

# Belief selection in point-based planning algorithms for POMDPs

Masoumeh T. Izadi<sup>1</sup>, Doina Precup<sup>1</sup>, Danielle Azar<sup>2</sup>

<sup>1</sup>McGill University, <sup>2</sup>American Lebanese University Byblos

**Abstract.** Current point-based planning algorithms for solving partially observable Markov decision processes (POMDPs) have demonstrated that a good approximation of the value function can be derived by interpolation from the values of a specially selected set of points. The performance of these algorithms can be improved by eliminating unnecessary backups or concentrating on more important points in the belief simplex. We study three methods designed to improve point-based value iteration algorithms. The first two methods are based on reachability analysis on the POMDP belief space. This approach relies on prioritizing the beliefs based on how they are reached from the given initial belief state. The third approach is motivated by the observation that beliefs which are the most overestimated or underestimated have greater influence on the precision of value function than other beliefs. We present an empirical evaluation illustrating how the performance of point-based value iteration (Pineau et al., 2003) varies with these approaches.

## 1 Introduction

Partially Observable Markov Decision Processes (POMDPs) are a standard framework for studying decision making under uncertainty. In POMDPs, the state of the system in which the decisions take place is never fully observed. Only observations that depend probabilistically on the hidden state are available. The best exact algorithms for POMDPs can be very inefficient in both space and time. Therefore a huge research effort has been devoted to developing approximation techniques in this field. Most planning algorithms attempt to estimate values for belief states, i.e. probability distributions over the hidden states of the system.

Recent research has been devoted to algorithms that take advantage of the fact that for most POMDP problems, a large part of the belief space is never experienced by the agent. Such approaches, which are known as point-based methods, consider only a finite set of belief points and compute values for the different actions only for these points. The plan generalization over the entire simplex is done based on the assumption that nearby points will have nearby values. Point-based value iteration methods (Pineau et al., 2003) have been very successful in solving problems which are orders of magnitude larger than classical POMDP problems. This algorithm performs point-based updates on a small set  $B$  of reachable points. The error of the approximation is proved to be bounded and it can be decreased by expanding the set of beliefs. However, value improvement depends to a large extent on which belief points are added to this set. Hence, the choice of belief points is a crucial problem in point-based value iteration, especially when dealing with large problems, and has been discussed by several authors. Spaan

and Vlassis (2004) explored the use of a large set of randomly generated reachable points. Pineau et al. (2003) discussed several heuristics for sampling reachable belief states. Smith and Simmons (2004) designed a heuristic search value iteration algorithm which maintains an upper and lower bound on the value function to guide the search for more beneficial beliefs to consider for backups.

In this paper we address the issue of dynamically generating a good ordering of beliefs in an efficient way. We explore the point-based value iteration algorithm in combination with several belief point selection heuristics. First, we make some corrections to the reachability metric proposed by Smith and Simmons (2005) which were discovered via private communications with the authors. This metric is designed to give more priority of being selected to points that are reachable in the near future. The intuition is that in discounted reward problems, belief points that are only reachable in many time steps do not play an important part in the computation of the value function approximation and we can ignore them. We compare this metric to the 1-norm distance metric previously suggested by Pineau et.al (2003) and study the applicability of this metric to the point-based value iteration algorithm. We also propose and investigate new methods for point selection in belief space for PBVI based on reachability analysis and belief value estimation error. Empirical results comparing these approaches is provided.

## 2 Background on Partially Observable Markov Decision Processes

Formally, a POMDP is defined by the following components: a finite set of hidden states  $S$ ; a finite set of actions  $A$ ; a finite set of observations  $Z$ ; a transition function  $T : S \times A \times S \rightarrow [0, 1]$ , such that  $T(s, a, s')$  is the probability that the agent will end up in state  $s'$  after taking action  $a$  in state  $s$ ; an observation function  $O : A \times S \times Z \rightarrow [0, 1]$ , such that  $O(a, s', z)$  gives the probability that the agent receives observation  $z$  after taking action  $a$  and getting to state  $s'$ ; an initial belief state  $b_0$ , which is a probability distribution over the set of hidden states  $S$ ; and a reward function  $R : S \times A \times S \rightarrow \mathfrak{R}$ , such that  $R(s, a, s')$  is the immediate reward received when the agent takes action  $a$  in hidden state  $s$  and ends up in state  $s'$ . Additionally, there can be a discount factor,  $\gamma \in (0, 1)$ , which is used to weigh less rewards received farther into the future.

The goal of planning in a POMDP environment is to find a way of choosing actions, or policy  $\pi$ , which maximizes the expected sum of future rewards

$$V^\pi(b) = E \left[ \sum_{t=0}^T \gamma^t r_{t+1} | b, \pi \right] \quad (1)$$

where  $T$  is the number of time steps in an episode (typically assumed finite) and  $r_{t+1}$  denotes the reward received at time step  $t + 1$ . The agent in a POMDP does not have knowledge of the hidden states, it only perceives the world through noisy observations as defined by the observation function  $O$ . Hence, the agent must keep a complete history of its actions and observations, or a sufficient statistic of this history, in order to act optimally. The sufficient statistic in a POMDP is the belief state  $b$ , which is a vector of length  $|S|$  specifying a probability distribution over hidden states. The elements of this vector,  $b(i)$ , specify the conditional probability of the agent being in state  $s_i$ , given the

initial belief  $b_0$  and the history (sequence of actions and observations) experienced so far.

After taking action  $a$  and receiving observation  $z$ , the agent updates its belief state using Bayes' Rule:

$$b'_{baz}(s') = P(s'|b, a, z) = \frac{O(a, s', z) \sum_{s \in S} b(s) T(s, a, s')}{P(z|a, b)} \quad (2)$$

where denominator is a normalizing constant and is given by the sum of the numerator over all values of  $s' \in S$ :

$$P(z|a, b) = \sum_{s \in S} b(s) \sum_{s' \in S} T(s, a, s') O(a, s', z)$$

We can transform a POMDP into a "belief state MDP" (Cassandra et al, 1997). Under this transformation, the belief state  $b$  becomes the (continuous) state of the MDP. The actions of the belief MDP are the same as in the original POMDP, but the transition and reward functions are transformed appropriately, yielding the following form of Bellman optimality equation for computing the optimal value function,  $V^*$ :

$$V^*(b) = \max_{a \in A} \sum_{z \in Z} P(z|a, b) \left[ \sum_{s \in S} b(s) \left( \sum_{s'} b'_{baz}(s') R(s, a, s') \right) + \gamma V^*(b'_{baz}) \right] \quad (3)$$

where  $b'_{baz}$  is the unique belief state computed based on  $b$ ,  $a$  and  $z$ , as in equation (2). As in MDPs, the optimal policy that the agent is trying to learn is greedy with respect to this optimal value function. The problem here is that there is an infinite number of belief states  $b$ , so solving this equation exactly is very difficult.

Exact solution methods for POMDPs take advantage of the fact that value functions for belief MDPs are piecewise-linear and convex, and thus can be represented using a finite number of hyperplanes in the space of beliefs YSondik1971. Value iteration updates can be performed directly on these hyperplanes. Unfortunately, exact value iteration is intractable for most POMDP problems with more than a few states, because the size of the set of hyperplanes defining the value function can grow exponentially with each step. For any fixed horizon  $n$ , the value function can be represented using a set of  $\alpha$ -vectors. The value function is the upper bound over all the  $\alpha$ -vectors:  $V_n(b) = \max_{\alpha} \sum_s \alpha(s) b(s)$ . Given  $V_{n-1}$ ,  $V_n$  can be obtained using the following backup operator:

$$V_n(b) \leftarrow \max_{a \in A} \left[ \sum_{z \in Z} P(z|a, b) \left( \sum_{s \in S} \sum_{s' \in S} b(s) b'_{baz}(s') R(s, a, s') + \gamma \max_{\alpha_{n-1}} \sum_{s' \in S} b'_{baz}(s') \alpha_{n-1}(s') \right) \right]$$

where  $\alpha_{n-1}$  are the  $\alpha$ -vectors used to represent  $V_{n-1}$ .

Exact value iteration algorithms (e.g. Sondik, 1971; Cassandra et al, 1997; Zhang & Zhang, 2001) perform this backup by manipulating directly the  $\alpha$ -vectors, using set projection and pruning operations. Although many  $\alpha$ -vectors can usually be pruned without affecting the values, this approach is still prohibitive for large tasks. Approximate methods attempt instead to approximate the value function in some way. These solution methods usually rely on maintaining hyperplanes only for a subset of the belief simplex. Different methods use different heuristics in order to define which belief points are of interest (e.g. Hauskrecht, 2000; Pineau et al, 2003; Smith and Simmons, 2004).

### 3 Point Based Value Iteration

The computational inefficiency of exact value updates leads to the exploration of various approximation methods that can provide good control solutions with less computational effort. Point-based value iteration (PBVI) is based on the idea of maintaining values, and  $\alpha$ -vectors for a selected set of belief points. This approach is designed based on the intuition that much of the belief simplex will not be reachable in general.

The algorithm starts with a set of beliefs, and computes  $\alpha$ -vectors only for these beliefs. The belief set  $B$  can then be expanded, in order to cover more of the belief space. New  $\alpha$ -vectors can then be computed for the new belief set, and the algorithm continues.

The update function with a fixed set of belief points  $B$  can be expressed as an operator  $H$  on the space of value functions, such that  $V_{i+1} = HV_i$ , where  $H$  is defined in (3). In order to show the convergence of such algorithms, we need to show that  $H$  is a contraction mapping, and that each estimate  $V_i$  is an upper bound on the optimal value function. If both of these conditions hold, the algorithm will converge to a fixed point solution,  $\bar{V}^* \geq V^*$ .

### 4 Belief Point Selection

In the PBVI algorithm, the selection of the belief points that will be used to represent the value function is done in an anytime fashion, with the goal of covering as densely as possible the set of reachable beliefs. The belief set  $B$  is initialized with just the initial belief state,  $b_0$ . Then, the space of reachable beliefs is sampled by a forward simulation, taking one action and receiving one observation. Different heuristics for sampling points have been proposed in (Pineau et al, 2003), but the Stochastic Simulation by Explorative Action heuristic (SSEA) is considered to perform best in general. In this approach, all possible actions at a given belief state in  $B$  are considered. One observation is sampled for each action and the new belief states are computed using (2). Then, the algorithm greedily picks the belief state that is farthest away from  $B$ , in the sense of the  $L_1$  distance (also called 1-norm). Hence, the number of points in  $B$  at most doubles at each iteration, because at most one extra belief point is added for each existing belief. This heuristic is motivated by an analytical upper bound on the approximation error, which depends on the maximum  $L_1$  distance from any reachable belief to  $B$ :

$$\epsilon_B = \max_{b' \in \bar{\Delta}} \min_{b \in B} \|b - b'\|_1 \quad (4)$$

where  $\bar{\Delta}$  is the set of all reachable beliefs. This heuristic attempts to greedily reduce  $\epsilon_B$  as quickly as possible. The authors also discuss other approaches for picking belief states, such as picking beliefs randomly (similarly to grid-based methods), or sampling reachable beliefs by using either random or greedy actions. In all of these cases (with the exception of random sampling), the space of reachable beliefs is covered gradually, as the algorithm progresses. This is due to the fact that at each point, the candidate beliefs that are considered are reachable in one time step from the current beliefs. Moreover, because PBVI is greedy in picking the next belief state to add, it can potentially overlook, at least for a few iterations, belief states that are important from the point

of view of estimating the value function. Spaan and Vlassis (2004) propose a different way of choosing belief points in the PERSEUS algorithm, which is aimed at addressing this problem. They sample a large set of reachable beliefs  $B$  during a random walk, but then only update a subset of  $B$ , which is sufficient to improve the value function estimate overall. The core belief selection heuristic proposed by Izadi et al. (2005) tries to start a point-based value iteration with a set of beliefs which span the whole simplex of reachable beliefs. Although this will help the algorithm converge to a good approximation in only a few iterations, finding this set of desired beliefs is computationally demanding, and the approach is problematic for large problems. Heuristic search value iteration (HSVI) (Smith & Simmons, 2004) keeps the value function bounded between an upper bound and a lower bound, an approach aimed at ensuring good performance for the candidate control policies. The algorithm was improved in (Smith & Simmons, 2005), by designing a tighter bound and much smaller size controllers, which makes HSVI better in terms of planning time and achieved values. However, the derivation contained in their paper has some problems, which we correct below.

#### 4.1 Selecting belief states based on reachability

In order to reason about the space of reachable beliefs, one can consider the initial belief vector,  $b_0$ , and all possible one-step sequences of actions and observations following it. Equation (2) then defines the set of all beliefs reachable in one step. By considering all one-step action-observation sequences that can occur from these beliefs, we can obtain all beliefs reachable from  $b_0$  in two steps, and so on. This will produce a tree rooted at the initial belief state  $b_0$ . The space of reachable beliefs consists of all the nodes of this tree (which is infinite in general, but cut to a finite depth in finite-horizon tasks). The discounted reachability  $\rho$  is a mapping from the space of reachable beliefs  $\Delta$  to the real numbers, defined as:  $\rho(b) = \gamma^L$  where  $L$  is the length of the shortest sequence of transitions from the initial belief state  $b_0$  to  $b$ . This definition implies that

$$\rho(b'_{baz}) \geq \gamma\rho(b) \quad (5)$$

because either  $b'_{baz}$  is obtained in one step from  $b$  (in which case we have equality), or if not, it may be obtained along a shorter path. Based on the definition of discounted reachability, Smith and Simmons (2005) define a generalized sample spacing measure  $\delta_P$  ( $0 \leq P < 1$ ). Their argument is that they want to give more weight to beliefs that are reachable in the near future, because their values influence the value function estimates more. To do this, they divide the  $L_1$  norm of the beliefs by  $(\rho(b))^P$ . However, this division actually has an opposite effect, emphasizing more beliefs that are in the distant future. The problem was discovered via private communications with the authors. To correct this problem, the sample spacing measure should be:

$$\delta_P(B) = \max_{b \in \Delta} \min_{b' \in B} \|b - b'\|_1 [\rho(b)]^P \quad (6)$$

where  $P$  is a parameter in  $[0, 1)$ . Note that if  $P = 0$ , we obtain exactly the heuristic described in (Pineau et al, 2003), and given here in equation (4). However, the theoretical development for that case has to be different than the one we will give here. In this case, an equal weight is given to beliefs that are at equal distance to the current set of

belief points, regardless how easy or hard they are to reach. The theoretical results and arguments stated in Smith and Simmons (2005) hold with the above definition of  $\delta_P(B)$ . For clarity, we present these results in here.

First we need to show that the update operator on the selected set of points by the above metric is a contraction mapping. To do this, consider the following weighted norm:

$$\|V - \bar{V}\|_{\xi} = \max_b |V(b) - \bar{V}(b)|\xi(b) \quad (7)$$

In other words, this is like a max norm but the elements are weighted by weights  $\xi$ .

**Theorem 1.** *The exact Bellman update is a contraction mapping under the norm (7) with contraction factor  $\gamma^{1-P}$ .*

**Proof:** For the proof, it is easier to consider action-value functions. We will use the same notation as (Smith and Simmons, 2005) in order to facilitate the comparison with their results. Let  $Q_a^V(b)$  be the value of executing action  $a$  in belief state  $b$ , given that the value function estimate for states is  $v$ :

$$Q_a^V(b) = R(b, a) + \gamma \sum_{b'} Pr(b'|b, a) V(b')$$

For any action  $a \in A$ , and for any value function estimators  $V$  and  $\bar{V}$  we have:

$$\begin{aligned} \|Q_a^V - Q_a^{\bar{V}}\|_{\rho^p} &= \max_b |Q_a^V(b) - Q_a^{\bar{V}}(b)| \times [\rho(b)]^p \\ &= \max_b \gamma \sum_{b'} Pr(b'|b, a) |V(b') - \bar{V}(b')| [\rho(b)]^p \\ &= \max_b \gamma \sum_{b'} Pr(b'|b, a) |V(b') - \bar{V}(b')| \left[ \frac{\gamma \rho(b)}{\gamma} \right]^p \\ &\leq \max_b \gamma \sum_{b'} Pr(b'|b, a) |V(b') - \bar{V}(b')| [\gamma^{-1} \rho(b')]^p \text{ (using (5))} \\ &\leq \max_b \gamma^{1-P} \sum_{b'} Pr(b'|b, a) \max_{b''} |V(b'') - \bar{V}(b'')| [\rho(b'')]^p \\ &= \gamma^{1-P} \sum_{b'} Pr(b'|b, a) \|V - \bar{V}\|_{\rho^p} \\ &= \gamma^{1-P} \|V - \bar{V}\|_{\rho^p} \end{aligned}$$

As a side note we need to mention that equation (10) in YSmith Simmons2005 and the contraction factor as stated there should be fixed as well.

Let  $HV$  be a ‘‘greedification’’ operator on the value function, defined as:  $HV(b) = \max_a Q_a^V(b), \forall b$ . Then, for any belief state  $b$  in the reachable belief space  $\Delta$  we have:

$$|HV(b) - H\bar{V}(b)| \leq \max_a |Q_a^V(b) - Q_a^{\bar{V}}(b)|$$

Multiplying both sides by  $[\rho(b)]^p$  we obtain:

$$|HV(b) - H\bar{V}(b)|[\rho(b)]^p \leq \max_a |Q_a^V(b) - Q_a^{\bar{V}}(b)|[\rho(b)]^p$$

Maximizing over  $b$ , we obtain:

$$\|HV - H\bar{V}\|_{\rho^p} \leq \max_a \|Q_a^V - Q_a^{\bar{V}}\|_{\rho^p} \leq \gamma^{1-p} \|V - \bar{V}\|_{\rho^p}$$

which completes the proof.  $\diamond$ .

The next theorem bounds the error of a policy based on an approximate value function  $\hat{V}$ .

**Theorem 2.** *The expected error introduced by a policy  $\hat{\pi}$  induced by an approximate value function  $\hat{V}$ , starting at the initial belief  $b_0$  is bounded by:*

$$\frac{2\gamma^{1-p}}{1-\gamma^{1-p}} \|V^* - \hat{V}\|_{\rho^p}$$

**Proof:** Let  $b \in \Delta$  be an arbitrary belief state and  $\pi^*$  be the optimal policy. Let  $V^{\hat{\pi}}$  be the value function of policy  $\hat{\pi}$ . Note that  $Q_{\hat{\pi}(b)}^{V^{\hat{\pi}}}(b) = V^{\hat{\pi}}(b)$ . Note that by the definition of the  $H$  operator,  $Q_{\hat{\pi}(b)}^{\hat{V}}(b) = H\hat{V}(b)$ . Note also that for the optimal value function, by its definition,  $V^* = HV^*$ . We have:

$$\begin{aligned} |V^{\pi^*}(b) - V^{\hat{\pi}}(b)| &= |V^*(b) - Q_{\hat{\pi}(b)}^{V^{\hat{\pi}}}(b)| \\ &= |V^*(b) - Q_{\hat{\pi}(b)}^{V^{\hat{\pi}}}(b) + Q_{\hat{\pi}(b)}^{\hat{V}}(b) - Q_{\hat{\pi}(b)}^{\hat{V}}(b)| \\ &\leq |V^*(b) - HV^*| + |Q_{\hat{\pi}(b)}^{\hat{V}}(b) - Q_{\hat{\pi}(b)}^{V^{\hat{\pi}}}(b)| \text{ (by grouping terms)} \\ &\leq |HV^*(b) - H\hat{V}(b)| + \gamma \sum_{b'} Pr(b'|b, \hat{\pi}(b)) |\hat{V}(b') - V^{\hat{\pi}}(b')| \end{aligned}$$

Multiplying both sides by  $[\rho(b)]^p$  we get:

$$\begin{aligned} |V^*(b) - V^{\hat{\pi}}(b)| [\rho(b)]^p &\leq |HV^*(b) - H\hat{V}(b)| [\rho(b)]^p \\ &\quad + \gamma \sum_{b'} Pr(b'|b, \hat{\pi}(b)) |\hat{V}(b') - V^{\hat{\pi}}(b')| \left[ \frac{\rho(b)\gamma}{\gamma} \right]^p \\ &\leq |HV^*(b) - H\hat{V}(b)| [\rho(b)]^p \\ &\quad + \gamma^{1-p} \sum_{b'} Pr(b'|b, \hat{\pi}(b)) |\hat{V}(b') - V^{\hat{\pi}}(b')| [\rho(b')]^p \\ &\leq |HV^*(b) - H\hat{V}(b)| [\rho(b)]^p + \gamma^{1-p} \|\hat{V} - V^{\hat{\pi}}\|_{\rho^p} \end{aligned}$$

By taking a max wrt  $b$  we obtain:

$$\begin{aligned} \|V^{\pi^*} - V^{\hat{\pi}}\|_{\rho^p} &\leq \|HV^* - H\hat{V}\|_{\rho^p} + \gamma^{1-p} \|\hat{V} - V^{\hat{\pi}}\|_{\rho^p} \\ &\leq \gamma^{1-p} \left( \|V^* - \hat{V}\|_{\rho^p} + \|\hat{V} - V^{\hat{\pi}}\|_{\rho^p} \right) \\ &\leq \gamma^{1-p} \left( \|V^* - \hat{V}\|_{\rho^p} + \|\hat{V} - V^*\|_{\rho^p} + \|V^* - V^{\hat{\pi}}\|_{\rho^p} \right) \\ &\leq \gamma^{1-p} \left( 2\|V^* - \hat{V}\|_{\rho^p} + \|V^* - V^{\hat{\pi}}\|_{\rho^p} \right) \end{aligned}$$

Solving this we obtain:

$$\|V^* - V^{\hat{\pi}}\|_{\rho^p} \leq \frac{2\gamma^{1-p}}{1 - \gamma^{1-p}} \|V^* - \hat{V}\|_{\rho^p}$$

Hence, the regret at  $b_0$  will be bounded as follows:

$$V^*(b_0) - V^{\hat{\pi}}(b_0) \leq \frac{2\gamma^{1-p}}{1 - \gamma^{1-p}} \|V^* - \hat{V}\|_{\rho^p} \quad \diamond$$

**Theorem 3.** *Let  $H_B$  be the update operator applied using only beliefs from set  $B$ . Then the error induced by a single application of  $H_B$  instead of the true operator  $H$  is bounded in  $\rho^p$  norm as:*

$$\|HV - H_B V\|_{\rho^p} \leq \frac{(R_{max} - R_{min})\delta_P(B)}{1 - \gamma^{1-p}}$$

**Proof:** We follow a similar argument to the one in (Pineau et al, 2003). Let  $b'$  be the reachable belief that is currently not included in the set  $B$  with the worst error in  $\rho^p$  norm. Let  $b \in B$  be the belief on the current simples that is closest to  $b'$ , in the sense of the  $\rho^p$  norm. The true optimal  $\alpha$ -vector at  $b'$  would be  $\alpha'$ , but instead we use the estimate  $\alpha$  that comes from  $b$ . Then, we have:

$$\begin{aligned} \|HV - H_B V\|_{\rho^p} &= (\alpha' b' - \alpha b') \rho(b') = (\alpha' b' - \alpha' b + \alpha' b - \alpha b') \rho(b') \\ &\leq [\alpha'(b' - b) + \alpha(b - b')] [\rho(b)]^p \text{ (because } \alpha \text{ is the optimal belief at } b) \\ &= (\alpha' - \alpha)(b' - b) [\rho(b')]^p \\ &\leq \|\alpha' - \alpha\|_{\infty} \max_{b'} \min_b \|b' - b\|_{\infty} [\rho(b')]^p = \|\alpha' - \alpha\|_{\infty} \delta_P(B) \\ &\leq \frac{R_{max} - R_{min}}{1 - \gamma^{1-p}} \delta_P(B) \end{aligned}$$

where  $R_{max}$  and  $R_{min}$  are the maximum and minimum rewards that can be achieved, and we used the result from Theorem 1 for the contraction factor in the denominator.  $\diamond$

**Theorem 4.** *The error  $\|V_t - V_t^B\|$  at any update step  $t$  is at most:*

$$\frac{(R_{max} - R_{min})\delta_P(B)}{(1 - \gamma^{1-p})^2}$$

**Proof:** The proof is identical to the one in (Pineau et al., 2003) but with the results from Theorem 1 and Theorem 3 plugged in.  $\diamond$

Algorithm 1 presents an approach for expanding the belief set using the reachability heuristic. Note that instead of looking at all reachable beliefs, we just sample, for each belief in the current set, one possible successor. Of course, this algorithm could be changed to take more samples, or to look farther into the future. However, farther lookahead is less likely to matter, because of the weighting used in the heuristic. The drawback of this approach is that after a few cycles the strategy will sample points from a rather restricted set of points and that could lead to smaller and smaller improvements. It must be noted that for instance for domains with deterministic observations

---

**Algorithm 1** Average-norm Belief Expansion (Initial belief set  $B$ )

---

```
for all  $b \in B$  do
  for all  $a \in A$  do
    Sample the current state  $s$  from  $b$ 
    Sample the next state  $s'$  from  $T(s, a, \cdot)$ 
    Sample the next observation  $z$  from  $O(a, s', \cdot)$ 
    Compute the next belief  $b'_{baz}$  reachable from  $b$ 
  end for
   $b^* = \operatorname{argmax}_{b'_{baz}} \min_{b'' \in B} \|b'_{baz} - b''\|_1 [\rho(b'_{baz})]^P$ 
   $B = B \cup \{b^*\}$ 
end for
return  $B$ 
```

---

and transitions this approach gives very similar results to the stochastic simulation with explorative actions (SSEA) heuristic, because of the narrow distribution of reachable beliefs. Considering that PBVI converges to a good approximation in just a few expansions, and that the factor  $\gamma$  is usually between 0.75 to 0.99, the effect of  $\delta_P(B)$  is not much different, in Algorithm 1, compared to the effect of  $\varepsilon(B)$  in SSEA-PBVI.

## 4.2 Breath first selection

A more extreme alternative to the reachability heuristics is to include *all* the beliefs that are reachable in the near future in the set of belief points. Theoretically, this should provide the best approximation in terms of the weighted norm that we are considering. But in many problems this is not feasible, because the size of the fringe of the belief tree grows exponentially. But, in order to get an estimate of how well we could do in this case, we consider adding one belief point for every possible action from every belief in the current set. The observations are still sampled. The main idea is that we typically expect the number of actions to be small, but the number of possible observations to be quite large. Obviously, this could be extended to sample  $k$  observations for each action. The algorithm using this idea is presented in Algorithm 2. In this case the size of  $B$  for the next set of point-based backups will be increased at most by a factor of  $|A|$ . Because this approach adds significantly more beliefs at each step, we would expect it to obtain a good approximation in a smaller number of expansions. But we want to ensure that the number of beliefs is also small enough to enable efficient value function backups.

## 4.3 Value-based selection

One interesting feature of point-based methods is that they can use the current estimate of the value function itself to decide which belief points to select next. So far, though, only one point-based algorithm, Stochastic Simulation with Greedy Action, PBVI-SSGA introduced in (Pineau et al, 2005), exploits this feature. Value-based methods attempt to include critical points in the set of selected beliefs, based on the current value approximation. We build upon the fact that the reachable states with highest and lowest expected rewards (as predicted by the current value function approximation) are

---

**Algorithm 2** Breadth First Belief Expansion (Initial belief set  $B$ )

---

```
for all  $b \in B$  do
  for all  $a \in A$  do
    Sample current state  $s$  from  $b$ 
    Sample next state  $s'$  from  $T(s, a, \cdot)$ 
    Sample next observation  $z$  from  $O(a, s', \cdot)$ 
    Compute the next belief  $b'_{baz}$  reachable from  $b$ 
    if  $b'_{baz} \notin B$  then  $B = B \cup \{b'_{baz}\}$ 
  end for
end for
return  $B$ 
```

---

the more desirable points for improving the precision of the value function. This method is presented in Algorithm 3. As seen here, the set of beliefs  $B$  at most triples in size with each expansion.

---

**Algorithm 3** Value-based belief expansion (Initial belief space  $B$ )

---

```
for all  $b \in B$  do
  for all  $a \in A$  do
    Sample  $s$  from  $b$ 
    Sample  $s'$  from  $T(s, a, \cdot)$ 
    Sample  $z$  from  $O(a, s', \cdot)$ 
    Compute the next belief  $b'_{baz}$  reachable from  $b$ 
  end for
   $max_b = \arg \max_{b'_{baz}} b'_{baz} \alpha$  and  $min_b = \arg \min_{b'_{baz}} b'_{baz} \alpha$  where  $\alpha$  is the best belief vector at the respective beliefs
   $B = B \cup \{max_b, min_b\}$ 
end for
return  $B$ 
```

---

## 5 Empirical evaluation

In order to compare the performance of the belief selection methods discussed in the previous sections, we selected a few standard domains previously used in the literature. Table 1 lists these problems with information about the problem size.

In each domain, we ran 250 trajectories starting from a fixed given initial belief following the approximately optimal policy generated by each method. We measure the the discounted sum of the rewards obtained on these trajectories. Table 2 shows this measure averaged over 10 independent runs, for the hallway, hallway2 and RockSample problems, and five runs for the tag domain (due to time restrictions). We present average performance and standard deviation over these runs. The first column in this table shows the standard PBVI algorithm in which Stochastic Simulation with Explorative Action has been used for belief set expansion. For the second algorithm, the parameter  $p$  is chosen such that the resulting value function fits best the true value function according to the average-norm heuristic. The second column reports these results with  $P = 0.99$ ;

**Table 1.** Domains used in the experiments

Domain	$ S $	$ A $	$ O $	$\mathcal{R}$
hallway	60	5	21	[0,1]
hallway2	90	5	17	[0,1]
tag	870	5	30	[-10,10]
RockSample[4,4]	257	9	2	[-100,10]

however, we experimented with many different settings of  $P$  and all results are very similar. The third and fourth columns contain results for the breadth-first and value-based heuristics.

We used 5 expansions of the set of beliefs  $B$  to reach an optimal solution. for all of the algorithms except for breadth-first. For the latter, we performed 3 expansions for the hallways and RockSample problems and 2 expansions for the tag domain. This is because the set  $B$  grows much faster for this algorithm. The complexity of the optimal value function can be measured by the number of  $\alpha$ -vectors used to represent it. PBVI keeps at most one  $\alpha$ -vector for each belief state in the set  $B$ . In the domains hallway, hallway2, and tag, there are 5 choices of actions and a high level of stochasticity. In the RockSample domain, actions are deterministic, and there is significantly less noise in the observations.

**Table 2.** Comparison of solution quality between different belief point selection strategies

Domain	PBVI (1-norm)	Average-norm	Breadth-First	Value-based
hallway	$0.51 \pm 0.03$	$0.52 \pm 0.03$	$0.52 \pm 0.03$	$0.51 \pm 0.03$
hallway2	$0.35 \pm 0.03$	$0.37 \pm 0.04$	$0.38 \pm 0.03$	$0.30 \pm 0.04$
tag	$-9.12 \pm 0.59$	$-8.16 \pm 0.8$	$-9.27 \pm .68$	$-8.18 \pm 1.27$
RockSample[4,4]	$17.78 \pm 1.08$	$19.36 \pm 2.5$	$15.05 \pm 3.13$	$8.57 \pm 0.21$

In the experiments, the set  $B$  almost always contains 32 points for the 1-norm and average-norm heuristics. The average size of  $B$  is 66 for the value-based method, but in the RockSample domain, only 15 belief points are selected on average by this heuristic. This is mainly due to the fact that the deterministic transitions make it difficult to explore a large enough part of the belief simplex using this method. The small size of the belief set results in poor performance on this domain, compared to the other approaches. We conjecture that a different way of facilitating exploration, perhaps by starting from different initial beliefs, would help. The breadth-first heuristic is much more aggressive in expanding the belief set, averaging 150 belief points for the hallway problems and 294 beliefs for the RockSample domain.

Overall, neither the breadth-first nor the value-based heuristic seem to help much. The average-norm heuristic is preferable to the 1-norm, as it uses roughly the same number of belief points but provides a better quality solution. In general the effect of these methods is difficult to narrow down, and further experimentation with different

domains is required. We believe the exploration-exploitation trade-off should also be considered in future experimentation, since it impacts significantly the quality of the solutions we obtain.

## 6 Conclusions and future work

The set of points selected for value iteration in point-based methods is very important for the quality of the computed approximate plan. In this paper, we introduced and evaluated several point selection criteria for point-based POMDP approximation methods. First, we studied the reachability metric as an alternative to 1-norm distance between belief states. This approach gives a higher priority beliefs in the immediate future. We also tried considering all beliefs which are only one step away from our current set. The number of backed up belief points is considerably larger in this case, so in principle this can allow a better approximation, although in the examples we studied the control quality does not improve much. We also tested the idea of using the value as a guide to select points. The empirical results do not show a clear winner among all these methods; the exploration-exploitation trade-off seems to play an important role, and should be taken into consideration in future studies.

The methods discussed in this paper focus on belief selection for the expansion phase of the PBVI. However, different methods can be adopted for choosing only belief points to perform backups in the value iteration phase of this algorithm as well. We expect such methods to have a greater influence on the speed of point-based methods in general, which is also suggested by the results of Spaan and Vlassis. We intend to study this further in the future.

## References

- A. R. Cassandra, M. L. Littman, and L. P. Kaelbling. A simple, fast, exact methods for partially observable Markov decision processes. In *Proceedings of UAI*, pages 54-61, 1997
- Masoumeh T. Izadi, Ajit Rajwade, and Doina Precup. Using core beliefs for point-based value iteration. In *Proceedings of IJCAI*, pages 1751-1753, 2005.
- M. Hauskrecht. Value-function approximations for Partially Observable Markov Decision Processes. In *Journal of Artificial Intelligence Research*, vol. 13, pages 33-94, 2000.
- Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithms for POMDPs. In *Proceedings of IJCAI*, pages 1025-1032, 2003.
- Trey Smith, and Ried Simmons Heuristic search value iteration for POMDPs. In *Proceedings of UAI* pages 520-527, 2004.
- Trey Smith, and Ried Simmons Point-based POMDP Algorithm: Improved Analysis and Implementation. In *Proceedings of ICML*, 2005.
- E.J. Sondik The optimal control of Partially Observable Markov Processes. *Ph.D. thesis, Stanford University*, 1971.
- M.T.J. Spaan, and N. Vlassis Perseus: Randomized point-based value iteration for POMDPs. In *Journal of Artificial Intelligence Research*, pages 195-220, 2005.
- N.L. Zhang, and W. Zhang Speeding up the convergence of value iteration in partially observable Markov decision processes. In *Journal of Artificial Intelligence Research*, vol.14, pages 29-51, 2001.