Using core beliefs for point-based value iteration

Masoumeh T. Izadi
McGill University
School of Computer Science
mtabae@cs.mcgill.ca

Ajit V. Rajwade
University of Florida
CISE Department
avr@cise.ufl.edu

Doina Precup
McGill University
School of Computer Science
dprecup@cs.mcgill.ca

Abstract

Recent research on point-based approximation algorithms for POMDPs demonstrated that good solutions to POMDP problems can be obtained without considering the entire belief simplex. For instance, the Point Based Value Iteration (PBVI) algorithm [Pineau *et al.*, 2003] computes the value function only for a small set of belief states and iteratively adds more points to the set as needed. A key component of the algorithm is the strategy for selecting belief points, such that the space of reachable beliefs is well covered. This paper presents a new method for selecting an initial set of representative belief points, which relies on finding first the basis for the reachable belief simplex. Our approach has better worst-case performance than the original PBVI heuristic, and performs well in several standard POMDP tasks.

1 Introduction

Partially Observable Markov Decision Processes (POMDPs) provide a general framework for planning under uncertainty [Kaelbling *et al.*, 1998; Sondik, 1971]. One of the standard algorithms used to provide solutions to POMDPs is value iteration, which associates values to probability distributions over states. Because exact value iteration is intractable, a lot of recent work has focused on approximate algorithms, which rely on compressing the belief space or considering a finite set of reachable beliefs. A representative algorithm for this class is point-based value iteration (PBVI) [Pineau *et al.*, 2003]. In PBVI, a finite set of reachable belief points is selected heuristically, and values are computed only for these points. The success of PBVI depends crucially on the selection of the belief points. In particular, these points should cover the space of reachable beliefs as evenly as possible. The main PBVI heuristic uses one-step simulated trajectories in order to find reachable beliefs. The belief that is "farthest away" from the points already included is greedily added to the set. We explore an alternative heuristic for choosing belief points. We aim to find *core beliefs*, i.e. beliefs which form a basis for the space of all reachable beliefs. This approach is aimed at covering more quickly the space of beliefs that are reachable, by looking farther into the future. This improves the worst-case behavior of the PBVI algorithm. We discuss how core beliefs can be found, and we illustrate the behavior of the algorithm on several standard POMDP benchmarks.

2 POMDPs and point-based value iteration

A POMDP is described by a tuple $\langle \mathcal{S}, \mathcal{A}, O, T, \Omega, R, \gamma, b_0 \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions and O is a finite set of observations. The probability of any state at $t = 0$, is described by the initial belief, b_0. The transition function $T(s,a,s')$ describes the probability of going from state s to state s' given that action a is chosen. The observation function $\Omega(o,s,a)$ describes the probability of observation o given that action a was taken and state s was reached. The reward function, $R(s,a)$, specifies the expected immediate reward, obtained after executing action a in state s. The discount factor $\gamma \in (0,1)$ is used to weigh less delayed rewards. The goal of an agent acting in a POMDP is to find a way of choosing actions (policy) maximizing the future return: $E[\sum_t \gamma^t R(s_t,a_t)]$. In order to achieve this goal, the agent must keep either a complete history of its actions and observations, or a sufficient statistic of the history. The sufficient statistic in a POMDP is the belief state, b, a vector with $|\mathcal{S}|$ components, which represents a probability distribution over states:

$$b_t(s) = P(s_t = s | b_0, a_0 o_1 \ldots a_{t-1} o_t)$$

On each time step, after taking action a and making observation o, the agent updates its belief state using Bayes rule:

$$b(s') \leftarrow P(s'|a,o,b) = \frac{\Omega(o,s',a)\sum_s b(s) T(s,a,s')}{P(o|a,b)}, \quad (1)$$

where the denominator is just a normalizing constant.

In value-based POMDP methods, the agent computes a value function, which is maintained over belief states. A lot of recent research has been devoted to computing such values approximately. Compression-based methods [Roy and Gordon, 2003; Poupart and Boutilier, 2004] aim to compress the belief space, and maintain a value function only on this compressed version. Point-based value iteration (PBVI)[Pineau *et al.*, 2003] maintains values and gradients (α-vectors, in POMDP terminology) for a selected set of belief points, B. The idea is that much of the belief simplex will not be reachable in general. The set B is expanded over time, in order to cover more of the reachable belief space, and these expansions are interleaved with computing new value estimates.

3 Belief selection

In order to extend the belief set B, the standard PBVI heuristic considers, for each $b \in B$, all possible actions a and samples one observation o for each action. Among these beliefs reachable from b, the belief b' that is farthest away from B is picked and added to B. 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:

$$\varepsilon_B = \max_{b' \in \bar{\Delta}} \min_{b \in B} ||b - b'||_1$$

where $\bar{\Delta}$ is the set of all reachable beliefs. PBVI does not keep an upper bound on the value function when expanding its belief set. However Smith and Simmons [Smith and Simmons, 2004] have developed an alternative anytime solution method, HSVI, based on maintaining an upper and a lower bound of the optimal value function to guide the local value updates. They choose the belief points to update using forward heuristic search. Vlassis and Spaan [Vlassis and Spaan, 2004] instead sample a large set of reachable beliefs B during a random walk, but then only update a subset of of B, sufficient to improve values overall.

We propose an approach for covering the space of reachable beliefs more quickly, by considering longer courses of action, as well as by removing the strictly greedy character of the PBVI algorithm. We will use core beliefs, which form a basis for the belief simplex. The idea of core beliefs is inspired by work on predictive state representations (PSRs) [Littman *et al.*, 2002], an alternative way of representing stochastic dynamical systems. In order to reason about the space of reachable beliefs, one can consider the initial belief vector, b_0, and all possible sequences of actions and observations following it. These sequences are called tests. James and Singh (2004) introduced the concept of an infinite matrix, whose rows correspond to all histories and whose columns correspond to all tests. In this paper we use the same infinite matrix but the rows correspond to all reachable beliefs from a given initial belief point. For a row b and a test $a_0 o_1, \ldots a_{n-1} o_n$, the content of the corresponding element in the matrix is the probability that $o_1, \ldots o_n$ is observed given that action sequence $a_0, \ldots a_{n-1}$ is executed starting from b. This matrix has finite rank, at most equal to the number of states in the POMDP. We define core beliefs as the set of linearly independent rows of this matrix.

The computation of core beliefs can be done exactly using the POMDP model, as shown in Algorithm 1. It is similar to the computation of core tests in [James and Singh, 2004], except that here we compute the elements of the matrix having the POMDP model while they estimate these elements from data without the model. But in large environments, considering all one-step extensions for all tests is prohibitive. We experimented with two heuristics for the extensions:

1. Random: pick a random action, but consider all observations.
2. ε-greedy: consider actions that are likely to be picked by a good policy. We generate first a subset of the core beliefs based on all one-step tests. Then, we run PBVI on this subset, for a fixed number of iterations, in order to get a first approximation of the value function. We extend the core belief matrix by picking one action, in ε-greedy fashion. The next observation can be sampled as well.

Algorithm 1 Finding core beliefs

Initialize the matrix with b_0 and rows corresponding to all beliefs reachable by one-step tests a, o. One column is all 1's and the others correspond to one-step tests a, o.
for all elements (b, ao) **do**
  Compute $P(o|b,a)$, using (1) and store it in row b, column a, o
end for
repeat
  Compute a set of linearly independent rows and columns
  Add the linearly independent rows to the set of core beliefs
  Eliminate all linearly dependent rows and columns
  Add all one-step extensions to the matrix and compute the rank of the extended matrix
until The rank is unchanged

Note that for PBVI, the maximum error value for ε_B that can be achieved is 2, attained for the case in which two beliefs are non-zero over different states. This case cannot occur when all core beliefs are part of B. Hence, the worst-case error when using core beliefs is strictly smaller.

4 Experimental results

We performed experiments on several standard POMDP domains used in the literature. For the smaller three domains (4x4 grid, 4x3 grid and Cheese) we computed the core beliefs exactly (column "CBVI-all" of Table 1). The proposed heuristics are run using all possible observations in the one-step extensions. We also ran standard PBVI with a similar number of initial beliefs. In all cases we performed 5 iterations of PBVI to obtain a value function. Then, we ran 251 trials, with each trial cut at 251 steps. We averaged the results obtained over these trials, and over 5 independent runs.

All versions of the CBVI algorithm yield improvements over the PBVI heuristic. Surprisingly, in the 4x4 and Cheese domains, the best result is achieved by the ε-greedy heuristic. We conjecture that the reason is that the ε-greedy heuristic actually focuses the computation on good actions, that are likely to be used in the optimal policy.

In the larger domains, we only used the ε-greedy heuristic with sampled observations, as the other versions are too expensive. The experimental setup is as above. In this case, we find a first set of core beliefs, then run PBVI for 5 iterations. We take the resulting belief set, and add one-step extensions using ε-greedy actions and sampled observations, as above. This results in a new matrix, which yields a new set of core beliefs. This process is repeated at most 5 times; if no more core beliefs are detected in a new iteration, the process stops. The results are presented in the columns CBVI(1) to CBVI(5) in Table 2. Note that for the Coffee domain, all core beliefs are found after two such repetitions. Hence, there is no data for the CBVI(3)-CBVI(5). The reward for standard

Domain	PBVI	CBVI-all	CBVI-rand	CBVI-egr
4x4 grid	3.17 ± 0.22	3.80 ± 0.11	3.45 ± 0.22	3.96 ± 0.28
4x3 grid	1.58 ± 0.35	2.45 ± 0.44	2.10 ± 0.09	2.07 ± 0.11
Cheese	3.53 ± 0.13	3.53 ± 0.08	3.65 ± 0.22	3.81 ± 0.35

Table 1: Small domains

Domain	PBVI-orig	VS	PBVI	CBVI(1)	CBVI(2)	CBVI(3)	CBVI(4)	CBVI(5)
Maze33	2.25	2.34	2.248	2.185 ± 0.3	2.285 ± 0.31	2.165 ± 0.38	2.234 ± 0.25	2.301 ± 0.26
Hallway1	0.53	0.51	0.503	0.535 ± 0.03	0.617 ± 0.08	0.556 ± 0.03	0.574 ± 0.04	0.582 ± 0.08
Hallway2	0.35	0.31	0.379	0.333 ± 0.04	0.334 ± 0.04	0.392 ± 0.05	0.403 ± 0.04	0.400 ± 0.03
Coffee	-3.00	-	-2.759 ± 0.93	0.024 ± 2.14	0.58 ± 2.72	-	-	-

Table 2: Large domains

PBVI using a similar number of initial beliefs is in the column PBVI in Table 2. The results reported in the original PBVI paper (PBVI-orig)[Pineau *et al.*, 2003] and the results by Vlassis and Spaan (VS)[Vlassis and Spaan, 2004] are in the columns PBVI-org and VS respectively. As shown in Table 2, for Maze33 all algorithms obtain very similar results. The Coffee domain is a very favorable case for CBVI, because it has a nice structure, and the reachable space of beliefs lies in a very low dimensional part of the entire belief simplex. Generating core beliefs can be done in 2-3 seconds for this domain, because there are only two core beliefs for this problem, although its POMDP representation has 32 states. Our experiments show a huge value improvemnet for this particular case over PBVI. For the larger domains, Hallway1 and Hallway2, CBVI also has a definite advantage, especially in the later iterations, when the initial subset of core beliefs is larger. The time complexity of CBVI is the same as PBVI if generating the core beliefs can be done in a preprocessing phase. In small domains and large domains with small intrinsic dimensionality (e.g. Coffee) this can be done in a few seconds. However, generating all core beliefs in large domains in general is very expensive. In our experiments, we reached the maximum number of core beliefs in at most five extensions of the matrix in algorithm 1; however, the computation time, taking into account the process of belief generation is two orders of magnitude longer than for PBVI. Although this approach improves the value function and the policy that is obtained, it is significantly more expensive, at the moment, unless the problem at hand has special structure. We are investigating different ways to compute a reasonably large portion of the space of core beliefs at a lower computation cost.

Acknowledgments

This research was supported in part by funding from NSERC and CFI. We thank Joelle Pineau for very helpful discussions and insights into point-based algorithms, as well as for constructive feedback. We thank Matthijs Spaan for providing code for his point-based value iteration algorithm.

References

[James and Singh, 2004] Michael James and Satinder Singh. Predictive State Representations: A New Theory for Modeling Dynamical Systems. *ICML 21*, pages 417-424, 2004.

[Kaelbling *et al.*, 1998] Leslie P. Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[Littman *et al.*, 2002] Michael Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *NIPS 14*, pages 1555–1561, 2002.

[Pineau *et al.*, 2003] Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithms for POMDPs. In *Proceedings of IJCAI*, pages 1025–1032, 2003.

[Poupart and Boutilier, 2004] Pascal Poupart and Craig Boutilier. VDCBPI: an approximate scalable algorithm for large scale POMDPs. In *NIPS 17*, 2004.

[Roy and Gordon, 2003] Nicholas Roy and Geoffrey Gordon. Exponential family PCA for belief compression in POMDPs. In *NIPS 15*, pages 1635–1642, 2003.

[Smith and Simmons, 2004] Tery Smith and Reid Simmons. Heuristic Search Value Iteration for POMDPs. In *UAI 20*, pages 520-527, 2004.

[Sondik, 1971] Edward J. Sondik. *The optimal control of Partially Observable Markov Decision Processes*. PhD Thesis, Stanford University, 1971.

[Vlassis and Spaan, 2004] Nikos Vlassis and Matthijs Spaan. A point-based POMDP algorithm for robot planning. In *Proceedings of ICRA*, 2004.