TASK-AGNOSTIC CONSTRAINING IN AVERAGE REWARD POMDPS

Guido Montúfar[1,2], **Johannes Rauh**[2], **Nihat Ay**[2,3,4]
[1]Department of Mathematics and Department of Statistics, UCLA, CA 90095;
[2]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany;
[3]Santa Fe Institute, NM 87501; [4]University of Leipzig, 04009 Leipzig, Germany

ABSTRACT

We study the shape of the average reward as a function over the memoryless stochastic policies in infinite-horizon partially observed Markov decision processes. We show that for any given instantaneous reward function on state-action pairs, there is an optimal policy that satisfies a series of constraints expressed solely in terms of the observation model. Our analysis extends and improves previous descriptions for discounted rewards or which covered only special cases.

Keywords: Partial observability, Markov decision process, stochastic policy, memoryless policy, optimal planning

1 INTRODUCTION

The problem of maximizing the expected long term reward in partially observable Markov decision processes (POMDPs) over the set of memoryless stochastic policies has been studied in a number of recent papers (see, e.g., Ross, 1983; Vlassis et al., 2012; Ay et al., 2013; Montúfar et al., 2015; Montúfar et al., 2015; Azizzadenesheli et al., 2016; Montúfar & Rauh, 2017; Rauh et al., 2019). An important question is how to characterize good priors, i.e., constraints, that could be imposed on a policy model, so as to reduce the complexity of learning, without incurring losses in terms of the achievable values of the objective. At one extreme, when there is a single problem (POMDP) under consideration, the full answer to this question corresponds to characterizing an optimal policy for this particular problem. At the other extreme, when all possible problems are under consideration, one can impose no constraints at all, since each possible policy will be the unique optimizer of one particular problem. We are interested in the situation where certain general properties of the POMDP are given, and how we might translate these properties into policy constraints as described above, independently of the specific instantaneous reward function.

2 DEFINITIONS

Formally a POMDP is a tuple $(W, S, A, \alpha, \beta, R)$, where W, S, A are finite sets of world states, sensor states, and actions, $\beta\colon W \to \Delta_S$ is a Markov kernel describing sensor measurements (observation model), $\alpha\colon W \times A \to \Delta_W$ is a Markov kernel describing world state transitions, and $R\colon W \times A \to \mathbb{R}$ is an instantaneous reward function. A policy is a mechanism for selecting actions. We consider memoryless (and thus time independent) stochastic policies, which are described by Markov kernels of the form $\pi\colon S \to \Delta_A$, and which we call simply *policies*. We denote the set of policies by $\Delta_{S,A}$. The world state and the instantaneous reward are updated at discrete time steps by iterating β, π, α, R as illustrated in Figure 1. When β is a deterministic injective map, the observations fully identify the world state, and the POMDP reduces to a Markov decision process (MDP).

The objective of learning is to find a policy that maximizes some form of expected reward. We will focus on the average reward over an infinite horizon. We make the standard assumption that for each fixed policy, the Markov chain of world states is irreducible and aperiodic (e.g., α is strictly positive). This implies that there is a unique stationary limit distribution $p_W^\pi \in \Delta_W$, which is independent of

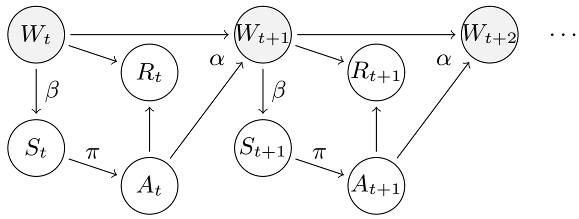

Figure 1: The graphical structure of a POMDP with a memoryless policy.

the initial state distribution $\mu \in \Delta_W$. In this case, the average reward can be written as

$$\mathcal{R}(\pi) = \lim_{T\to\infty} \mathbb{E}\Big[\sum_{t=0}^{T-1} \frac{1}{T} R(W_t, A_t)\Big|\pi, W_0 \sim \mu\Big] = \sum_w p_W^\pi(w) \sum_a \sum_s \pi(a|s)\beta(s|w)R(w,a). \tag{1}$$

We will also write $\xi(a|w) = \sum_s \beta(s|w)\pi(a|s)$ for the world state policy. The setting of discounted rewards replaces $\frac{1}{T}$ by a discount factor γ^t with $\gamma \in (0,1)$, which leads to a function that depends on the start state distribution.

We denote as a *task* any particular choice of R in a POMDP. Of course other definitions might be meaningful too, depending on the context.

3 CONSTRAINTS SATISFIABLE BY OPTIMAL POLICIES

We are interested in the following theorem, which provides a certain type of extension of the well known fact that any MDP has an optimal policy which is memoryless and deterministic. The theorem can be interpreted as saying that *determinism* can be used as a task-agnostic prior for solving POMDPs.

Theorem 1 (Montúfar & Rauh 2017, Theorem 1)**.** *Consider a POMDP* $(W, S, A, \alpha, \beta, R)$*. Then there is a policy* $\pi^* \in \Delta_{S,A}$ *which satisfies*

$$|\operatorname{supp}(\pi^*(\cdot|s))| \leq |\operatorname{supp}(\beta(s|\cdot))|, \quad \textit{for all } s \in S,$$

and $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ *for all* $\pi \in \Delta_{S,A}$*.*

There exists an optimal policy that randomizes only as many actions as there are states compatible with the current observation. The existence of fully deterministic optimal policies for MDPs, assigning positive probability to only one action at each observation, follows immediately, since $|\operatorname{supp}(\beta(s|\cdot))| = 1$ for MDPs. We note that it is possible to construct examples of POMDPs for which each optimal memoryless policy attains the bounds specified in the theorem with equality. One might consider the requirements on the observation model to be somewhat restrictive. However, recent work (Rauh et al., 2019) shows (for discounted rewards), that if β nearly satisfies these requirements, then there is a nearly optimal policy that satisfies the specified support constraints. This means that we can obtain generally applicable task-agnostic priors for approximately optimal policies.

Theorem 1 was shown using a notion of policy improvement cones for expected discounted rewards. The average reward case was then obtained by means of limit arguments over the discount factor. We would like to deduce it based solely on the geometry of the optimization problem. A geometric approach in this spirit was pursued by Montúfar et al. (2015) based on a decomposition of the average reward function into a continuum of linear pieces, but obtaining only the constraints for $s \in S$ with $|\operatorname{supp}(\beta(s|\cdot))| \leq 1$.

We present a geometric analysis based on a notion of policy improvement cones for average rewards. We obtain the following refinement of Theorem 1:

Theorem 2. *Consider a POMDP* $(W, S, A, \alpha, \beta, R)$*. Then there is a policy* $\pi^* \in \Delta_{S,A}$ *which satisfies*

$$\sum_{s' \in S'} |\operatorname{supp}(\pi^*(\cdot|s'))| \leq |\bigcup_{s' \in S'} \operatorname{supp}(\beta(s'|\cdot))| + |S'| - 1, \quad \textit{for all } S' \subseteq S,$$

and $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ *for all* $\pi \in \Delta_{S,A}$.

The statement of Theorem 1 corresponds to the inequalities for $S' \subseteq S$, $|S'| = 1$. In the case of a deterministic β, the sets $\operatorname{supp}(\beta(s'|\cdot))$ are disjoint for all $s' \in S$. In this special case, the constraints for $S' \subseteq S$, $|S'| \geq 2$ in Theorem 2 are already implied by the constraints for $S' \subseteq S$, $|S'| = 1$ in Theorem 1.

4 Policy improvement cones for average rewards

Definition 3 (World policy improvement cones)**.** Fix a world policy $\xi \in \Delta_{W,A}$. We write $\xi_w = (\xi(a|w))_{a\in A} \in \Delta_{\{w\},A}$ and $\nabla_{\xi_w} = (\partial_{\xi(a|w)})_{a\in A}$. For any $w \in W$ define

$$l^\xi_w = \nabla_{\xi_w}\mathcal{R}(\xi) \in T_\xi\Delta_{\{w\},A}. \tag{2}$$

The world policy improvement cone at ξ for a given set $W' = \{w_1, \dots, w_k\} \subseteq W$ is

$$L^{\xi,W'} = \{\xi' \in \Delta_{W,A} \colon \langle(\xi'_w - \xi_w), l^\xi_w\rangle \geq 0,\ \text{for } w \in W',\ \text{and } \xi'_w = \xi_w,\ \text{for } w \in W \setminus W'\}. \tag{3}$$

This is an intersection of $|W'|$ half-spaces in $\Delta_{W',A}$, with fixed values in $\Delta_{(W\setminus W'),A}$.

Lemma 4 (World policy improvement cones)**.** *For any* $\xi \in \Delta_{W,A}$*,* $W' \subseteq W$*, and* $\xi' \in L^{\xi,W'}$*, we have* $\mathcal{R}(\xi') \geq \mathcal{R}(\xi)$*.*

Proof. Given any world policy $\xi \in \Delta_{W,A}$, we write $p^\xi_W \in \Delta_W$ for the corresponding stationary world state distribution, and $p^\xi \in \Delta_{W\times A}$ for the corresponding joint distribution with $p^\xi(w,a) = p^\xi_W(w)\xi(a|w)$. The average reward of ξ is $\mathcal{R}(\xi) = \langle p^\xi, R\rangle = \sum_{w,a} p^\xi(w,a)R(w,a)$. Let $\xi' = \xi + \sum_i \lambda_i l^\xi_i \in L^{\xi,W'}$, where i indexes the elements of W'. By Proposition 5 below, $p^{\xi'} = p^\xi + \sum_i \mu_i r^\xi_i$, where $r^\xi_i = \frac{d}{d\epsilon}\big|_{\epsilon=0} p^{\xi+\epsilon l^\xi_i}$ and $\mu_i \geq 0$. Since $\xi + \epsilon l^\xi_i \in L^{\xi,W'}$ for any i and $\epsilon > 0$ small enough,

$$0 \leq \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \mathcal{R}(\xi + \epsilon l^\xi_i) = \frac{d}{d\epsilon}\bigg|_{\epsilon=0} \langle p^{\xi+\epsilon l^\xi_i}, R\rangle = \langle r^\xi_i, R\rangle,$$

and hence $\mathcal{R}(\xi') = \langle p^\xi, R\rangle + \sum_i \mu_i\langle r^\xi_i, R\rangle \geq \mathcal{R}(\xi)$. ☐

Proposition 5 (Cones of world policies and stationary joint distributions)**.** *Consider the map* $f\colon \Delta_{W,A} \to \Delta_{W\times A};\ \xi \mapsto p^\xi$ *that maps a world policy* ξ *to the corresponding stationary joint distribution* $p^\xi(w,a) = p^\xi_W(w)\xi(a|w)$*. Let* $W' = \{w_1, \dots, w_k\}$*. A cone of the form* $L^{W'} = \{\xi + \sum_i \lambda_i l_i \colon \lambda_i \geq 0, i = 1, \dots, k\} \subseteq \Delta_{W,A}$*, where* $l_i \in T_\xi\Delta_{\{w_i\},A}$*,* $i = 1, \dots, k$*, maps to a cone of the form* $f(L^{W'}) = \{p^\xi + \sum_i \mu_i r_i \colon \mu_i \geq 0, i = 1, \dots, k\} \subseteq \Delta_{W\times A}$*, where one may choose* r_i *as* $r_i = p^{\xi+l_i} - p^\xi$ *or as* $r_i = \frac{d}{d\epsilon}\big|_{\epsilon=0} p^{\xi+\epsilon l_i}$*,* $i = 1, \dots, k$*.*

Proof. We first show that $f(L^{W'})$ is a convex cone with extreme rays $M^i = \{p^\xi + \mu_i(p^{\xi+l_i} - p^\xi) \colon \mu_i \geq 0\}$, $i = 1, \dots, k$. For each i, the ray $L^i = \{\xi + \lambda_i l_i \colon \lambda_i \geq 0\}$ is a product of convex sets (it consists of vectors with all coordinates fixed except one coordinate which is a ray), and hence its image $f(L^i)$ is convex (see Proposition 6 below), which implies $f(L^i) = \{\xi + \mu_i r_i \colon \mu_i \geq 0\} = M^i$, $r_i = (p^{\xi+l_i} - p^\xi)$. The cone $L^{W'}$ is also a product of convex sets, and hence $f(L^{W'})$ is also a convex set. In fact, any subset $W'' \subseteq W'$ will produce a convex set. Assuming that f is injective on $L^{W'}$, this implies that M^i are the extreme rays. But the only way f can be non-injective is if $p^\xi_W(w) = 0$ for some w, in which case $p^\xi(w,a) = 0$ for all $a \in A$, and w can be excluded.

It remains to show that it is possible to replace the cone generator $r_i = p^{\xi+l_i} - p^\xi$ by $r_i = \frac{d}{d\epsilon}\big|_{\epsilon=0} p^{\xi+\epsilon l_i}$. This follows since the statement of the proposition holds true for any rescaling of the l_i. ☐

Proposition 6 (Convex sets of world policies and stationary joint distributions)**.** *Consider a set* $G \subseteq \Delta_{W,A}$ *of world policies. Assume that* G *is a Cartesian product of convex sets,* $G = \times_w G_w$*, where each* $G_w \in \Delta_{\{w\},A}$ *is convex. Then the set* $K \subseteq \Delta_{W\times A}$ *of stationary joint distributions corresponding to* G *is convex.*

Proof. The set $K \subseteq \Delta_{W\times A}$ consists of all joint distributions of the form $p^\xi(w)\xi(a|w)$, where $p^\xi(w)$ is the stationary distribution generated by a world policy ξ, and $\xi(a|w) \in G$. This set can be described as the intersection of two sets, $K = F \cap J$. If both J and F are convex, then clearly $K = J \cap F$ is convex. The first set, J, consists of all joint distributions $p(w,a)$ for which the image $p(w,w') = \sum_a p(w,a)\alpha(w'|w,a)$ is an element of the Kirchhoff polytope in $\Delta_{W\times W}$. This corresponds to requiring that the marginal is a stationary distribution of the conditional. Since J is the preimage of a polytope by a linear map, it is an affine set. The second set, F, consists of all joint distributions whose conditionals have the form $p(a|w) = \xi(a|w)$ for some $\xi \in G$. Now Proposition 7 below shows that F is convex whenever G is a Cartesian product of convex sets. $\square$

Proposition 7 (Montúfar et al. 2015, Proposition 8)**.** *Consider a set $G \subseteq \Delta_{W,A}$ of world policies. Assume that G is a Cartesian product of convex sets, $G = \times_w G_w$, where each $G_w \subseteq \Delta_A$ is a convex set. Then the set $F = \{p(w)p(a|w) \colon p(w) \in \Delta_W, p(a|w) \in G\} \subseteq \Delta_{W\times A}$ of all joint distributions with conditional distributions from G is convex.*

Now we translate the improvement cones for world policies to sensor policies. Consider a policy $\pi \in \Delta_{S,A}$ and its corresponding world policy $\xi = f_\beta(\pi) \in \Delta_{W,A}$, where f_β is the linear map $\Delta_{S,A} \to \Delta_{W,A}$; $\pi(a|s) \mapsto \sum_s \beta(s|w)\pi(a|s)$. We can define sensor policy improvement cones as follows.

Definition 8 (Policy improvement cones)**.** For each sensor state $s \in S$, define

$$L^{\pi,s} = \{\pi' \in \Delta_{S,A} \colon f_\beta(\pi') \in L^{\xi,W_s}, \text{ and } \pi'_{s'} = \pi_{s'}, \text{ for } s' \neq s\}, \tag{4}$$

where $W_s = \operatorname{supp}(\beta(s|\cdot))$. This is an intersection of $|W_s|$ half-spaces in $\Delta_{\{s\},A}$, with fixed values in $\Delta_{(S\setminus\{s\}),A}$.

Lemma 9 (Policy improvement cones)**.** *For any $\pi \in \Delta_{S,A}$, $s \in S$, and $\pi' \in L^{\pi,s}$, we have $\mathcal{R}(\pi') \geq \mathcal{R}(\pi)$.*

Proof. This follows immediately by the way the cones $L^{\pi,s}$ are defined, and in view of the world policy improvement cone Lemma 4. $\square$

Lemma 9, together with Lemma 10 below, implies Theorem 1.

Lemma 10 (Montúfar & Rauh 2017, Lemma 5)**.** *Let P be a polytope with affine hull V, and let $l_1,\dots,l_k$ be vectors in V. For any $p \in P$, let $L_{i,+} = \{q \in P \colon \langle l_i, q-p\rangle \geq 0\}$. Then $\bigcap_{i=1}^k L_{i,+}$ contains an element q that belongs to a face of P of dimension at most $k-1$.*

We see that across all the policy improvement cones, there are only a total of $|W|$ inequalities. Instead of working with individual cones for each s, we can consider any set of the form $W_{S'} = \cup_{s'\in S'} W_{s'}$, and

$$L^{\pi,W_{S'}} = \{\pi' \in \Delta_{S,A} \colon f_\beta(\pi') \in L^{\xi,W_{S'}}, \text{ and } \pi'_s = \pi_s, \text{ for } s \in W \setminus W_{S'}\}. \tag{5}$$

In this case, Lemma 4, together with Lemma 10, implies that there is an optimal policy π^* with $\sum_{s'\in S'} |\operatorname{supp}(\pi(\cdot|s'))| \leq |S'| + |W_{S'}| - 1$ for all $S' \subseteq S$. Note that the $(d-1)$-dimensional faces of $\Delta_{S,A}$ are policies with at most $|S| + d - 1$ non-zero entries. This proves Theorem 2.

5 Discussion

We introduced a notion of policy improvement cones for the average reward in infinite-horizon POMDPs with memoryless stochastic policies. This allows us to study the average reward optimization problem globally (vs. policy gradients which only formulate local descriptions). We prove the existence of optimal policies that satisfy a series of constraints. These constraints are independent of the instantaneous reward function at hand, and hence they can be regarded as task-independent priors for POMDPs.

The results are formulated in terms of certain properties of the observation model that might be considered to be restrictive. In order to obtain generally applicable priors for approximate optimal policies, future work could explore extensions of the stability analysis of Rauh et al. (2019) from discounted to average rewards.

ACKNOWLEDGMENT

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

REFERENCES

Nihat Ay, Guido Montúfar, and Johannes Rauh. Selection criteria for neuromanifolds of stochastic dynamics. In Yoko Yamaguchi (ed.), *Advances in Cognitive Neurodynamics (III)*, pp. 147–154. Springer, 2013.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Open problem: Approximate planning of POMDPs in the class of memoryless policies. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *JMLR W&CP*, pp. 1639–1642. JMLR.org, 2016.

Guido Montúfar and Johannes Rauh. Geometry of policy improvement. In *Geometric Science of Information, LNCS 10589*, pp. 282–290. Springer, 2017.

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

Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. Geometry and determinism of optimal stationary control in POMDPs. *arXiv:1503.07206*, 2015.

Johannes Rauh, Nihat Ay, and Guido Montúfar. A continuity result for optimal memoryless planning in POMDPs. *Accepted for presentation at the 4th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*, 2019. URL `https://www.researchgate.net/publication/331824549_A_continuity_result_for_optimal_memoryless_planning_in_POMDPs`.

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

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

A Illustration

We visualize the geometry of the optimization problem in a small example. Let $W = \{1, 2\}$, $S = \{1, 2\}$, $A = \{1, 2\}$. Then $\Delta_{S,A}$ and $\Delta_{W,A}$ are squares, and $\Delta_{W\times A}$ is a tetrahedron. In Figure 2 we plot these sets, alongside with the values of the average reward, for a fixed random choice of the world state transition kernel α, a fixed random choice of the reward function R, and various choices of the observation kernel β.

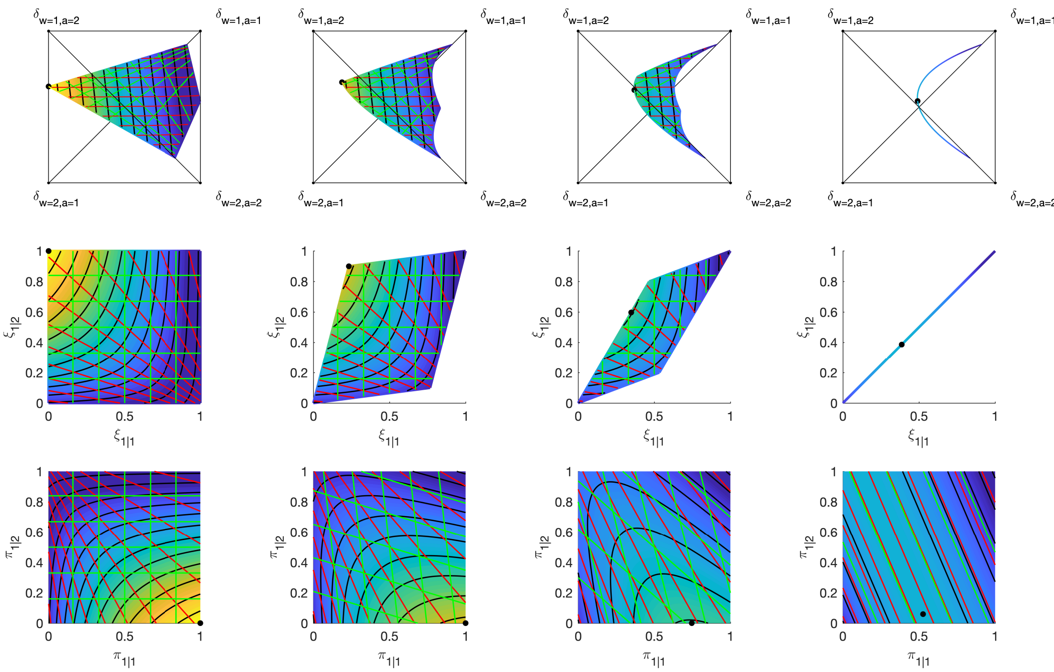

Figure 2: Top row: Simplex $\Delta_{W\times A}$ containing the set of feasible stationary joint distributions. Middle row: Polytope $\Delta_{W,A}$ containing the set of feasible world state policies. Bottom row: Policy polytope $\Delta_{S,A}$. Here R and α are fixed and β is ranging from full observability $[1, 0; 0, 1]$ (left) to blind $[1, 1; 0, 0]$ (right). Color indicates the average reward (darker is lower). Shown are also the level sets of the average reward (black), stationary distribution (red), and rows of the world policy (green). The red and green level sets are linear over $\Delta_{W\times A}$, $\Delta_{W\times A}$, and $\Delta_{S,A}$.