Radboud University Nijmegen

Faculty of Science

Belief-dependent Rewards for POMDPs

-

Thesis BSc Mathematics

Author:
Serena Rietbergen

Supervisor:
dr. Nils Jansen

Second reader:
- -

-

Contents

Chapter 1

Introduction

Chapter 2

Description

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

- The state space S;
- The action space A;
- The transition function $T : S \times A \times S \to R$
  $T(s, a, s') = Pr(s' \mid s, a)$
- Observation space Ω
- Observation function $O : S \times A \times \Omega \to R$
  $O(s', a, o) = Pr(o \mid s', a)$
- A reward function $r : S \times A \to R$
- Initial probability distribution over states b_0

TO WRITE: how/why belief mdp are nice

TO WRITE: belief states and how they're updated

TO WRITE: definition of $b^{a,o}$

Definition 2.2. A POMDP $(S, A, \Omega, T, O, r, b_0)$ can be written as a belief MDP (Δ, A, τ, ρ), where

- $\Delta = \Pi(S)$, the set of probability distributions over S;
- action space A;
- $\tau : \Delta \times A \times \Delta \to R$, new transition function
- $\rho : \Delta \times A \to R$, new reward function.

TO WRITE: what is their connection to each other

the objective is to maximize the cumulative reward by looking for a policy taking the current belief state as input, i.e.

$$\pi^* = \arg\max_{\pi \in A^{\Delta}} J^{\pi}(b_0) \tag{2.1}$$

$$J^{\pi}(b_0) = E\big[\sum_{t=0}^{\infty} \gamma \rho_t \mid b_0, \pi\big] \tag{2.2}$$

$$V_n(b) = \max_{a \in A} \big[\rho(b,a) + \gamma \sum_o Pr(o \mid a,b) V_{n-1}(b^{a,o})\big] \tag{2.3}$$

$J^{\pi^*}(b) = V_{n=H}(b)$ with H the (possibly infinite) horizon of the problem.

POMDP	$r(s,a)$
MDP	$\rho(b,a) = \sum_s b(s) r(s,a)$

The recursive computation of V_n has the property to generate piecewise-linear and convex value functions for each horizon.

If Γ_n is the set of vectors representing the value function for horizon n, then $V_n(b) = \max_{\alpha \in \Gamma_n} \sum_s b(s)\alpha(s)$ where $\alpha(s)$ is the expected reward for state s.

$$V_n(b) = \max_{a \in A} \sum_o \sum_s b(s) \big[\frac{r(s,a)}{|\Omega|} \sum_{s'} T(s,a,s') O(s',a,o) \chi_{n-1}(b^{a,o}, s')\big] \tag{2.4}$$

$$\overline{\Gamma_n}^{a,o} = \{\frac{r^a}{|\Omega|} + P^{a,o} \cdot \alpha_{n-1} \mid \alpha_{n-1} \in \Gamma_{n-1}\} \tag{2.5}$$

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:

$$\rho(b,a) = \max_{\alpha \in \Gamma_\rho^a} \big[\sum_s b(s)\alpha(s)\big] \tag{2.6}$$

Changes into a new value function

$$V_n(b) = \max_{a \in A} \sum_s b(s) \big[\arg\max_{\alpha \in \Gamma^a} \rho(b \cdot \alpha) + \sum_o \sum_{s'} T(s.a.s') O(s',a,o) \chi_{n-1}(b^{a,o}, s')\big] \tag{2.7}$$

$$\overline{\Gamma_n}^{a,o} = \{P^{a,o} \cdot \alpha_{n-1} \mid \alpha_{n-1} \in \Gamma_{n-1}\} \tag{2.8}$$

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.