize the information held at each time step about the target variable. Fortunately, information theory (Cover & Thomas, 1991) provides the theory and tools to quantify information of probabilistic models.

In the POMDP framework, the information about the state of the system is represented by belief-states. Formally, if $S_t \in \mathcal{S}$ is the random variable with probability distribution $Pr(S_t) = \mathbb{P}$ that represents the belief-state at time t, Shannon's measure of the uncertainty held by the probability distribution $\mathbb{P}$ can be used to evaluate different belief-states Shannon (1948) :

$$\mathcal{H}(S_t) = \mathcal{H}(\mathbb{P}) = \mathbb{E}_{S_t}\left[-log(Pr(S_t))\right].$$

In the parametric case, Shannon's entropy consists in the expectation of the negative log-likelihood function. As the logarithm is a monotonically increasing function, the entropy can be interpreted as the mean of how much *unlikely* is that a parameter θ explains a specific state. For example, for categorical distributions like the ones used for discrete POMDP beliefs, the uncertainty can be measured as follows,

$$\mathcal{H}(\mathbf{b}) = -\sum_{s\in\mathcal{S}} \mathbf{b}(s)\log(\mathbf{b}(s)).$$

For simplicity, this paper assumes that uncertainty is inversely proportional to information, even though more refined relative versions provide a better way to explain this relationship (Shannon, 1948; Kullback & Leibler, 1951).

3.2 Information-based Performance Criteria

In general, the finite horizon performance criterion of a belief-MDP is

$$V_H^\pi(\mathbf{b}_0) = \mathbb{E}_{B_{1:H}}\left[\sum_{t=1}^{H}\gamma^{t-1}\rho(B_t,\pi(B_t,t))\middle| B_0=\mathbf{b}_0,\pi\right]. \tag{5}$$

For ρPOMDPs, this criterion sums each step's information measure, which means that the objective is to always be in a high-valued information-state if $\gamma = 1$. On the other hand, if $\gamma < 1$ then the objective is to obtain information as quickly as possible.

However, in active diagnosis the information gathering in intermediate steps is only a means to arrive to a high-valued information-state. Then, the objective is to optimize the information *at the end* of the process, meaning that the information measure is a single reward that is given to the agent in the final step. Therefore, the same criterion of Equation 5 can be interpreted as the sum of a non-stationary belief-reward with zero reward for all steps except the final one. This criterion can be written for the finite horizon as follows :

$$V_H^\pi(\mathbf{b}_0) = \mathbb{E}_{B_{1:H}}\left[\rho(B_H,\pi(B_H))|B_0=\mathbf{b}_0,\pi\right].$$

Consider now the case when the horizon is not known and cannot be controlled. Here, the expectation over a *stopping probability* can be used to reason about the future steps, meaning that at each step the process may stop with a probability $1-\gamma$. Then, as the probability of continuation is γ, the probability of the system to stop at horizon h is

$$Pr(H=h) = \gamma^{h-1}(1-\gamma),$$

which gives

$$\begin{aligned}
V^\pi(\mathbf{b}_0) &= \mathbb{E}_H[V_H^\pi(\mathbf{b}_0)] \\
&= \lim_{n\to\infty}\sum_{h=1}^{n} Pr(H=h)\mathbb{E}_{B_{1:h}}\left[\rho(B_h,\pi(B_h))|B_0=\mathbf{b}_0,\pi\right] \\
&= \lim_{n\to\infty}\sum_{h=1}^{n} \gamma^{h-1}(1-\gamma)\mathbb{E}_{B_{1:h}}\left[\rho(B_h,\pi(B_h))|B_0=\mathbf{b}_0,\pi\right] \\
&= (1-\gamma)\lim_{n\to\infty}\mathbb{E}_{B_{1:h}}\left[\sum_{h=1}^{n}\gamma^{h-1}\rho(B_h,\pi(B_h))\middle| B_0=\mathbf{b}_0,\pi\right],
\end{aligned}$$

which is equivalent to a discounted infinite-horizon criterion for optimization purposes.

If we choose $\rho(\mathbf{b},a)$ to be the opposite positive value of the entropy (i.e., $\log(|\mathcal{S}|) - H(\mathbf{b})$), then a proper information-based criterion for active diagnosis can be constructed, based on the axiomatic derivation of this information measure by Shannon (1948).

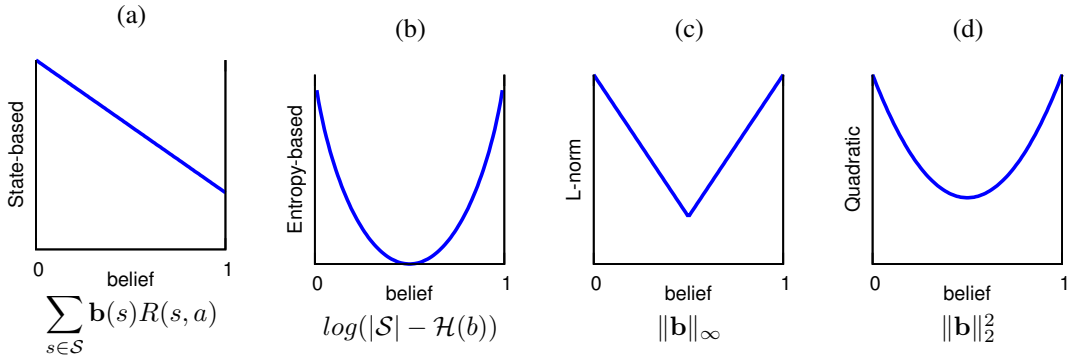

FIGURE 1: Information-based rewards in a 1-simplex : (a) state-based, (b) negentropy, (c) L_∞-norm and (d) squared L_2-norm.

3.3 Information-based Rewards

In Section 2.1 the $\rho(\mathbf{b}, a)$ function was defined as the expected value of the reward function $R(s, a)$, which gives the *state-dependent reward*

$$\rho_R(\mathbf{b}, a) = \sum_{s\in\mathcal{S}} \mathbf{b}(s)R(s, a),$$

(see Figure 1(a)).

This type of reward will clearly not comply with the performance criterion defined in the previous section. Instead, we need an *negentropy reward* :

$$\rho_H(\mathbf{b}, a) = \log(|S|) + \sum_{s\in\mathcal{S}} \mathbf{b}(s)log(\mathbf{b}(s)).$$

This reward function is a convex function that represents the opposite of the entropy, where beliefs with high uncertainty produce low rewards, and beliefs with low uncertainty produces high rewards (see Figure 1(b)).

However, entropy is a complex function to work with, so other types of rewards can be defined by using approximations of this negentropy reward. For example the L-norms of the belief-state are conventional statistics that are easier to analyze and compute. In particular, the *quadratic reward* (i.e. the squared L_2-norm of Figure 1(d)) is defined as

$$\rho_Q(\mathbf{b}, a) = \|\mathbf{b}\|_2^2 = \sum_{s\in\mathcal{S}} \mathbf{b}(s)^2,$$

and corresponds to the most significant term of the Taylor expansion of the ρ_H reward. Another interesting approximation is the *linear reward* (i.e., the L_∞-norm of Figure 1(c)), which is a *PWLC reward* consisting in the maximum of $|\mathcal{S}|$ hyperplanes, and can be trivially defined as

$$\rho_L(\mathbf{b}, a) = \|\mathbf{b}\|_\infty = \max_s \mathbf{b}(s),$$

and corresponds to an upper linear approximation of ρ_H. Therefore, using these reward functions indirectly reduces the entropy with less computational effort, at the price of not being optimal.

3.4 Solving ρPOMDPs

In this section we quickly review how to solve ρPOMDPs. For more details please refer to (Araya-López *et al.*, 2010a; Araya-López, 2013).

3.4.1 Convex Reward Function

In ρPOMDPs the rewards are defined directly over belief-states, making the belief-MDP harder to solve. Indeed, the value function is not always PWLC because it depends on the structure of the $\rho(\mathbf{b}, a)$ function. Fortunately, the convexity property also holds if the reward function $\rho(b, a)$ is convex, as shown in

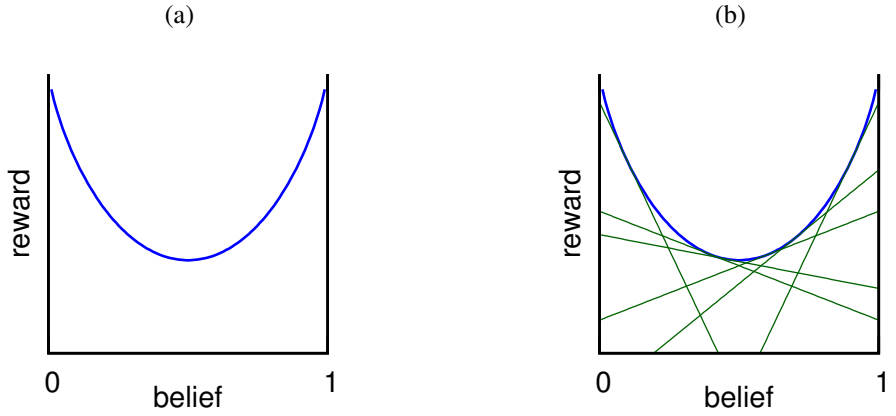

FIGURE 2: A PWLC approximation of a convex non-linear reward using tangent hyperplanes. In (a) a set of points in the belief-space and their projections to the reward function. In (b) the tangent hyperplanes of the reward function at each belief-point.

(Araya-López *et al.*, 2010a). Convexity is a property commonly found in information measures, because the objective is to avoid belief distributions that do not give much information on which state the system is in, and to assign higher rewards to those beliefs that give higher probabilities of being in a specific state. Specifically, the three reward functions defined in the Section 3.3 are convex functions.

3.4.2 Solutions for PWLC Reward Functions

When the information-based reward function is PWLC, only small adaptations to exact POMDP algorithms are needed to compute an optimal value function using exact algorithms like Incremental Pruning (Cassandra *et al.*, 1997) Similarly, point-based algorithms can be applied by considering a reward function representation as an envelope of hyperplanes.

Point-based algorithms select the hyperplane that maximizes the value function at each belief-point, so the same can be applied to a set Ψ^a of reward vectors. Then, for the *backup* computations of Section 2.3, the Γ-set computations can be modified as follows :

$$\begin{aligned} V_t^*(\mathbf{b}) &= \max_{a\in\mathcal{A}} \left\{ \max_{\beta\in\Psi^a} \left\{\mathbf{b}^\top \beta\right\} + \gamma \sum_{z\in\mathcal{Z}} \max_{\alpha\in\Gamma_{t-1}} \left\{\mathbf{b}^\top P^{a,z\top}\alpha\right\} \right\} \\ &= \max_{a\in\mathcal{A}} \left\{ \mathbf{b}^\top \operatorname*{argmax}_{\beta\in\Psi^a} \left\{\mathbf{b}^\top \beta\right\} + \gamma \sum_{z\in\mathcal{Z}} \mathbf{b}^\top P^{a,z\top} \operatorname*{argmax}_{\alpha\in\Gamma_{t-1}} \left\{\mathbf{b}^\top P^{a,z\top}\alpha\right\} \right\} \\ &= \max_{\alpha\in\Gamma_t^{\mathbf{b}}} \mathbf{b}^\top \alpha, \text{ where} \\ \Gamma_t^{\mathbf{b}} &= \bigcup_{\in\mathcal{A}} \left\{ \operatorname*{argmax}_{\beta\in\Psi^a} \left\{\mathbf{b}^\top \beta\right\} + \gamma \sum_{z\in\mathcal{Z}} P^{a,z\top} \operatorname*{argmax}_{\alpha\in\Gamma_{t-1}} \left\{\mathbf{b}^\top P^{a,z\top}\alpha\right\} \right\}. \end{aligned}$$

As all point-based methods use the same *backup* function, any point-based algorithm can potentially be used for solving ρPOMDPs.

3.4.3 Generalizing to non-PWL Reward Functions

Even though some interesting results can be obtained using PWLC rewards, most information measures are not piecewise linear functions. In theory, each step of value iteration can be analytically computed using non-piecewise-linear functions, but the expressions are not closed as in the linear case, growing in complexity and becoming unmanageable after a few steps. However, convex functions can be efficiently approximated by piecewise linear functions, making it possible to apply exact or PB methods with a bounded error, as long as the approximation of ρ is bounded.

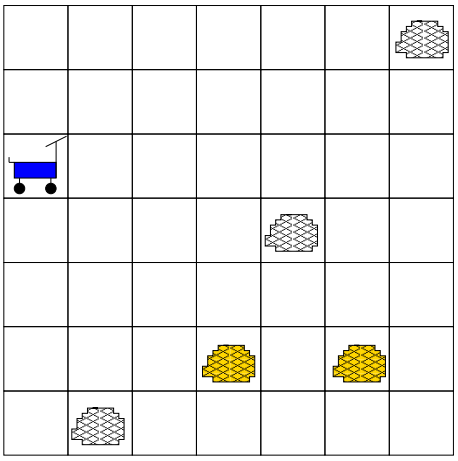

FIGURE 3: Graphical representation of the Rock Diagnosis problem with 5 rocks and a grid side of 7. The yellow rocks are the "good" rocks (with scientific value) and the others are the "bad" rocks (no scientific value), yet each rock type is unknown to the agent. Please note that, for diagnosis, a "good" rock is not more important than a "bad" one, because the objective is to disambiguate all of them, and not sampling only "good" ones like in the original Rock Sampling problem.

Consider then a continuous, convex and piecewise differentiable reward function $\rho(\mathbf{b})$, and an arbitrary (and finite) set of points $\mathcal{B} \subset \Delta$ where the gradient $\nabla\rho$ is well defined (see Figure 2(a)). A lower PWLC approximation of $\rho(\mathbf{b})$ can be obtained by using each element $\mathbf{b}' \in \mathcal{B}$ as a base point for constructing a tangent hyperplane which is always a lower bound of $\rho(\mathbf{b})$ (see Figure 2(b)). Concretely, $\omega_{\mathbf{b}'}(\mathbf{b}) = \rho(\mathbf{b}') + (\mathbf{b} - \mathbf{b}')^\top \nabla\rho(\mathbf{b}')$ is the linear function that defines the tangent hyperplane, which leads to a lower approximation of $\rho(\mathbf{b})$ using a set $\mathcal{B}$ of

$$\rho_{\mathcal{B}}(\mathbf{b}) = \max_{\mathbf{b}' \in \mathcal{B}} \omega_{\mathbf{b}'}(\mathbf{b}).$$

It is well known that the error of a piecewise linear approximation of a Lipschitz function is bounded because the gradient $\nabla\rho(\mathbf{b}')$ that it is used to construct the hyperplane has bounded norm (Saigal, 1979). Unfortunately, some interesting functions like the negentropy reward are not Lipschitz ($f(x) = x\log(x/c)$ has an infinite slope when $x \to 0$). Yet, under certain mild assumptions, a proper error bound can still be found, as we showed in (Araya-López *et al.*, 2010a).

Knowing now that the approximation of ρ is bounded for a wide family of functions, exact and PB techniques can be applied by considering $\rho_{\mathcal{B}}(\mathbf{b})$ as the PWLC reward function. These algorithms can be safely used because the propagation of the error due to the exact or PB updates are bounded, as also shown in (Araya-López *et al.*, 2010a). In the case of point-based algorithms, the selection of $\mathcal{B}$, the set of points for the PWLC approximation, and the set of points for the algorithm, can be shared. However, please be aware that, for the negentropy reward the gradient is not defined on the boundary, so the collected points in this zone must be modified or removed.

4 Experiments : The Rock Diagnosis Problem

Here, we introduce the *rock diagnosis* problem, which is a variation of the *rock sampling* problem proposed in (Smith & Simmons, 2004). In the original problem a grid-map of rock positions is given to a rover, whose objective is to perform sampling procedures (that destroy the rocks) to those rocks that have a "good" scientific value while traversing the map. If the rover samples a "good" rock it receives a positive reward. In the contrary, it receives a negative reward for sampling "bad" rocks because the sampling procedure is expensive. Obviously, the types of the rocks are not given to the rover, but the rover is equipped with a noisy long-range sensor to query them. The efficiency of the long-range sensor (i.e., the probability of correctly identifying the rock type) decreases exponentially with the Euclidean distance to the rock.

The rock diagnosis variation (see Figure 3) consists in removing the sampling action from the rover, meaning that the rover can only move through the grid and use its long-range sensor to query rocks. The objective is to reduce the uncertainty about the rocks' nature, whatever their type. This information may be used later by a human to analyze the distribution of rocks, or to perform any other procedure that needs a highly-confident knowledge about the nature of the rocks. This same kind of setup can be used for industrial applications such as searching for oil wells, detecting plagues in crop fields or analyzing geologic veins for mining.

The rock diagnosis problem was selected because it needs smart long-term exploration, where a suitable path of uninformative actions (i.e., cardinal moves) leads to highly-informative observations in the future.

4.1 ρPOMDP Model of the Problem

Let us model the state of the system by the pair $s = \langle loc, rtype \rangle$, where loc is the localization of the rover in the grid, and $rtype = \langle r_1, r_2, \ldots, r_n \rangle$ contains the type of each rock in the map, where $r_i \in \{good, bad\}$. This forms a state space of size $|\mathcal{S}| = l^2 \times 2^n$. The action space corresponds to $\mathcal{A} = \{north, east, south, west, check_1, check_2, \ldots, check_n\}$, where the first 4 actions are the cardinal moves, and a $check_i$ action uses the long-range sensor to query the rock i.

The transition function is completely deterministic : under the cardinal moves the rover deterministically changes position according to the selected direction and the grid boundaries, and $check_i$ actions do not modify the state in any way. On the contrary, the observation function is stochastic, because observations depend on the noisy long-range sensor. The observation space is $\mathcal{Z} = \{none, good, bad\}$, where $none$ is obtained if and only if the agent executes a cardinal move action, and $good$ or bad are obtained depending on the queried rock type and the distance to the target rock.

Let r_i be the type and p_i be the position of rock i in the grid, then the efficiency of the long-range sensor is

$$efficiency = Pr(Z = r_i | A = check_i, S = \langle loc, rtype \rangle) = \frac{1 + e^{-\|loc - p_i\|_2}}{2}.$$

As the $none$ observation can only be obtained by cardinal move actions, the probability of obtaining an incorrect observation by the long-range sensor is $1 - efficiency$. Please note that the p_i positions are given to the agent, so they can be implicitly encoded in the observation function.

The objective is to reduce the uncertainty of the target variable $rtype$ at the end of the process, but as the horizon is unknown to the agent, the information-based discounted infinite-horizon criterion of Section 3.2 is used. Under this criterion, the maximum final reward corresponds to $n \log(2)$.

4.2 Experimental Setup

For the experiments we consider a number of rocks n and only square grids of side l. Figure 4 shows the specific map configurations used in this section. These configurations will be identified by the numbers above each map, which are in the form n-l.

For each instance of a problem, the following algorithms were used :

- **Random policy**. At each step, the algorithm chooses a random action to execute. Any smart algorithm should perform better than (or at least equal to) this policy.
- **Myopic strategy**. At each step, the algorithm selects the action with the greatest expected next step reward. In simpler words, it is an *information-greedy* approach that chooses the immediate most informative action.
- **Information-lookahead approach**. Using the ρPOMDP definition and PWLC approximations if needed, a policy is obtained using the PERSEUS algorithm. Here, not only the negentropy reward was tested, but also the quadratic and linear approximations were used.

The PERSEUS algorithm was selected due to its time efficiency, performing asynchronous backups and using a constant set of belief-points gathered using random simulations. The asynchronous value iterations are stopped when the infinite norm of the difference between two successive iterations is below some threshold. Concretely,

$$\|V_i - V_{i-1}\|_\infty \leq \frac{(R_{max} - R_{min})\epsilon}{1 - \gamma},$$

where ϵ is an accuracy parameter given to the algorithm. For all the experiments we used $\gamma = 0.95$ and $\epsilon = 0.1$.

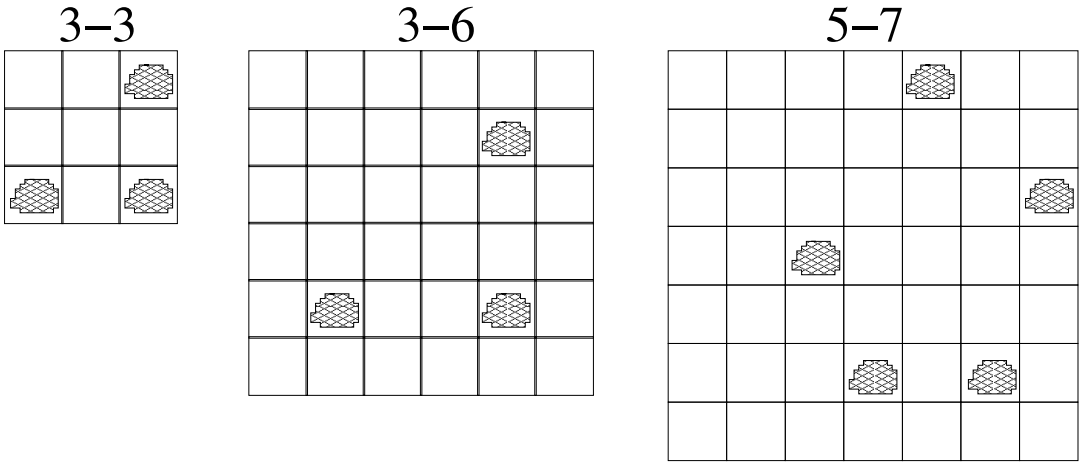

FIGURE 4: The three specific Rock Diagnosis maps used in this section. The numbers above the figures are in the form n-l, where n is the number of rocks and l is the side of the grid.

4.3 Results

The experiments for each map are summarized on Table 1. For each experiment 10 policies were generated, each one evaluated on-line using 100 trajectories. Several sizes for the belief-point set were tested, yet in Table 1 the results are shown only for 1000 and 5000 points. For each trajectory, the type of each rock is sampled from the initial belief, which corresponds to the uniform distribution. The horizon of the trajectories was fixed to $H = 100$, yet results for shorter horizons can be obtained by truncating the simulation output.

The performance results are conclusive : PB approaches are always significantly better than the myopic approach. Unsurprisingly, PB strategies with 5000 points provide better results than with 1000 points, and even these last ones outperform largely the myopic strategy. However, the random strategy is highly competitive, usually outperforming the results with 1000 points and sometimes even with 5000 points as it is highlighted in the table.

In the 5-7 problem, the random strategy obtains in general a better mean value than PB strategies except for the PB-Linear one with 5000 points. However, the standard deviation shows that the variability of the random strategy is very high with respect to the value, so very poor performance can be sometimes expected.

By comparing the different reward functions used for the PB methods, it can be seen that the quadratic reward gives usually lower results than the other two, and the linear reward provides very good results. Probably, the negentropy reward might exceed the linear approximation if a considerable number of points were used, but the overwhelming amount of time needed for such experiment is prohibitive. The poor results in the 7-5 map can be explained by the curse of dimensionality : with the same amounts of points (1000 or 5000) a higher dimensional belief-space is sampled, so a much worse approximation is obtained [2]. Indeed, the linear reward is an easier function to approximate than the other two, and therefore it provides the best results. This can be seen not only in the 5-7 map, but also in the two smaller maps. In the 3-6 map the linear reward provides a slightly better result than the others, and in the 3-3 map it succeeds gathering 99% of the perfect information.

The problem with increasing the number of points to achieve better performance is always related to the computational effort. Increasing the amount of points always increases the mean value but it also always increases the computational effort required. Unfortunately, the off-line time cannot be predicted, and its variability can be enormous. For example, consider the worst off-line time of Table 1, which corresponds to PB-Linear with 1000 points for the 3-3 map. The mean is near 3 hours, but its standard deviation exceeds 1 hour, meaning that some repetitions can take several hours or only a few minutes. Even in the best off-line time scenario, which is the PB-Quadratic strategy with 1000 points for the 3-6 problem, the standard deviation is more than a quarter of the mean value. However, as this depends on the belief-point collection, the off-line time might be more stable if a smarter selection of points is used.

2. An ad-hoc amount of points for each map will imply an exponential growth of them with the dimensionality

algorithm	**rocks-side**	$\lvert\mathcal{B}\rvert$	**total return** $[nats]$	**on-line time** $[ms]$	**off-line time** $[s]$
Random	3-3	—	1.55 ± 0.40	0.82 ± 2.24	—
Myopic	3-3	—	0.69 ± 0.30	$\mathbf{129.16 \pm 23.60}$	—
PB-Entropy	3-3	1000	1.17 ± 0.27	157.80 ± 53.28	$\mathbf{20.69 \pm 8.06}$
PB-Quadratic	3-3	1000	1.39 ± 0.29	295.80 ± 147.07	65.75 ± 61.27
PB-Linear	3-3	1000	1.55 ± 0.43	450.38 ± 70.21	579.53 ± 192.38
PB-Entropy	3-3	5000	1.58 ± 0.25	675.44 ± 154.21	315.34 ± 180.10
PB-Quadratic	3-3	5000	1.58 ± 0.24	1453.35 ± 503.57	1105.48 ± 521.32
PB-Linear	3-3	5000	$\mathbf{2.06 \pm 0.03}$	957.26 ± 221.96	10076.82 ± 4445.40
Random	3-6	—	0.59 ± 0.41	0.14 ± 0.32	—
Myopic	3-6	—	0.01 ± 0.01	434.38 ± 30.68	—
PB-Entropy	3-6	1000	0.30 ± 0.03	$\mathbf{116.29 \pm 26.86}$	19.79 ± 5.49
PB-Quadratic	3-6	1000	0.50 ± 0.01	$\mathbf{131.69 \pm 34.44}$	$\mathbf{12.34 \pm 5.61}$
PB-Linear	3-6	1000	0.69 ± 0.02	403.48 ± 167.61	56.68 ± 42.71
PB-Entropy	3-6	5000	$\mathbf{0.76 \pm 0.09}$	211.70 ± 40.73	81.83 ± 13.39
PB-Quadratic	3-6	5000	$\mathbf{0.74 \pm 0.06}$	279.04 ± 77.80	62.06 ± 14.66
PB-Linear	3-6	5000	$\mathbf{0.79 \pm 0.08}$	2749.77 ± 1259.17	5178.52 ± 4463.55
Random	5-7	—	$\mathbf{0.50 \pm 0.42}$	0.16 ± 0.36	—
Myopic	5-7	—	0.03 ± 0.04	2869.11 ± 206.44	—
PB-Entropy	5-7	1000	0.08 ± 0.02	$\mathbf{424.99 \pm 80.04}$	59.40 ± 13.16
PB-Quadratic	5-7	1000	0.11 ± 0.03	$\mathbf{438.97 \pm 75.57}$	$\mathbf{15.08 \pm 2.96}$
PB-Linear	5-7	1000	0.23 ± 0.06	$\mathbf{495.03 \pm 104.83}$	25.01 ± 8.53
PB-Entropy	5-7	5000	0.37 ± 0.09	800.96 ± 247.22	281.08 ± 75.29
PB-Quadratic	5-7	5000	0.12 ± 0.03	625.34 ± 117.43	35.61 ± 3.39
PB-Linear	5-7	5000	$\mathbf{0.53 \pm 0.03}$	846.30 ± 175.76	67.60 ± 17.41

TABLE 1: Rock Diagnosis Results. Total return, on-line time and off-line time over 10 repetitions of 100 trajectories of 100 steps (mean values plus standard deviations). The bold and red values stand for the best and worst results respectively.

It is important to notice that, for PB strategies, good performances are correlated with high computational effort in Table 1. This can be explained due to the several informative actions available at each step, generating several vectors that are not easily dominated in the whole belief-space. If a dense approximation is used, these vectors will be propagated to produce check and move actions when suitable. If only a sparse approximation is available, these vectors will be easily dominated by a small set of vectors, generating useless policies like checking always the same rock, or moving towards a rock but not checking it. These poor policies can be obtained very fast, explaining the counter intuitive results that the computational effort of higher-dimensional maps is less important than for lower-dimensional ones.

Fortunately, for this same reason it is much more affordable to increment the numbers of points for the 5-7 map than for the 3-3. This is shown in Figure 5 where the reward evolution for 5000 and 20000 points is plotted. Notice that even though the final reward ($H = 100$) of the random strategy is near the values of information-lookahead strategies for 5000 points, this is not true for shorter horizons. Indeed, information-lookahead strategies gather almost all the information in the first 10 steps, yet later they acquire information no faster than the myopic strategy. As expected, there is an improvement by going from 5000 to 20000 points. However, 20000 points are still not enough to densely sample the belief-space for this random collection strategy, meaning that the best corresponding policies are still far away from the optimal one.

Figure 6 shows at which rate the total return and the computational effort increase with more points for the 5-7 map. The time results show that the myopic strategy is slower than the information-lookahead ones for all but the linear reward with 20000 points, but this phenomenon occurs only due to the sparse approximation of the value function as the performance results show. These results confirm that myopic strategies may not only produce very poor results, but also may be significantly slower too [3].

As expected, Figure 6 shows that the apparent good performance of the random strategy for the 5-7 map in Table 1 was only due to the relatively poor lookahead policies obtained with 5000 points. Indeed, a fast

3. This is only true for the worst case scenario where no analytical expression can be used to speed up the computations. In practice, ad-hoc myopic strategies will be significantly faster than lookahead ones.

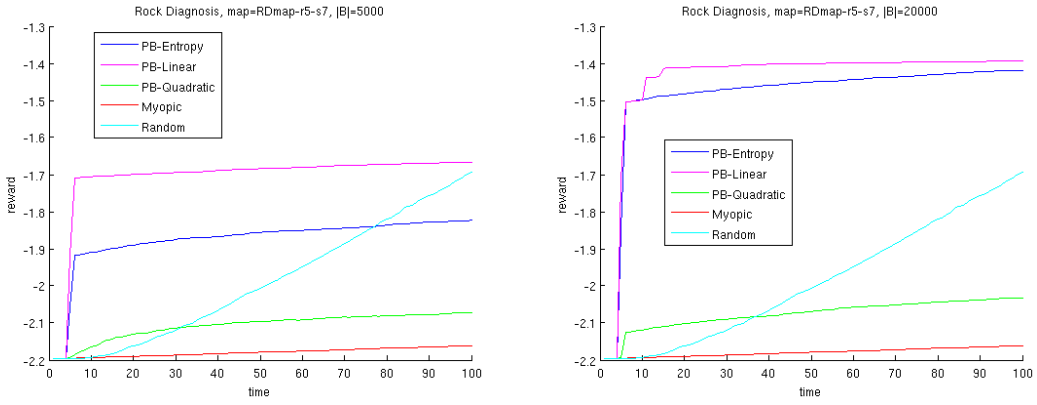

FIGURE 5: Rock Diagnosis reward evolution for the 5-7 map over the first 100 steps. The left-hand figure shows the results for $|\mathcal{B}| = 5000$ and the right-hand one shows the results for $|\mathcal{B}| = 20000$.

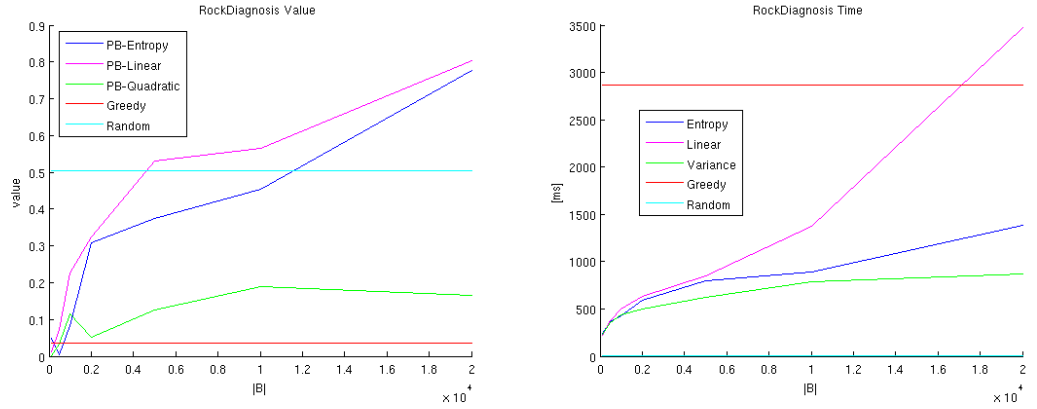

FIGURE 6: Rock Diagnosis value and time performance for the 5-7 map. The left-hand figure shows the value performance depending on the number of belief-points (between 100 and 20000), and the right-hand one shows on-line time performance for the same belief-point scale.

inspection of the trajectories generated by these policies confirms that they efficiently gather information for a first rock, but then they fail to examine the rest. Increasing the number of points obviously needs more computational effort, but for example the entropy reward with 20000 points outperforms random, and is faster than myopic. On the other hand, the quadratic reward fails to increase its performance with more points, and indeed this performance decreases with 20000 points. This result is unlikely explained by the variability of a random collection of points, because through all the set sizes the quadratic reward fails to provide a competitive results.

5 Conclusions and Future Work

In this paper, we have modeled and solved an active diagnosis problem using an negentropy performance criterion, as an practical example of the utility of ρPOMDPs. Our method allows solving active diagnosis and similar problems that need smart exploration generically through an information-lookahead strategy. In terms of empirical results, three different information-based rewards were tested, all of them outperforming in general the myopic and random approaches. Even though the myopic approach seems a very naive approach, it is a common technique in several communities that have addressed this problem. Regarding the performance comparison between the three information-based rewards, it seems that the negentropy one is suitable if dense approximations can be achieved, and the linear one is more efficient if only sparse approximations are available.

The performance of PB algorithms is highly dependent on the belief-point collection. We showed that the belief-set size plays an important role on the performance of information-lookahead strategies. Indeed, in POMDPs the collection methods are often the key difference between PB solvers (Shani *et al.*, 2012), so an obvious extension is to use more elaborated collection methods than the static randomized belief-set, for example dynamic set sizes and an error-based collection.

Also, further research is needed on the counter-intuitive result that a harsh linear approximation offers a better performance than the negentropy reward when only a small number of belief-points is available. Understanding why this phenomenon occurs may help to build faster and simpler algorithms to approximately solve active diagnosis problems.

A major issue in POMDP solvers is scalability, because the computational time usually grows exponentially with the dimensionality. Fortunately, real-world problems are usually highly structured, meaning that the transition and observation functions can often be factored to exploit this structure. This allows solving problems with less computational effort than when using a plain representation (Poupart, 2005). In particular, for most of the active diagnosis problems that need smart exploration, the state can be factored in a *visible* and a *hidden* part, which leads to use the Mixed Observable MDP formalism (Ong *et al.*, 2009; Araya-López *et al.*, 2010b).

Références

ARAYA-LÓPEZ M. (2013). *Near-Optimal Algorithms for Sequential Information-Gathering Decision Problems*. PhD thesis, Lorraine University, Nancy, Lorraine, France.

ARAYA-LÓPEZ M., BUFFET O., THOMAS V. & CHARPILLET F. (2010a). A POMDP extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems 23 (NIPS)*.

ARAYA-LÓPEZ M., THOMAS V., BUFFET O. & CHARPILLET F. (2010b). A closer look at MOMDPs. In *Proceedings of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*.

ASTROM K. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, **10**(1), 174 – 205.

BELLMAN R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, **60**, 503–516.

CASSANDRA A. R. (1998). *Exact and approximate algorithms for partially observable Markov decision processes*. PhD thesis, Brown University, Providence, RI, USA.

CASSANDRA A. R., LITTMAN M. L. & ZHANG N. L. (1997). Incremental pruning : A simple, fast, exact method for partially observable Markov decision processes. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI)*, p. 54–61.

COVER T. M. & THOMAS J. A. (1991). *Elements of Information Theory*. Wiley-Interscience.

HAUSKRECHT M. (2000). Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research (JAIR)*, **13**, 33–94.

HOWARD R. A. (1960). *Dynamic Programming and Markov Processes*. Cambridge, Massachusetts : MIT˜Press.

ISHIDA Y. (1997). Active diagnosis by self-organization : An approach by the immune network metaphor. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 1084–1091.

JI S. & CARIN L. (2007). Cost-sensitive feature acquisition and classification. *Pattern Recognition*, **40**(5), 1474–1485.

KRAUSE A. & GUESTRIN C. (2005). Near-optimal nonmyopic value of information in graphical models. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

KRISHNAMURTHY V. (2002). Algorithms for optimal scheduling and management of hidden Markov model sensors. *IEEE Transactions on Signal Processing*, **50**(6), 1382–1397.

KULLBACK S. & LEIBLER R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 49–86.

KURNIAWATI H., HSU D. & LEE W. S. (2008). SARSOP : Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics : Science and Systems IV*.

LITTMAN M. L. (1996). *Algorithms for sequential decision-making*. PhD thesis, Providence, RI, USA.

LOVEJOY W. S. (1991). Computationally feasible bounds for partially observed Markov decision processes. *Operations Research*, **39**(1), 162–175.

MADANI O., HANKS S. & CONDON A. (2003). On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence*, **147**(1-2), 5–34.

NEMHAUSER G. L., WOLSEY L. A. & FISHER M. L. (1978). An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming*, **14**(1), 265–294.

ONG S. C., PNG S. W., HSU D. & LEE W. S. (2009). POMDPs for robotic tasks with mixed observability. In *Proceedings of Robotics : Science and Systems V (RSS)*.

PELLEGRINI J. & WAINER J. (2003). On the use of POMDPs to model diagnosis and treatment of diseases. In *IV Encontro Nacional de Inteligencia Artificial, 2003, Campinas. IV Encontro Nacional de Inteligencia Artificial.*

PINEAU J., GORDON G. & THRUN S. (2003). Point-based value iteration : An anytime algorithm for POMDPs. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, p. 1025–1032.

POUPART P. (2005). *Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes*. PhD thesis, University of Toronto.

POUPART P., KIM K.-E. & KIM D. (2011). Closing the gap : Improved bounds on optimal POMDP solutions. In F. BACCHUS, C. DOMSHLAK, S. EDELKAMP & M. HELMERT, Eds., *Proceeding of International Conference on Automated Planning and Scheduling (ICAPS)*.

PUTERMAN M. (1994). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. Wiley-Interscience.

REZAEIAN M. (2007). Sensor scheduling for optimal observability using estimation entropy. In *Proceedings of the 5th IEEE International Conference on Pervasive Computing and Communications Workshops*, p. 307–312, Washington, DC, USA : IEEE Computer Society.

SAIGAL R. (1979). On piecewise linear approximations to smooth mappings. *Mathematics of Operations Research*, **4**(2), 153–161.

SAIGOL Z. A., DEARDEN R. W., WYATT J. L. & MURTON B. J. (2009). Information-lookahead planning for AUV mapping. In *Proceedings of the International Joint Conference on Artifical intelligence (IJCAI)*, p. 1831–1836, San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

SHANI G., BRAFMAN R. I. & SHIMONY S. E. (2007). Forward search value iteration for POMDPs. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

SHANI G., PINEAU J. & KAPLOW R. (2012). A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System technical journal*, **27**, 379–423.

SINGH A., KRAUSE A., GUESTRIN C. & KAISER W. J. (2009). Efficient informative sensing using multiple robots. volume 34, p. 707–755.

SMALLWOOD R. D. & SONDIK E. J. (1973). The optimal control of partially observable Markov decision processes over a finite horizon. *Operation Research*, **21**, 1071–1088.

SMITH T. & SIMMONS R. G. (2004). Heuristic search value iteration for POMDPs. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*.

SMITH T. & SIMMONS R. G. (2005). Point-based POMDP algorithms : Improved analysis and implementation. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.

SPAAN M. & VLASSIS N. (2005). Perseus : Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research (JAIR)*, **24**, 195–220.

VOGEL J. & MURPHY K. (2007). A non-myopic approach to visual search. In *Proceedings of the 4th Canadian Conference on Computer and Robot Vision*, p. 227–234, Washington, DC, USA : IEEE Computer Society.

WILLIAMS J. L. (2007). *Information theoretic sensor management*. PhD thesis, Cambridge, MA, USA.

ZHENG A. X., RISH I. & BEYGELZIMER A. (2005). Efficient test selection in active diagnosis via entropy approximation. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI)*.