The geometry of discounted stationary distributions of 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

Notation of Markov decision processes

- *State, oberservation, action spaces.* finite sets $\mathcal{S}, \mathcal{O}$ and $\mathcal{A}$.
- *Observation mechanism.* Markov kernel $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}}$.
- *Action mechanism.* Markov kernel $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$.
- *Policies and effective policies.* Markov kernels $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$; every policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ induces an *effective policy* $\pi \circ \beta \in \Delta_{\mathcal{A}}^{\mathcal{S}}$.
- *Effective policy polytope.* $\Pi_\beta := \{\pi \circ \beta \mid \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}\} \subseteq \Delta_{\mathcal{A}}^{\mathcal{S}}$.
- *State action and state transition kernels.* a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ induces transition kernels $P_\pi \in \Delta_{\mathcal{S}\times\mathcal{A}}^{\mathcal{S}\times\mathcal{A}}$ and $p_\pi \in \Delta_{\mathcal{S}}^{\mathcal{A}}$.
- *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

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

Factorisations of the reward function

An application of Fubini's theorem to the definition of R_γ^μ shows $R_\gamma^\mu(\pi) = \langle r, \eta_\gamma^{\pi,\mu}\rangle_{\mathcal{S}\times\mathcal{A}}$, where we call

$$\eta_\gamma^{\pi,\mu} := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi,\mu}(s_t = \cdot, a_t = \cdot) \in \Delta_{\mathcal{S}\times\mathcal{A}}$$

the *discounted stationary distribution* of π. Hence, the reward function R_γ^μ factorises in non linear and linear parts according to

$$\pi \mapsto \pi \circ \beta \mapsto \eta_\gamma^{\pi,\mu} \mapsto \langle r, \eta_\gamma^{\pi,\mu}\rangle_{\mathcal{S}\times\mathcal{A}}.$$

Objective

Study the algebraic and geometric properties of the set of discounted stationary distributions and of the mapping $\pi \mapsto \eta_\gamma^{\pi,\mu}$ since they encode the complexity of reward maximisation.

The rational degree of of discounted stationary distributions

Proposition 1 (Characterisation of discounted stationary distributions). *Let $\rho_\gamma^{\pi,\mu}$ denote the state marginal of $\eta_\gamma^{\pi,\mu}$. It holds that*

$$\eta_\gamma^{\pi,\mu} = (1-\gamma)(I-\gamma P_\pi^T)^{-1}(\mu * (\pi\circ\beta)) \quad \textit{and } \rho_\gamma^{\pi,\mu} = (1-\gamma)(I-p_\pi^T)^{-1}\mu. \tag{1}$$

Applying Cramer's rule to (1) yields

$$\eta_\gamma^{\pi,\mu}(s,a) = \pi(a|s)\rho_\gamma^{\pi,\mu}(s) = \pi(a|s)\cdot\frac{\det(I-\gamma p_\pi^T)_s^\mu}{\det(I-\gamma p_\pi^T)},$$

where $(I-\gamma p_\pi)_s^\mu$ is the matrix that is obtained by replacing the s-th column of $(I-\gamma p_\pi)$ by μ. Computing the degree of multivariate determinantal polynomials gives the following result.

Theorem 1 (Rational degree). *The reward function, the value function and the discounted stationary distribution $\eta_\gamma^{\pi,\mu}$ are rational functions in the entries of the policies. Restricted to the subset $\Pi \subseteq \Delta_{\mathcal{A}}^{\mathcal{O}}$ of policies which agree with a fixed policy on all states outside of $O \subseteq \mathcal{O}$ their degree is upper bounded by*

$$\left|\{s \in \mathcal{S} \mid \beta(o|s) > 0 \textit{ for some } o \in O\}\right|.$$

The polytope of discounted stationary distributions

We consider the fully observable case now, i.e. the case where β admits a left inverse and let us denote the set of all discounted stationary distributions with $\mathcal{N}_\gamma^\mu$.

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

$$\mathcal{N}_\gamma^\mu = \left(\eta_\gamma^\mu + \{w_\gamma^s \mid s \in \mathcal{S}\}^\perp\right) \cap \Delta_{\mathcal{S}\times\mathcal{A}},$$

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

Theorem 2 (Combinatorial equivalence of $\mathcal{N}_\gamma^\mu$ and $\Delta_{\mathcal{A}}^{\mathcal{S}}$). *The mapping $\pi \mapsto \eta_\gamma^{\pi,\mu}$ induces an order preserving morphism of the face lattices of $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\mathcal{N}_\gamma^\mu$. If further $\rho_\gamma^{\pi,\mu} > 0$ holds entrywise for all policies $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$, then this is an isomorphism.*

Acknowledgements

The authors thank Alex Lin and Tom Merkh 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 International Max Planck Research School for Mathematics in the Sciences and the Evangelisches Studienwerk Villigst.

The case of partial observability

In the partially observable case, the set of discounted stationary distributions is typically not a polytope, but by the Tarski-Seidenberg theorem, it still is a semi-algebraic set. In order to understand its defining inequalities, it is necessary 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, a linear inequality of the form

$$\sum_{s\in S, a\in\mathcal{A}} b_{sa}\pi_{sa} \le c,$$

corresponds to the polynomial inequality

$$c\prod_{s'\in S}\sum_{a'\in\mathcal{A}}\eta_{s'a'} \ge \sum_{s\in S}\sum_{a\in\mathcal{A}} b_{sa}\eta_{sa}\prod_{s'\in S\setminus\{s\}}\sum_{a'\in\mathcal{A}}\eta_{s'a'},$$

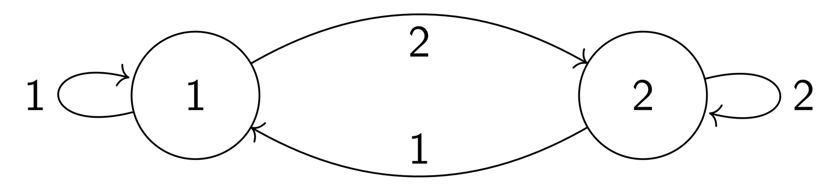
which is a polynomial inequality of degree at most $|S|$. Computing the defining inequalities of the effective policy polytope yields the following result.

Theorem 3 (Defining polynomial inequalities). *Let β be invertible and set $S_o := \{s \in \mathcal{S} \mid \beta_{os}^{-1} \neq 0\}$. Then $\eta \in \mathcal{N}_\gamma^\mu$ is a discounted stationary distrbution of the POMDP if and only if*

$$-\sum_{s\in S_o}\left(\beta_{os}^{-1}\eta_{sa}\cdot\prod_{s'\in S_o\setminus\{s\}}\sum_{a'}\eta_{s'a'}\right) \le 0 \quad \textit{for all } a \in \mathcal{A}, o \in \mathcal{O}.$$

A toy example

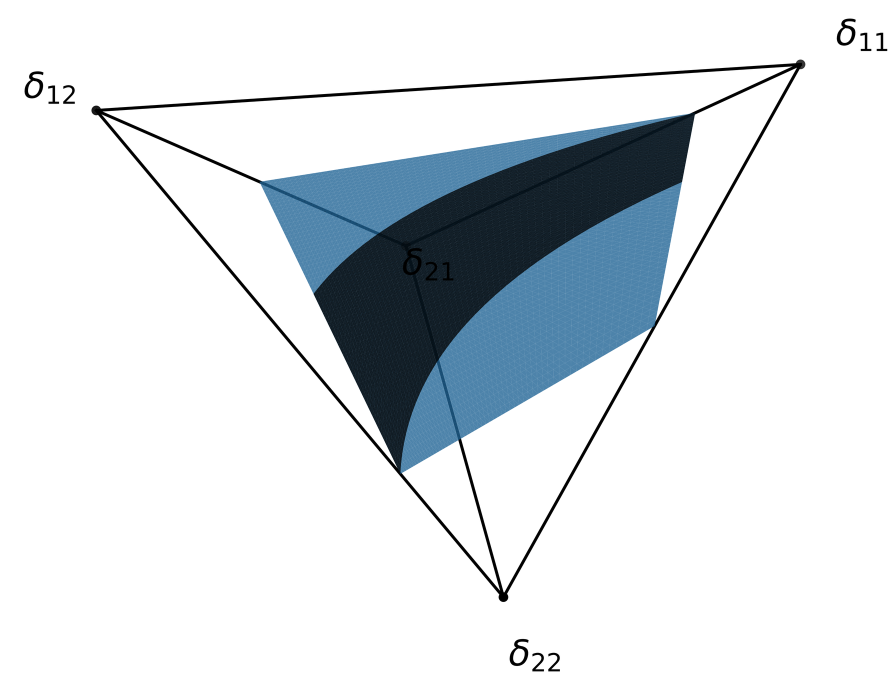
Let us consider a toy example with $\mathcal{S} = \mathcal{A} = \{1, 2\}$ and the following transition model α.



Further, we consider the observation space $\mathcal{O} = \{1, 2\}$ and the observation mechanism $\beta(1|s_i) = 1 - \delta_{i2}/2$. The defining two quadratic inequalities in the discounted state action polytope $\mathcal{N}_\gamma^\mu$ are given by

$$\eta_{11}\eta_{22} - \eta_{21}\eta_{11} - 2\eta_{21}\eta_{12} \le 0$$
$$\eta_{12}\eta_{21} - 2\eta_{22}\eta_{11} - \eta_{22}\eta_{12} \le 0.$$

In the following plot, the entire polytope of discounted stationary distributions for the fully observable case and its subset corresponding to the observation mechanism β are shown. The black lines show a three dimensional projection of the probability simplex $\Delta_{\mathcal{S}\times\mathcal{A}} \cong \Delta_3$.



Conclusion and outlook

- The degree of observability directly relates rational degree of the discounted stationary distributions.
- The set of discounted stationary distributions is a semi-algebraic subset of $\Delta_{\mathcal{S}\times\mathcal{A}}$ defined by a set of linear equalities $A\eta = b$ and polynomial inequalities $p(\eta) \le 0$, where
  - A depends on γ and α,
  - b depends on μ, γ and α,
  - p depends only on β and is homogeneous and square free.

References

[MGZA15] Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay. Geometry and determinism of optimal stationary control in partially observable Markov decision processes. *arXiv:1503.07206*, 2015.

[NT12] Tim Netzer and Andreas Thom. Polynomials with and without determinantal representations. *Linear algebra and its applications*, 437(7):1579–1595, 2012.

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

[Zie12] Günter M Ziegler. *Lectures on polytopes*, volume 152. Springer Science & Business Media, 2012.