Radboud University Nijmegen

Faculty of Science

Belief-dependent Rewards for POMDPs

-

Thesis BSc Mathematics

Supervisor:
dr. Nils Jansen

Author:
Serena Rietbergen

Second reader:
- -

-

Contents

Chapter 1

Introduction

Chapter 2

Description

Definition 2.1. A POMDP is a tuple $(S, \mathcal{A}, \Omega, T, O, r, b_0)$, where

- S, the state space;
- A, the action space;
- Ω, the observation space;
- $T : S \times \mathcal{A} \times S \to [0, 1]$, the transition function where $T(s, a, s') = Pr(s' \mid s, a)$;
- $O : S \times \mathcal{A} \times \Omega \to [0, 1]$, the observation function where $O(s', a, o) = Pr(o \mid s', a)$
- $r : S \times \mathcal{A} \to \mathbb{R}$, the reward function, and
- $b_0 : S \to [0, 1]$, the initial probability distribution over states.

TO WRITE: how/why belief mdp are nice

TO WRITE: belief states and how they're updated

A belief state $b \in \Delta = \Pi(S)$.

Definition 2.2. Let $b^{a,o}$ be the probability distribution after performing an action a and then observing observation o. The probability of ending in state s' after performing a and observing o then can be calculated as follows, where $Pr(o|a, b) = \sum_{s,s' \in S} O(s'', a, o)T(s, a, s'')b(s)$,

$$b^{a,o}(s') = \frac{O(s, a, o)}{Pr(o|a, b)} \sum_{s \in S} T(s, a, s')b(s) \tag{2.1}$$

Definition 2.3. A POMDP $(S, A, \Omega, T, O, r, b_0)$ can be written as a belief MDP $(\Delta, \mathcal{A}, \tau, \rho)$, where

- $\Delta = \Pi(S)$, the set of probability distributions over S;
- $\mathcal{A}$, the action space;
- $\tau : \Delta \times \mathcal{A} \times \Delta \to [0, 1]$, the new transition function
- $\rho : \Delta \times \mathcal{A} \to \mathbb{R}$, the new reward function.

A belief MDP is continuous, so

TO WRITE: what does continututity have to do w/ it?

TO WRITE: why wanna have optimal expected reward

The objective is to maximize the expected cumulative reward, which is done by looking for an optimal policy. Let γ be some discount factor and ρ_t be the expected immediate reward obtained at a certain step t, then the expected cumulative reward is defined as

$$J^{\pi}(b) = E\big[\sum_{t=0}^{\infty} \gamma \rho_t \mid b, \pi\big] \tag{2.2}$$

Then the optimal policy can be calculated as

$$\pi^* = \arg\max_{\pi \in \mathcal{A}^{\Delta}} J^{\pi}(b_0) \tag{2.3}$$

Definition 2.4. For $b \in \Delta$, let the value function be defined as

$$\begin{aligned} V_0(b) &= 0 \\ V_n(b) &= \max_{a \in \mathcal{A}} \big[\rho(b,a) + \gamma \int_{\Delta} \tau(b,a,b')V_{n-1}(b')db'\big] \\ &= \max_{a \in \mathcal{A}} \big[\rho(b,a) + \gamma \sum_{o \in \Omega} Pr(o \mid a,b) V_{n-1}(b^{a,o})\big] \end{aligned} \tag{2.4}$$

The function J^{π^*} can be computed recursively through this value function, due to Bellman's principle of optimality[1]. Let H be the possible infinite horizon of the problem, then$J^{\pi^*}(b) = V_{n=H}(b)$.

Observe that the reward function for a POMDP is based on a state and action through $r(s,a)$, while the reward function for a belief MDP depends on the a belief state and action through $\rho(b,a)$. The belief MDP reward function ρ can be derived from the POMDP reward function r :

$$\rho(b,a) = \sum_{s \in S} b(s) r(s,a) \tag{2.5}$$

When the ρ is defined as in Equation 2.5, the recursive computation in Equation 2.4 has the property to generate piecewise-linear and convex (PWLC) value functions for each horizon[2].

Definition 2.5. A function is piecewise-linear if it consists of n linear segments defined over n intervals.

Definition 2.6. A function $f : A \to B$ is convex if for all $\theta \in [0,1]$ and $x_1, x_2 \in A$

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta) f(x_2)$$

Let Γ_n be the set of vectors representing the value function for horizon n, then the value function can be rewritten to $V_n(b) = \max_{\alpha \in \Gamma_n} \sum_{s \in S} b(s)\alpha(s)$.

Using the PWLC property, Equation 2.4 can be refactored to be able to perform the Bellman update. Let $\chi_n(b) = \arg\max_{\alpha \in \Gamma_n} b\dot{\alpha}$ and $\chi_n(b,s) = (\chi_n(b))(s)$ in

$$V_n(b) = \max_{a \in \mathcal{A}} \sum_{o \in O} \sum_{s \in S} b(s) \big[\frac{r(s,a)}{|\Omega|} + \sum_{s' \in S} T(s,a,s') O(s',a,o) \chi_{n-1}(b^{a,o}, s')\big] \tag{2.6}$$

$$\overline{\Gamma_n}^{a,o} = \{ \frac{r^a}{|\Omega|} + P^{a,o} \cdot \alpha_{n-1} \mid \alpha_{n-1} \in \Gamma_{n-1} \} \tag{2.7}$$

Problem: a special kind of problem is when the performance criterion incorporates an explicit measure of the agent's knowledge about the system, which is based on the beliefs rather than states.

the reward of a pomdp cannot model this since it's only based on the current state and action. we prefer to consider a new way of defining rewards based on the acquired knowledge represented by belief states.

the direct link with pomdps is broken, but we can fix this by generalizing the pomdp framework to a *rho*-baes POMDP ρPOMDP, where the reward is not defined as a function $r(s, a)$ but directly as a function $\rho(b, a)$

$\rho(b, a)$ is not restricted to be only an uncertainty measurement, but can be a combination of the expected state-action rewards and an uncertainty or error measurement.

A belief-based value function is convex.

If ρ and V_0 are convex functions over $\Delta = \Pi(s)$, then the value function V_n of the belief MDP is convex over Δ at any time step n.

Last theorem is based on $\rho(b, a)$ being a convex function over b, which is a natural property. A reward function meant to reduce the uncertainty mush provide high payloads near the corners of the simplex and low payloads near its center. Which is why we're only focusing on reward functions that comply with convexity. V_0 might be any convex function for infinite-horizon problems, but $V_0 = 0$ for finite-horizon problems.

From now on ρ is a PWLC function, and can thus be represented as several Γ-sets, one Γ_ρ^a for each a:

$$\rho(b, a) = \max_{\alpha \in \Gamma_\rho^a} \big[\sum_s b(s) \alpha(s) \big] \tag{2.8}$$

Changes into a new value function

$$V_n(b) = \max_{a \in A} \sum_s b(s) \big[\operatorname*{arg\,max}_{\alpha \in \Gamma^a} \rho(b \cdot \alpha) + \sum_o \sum_{s'} T(s.a.s') O(s', a, o) \chi_{n-1}(b^{a,o}, s') \big] \tag{2.9}$$

$$\overline{\Gamma_n}^{a,o} = \{ P^{a,o} \cdot \alpha_{n-1} \mid \alpha_{n-1} \in \Gamma_{n-1} \} \tag{2.10}$$

Bibliography

[1] Mauricio Araya, Olivier Buffet, Vincent Thomas, and Françcois Charpillet. A pomdp extension with belief-dependent rewards. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 64–72. Curran Associates, Inc., 2010.

[2] Richard D. Smallwood and Edward J. Sondik. The optimal control of partially observable markov processes over a finite horizon. *Operations Research*, 21(5):1071–1088, 1973.