The Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon Partially Observable Markov Decision Processes

Johannes Müller[1], Guido Montúfar[1, 2]

[1] Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

[2] Department of Mathematics and Department of Statistics, UCLA, Los Angeles, USA

Max Planck Institute for Mathematics in the Sciences

UCLA

European Research Council
Established by the European Commission

Partially observable Markov decision processes (POMDPs)

- A POMDP is a tupel $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ and a MDP is a tupel $(\mathcal{S}, \mathcal{A}, \alpha, r)$, where:
- *State, oberservation, action spaces.* Finite sets $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$.
- *Observation mechanism.* Markov kernel $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}}$ from $\mathcal{S}$ to $\mathcal{O}$.
- *Action mechanism.* Markov kernel $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$ from $\mathcal{S}\times\mathcal{A}$ to $\mathcal{S}$.
- *Policies and effective policies.* Markov kernels $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ from $\mathcal{O}$ to $\mathcal{A}$; every policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ induces an *effective policy* $\tau = \pi \circ \beta \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ from $\mathcal{S}$ to $\mathcal{A}$.
- *Effective policy polytope.* The polytope $\Delta_{\mathcal{A}}^{\mathcal{S},\beta} := \{\pi \circ \beta \mid \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$.
- *State transition kernels.* a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ induces a transition kernel $p_\pi \in \Delta_{\mathcal{S}}^{\mathcal{A}}$.
- *Induced state action Markov process.* An initial distribution $\mu \in \Delta_{\mathcal{S}}$ and policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ define a Markov process $\mathbb{P}^{\pi,\mu}$ on $\mathcal{S}\times\mathcal{A}$.
- *Discounted reward.* We fix $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ and for $\gamma \in [0,1)$ we define

$$R_\gamma^\mu(\pi) := \mathbb{E}_{\mathbb{P}^{\pi,\mu}}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)\right].$$

Objective

Study the algebraic and geometric properties of the maximization of the reward R_γ^μ over $\Delta_{\mathcal{A}}^{\mathcal{O}}$.

The rational degree of the reward function

The following result describes the reward as the fraction of two determinantal polynomials.

Theorem 1. *It holds that*[a]

$$R_\gamma^\mu(\pi) = (1-\gamma)\cdot\frac{\det(I - \gamma p_\pi + r_\pi \otimes \mu)}{\det(I-\gamma p_\pi)} - 1 + \gamma,$$

where $r_\pi \in \mathbb{R}^{\mathcal{S}}$ is the one step reward *defined by $r_\pi(s) := \sum_a (\pi\circ\beta)(a|s) r(s,a)$. In particular, the reward function is a rational function in the entries of the policy. If restricted to the subset $\Pi \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$ of policies which agree with a fixed policy π_0 on all observations outside of $O \subseteq \mathcal{O}$, the rational degree of the reward function can be upper bounded by*

$$\deg(R_\gamma^\mu|_\Pi) \leq \max_{\pi\in\Pi} \operatorname{rank}(p_\pi - p_{\pi_0}) \leq \left|\left\{s \in \mathcal{S} \mid \beta(o|s) > 0 \text{ for some } o \in O\right\}\right|. \qquad (1)$$

Analogue statements can be made for the value function and the state-action frequencies, which are both important objects in Markov decision processes.

[a] Here, $\otimes$ denotes the Kronecker product.

State-action frequencies

A short calculation shows $R_\gamma^\mu(\pi) = \langle r, \eta_\gamma^{\pi,\mu}\rangle_{\mathcal{S}\times\mathcal{A}}$, where

$$\eta_\gamma^{\pi,\mu}(s,a) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi,\mu}(s_t = s, a_t = a)$$

is called the *state-action frequency* of π. We denote the set of all state-action frequencies in the fully and in the partially observable case by

$$\mathcal{N}_\gamma^\mu := \left\{\eta_\gamma^{\pi,\mu} \in \Delta_{\mathcal{S}\times\mathcal{A}} \mid \pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}\right\} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}} \text{ and } \mathcal{N}_\gamma^{\mu,\beta} := \left\{\eta_\gamma^{\pi,\mu} \in \Delta_{\mathcal{S}\times\mathcal{A}} \mid \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}\right\} \subseteq \mathcal{N}_\gamma^\mu.$$

The reward maximization problem is equivalent to the maximization of a linear function over the set $\mathcal{N}_\gamma^{\mu,\beta}$ of state-action frequencies. It is well known that the state-action frequencies $\mathcal{N}_\gamma^\mu$ of a fully observable Markov decision process form a polytope.

Proposition 2 (Characterisation of $\mathcal{N}_\gamma^\mu$)**.** *It holds that*

$$\mathcal{N}_\gamma^\mu = [0,\infty)^{\mathcal{S}\times\mathcal{A}} \cap \left\{\eta \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \mid \langle w_\gamma^s, \eta\rangle_{\mathcal{S}\times\mathcal{A}} = (1-\gamma)\mu_s \text{ for } s \in \mathcal{S}\right\},$$

where $w_\gamma^s = \delta_s \otimes \mathbb{1} - \gamma\alpha(s|\cdot,\cdot)$. In particular, $\mathcal{N}_\gamma^\mu$ is a subpolytope of $[0,\infty)^{\mathcal{S}\times\mathcal{A}}$, which is contained in an affine subspace with orientation only depending on γ and α.

Remark 3. By Theorem 1 the set $\mathcal{N}_\gamma^{\mu,\beta}$ of state-action distributions possesses a rational parametrization and is therefore a semialgebraic set by the Tarski-Seidenberg theorem and we will describe the defining polynomial inequalities in Theorem 4 and (2).

State-action frequencies of partially observable models

In partially observable models only the policies in the effective policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S},\beta} \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$ can be realized. In order to understand the polynomial inequalities defining the state-action frequencies, we need to understand how linear inequalities in the policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ behave in $\mathcal{N}_\gamma^\mu$. Since the inverse of $\pi \mapsto \eta_\gamma^{\pi,\mu}$ is given by conditioning it holds that

$$\sum_{s\in S, a\in\mathcal{A}} b_{sa}\pi_{sa} \geq c \Leftrightarrow \sum_{s\in S}\sum_{a\in\mathcal{A}} b_{sa}\eta_{sa} \prod_{s'\in S\setminus\{s\}}\sum_{a'\in\mathcal{A}} \eta_{s'a'} \geq c \prod_{s'\in S}\sum_{a'\in\mathcal{A}} \eta_{s'a'}. \qquad (2)$$

Using this principle we can derive the following abstract result.

Theorem 4. *Under mild conditions we have $\mathcal{N}_\gamma^{\mu,\beta} = \mathcal{N}_\gamma^\mu \cap \mathcal{V} \cap \mathcal{B}$, where $\mathcal{V}$ is a variety generated by multi-homogeneous polynomials and $\mathcal{B}$ is a basic semialgebraic set described by multi-homogeneous polynomial inequalities. The face lattices of $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ and $\mathcal{N}_\gamma^{\mu,\beta}$ are isomorphic.*

Using (2) one can obtain an explicit polynomial expression for the set $\mathcal{N}_\gamma^{\mu,\beta}$. If for example β has independent columns and if we set $S_o := \{s \in \mathcal{S} \mid \beta_{os}^+ \neq 0\}$, the semialgebraic set $\mathcal{B}$ from Theorem 4 is described by the multihomogeneous inequalities

$$p_{ao}(\eta) := \sum_{s\in S_o}\left(\beta_{os}^+\eta_{sa}\cdot\prod_{s'\in S_o\setminus\{s\}}\sum_{a'}\eta_{s'a'}\right) \geq 0 \quad \text{for all } a \in \mathcal{A}, o \in \mathcal{O}.$$

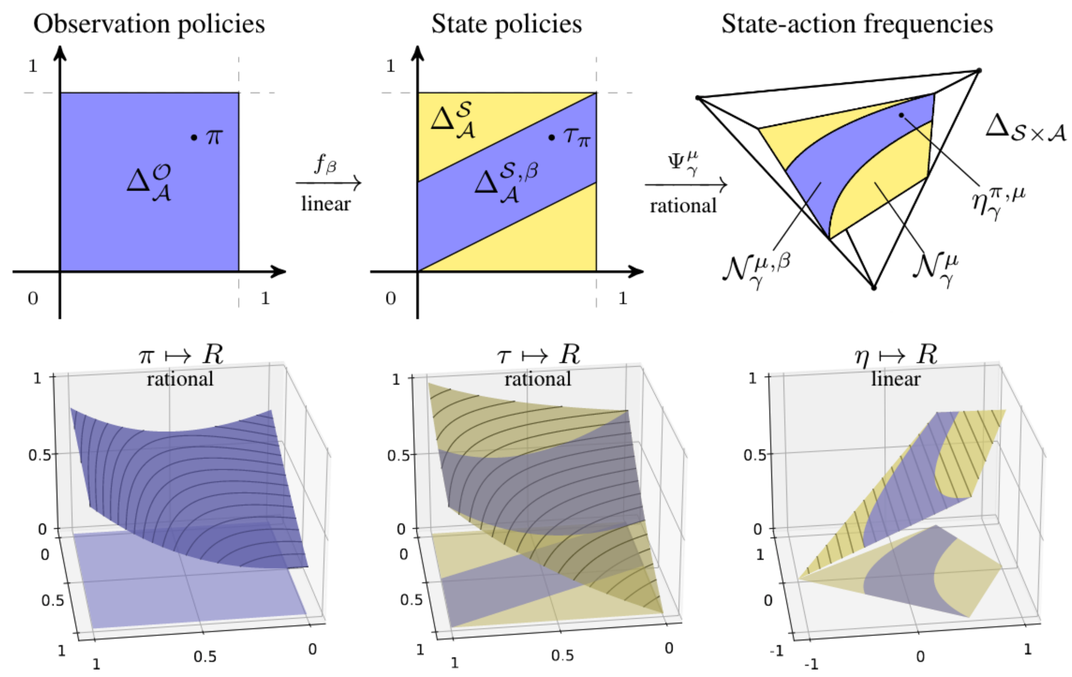
Number of critical points of reward maximization

By Theorem 4, the problem of reward maximization is equivalent to the maximization of a linear function subject to polynomial constraints. Hence, one can use the general theory of algebraic degrees of polynomial optimization to bound the number of critical points of the reward function, see [NR09, MM21]. In special cases one can refine those general bounds using the special algebraic structure of the polynomial constraints to obtain the following result.

Proposition 5. *Let $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \alpha, \beta, r)$ be a POMDP describing a blind controller with two actions, i.e., $\mathcal{O} = \{o\}$ and $\mathcal{A} = \{a_1, a_2\}$ and let r, α and μ be generic and let $\gamma \in (0,1)$. Then the reward function R_γ^μ has at most $|\mathcal{S}|$ critical points in the interior* $\operatorname{int}(\Delta_{\mathcal{A}}^{\mathcal{O}}) \cong (0,1)$.

Illustration of the results

We want to illustrate our results for a POMDP with two states, two actions and two observations, for details see [MM21].



The top row shows the observation policy polytope $\Delta_{\mathcal{A}}^{\mathcal{O}}$; the associated state policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S}}$ (yellow) and the subset of effective policies $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ (blue); and the corresponding sets of discounted state-action frequencies in the simplex $\Delta_{\mathcal{S}\times\mathcal{A}}$ (a tetrahedron). The bottom shows the graph of the discounted reward R_γ^μ as a function of the observation policy π; the state policy τ; and the discounted state-action frequencies η.

Conclusion

- The degree of observability directly relates to the rational degree of the reward function.
- The state-action frequencies form a basic semi-algebraic set.
- Reward maximization is equivalent to a polynomially constraint optimization problem with linear objective.

References

[MM21] Johannes Müller and Guido Montúfar. The Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon POMDPs, 2021.

[NR09] Jiawang Nie and Kristian Ranestad. Algebraic degree of polynomial optimization. *SIAM Journal on Optimization*, 20(1):485–502, 2009.

[Put14] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Acknowledgements

The authors thank Alex Lin and Tom Merkh and Bernd Sturmfels for valuable discussions and acknowledge support by the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement no 757983). JM acknowledges support by the IMPRS for Mathematics in the Sciences and the Evangelisches Studienwerk Villigst.