Max-Planck-Institut
für Mathematik
in den Naturwissenschaften
Leipzig

A continuity result for optimal memoryless planning in POMDPs

by

Johannes Rauh, Nihat Ay, and Guido Montúfar

Preprint no.: 5 2021

A continuity result for optimal memoryless planning in POMDPs

Johannes Rauh
MPI MIS
jarauh@mis.mpg.de

Nihat Ay
MPI MIS / Uni Leipzig / Santa Fe Institute
nay@mis.mpg.de

Guido Montúfar
UCLA Math and Stat / MPI MIS
montufar@math.ucla.edu

Abstract

Consider an infinite horizon partially observable Markov decision process. We show that the optimal discounted reward under memoryless stochastic policies is continuous under perturbations of the observation channel. This implies that we can find approximately optimal memoryless policies by solving an approximate problem with a simpler observation channel.

Keywords: POMDPs, memoryless stochastic policy, optimal policy

1 Introduction

Policy optimization in partially observable Markov decision processes (POMDPs) is known to be a difficult problem. In order to better understand this problem, we can study special cases where the system has some additional structure (e.g., the observations identify the world state to within a few possibilities), or we can also restrict the optimization problem to policies with some additional structure (e.g., memoryless policies). The optimization problem over memoryless policies has been discussed in various works (see, e.g., Ross, 1983; Vlassis et al., 2012; Azizzadenesheli et al., 2016).

In this context, the connections between information, memory and value are of particular interest (see Kaelbling et al., 1998). We are interested in the relations that exist between the observation channel, on the one hand, and the structure of the optimal memoryless policies, on the other hand. In particular, we are interested in whether certain types of POMDP optimization problems allow for optimal or nearly optimal policies that have a particularly simple structure.

Previous work in this direction has characterized families of policies that contain optimal memoryless policies for any POMDP of a particular type (see Montúfar and Rauh, 2017; Montúfar et al., 2015; Montúfar et al., 2015). In particular, these works consider the number of actions that a memoryless policy needs to randomize at a given observation, depending on the number of world states that are compatible with that observation. In other words, depending on the properties of the observation channel, they conclude that there exists a simple optimal memoryless policy. A natural question is: If the observation channel nearly satisfies the conditions under which it is known that a simple optimal policy exists, can we conclude that there exists a simple policy that is nearly optimal? In this short article, we show that this is indeed the case. Thereby, we contribute to the understanding of approximate memoryless stochastic planning in POMDPs.

2 POMDPs and localization of optimal policies

We briefly introduce the definitions and settings.

Definition 1. A *Partially Observed Markov Decision Process* (POMDP) is a tuple $(\mathcal{W}, \mathcal{S}, \mathcal{A}, \alpha, \beta, R)$ consisting of

1. finite sets $\mathcal{W}$ (world states), $\mathcal{S}$ (sensor states/observations) and $\mathcal{A}$ (actions),
2. Markov kernels/channels $\alpha : \mathcal{W} \times \mathcal{A} \to \mathcal{W}$ (world state transition) and $\beta : \mathcal{W} \to \mathcal{S}$ (observation channel),
3. and a reward function $R : \mathcal{W} \times \mathcal{A} \to \mathbb{R}$.

A Markov decision process (MDP) is the special case where β conveys full information about the world state.

We consider time independent memoryless stochastic policies.

Definition 2. A *policy* is a Markov kernel $\pi : \mathcal{S} \to \mathcal{A}$. Denote $\Delta_{S,A}$ the set of all such policies.

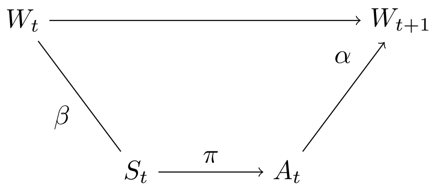

Figure 1: The graphical structure of a POMDP.

A POMDP, a policy π, and a starting distribution μ on $\mathcal{W}$, together, define a stochastic process $(W^t, S^t, A^t)_t$, which is a sequence of random variables $(W^t)_t$ (world states), $(S^t)_t$ (sensor states), and $(A^t)_t$ (actions). The graphical structure is shown in Figure 1.

We consider infinite horizon discounted rewards.

Definition 3. Given a POMDP and a *policy* π, the *discounted reward* with discount factor γ is

$$\mathcal{R}_\gamma(\pi) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \sum_{w,a} R(w,a)P(W^t = w, A^t = a).$$

We denote $\mathcal{R}^*_\gamma = \sup_{\pi\in\Delta_{S,A}} \mathcal{R}_\gamma(\pi)$ the optimal value over the set of memoryless stochastic policies.

We are interested in the following theorem by Montúfar and Rauh (2017), which provides a type of extension of the well known fact that any MDP has an optimal policy over the set of memoryless stochastic policies which is deterministic. Given $\pi\colon \mathcal{S}\to\mathcal{A}$, let $\operatorname{supp}(\pi(\cdot|s)) = \{a\in\mathcal{A}\colon \pi(a|s) > 0\}$.

Theorem 4. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$. Then there is a policy $\pi^*\in\Delta_{S,A}$ with $|\operatorname{supp}(\pi^*(\cdot|s))| \leq |\operatorname{supp}(\beta(s|\cdot))|$ for all $s\in S$, and $\mathcal{R}_\gamma(\pi^*) \geq \mathcal{R}_\gamma(\pi)$ for all $\pi\in\Delta_{S,A}$.*

This result implies that, in order to optimize a POMDP over the set of memoryless stochastic policies, it suffices to consider a subset of policies. This can be translated into a choice of a policy model with few parameters or also into heuristics for the policy optimization. The result is optimal in the sense that there are POMDPs $(W, S, A, \alpha, \beta, R)$ where each policy $\pi^*\in\Delta_{S,A}$ with $\mathcal{R}_\gamma(\pi^*) \geq \mathcal{R}_\gamma(\pi)$ for all $\pi\in\Delta_{S,A}$ satisfies $|\operatorname{supp}(\pi^*(\cdot|s))| \geq |\operatorname{supp}(\beta(s|\cdot))|$.

A problem with Theorem 4 is that, in concrete situations, we might not be able to exclude world states with absolute certainty, meaning that $|\operatorname{supp}(\beta(s|\cdot))| = |\mathcal{W}|$. Moreover, the number of world states might be as large or larger than the number of possible actions, $|\mathcal{W}| \geq |\mathcal{A}|$, in which case the statement of the theorem is vacuous.

3 Continuity of the reward and localization of nearly optimal policies

In order to remedy the shortcomings of Theorem 4, we need a continuous version of the characterization. Our continuous extension is as follows. We show that if the observation channel β is close to some other channel β', in an appropriate sense, then we can find a near to optimal policy π with $|\operatorname{supp}(\pi(\cdot|s))| \leq |\operatorname{supp}(\beta'(s|\cdot))|$ for all $s\in S$.

Theorem 5. *Consider a POMDP with $\gamma < 1$ and $\|R\|_\infty := \max_{w,a}|R(w,a)|$. Let β' be a Markov kernel $\mathcal{W}\to\mathcal{S}$ that satisfies*

$$\|\beta(\cdot|w) - \beta'(\cdot|w)\|_{TV} = \frac{1}{2}\sum_s \big|\beta(s|w) - \beta'(s|w)\big| \leq \epsilon \quad \textit{for all } w\in\mathcal{W}.$$

Then there is a policy π that satisfies $|\operatorname{supp}(\pi(\cdot|s))| \leq |\operatorname{supp}(\beta'(s|\cdot))|$ for all s and $\mathcal{R}_\gamma(\pi) \geq \mathcal{R}^*_\gamma - 2\frac{\epsilon}{1-\gamma}\|R\|_\infty$.

We prove this theorem based on a continuity result for the discounted reward function with respect to the observation channel.

Theorem 6. *Consider two POMDPs $(\mathcal{W}, \mathcal{S}, \mathcal{A}, \alpha, \beta, R)$ and $(\mathcal{W}, \mathcal{S}, \mathcal{A}, \alpha, \beta', R)$ that satisfy*

$$\|\beta(\cdot|w) - \beta'(\cdot|w)\|_{TV} = \frac{1}{2}\sum_s \big|\beta(s|w) - \beta'(s|w)\big| \leq \epsilon \quad \textit{for all } w\in\mathcal{W}.$$

Then the discounted reward functions $\mathcal{R}_\gamma$, $\mathcal{R}'_\gamma$ of the two POMDPs satisfy

$$|\mathcal{R}_\gamma(\pi) - \mathcal{R}'_\gamma(\pi)| \leq \frac{\epsilon}{1-\gamma}\|R\|_\infty \quad \textit{for any policy } \pi\in\Delta_{S,A} \textit{ and any } \gamma < 1,$$

where $\|R\|_\infty := \max_{w,a}|R(w,a)|$. This implies, in particular, $|\mathcal{R}^*_\gamma - \mathcal{R}'^*_\gamma| \leq \frac{\epsilon}{1-\gamma}\|R\|_\infty$.

We present the proofs in the following section.

4 Proofs of the continuity result

Lemma 7. *Under the assumptions of Theorem 6, denote by A^t, W^t the action and world process of the POMDP with β, and denote by A'^t, W'^t the action and world process of the POMDP with β', where both POMDPs are controlled by the same policy π. Then,*

$$|\Pr(A^t = a|W^t = w) - \Pr(A'^t = a|W'^t = w)| \leq \epsilon \qquad \textit{for all } a, w.$$

Proof. (Proof of Lemma 7) The inequality follows from

$$\begin{aligned}
\Big|\Pr(A^t = a|W^t = w) &- \Pr(A'^t = a|W'^t = w)\Big| \\
&= \Big|\sum_s \pi(a|s)(\beta(s|w) - \beta'(s|w))\Big| \\
&= \Big|\sum_{s:\beta(s|w)\geq\beta'(s|w)} \pi(a|s)(\beta(s|w) - \beta'(s|w)) - \sum_{s:\beta'(s|w)>\beta(s|w)} \pi(a|s)(\beta'(s|w) - \beta(s|w))\Big| \\
&\leq \max\Big\{\sum_{s:\beta(s|w)\geq\beta'(s|w)} \pi(a|s)(\beta(s|w) - \beta'(s|w)), \sum_{s:\beta'(s|w)>\beta(s|w)} \pi(a|s)(\beta'(s|w) - \beta(s|w))\Big\} \\
&\leq \max\Big\{\sum_{s:\beta(s|w)\geq\beta'(s|w)} (\beta(s|w) - \beta'(s|w)), \sum_{s:\beta'(s|w)>\beta(s|w)} (\beta'(s|w) - \beta(s|w))\Big\} \\
&= \|\beta'(\cdot|w) - \beta(\cdot|w)\|_{\mathrm{TV}}.
\end{aligned}$$

□

Lemma 8. *Under the assumptions of Lemma 7, for all $t \geq 0$,*

$$\sum_{aw} |\Pr(A^t W^t = aw) - \Pr(A'^t W'^t = aw)| \leq (t+1)\epsilon, \tag{1a}$$

$$\sum_{w} |\Pr(W^t = w) - \Pr(W'^t = w)| \leq t\epsilon. \tag{1b}$$

Proof. (Proof of Lemma 8) The proof is by induction. For $t = 0$, $\Pr(W^0 = w) = \Pr(W'^0 = w)$, so (1b) holds for $t = 0$. Assuming that (1b) holds for some t,

$$\begin{aligned}
\sum_{aw} &|\Pr(A^t W^t = aw) - \Pr(A'^t W'^t = aw)| \\
&\leq \sum_w |\Pr(W^t = w) - \Pr(W'^t = w)| \sum_a |\Pr(A^t = a|W^t = w)| \\
&\quad + \sum_{aw} |\Pr(A^t = a|W^t = w) - \Pr(A'^t = a|W'^t = w)||\Pr(W'^t = w)| \\
&\leq \sum_w |\Pr(W^t = w) - \Pr(W'^t = w)| \\
&\quad + \sup_w \sum_a |\Pr(A^t = a|W^t = w) - \Pr(A'^t = a|W'^t = w)| \\
&\leq t\epsilon + \epsilon = (t+1)\epsilon.
\end{aligned}$$

Assuming that (1a) holds for $t-1$,

$$\begin{aligned}
\sum_w &|\Pr(W^t = w) - \Pr(W'^t = w)| \\
&= \sum_w \Big|\sum_{a,w'} \alpha(w|a,w')\big(\Pr(A^{t-1}W^{t-1} = aw') - \Pr(A'^{t-1}W'^{t-1} = aw')\big)\Big| \\
&\leq \sum_{a,w'} \sum_w \alpha(w|a,w')\Big|\Pr(A^{t-1}W^{t-1} = aw') - \Pr(A'^{t-1}W'^{t-1} = aw')\Big| \\
&\leq \sum_{a,w'} \Big|\Pr(A^{t-1}W^{t-1} = aw') - \Pr(A'^{t-1}W'^{t-1} = aw')\Big| \leq t\epsilon.
\end{aligned}$$

□

Proof. (Proof of Theorem 6) Using

$$\sum_{t=0}^{\infty}(t+1)\gamma^t = \frac{1}{\gamma}\sum_{t=1}^{\infty}t\gamma^t = \frac{\partial}{\partial\gamma}\sum_{t=0}^{\infty}\gamma^t = \frac{\partial}{\partial\gamma}\frac{1}{1-\gamma} = \frac{1}{(1-\gamma)^2}$$

and Lemma 8,

$$\begin{aligned}
&|\mathcal{R}(\pi) - \mathcal{R}(\pi')| \\
&\leq (1-\gamma)\sum_{t=0}^{\infty}\sum_{w,a}\gamma^t R(w,a)|P(W^t, A^t = w,a) - P(W'^t, A'^t = w,a)| \\
&\leq (1-\gamma)\frac{1}{\gamma}\sum_{t=1}^{\infty}\gamma^t\|R\|_\infty t\epsilon = \frac{\epsilon}{1-\gamma}\|R\|_\infty.
\end{aligned}$$

□

Let $\mathcal{R}(\beta,\pi)$ be the expected reward for observation kernel β and policy π.

Lemma 9. *Let $\mathcal{R}_\gamma$, $\mathcal{R}'_\gamma$ be the discounted reward functions of two POMDPs $(\mathcal{W},\mathcal{S},\mathcal{A},\alpha,\beta,R)$, $(\mathcal{W},\mathcal{S},\mathcal{A},\alpha,\beta',R)$, and suppose that there exists $c > 0$ with $|\mathcal{R}_\gamma(\pi) - \mathcal{R}'_\gamma(\pi)| \leq c$ for all policies π. If π'^* is the optimal policy for $\mathcal{R}'_\gamma$, then $\mathcal{R}^*_\gamma \geq \mathcal{R}_\gamma(\pi'^*) \geq \mathcal{R}^*_\gamma - 2c$.*

Proof. (Proof of Lemma 9) The first inequality is by definition of $\mathcal{R}^*_\gamma$. For any policy π for $\mathcal{R}_\gamma$,

$$\mathcal{R}_\gamma(\pi'^*) \geq \mathcal{R}'(\pi'^*) - c \geq \mathcal{R}'(\pi) - c \geq \mathcal{R}_\gamma(\pi) - 2c.$$

The second inequality follows when π is an optimal policy for $\mathcal{R}_\gamma$. □

Proof. (Proof of Theorem 5) This follows from Theorem 6 and Lemma 9 and Theorem 4. □

5 Discussion

We presented a continuity result that extends the applicability of previous theoretical results on the structure of optimal policies of POMDPs, and allows us to discuss approximately optimal policies. Continuity, in the way that we studied here, could be investigated not only in terms of the observation channel, but also in terms the state transition kernel. These continuity results might also serve to make statements about consistency in policy optimization in reinforcement learning, when the agent needs to estimate the world model (i.e., the kernels β and α).

Acknowledgment This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement no 757983).

References

K. Azizzadenesheli, A. Lazaric, and A. Anandkumar. Open problem: Approximate planning of pomdps in the class of memoryless policies. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1639–1642. PMLR, 2016.

L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1):99–134, 1998.

G. Montúfar and J. Rauh. Geometry of policy improvement. In *Geometric Science of Information, LNCS 10589*, pages 282–290. Springer, 2017.

G. Montúfar, K. Ghazi-Zahedi, and N. Ay. A theory of cheap control in embodied systems. *PLoS Computational Biology*, 11(9):1–22, 2015.

G. Montúfar, K. Ghazi-Zahedi, and N. Ay. Geometry and determinism of optimal stationary control in POMDPs. *arXiv:1503.07206*, 2015.

S. M. Ross. *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, Inc., 1983.

N. Vlassis, M. L. Littman, and D. Barber. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory*, 4(4):12:1–12:8, 2012.