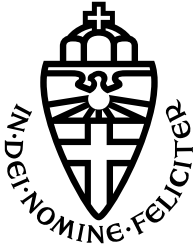


RADBOD UNIVERSITY NIJMEGEN



FACULTY OF SCIENCE

---

# Belief-dependent Rewards for POMDPs

---

THESIS BSc MATHEMATICS

*Author:*  
Serena RIETBERGEN

*Supervisor:*  
dr. Nils JANSEN

*Second reader:*  
- -

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Description</b>	<b>3</b>

## Chapter 1

# Introduction

# Chapter 2

## Description

**Definition 2.1** (Partially Observable Markov Decision Process). A Partially Observable Markov Decision Process, or POMDP, is a tuple  $(S, \mathcal{A}, \Omega, T, O, r, b_0)$ , where

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

A POMDP is only partially observable, which indicates that we do not know everything about the model. We specifically are unaware of what state a POMDP currently is in. To create a notion of what state we believe to be in, we use belief states.

**Definition 2.2** (Belief state). Given a POMDP  $(S, \mathcal{A}, \Omega, T, O, r, b_0)$ , a belief state  $b$  is a probability distribution over  $S$  called  $\Delta$ . The probability of the POMDP being in state  $s$  according to belief  $b$ , is denoted by  $b(s)$ . Logically  $\sum_{s \in S} b(s) = 1$  and for all states  $s$  we have that  $b(s) \in [0, 1]$ .

A belief state evolves over time and a POMDP will start in their initial belief state  $b_0$ . The belief state will be updated according to the action that is undertaken and an observation that will be observed after taking said action.

**Definition 2.3** (Updated belief state). Let  $Pr(o|a, b) = \sum_{s, s' \in S} O(s', a, o)T(s, a, s')b(s)$ . The probability of ending in state  $s'$  after performing an action  $a$  and then observing observation  $o$  is calculated as

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

With the help of belief states, we can create a fully observable continuous space belief MDP representing a POMDP.

**Definition 2.4** (Belief MDP). A POMDP  $(S, \mathcal{A}, \Omega, T, O, r, b_0)$  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;
- $\tau : \Delta \times \mathcal{A} \times \Delta \rightarrow [0, 1]$ , the new transition function
- $\rho : \Delta \times \mathcal{A} \rightarrow \mathbb{R}$ , the new reward function.

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

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

With this information, we can calculate the optimal policy

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

The value  $J^{\pi^*}(b_0)$  can be computed recursively through a value function  $V_n$ , due to Bellman's principle of optimality[1]. If we let  $H$  be the possible infinite horizon of the problem, the optimal value can be concluded from  $J^{\pi^*}(b) = V_{n=H}(b)$ .

**Definition 2.5** (Value function for POMDPs). For  $b \in \Delta$ , let the value function be defined as

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

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  $\rho$  can be derived from the POMDP reward function  $r$  :

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

When the  $\rho$  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[3].

**Definition 2.6** (Piecewise-linear function). A function is piecewise-linear if it consists of  $n$  linear segments defined over  $n$  intervals.

**Definition 2.7** (Convex function). A function  $f : A \rightarrow B$  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  $\Gamma_n$  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  $\chi_n(b, s) = (\chi_n(b))(s)$  where  $\chi_n(b) = \arg \max_{\alpha \in \Gamma_n} b \cdot \alpha$  in

what write  
about  
gamma  
and over-  
line gamma

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

TO WRITE: connection

Let  $r^a(s) = r(s, a)$  and  $P^{a,o}(s, s') = T(s, a, s') O(s', a, o)$  in

$$\bar{\Gamma}_n^{a,o} = \left\{ \frac{r^a}{|\Omega|} + P^{a,o} \cdot \alpha_{n-1} \mid \alpha_{n-1} \in \Gamma_{n-1} \right\} \quad (2.7)$$

$$\bar{\Gamma}_n = \bigcup_{a \in \mathcal{A}} \bigoplus_{o \in \Omega} \bar{\Gamma}_n^{a,o} \quad (2.8)$$

Where  $U \oplus W = \{u + w \mid u \in U, w \in W\}$ .

All problems who only have partial observability tackle a certain issue through getting information about the model or achieving some goal. Most of the time, the concerning problem is tackled through obtaining enough information get an optimal reward. However, some *active sensing* problems (where the objective is acting in a way to acquire knowledge about a part of the model) cannot be resolved this way. For these problems, the performance of an agent incorporates their obtained knowledge of the system. This is based upon beliefs, and not merely on states.

In Equation 2.5, we observe that the reward of a POMDP is solely based on the current state and action and does not incorporate the obtained knowledge. However, simply putting the entire history of the model in the state, would lead to a combinatorial explosion.

To tackle this issue, we propose a new method of defining rewards based on the acquired knowledge through the usage of belief states. Since we will use a different definition than the one described in Equation 2.5, the direct link with POMDPs is broken. Luckily, a lot of the components of a POMDP are still usable. For this we use a generalization of the POMDP framework, where the reward is directly defined as a function  $\rho(b, a)$ .

**Definition 2.8** ( $\rho$ POMDP). A  $\rho$ -based POMDP ( $\rho$ POMDP) is a tuple  $(S, \mathcal{A}, T, O, \rho, b_0)$ , where

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

We see that  $\rho$  is often just some uncertainty or error measure, but it can also be combined with some expected state-reward.

Observe that the belief-based value function for POMDPs is convex (Equation 2.4), because  $r(s, a)$  is linear in Equation 2.5. The value function for  $\rho$ -POMDPs is also convex, if  $\rho(b, a)$  is convex.

**Theorem 2.9.** *If  $\rho$  and  $V_0$  are convex functions over  $\delta$ , then the value function  $V_n$  of the belief MDP is convex over  $\Delta$  at any time step  $n$ .*

Note that  $\rho(b, a)$  being convex over  $\Delta$  is due to the fact that uncertainty measures are naturally convex[2]. We will only be focusing on reward functions that are convex. The initial value function  $V_0 = 0$  for finite-horizon problems, but can be any convex function for infinite-horizon problems.

Assume that  $\rho$  is a PWLC function. Therefore it can be represented as several  $\Gamma$ -sets; one  $\Gamma_\rho^a$  for each action in  $\mathcal{A}$ :

$$\rho(b, a) = \max_{\alpha \in \Gamma_\rho^a} \left[ \sum_s b(s) \alpha(s) \right] \quad (2.9)$$

Let  $\chi_\rho^a(b) = \arg \max_{\alpha \in \Gamma_\rho^a} (b \cdot \alpha)$  in  $\chi_\rho^a(b, s) = (\chi_\rho^a(b))(s)$ . This creates a new value function for  $\rho$ POMDPs as follows:

$$V_n(b) = \max_{a \in \mathcal{A}} \sum_{s \in S} b(s) \left[ \chi_\rho^a(b, s) + \sum_{o \in \Omega} \sum_{s' \in S} T(s, a, s') O(s', a, o) \chi_{n-1}(b^{a,o}, s') \right] \quad (2.10)$$

This new value function uses  $\Gamma_\rho^a$  and generates  $|\Omega| \times |\mathcal{A}|$   $\Gamma$ -sets, where  $P^{a,o}(s, s') = T(s, a, s') O(s', a, o)$ :

$$\bar{\Gamma}_n^{a,o} = \{P^{a,o} \cdot \alpha_{n-1} \mid \alpha_{n-1} \in \Gamma_{n-1}\} \quad (2.11)$$

Just like for POMDPs, we can use exact algorithms for  $\rho$ POMDPs. The difference is that now the cross-sum to calculate  $\bar{\Gamma}_n$  now includes

- one  $\alpha^{a,o}$  for each observation  $\Gamma$ -set  $\bar{\Gamma}_n^{a,o}$ , and
- one  $\alpha_\rho$  from the  $\Gamma$ -set  $\Gamma_\rho^a$  corresponding to the reward.

This results in the new cross-sum

$$\bar{\Gamma}_n = \bigcup_{a \in \mathcal{A}} \left[ \bigoplus_{o \in \Omega} \bar{\Gamma}_n^{a,o} \oplus \Gamma_\rho^a \right] \quad (2.12)$$

This equates for  $|R|$  times more vectors than with a regular POMDP. Let  $|R|$  be the number of  $\alpha$ -vectors specifying the  $\rho(b, a)$  functions. Note that  $|R|$  depends on the considered action taken.

# Bibliography

- [1] Mauricio Araya, Olivier Buffet, Vincent Thomas, and François 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] Mauricio Araya, Olivier Buffet, Vincent Thomas, and François Charpillet. A pomdp extension with belief-dependent rewards. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 64–72. Curran Associates, Inc., 2010.
- [3] 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.