ρ-POMDPs have Lipschitz-Continuous ϵ-Optimal Value Functions

Mathieu Fehr[1], Olivier Buffet[2], Vincent Thomas[2], Jilles Dibangoye[3]

[1] École Normale Supérieure de la rue d'Ulm, Paris, France – firstname.lastname@ens.fr
[2] Université de Lorraine, CNRS, Inria, LORIA, Nancy, France – firstname.lastname@loria.fr
[3] Université de Lyon, INSA Lyon, Inria, CITI, Lyon, France – firstname.lastname@inria.fr

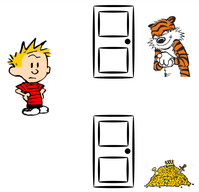
Introduction

Many state-of-the-art algorithms for solving Partially Observable Markov Decision Processes (POMDPs) rely on turning the problem into a "fully observable" problem—a belief MDP—and exploiting the piece-wise linearity and convexity (PWLC) of the optimal value function in this new state space (the belief simplex Δ). This approach has been extended to solving ρ-POMDPs—*i.e.*, for information-oriented criteria—when the reward ρ is **convex** in Δ.

Here: focus on ρ-POMDPs with $\boldsymbol{\lambda}_\rho(b)$**-Lipschitz-Continuous** reward function.

POMDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, P, r, \gamma, b_0 \rangle$

$\mathcal{S}, \mathcal{A}, \mathcal{Z}$ = finite sets of states, actions & observations
$P_{a,z}(s,s') = Pr(s', z | s, a)$
$r(s,a)$ = reward associated to $s \xrightarrow[(z=?)]{a} (s' = ?)$
γ = discount factor ($\in [0;1)$)
b_0 = initial belief state ($\in \Delta = \Pi(\mathcal{S})$)

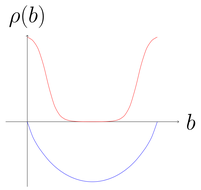


Information-Oriented Control : ρ-POMDPs

Information-oriented performance criterion?
Replace $r(s,a)$ **with** $\rho(b)$ [Araya-López et al., 2010]
Examples:

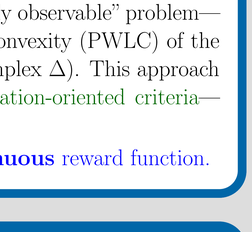
$\rho(b) \doteq -H(b) = \sum_s b(s) \log b(s)$ (neg-Entropy) → convex

$\rho(b) \doteq \sigma(\alpha(\|b_X\|_\infty - \beta))$ (with $\sigma(\cdot)$ the sigmoid function) → LC



Solving (ρ-)POMDPs as belief MDPs

Solving a POMDP = Finding a *policy* π
so as to maximize the expected discounted sum of rewards.
Common approach:
(i) **Turn** POMDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{Z}, P_{z,a}(s'|s), r(s,a), \gamma, b_0 \rangle$
↓
into belief MDP $\langle \Delta, \mathcal{A}, P(b'|b,a), \overbrace{r(b,a)}^{\mathbb{E}_b[r(S,a)]}, \gamma, b_0 \rangle$
(ii) **Compute** V^*, fixed point of Bellman's *optimality* operator ($\mathcal{H}$),
$\mathcal{H}V : b \mapsto \max_a [r(b,a) + \gamma \sum_z \|P_{a,z} b\|_1 V(b^{a,z})]$.
(iii) **Act** greedily with respect to V^*.
For ρ-POMDPs, the reward function ρ is given *a priori*.



⚠ Lipschitz Continuities (LC) ⚠

Here, usual **uniform+scalar LC** $\xrightarrow{\text{becomes}}$ **local+vector LC**:

Definition 1 *Let $f : X \to Y$ be a function, with X and Y two normed spaces. f is* local+vector Lipschitz-continuous *if*
$$\forall \boldsymbol{x}, \exists \boldsymbol{\lambda}_f(\boldsymbol{x}) \in \mathbb{R}^+ \text{ s.t., } \forall \boldsymbol{x}', \|f(\boldsymbol{x}) - f(\boldsymbol{x}')\| \leq \boldsymbol{\lambda}_f(\boldsymbol{x}) |\boldsymbol{x} - \boldsymbol{x}'|.$$

Properties of $\mathcal{H}$ and $V^*_H(b)$ (H finite)

Proposition 1 ($\mathcal{H}$ **preserves Piece-wise Linearity and Convexity** (PWLC))

Proposition 2 ($\mathcal{H}$ **preserves Lipschitz-Continuity** (LC))
If ρ is $\boldsymbol{\lambda}_\rho(\cdot,\cdot)$-LC, and V is $\boldsymbol{\lambda}_V(\cdot)$-LC, then $\mathcal{H}V$ is $\boldsymbol{\lambda}_{\mathcal{H}V}(\cdot)$-LC with, $\forall b$,
$$\boldsymbol{\lambda}_{\mathcal{H}V}(b) = \vec{\max}_a \Big[\boldsymbol{\lambda}_\rho(b,a) + \gamma \sum_z \big[(|V(b^{a,z})| + \boldsymbol{\lambda}_V(b^{a,z}) b^{a,z}) \mathbf{1} + \boldsymbol{\lambda}_V(b^{a,z}) \big] P_{a,z} \Big].$$

Properties of $V^*_H(b)$

$\rho(b)$ is	$V^*_H(b)$ is	(note)
linear	PWLC	(= POMDP) [Smallwood and Sondik, 1973]
PWLC	PWLC	(⇔ IR-POMDP [Satsangi et al., 2015])
$\boldsymbol{\lambda}_\rho(\cdot,\cdot)$-LC	$\boldsymbol{\lambda}'_H(\cdot)$-LC	(new)

Bounding $V^*(b)$ ($H = \infty$)

	PWLC	LC	(pointwise - PW)
$U(b)$	sawtooth approx.	downward cones	(points)
$L(b)$	upper envelope of hyperplanes	upward cones	(points)

Here: U and L computed using **HSVI** [Smith, 2007].

Problem: Theoretical $\boldsymbol{\lambda}(\cdot)$ values overly pessimistic → poor approximations
Solution: Incremental search for a good (uniform+scalar) constant λ
(despite loss of theoretical guarantees) → **inc-lc-HSVI**
POMDP example: $U(b)$ and $L(b)$ for the Tiger problem:

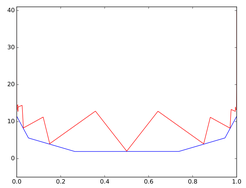


Fig. 1: pwlc-HSVI

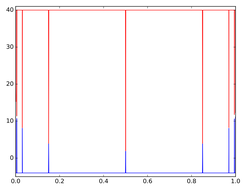


Fig. 2: lc-HSVI (theoretical $\boldsymbol{\lambda}$)

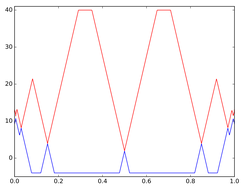


Fig. 3: inc-lc-HSVI (best guess)

Other topics: criteria ($-H(b)$, $var(b)$, ...), initialization, MCTS/POMCP, POSGs, LC MDPs, ...

Comparing variants of HSVI

Setting: Experiments conducted on (i) several standard POMDPs + (ii) a ρ-POMDP on a grid (Fig. 4) with the goal of (not) knowing the x or y position.

Not shown: results evaluating the two (complementary) criteria used to detect invalid λ values in inc-lc-HSVI.

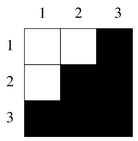


Fig. 4: `grid-info`

x-HSVI	pwlc				pw				lc				inc-lc(noNUI)			
	t (s)	(#it)	$gap(b_0)$	λ	t (s)	(#it)	$gap(b_0)$		t (s)	(#it)	$gap(b_0)$	λ	t (s)	(#it)	$gap(b_0)$	λ
4x3.95	**1**	(134)	0.10	1.19	600	(447)	0.94		600	(214)	3.27	1.9e+05	**1**	(254)	0.10	1
4x4.95	**1**	(120)	0.10	0.66	**1**	(134)	0.10		**6**	(134)	0.10	2.4e+05	**0**	(125)	0.10	1
hallway	600	(414)	0.35	0.70	600	(683)	1.30		600	(203)	1.33	1.8e+00	600	(611)	1.12	1
hallway2	600	(385)	0.67	0.63	600	(690)	1.04		600	(208)	1.06	4.0e+00	600	(668)	1.00	1
milos-aaai97	600	(1152)	29.55	89.49	600	(1725)	49.05		600	(595)	52.87	2.3e+03	600	(1797)	43.78	64
network	**498**	(7703)	0.10	168.22	600	(3021)	453.62		600	(941)	510.76	5.0e+16	**34**	(2819)	0.10	128
paint.95	**2**	(143)	0.10	1.00	600	(3695)	3.55		600	(1008)	4.37	1.7e+281	**0**	(84)	0.10	1
pentagon	601	(27)	0.31	1.00	600	(89)	0.83		600	(32)	0.83	2.8e+01	600	(60)	0.73	1
shuttle.95	**0**	(23)	0.10	22.77	**0**	(42)	0.09		**0**	(42)	0.09	7.5e+00	**0**	(47)	0.08	4
tiger85	**0**	(15)	0.09	55.00	**0**	(15)	0.08		**0**	(15)	0.08	2.2e+02	**0**	(15)	0.07	64
grid-info kx	–	(–)	–	–	600	(709)	0.24		600	(358)	1.82	3.6e+01	**1**	(279)	0.10	4
grid-info ky	–	(–)	–	–	**27**	(432)	0.10		**344**	(426)	0.10	4.0e+01	**1**	(279)	0.10	4
grid-info ¬kx	–	(–)	–	–	600	(2319)	9.15		600	(889)	13.74	3.3e+01	**4**	(695)	0.10	4
grid-info ¬ky	–	(–)	–	–	600	(1393)	6.66		600	(604)	8.52	3.1e+01	**4**	(707)	0.10	4

- inc-lc-HSVI interrupted on LXU (L crosses U), ~~NUI (non-unif. improvement)~~, UR (unstable results)
- lc-HSVI scales poorly due to (i) large $\lambda_{\text{lc-HSVI}}$ + (ii) high CPU cost
- inc-lc-HSVI much better, sometimes competes with pwlc-HSVI
- $\lambda_{\text{inc-lc-HSVI}}$ same order of magnitude as $\lambda_{\text{pwlc-HSVI}}$

$L(b_0)$ and $U(b_0)$ as a function of *time*

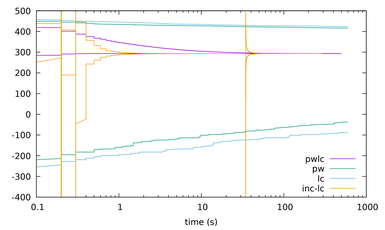


Fig. 5: network

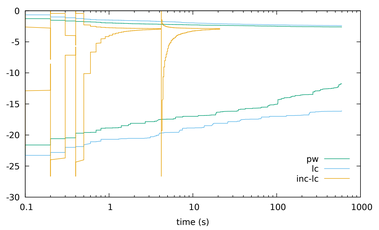


Fig. 6: grid-info ¬kx

M. Araya-López, O. Buffet, V. Thomas, and F. Charpillet. A POMDP extension with belief-dependent rewards. In *Advances in Neural Information Processing Systems 23 (NIPS-10)*, 2010.

Y. Satsangi, S. Whiteson, and M. T. J. Spaan. An analysis of piecewise-linear and convex value functions for active perception POMDPs. Technical Report IAS-UVA-15-01, IAS, Universiteit van Amsterdam, 2015.

R. Smallwood and E. Sondik. The optimal control of partially observable Markov decision processes over a finite horizon. *Operation Research*, 21, 1973.

T. Smith. *Probabilistic Planning for Robotic Exploration*. PhD thesis, Carnegie Mellon University, 2007.