A POMDP Extension with Belief-dependent Rewards

Mauricio Araya[1] Olivier Buffet[2] Vincent Thomas[1] François Charpillet[2]

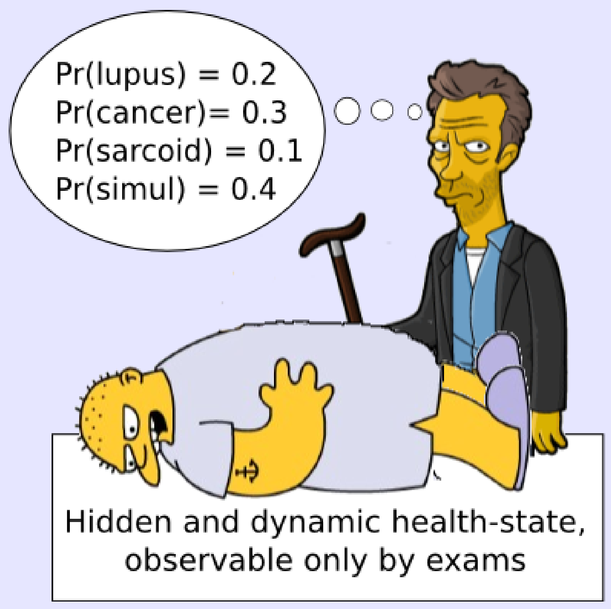
Nancy Université[1]– INRIA[2]

LORIA – Campus Scientifique – BP 239 – 54506 Vandoeuvre-lès-Nancy Cedex – France

Motivation Example

Dr. House must perform actions (exams) to infer the health status of a patient. His job is not to treat the patient, but to perform the correct exams to reduce the uncertainty.

This and other problems such as surveillance [1], are usually modelled as Partially Observable Markov Decision Processes (POMDPs), but this framework do not support **reward depending on the belief** and not on the state.

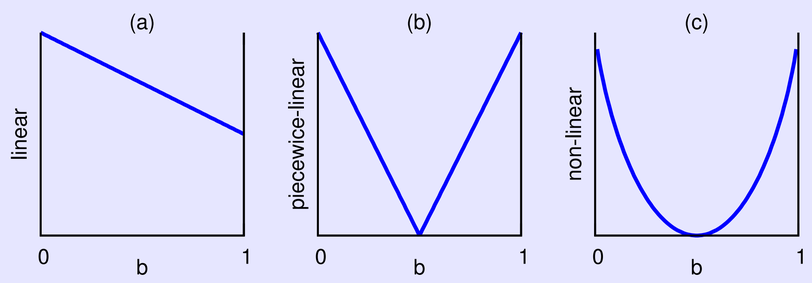


Belief-dependent Rewards

Our proposal is to **extend POMDPs** to a more general framework allowing arbitrary (convex) **belief-dependent rewards** (ρPOMDP).

State-dependent rewards are always linear functions in the belief-state space (a), but more complex functions can be defined directly as belief-dependent rewards, such as piecewise linear function (b) or non-linear functions (c).

Examples of belief-dependent rewards



Introducing ρPOMDPs

ρPOMDPs consist in a relaxation of the POMDPs definition of the reward function. Instead of defining the reward over the state space $\mathcal{S}$, we define the reward as a function on the belief-state space $\mathbf{\Delta}$.

POMDP	ρPOMDP
$\langle \mathcal{S}, \mathcal{A}, \Omega, T, O, r, b_0 \rangle$	$\langle \mathcal{S}, \mathcal{A}, \Omega, T, O, func(b), b_0 \rangle$
$\downarrow$	$\downarrow$
belief MDP	belief MDP
$\langle \Delta, \mathcal{A}, \tau, \rho = \sum_s b(s) r(s,a), b_0 \rangle$	$\langle \Delta, \mathcal{A}, \tau, \rho = func(b), b_0 \rangle$

Solution techniques for POMDPs rely on the convexity of the value function [2], based on the linearity of the state-dependent reward $r(s,a)$. Yet, this property holds for a wider class of ρ-functions.

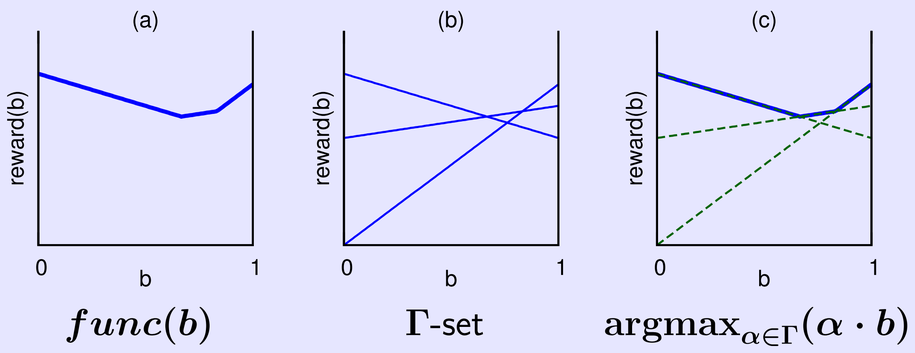
Theorem (Convexity)

If ρ and V_0 are convex functions over Δ, then the value function V_n of the belief MDP is convex over Δ at any time step n.

Proof in NIPS supplementary material [3]

Solving ρPOMDPs

If $\rho = func(b)$ is piecewise-linear and convex (PWLC), it can be represented by a set of hyperplanes.



$func(b)$ Γ-set $\mathbf{argmax}_{\alpha\in\Gamma}(\alpha \cdot b)$

Few modifications to **Value Iteration** are needed to support ρPOMDPs.

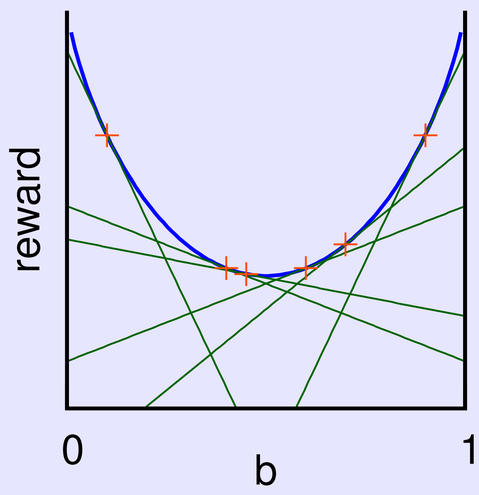
$$V_n(b) = \max_{a\in\mathcal{A}} \left\{ b \cdot \left[\mathbf{argmax}_{\alpha\in\Gamma^a_\rho}(\alpha \cdot b) + \gamma \sum_o \mathbf{argmax}_{\alpha\in\Gamma^{a,o}_n}(\alpha \cdot b) \right] \right\},$$

where Γ^a_ρ represents the reward function for a given action a, and $\Gamma^{a,o}_n$ is the set of projections for a and o of the last value function.

If ρ is not piecewise-linear (but convex), then we can build an approximation using PWLC functions and then solve value iteration.

Approximating non-linear ρ-functions

Idea: Use a lower piecewise-linear approximation of $\rho(b)$ using a set of pivot points $b' \in B \subset \Delta$.



Is this approximation bounded?

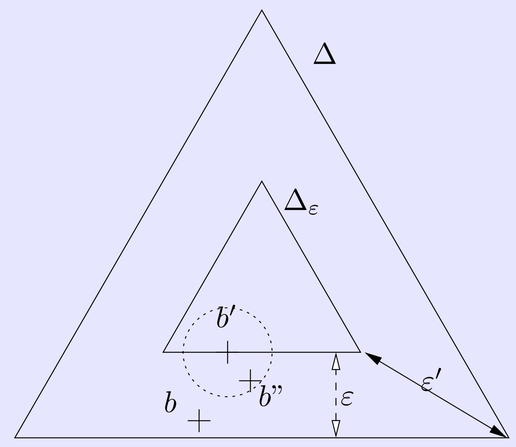
If $\rho(b)$ is Lipschitzian, then the aproximation is bounded. But, some uncertainty measurements (such as entropy) have undefined gradient at the simplex boundary. For these cases, a more generic result can be proven for α-Hölderian functions:

$$\exists \alpha \in (0,1],\ \exists K_\alpha > 0,\ \text{s.t. } |f(x) - f(y)| \le K_\alpha \|x - y\|_1^\alpha.$$

Theorem (ρ-bound)

Let ρ be a continuous and convex function over Δ, differentiable everywhere in Δ^o (the interior of Δ), and satisfying the α-Hölder condition with constant K_α. The error of an approximation ω_B can be bounded by $C\delta_b^\alpha$, where C is a scalar constant that depends on K_α, and δ_b is the density of the set B.

Proof in NIPS supplementary material [3]



This theorem uses the fact that for each b, there is always a $b'' \in B$ far enough from the boundary of the simplex but within a bounded distance to b.

Value Function Bound

If ρ is α-Hölderian, then it can be proven that:

$$\|V_t - V_t^*\|_\infty \le \frac{C\delta_B^\alpha}{1-\gamma}$$

meaning that the error of the value function V_t using the ρ-approximation is bounded for **exact solving algorithms** [4].

For **point-based algorithms** [5] a proper bound can also be found.

$$\|\hat{V}_t - V_t\|_\infty = \frac{(R_{max} - R_{min} + C\delta_B^\alpha)\delta_B}{1-\gamma}$$

[1] M. Spaan, "Cooperative active perception using POMDPs," in *AAAI 2008 Workshop on Advancements in POMDP Solvers*, July 2008.

[2] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov decision processes over a finite horizon," *Operation Research*, vol. 21, pp. 1071–1088, 1973.

[3] M. Araya-López, O. Buffet, V. Thomas, and F. Charpillet, "A POMDP extension with belief-dependent rewards – extended version," Tech. Rep. RR-7433, INRIA, Oct 2010. (See also NIPS supplementary material).

[4] A. Cassandra, *Exact and approximate algorithms for partially observable Markov decision processes*. PhD thesis, Providence, RI, USA, 1998.

[5] J. Pineau, G. Gordon, and S. Thrun, "Anytime point-based approximations for large POMDPs," *Journal of Artificial Intelligence Research (JAIR)*, vol. 27, pp. 335–380, 2006.

http://maia.loria.fr/ {mauricio.araya, olivier.buffet, vincent.thomas, francois.charpillet}@loria.fr