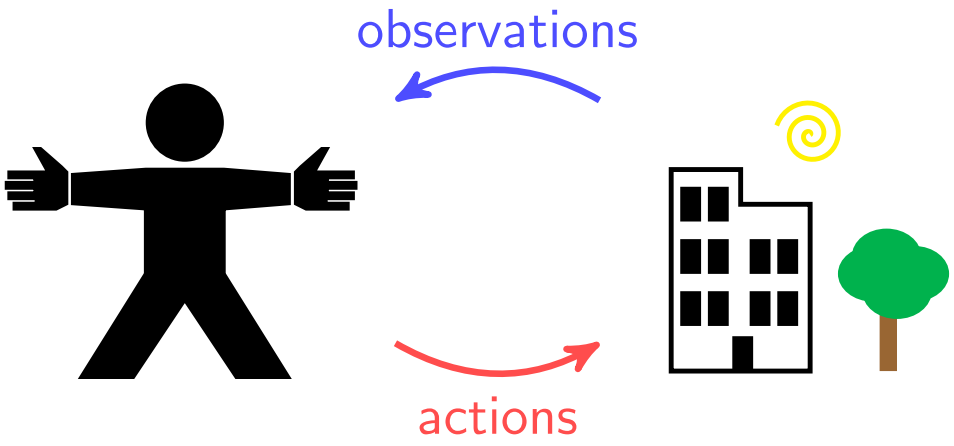
Geometry of Memoryless Stochastic Policy Optimization in Infinite-Horizon POMDPs

Guido Montúfar
UCLA & MPI MiS

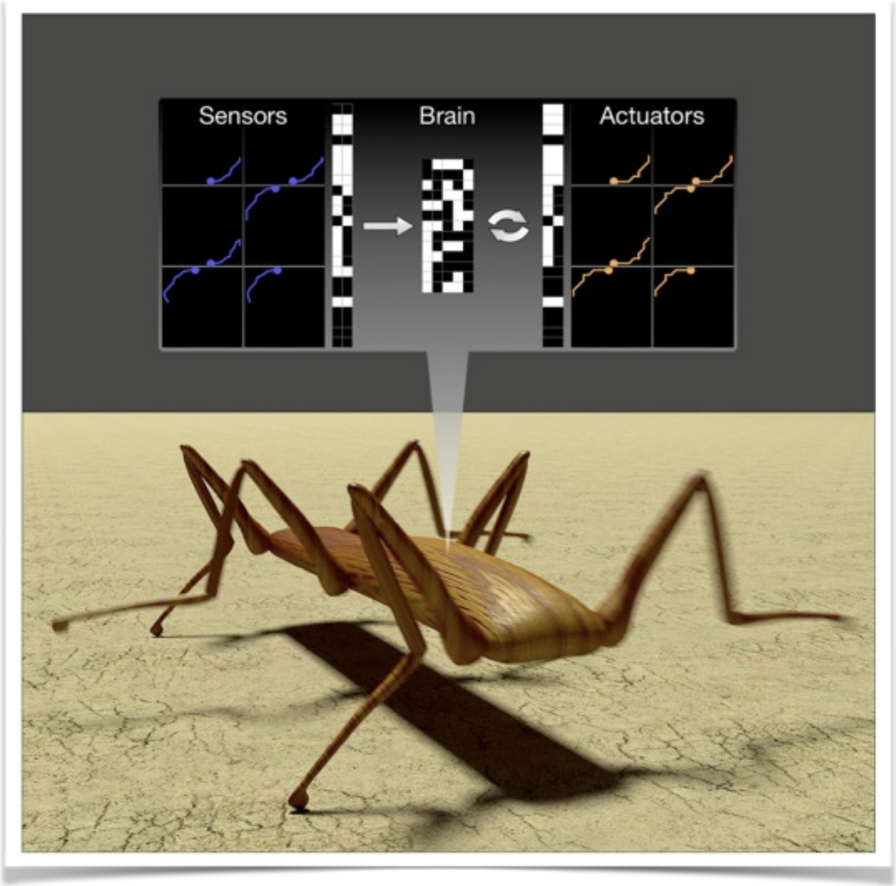
Joint work with
J. Müller

Previous works with
J. Rauh K. Zahedi N. Ay

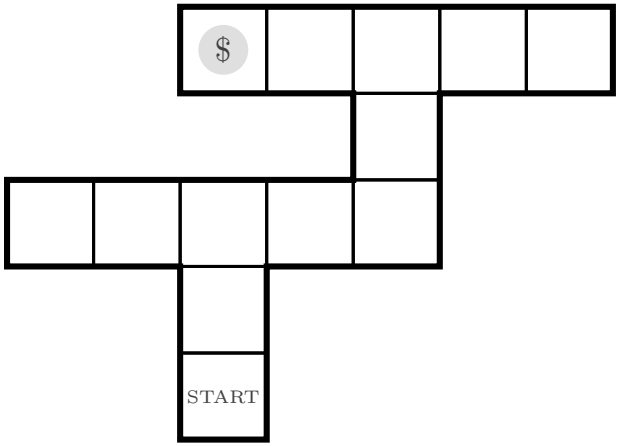
International Conference on Information Geometry for Data Sciences (IG4DS), Sep 2022



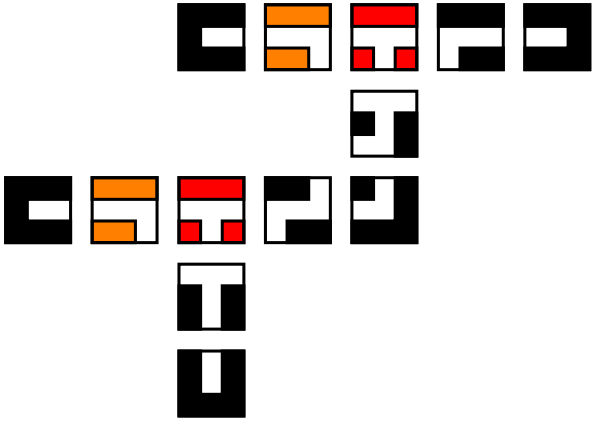
Want to optimize the action selection mechanism (policy)



World



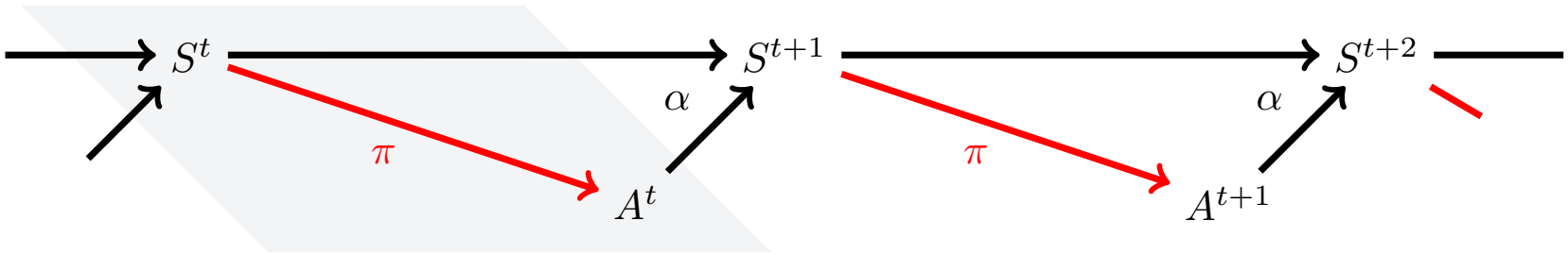
Observations



[Montúfar et al., 2015, Montúfar et al., 2015]

Markov process $\mathbb{P}^{\pi,\mu}$ on $\mathcal{S} \times \mathcal{A}$ with transition kernel

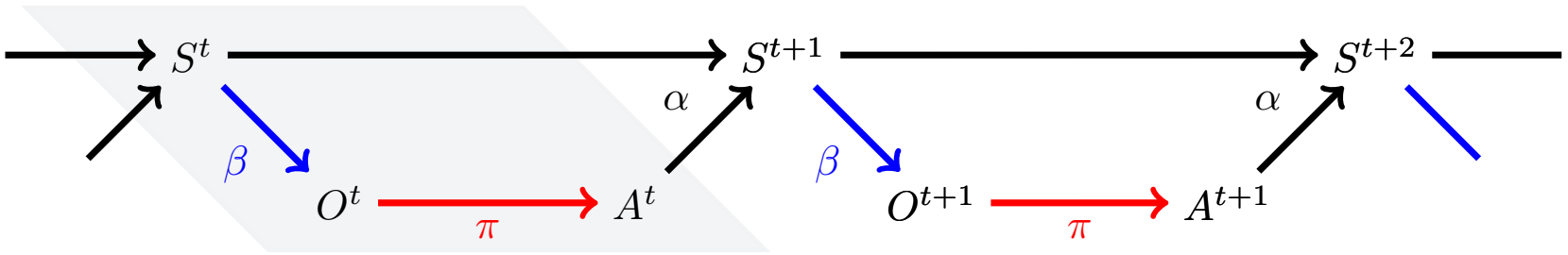
$$P_\pi(s', a'|s, a) = \alpha(s'|s, a)\pi(a'|s')$$



Markov Decision Process

Markov process $\mathbb{P}^{\pi,\mu}$ on $\mathcal{S} \times \mathcal{A}$ with transition kernel

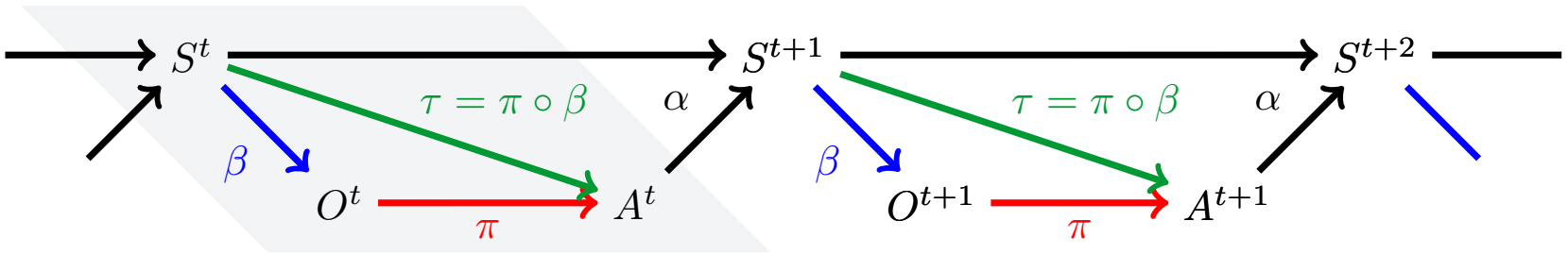
$$P_\pi(s', a'|s, a) = \alpha(s'|s, a) \sum_{o'} \pi(a'|o')\beta(o'|s')$$



Partially Observable Markov Decision Process

Markov process $\mathbb{P}^{\pi,\mu}$ on $\mathcal{S} \times \mathcal{A}$ with transition kernel

$$P_\pi(s', a'|s, a) = \alpha(s'|s, a)(\pi \circ \beta)(a'|s')$$



Partially Observable Markov Decision Process

Agent receives instantaneous reward $r(s, a)$ for taking action a at state s

Our objective is to optimize the expected long term reward:

Expected discounted reward

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

Expected mean reward

$$R_1^\mu(\pi) := \lim_{T\to\infty}\mathbb{E}_{\mathbb{P}^{\pi,\mu}}\left[\frac{1}{T}\sum_{t=0}^{T-1} r(s_t, a_t)\right]$$

In this case we will assume P_π has a unique stationary distribution

Some years ago...

Known:

- MDPs have optimal policies that are memoryless and deterministic.
- POMDPs may have memoryless stochastic policies that are better than any memoryless deterministic policy.

J. Rauh

K. Zahedi

N. Ay

Some years ago...

Known:

- MDPs have optimal policies that are memoryless and deterministic.
- POMDPs may have memoryless stochastic policies that are better than any memoryless deterministic policy.

Question: Do POMDPs have optimal memoryless policies that are more or less deterministic?

J. Rauh

K. Zahedi

N. Ay

Some years ago...

Known:

- MDPs have optimal policies that are memoryless and deterministic.
- POMDPs may have memoryless stochastic policies that are better than any memoryless deterministic policy.

Question: Do POMDPs have optimal memoryless policies that are more or less deterministic? **Yes,**

- deterministic at any observation that fully identifies the state [Montúfar et al., 2015]

J. Rauh

K. Zahedi

N. Ay

Some years ago...

Known:

- MDPs have optimal policies that are memoryless and deterministic.
- POMDPs may have memoryless stochastic policies that are better than any memoryless deterministic policy.

Question: Do POMDPs have optimal memoryless policies that are more or less deterministic? **Yes,**

- deterministic at any observation that fully identifies the state [Montúfar et al., 2015]
- randomize at most k actions at any observation that identifies the state up to k possibilities [Montúfar and Rauh, 2017]

J. Rauh

K. Zahedi

N. Ay

Some years ago...

Known:

- MDPs have optimal policies that are memoryless and deterministic.
- POMDPs may have memoryless stochastic policies that are better than any memoryless deterministic policy.

Question: Do POMDPs have optimal memoryless policies that are more or less deterministic? **Yes,**

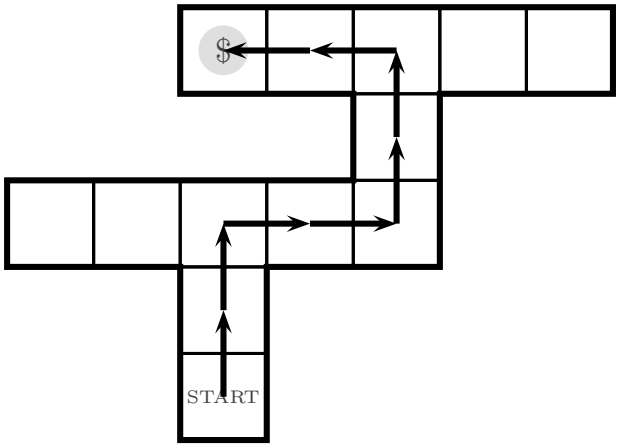
- deterministic at any observation that fully identifies the state [Montúfar et al., 2015]
- randomize at most k actions at any observation that identifies the state up to k possibilities [Montúfar and Rauh, 2017]
- aggregated randomization over subsets of observations bounded by union of possible states [Montúfar et al., 2019]

J. Rauh

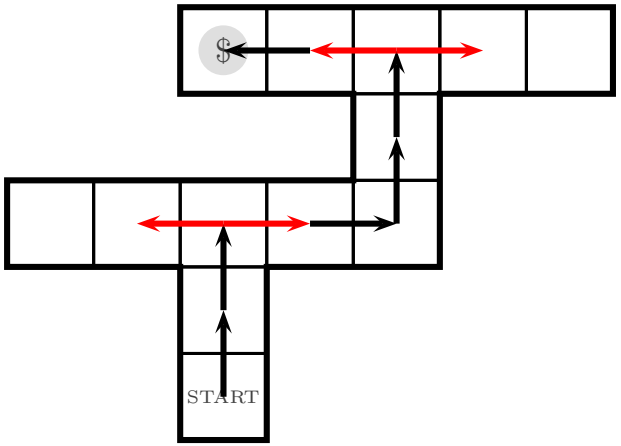
K. Zahedi

N. Ay

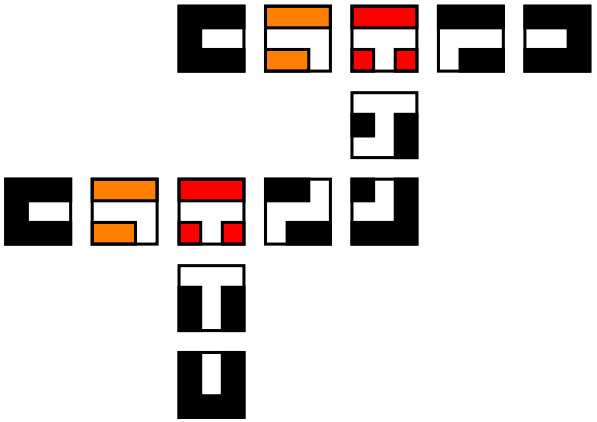
Policy based on states (MDP)



Policy based on observations (POMDP)



Observations



In general the randomized actions are not those taken if states were known

A short calculation shows that

$$R_\gamma^\mu(\pi) = \sum_{s,a} r(s,a)\eta_\gamma^{\pi,\mu}(s,a) = \langle r, \eta_\gamma^{\pi,\mu}\rangle_{\mathcal{S}\times\mathcal{A}},$$

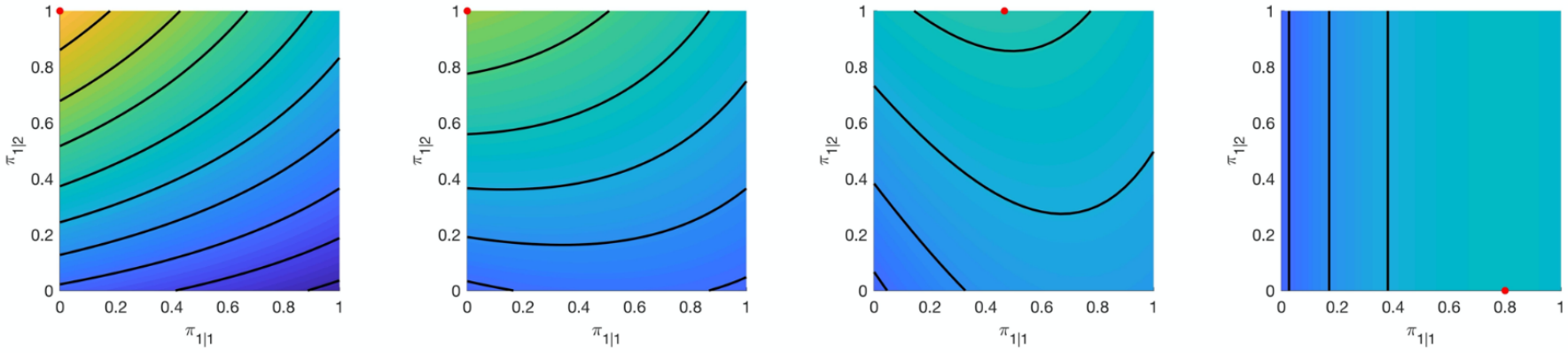
where the expected (discounted) state-action frequency is

$$\eta_\gamma^{\pi,\mu}(s,a) := \begin{cases} (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}^{\pi,\mu}(s_t = s, a_t = a), & \text{if } \gamma \in (0,1) \\ \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{P}^{\pi,\mu}(s_t = s, a_t = a), & \text{if } \gamma = 1. \end{cases}$$

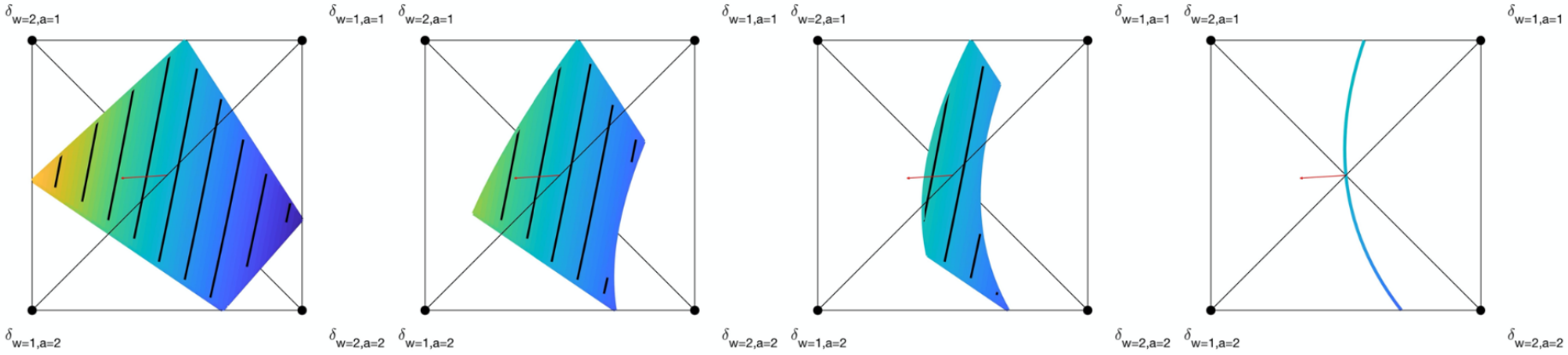
Observation: The optimization problem is linear in η

Idea: Study set of feasible η and how η (and R) depends on π

Policies π

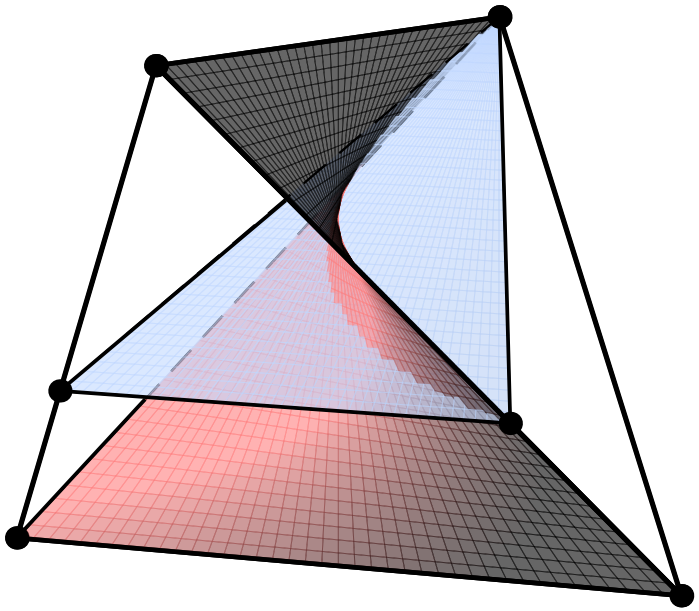


State-action frequencies η



full observability → no observability

Linear optimization over a constrained set of feasible η



Stationarity

Partial Observability

We showed:

- set of feasible η decomposes into certain (infinitely many) convex sets
- at any π there exists a global policy improvement cone of R_γ^μ

More recently...

Question:

- Structure of the objective over π?
- Describe the set of feasible state-action frequencies η?

J. Müller

[Müller and Montúfar, 2022a]

More recently...

Question:

- Structure of the objective over π?
- Describe the set of feasible state-action frequencies η?

We show:

- η and R_γ^μ are rational in π of degree directly related to observability.

J. Müller

[Müller and Montúfar, 2022a]

More recently...

Question:

- Structure of the objective over π?
- Describe the set of feasible state-action frequencies η?

We show:

- η and R_γ^μ are rational in π of degree directly related to observability.
- Explicit semialgebraic description of feasible state-action frequencies.

J. Müller

[Müller and Montúfar, 2022a]

More recently...

Question:

- Structure of the objective over π?
- Describe the set of feasible state-action frequencies η?

We show:

- η and R_γ^μ are rational in π of degree directly related to observability.
- Explicit semialgebraic description of feasible state-action frequencies.
- POMDP generalization of MDP dual linear programming.

J. Müller

[Müller and Montúfar, 2022a]

More recently...

Question:

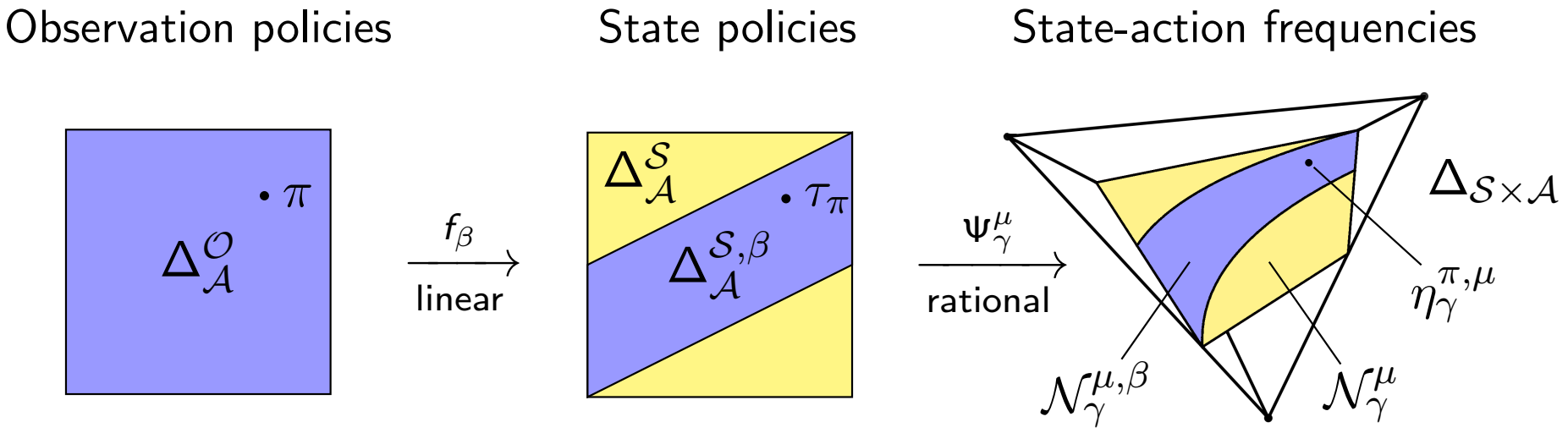
- Structure of the objective over π?
- Describe the set of feasible state-action frequencies η?

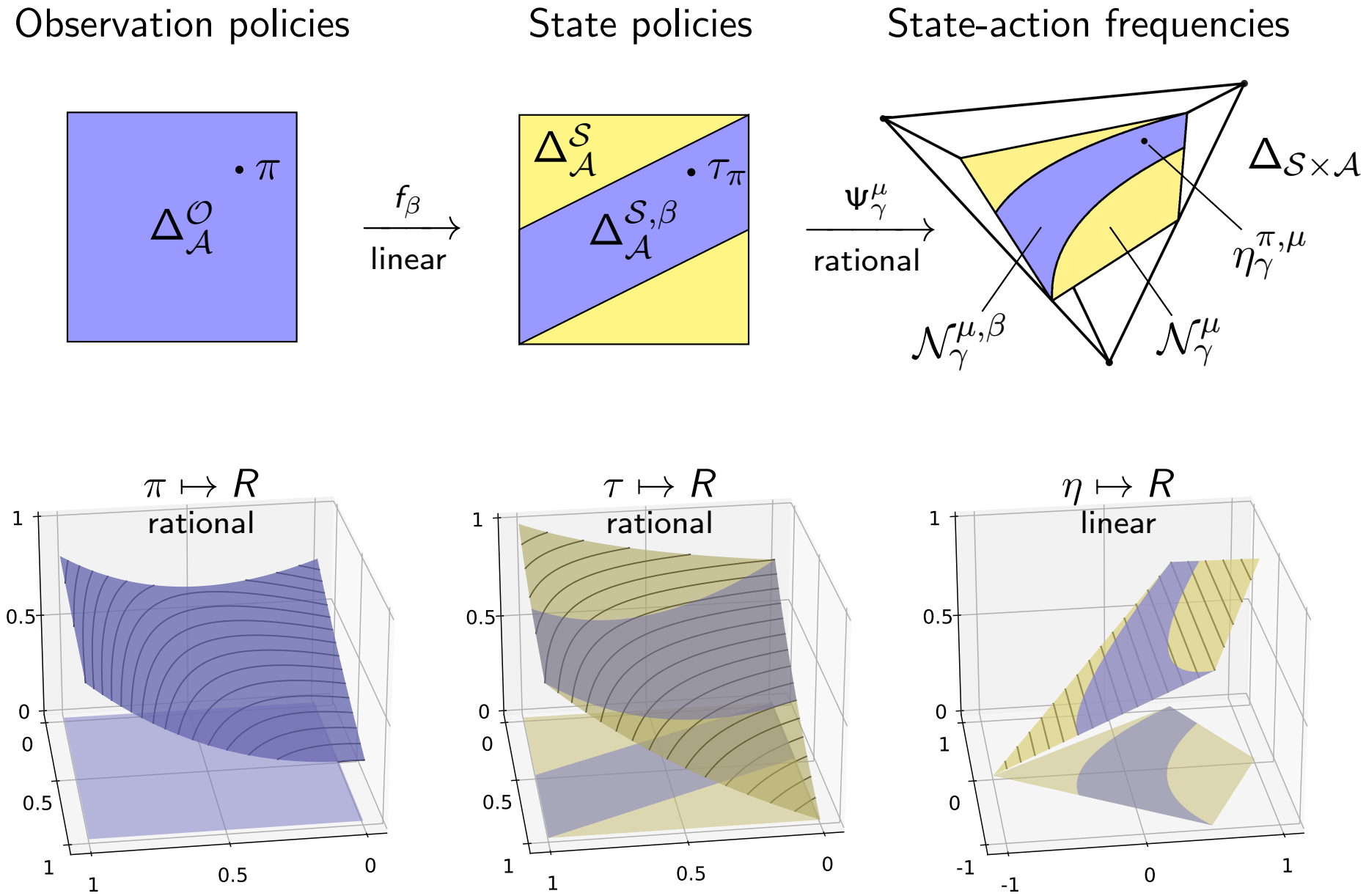
We show:

- η and R_γ^μ are rational in π of degree directly related to observability.
- Explicit semialgebraic description of feasible state-action frequencies.
- POMDP generalization of MDP dual linear programming.
- Bound number of critical points and components of super-level sets.

J. Müller

[Müller and Montúfar, 2022a]


Observation policies
State policies
State-action frequencies
π
$\Delta_{\mathcal{A}}^{\mathcal{O}}$
f_β
linear
$\Delta_{\mathcal{A}}^{\mathcal{S}}$
τ_π
$\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$
Ψ_γ^μ
rational
$\Delta_{\mathcal{S}\times\mathcal{A}}$
$\eta_\gamma^{\pi,\mu}$
$\mathcal{N}_\gamma^{\mu,\beta}$
$\mathcal{N}_\gamma^{\mu}$


Observation policies
State policies
State-action frequencies
$\Delta_{\mathcal{A}}^{\mathcal{O}}$
π
f_β
linear
$\Delta_{\mathcal{A}}^{\mathcal{S}}$
$\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$
τ_π
Ψ_γ^μ
rational
$\Delta_{\mathcal{S}\times\mathcal{A}}$
$\eta_\gamma^{\pi,\mu}$
$\mathcal{N}_\gamma^{\mu,\beta}$
$\mathcal{N}_\gamma^{\mu}$
$\pi \mapsto R$
rational
$\tau \mapsto R$
rational
$\eta \mapsto R$
linear

1 The Parametrization

2 The Feasible Set

3 Number and Location of Critical Points

4 A few Experiments

Theorem 1 (Degree of POMDPs)

- $\eta_\gamma^{\pi,\mu}$, $\rho_\gamma^{\pi,\mu}$, V_γ^π *and* $R_\gamma^\mu(\pi)$ *are rational functions with common denominator in the entries of* π.
- *Further, if restricted to policies which agree with a fixed* π_0 *on all states outside of* $O \subseteq \mathcal{O}$, *they have degree at most*

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

We also obtain several refinements of this.

By Theorem 1 and the Tarski-Seidenberg theorem:

Corollary 2

Let $\gamma \in (0,1)$. The set of feasible $\eta_\gamma^{\pi,\mu}$ and V_γ^π are semialgebraic sets[1]*. The level sets and super-level sets are semialgebraic sets.*

We also compute the linear and polynomial (in)equalities.

[1] A set defined by polynomial inequalities or a finite union of such sets.

By Theorem 1, the cumulative reward of (PO)MDPs is a degree-one rational function in every row of the effective policy. This yields:

Corollary 3 (Determinism of optimal policies)

There is an optimal memoryless policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}}$ that is deterministic on every o with $|\{s \in \mathcal{S} \mid \beta(o|s) > 0\}| \leq 1$.

This was shown in [Montúfar et al., 2015] using different methods.

Moreover, we obtain a refinement of [Dadashi et al., 2019] Line Theorem:

Proposition 4 (Linear interpolation and speed)

- *Let* $\gamma \in (0,1)$ *and let* $\pi_0, \pi_1 \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ *differ on at most* k *states. Then interpolations* $\lambda \mapsto V_\lambda$ *and* $\lambda \mapsto \eta_\lambda$ *have rational degrees at most* k*.*
- *If they differ on at most one state* $\tilde{s} \in \mathcal{S}$ *then*

$$V_\lambda = V_0 + c(\lambda)\cdot(V_1 - V_0) \;\; \textit{and} \;\; \eta_\lambda^\mu = \eta_0^\mu + c(\lambda)\cdot(\eta_1^\mu - \eta_0^\mu) \;\; \textit{for all } \lambda \in [0,1],$$

where

$$c(\lambda) = \frac{\det(I - \gamma p_1)\lambda}{\det(I - \gamma p_\lambda)} = \lambda \cdot \frac{\rho_\lambda^\mu(\tilde{s})}{\rho_1^\mu(\tilde{s})}.$$

1 The Parametrization

2 The Feasible Set

3 Number and Location of Critical Points

4 A few Experiments

MDP state-action frequencies

Proposition 5 (Characterization of $\mathcal{N}_\gamma^\mu$)

Let $(\mathcal{S}, \mathcal{A}, \alpha, r)$ *be an MDP,* $\mu \in \Delta_\mathcal{S}$ *be an initial distribution and* $\gamma \in (0, 1]$*. It holds that*

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

where $w_\gamma^s := \delta_s \otimes \mathbb{1}_\mathcal{A} - \gamma\alpha(s|\cdot,\cdot)$*.*

For $\gamma \in (0, 1)$*,* $\Delta_{\mathcal{S}\times\mathcal{A}}$ *can be replaced by* $[0, \infty)^{\mathcal{S}\times\mathcal{A}}$ *in* (1)*.*

Effective policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$

- Inequality descriptions of linear-map images of polytopes are non-trivial but can be approached algorithmically.
- To obtain closed-formulas, we make an assumption on β, but our subsequent analysis can handle any inequalities.

Effective policy polytope $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$

- Inequality descriptions of linear-map images of polytopes are non-trivial but can be approached algorithmically.
- To obtain closed-formulas, we make an assumption on β, but our subsequent analysis can handle any inequalities.

Theorem 6 (Characterization of the effective policy polytope)
Assume $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{O}}$ has linearly independent columns. Then

$$\Delta_{\mathcal{A}}^{\mathcal{S},\beta} = \Delta_{\mathcal{A}}^{\mathcal{S}} \cap \mathcal{U} \cap \mathcal{C}, \quad \textit{where} \tag{2}$$

- $\mathcal{U} = \{\pi \circ \beta \mid \pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}}\} = \ker(\beta^T)^{\perp}$ *is a subspace,*
- $\mathcal{C} = \{\tau \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \mid \beta^{+}\tau \geq 0\}$ *is a pointed polyhedral cone.*

Further, the face lattices of $\Delta_{\mathcal{A}}^{\mathcal{O}}$ and $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ are isomorphic.

Also descriptions for deterministic β.

POMDP state-action frequencies

Theorem 7 (Characterization of $\mathcal{N}_\gamma^{\mu,\beta}$)

Let $\gamma \in (0,1]$ and assume that $\rho_\gamma^{\pi,\mu}(s) > 0$, $s \in \mathcal{S}$, for all π. Then

$$\mathcal{N}_\gamma^{\mu,\beta} = \mathcal{N}_\gamma^{\mu} \cap \mathcal{V} \cap \mathcal{B}, \quad \textit{where}$$

- *$\mathcal{V}$ is defined by multi-homogeneous polynomial eqs. (from $\mathcal{U}$),*
- *$\mathcal{B}$ is defined by multi-homogeneous polynomial ineqs. (from $\mathcal{C}$).*

Further, the face lattices of $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ and $\mathcal{N}_\gamma^{\mu,\beta}$ are isomorphic.

POMDP state-action frequencies

Explicit (in)equalities of $\mathcal{N}_\gamma^{\mu,\beta}$

- By Theorem 6 and Proposition 9, the defining polynomials of the basic semialgebraic set $\mathcal{B}$ are indexed by $a \in \mathcal{A}, o \in \mathcal{O}$ and are given by

$$p_{ao}(\eta) := \sum_{s \in S_o} \left(\beta^+_{os} \eta_{sa} \prod_{s' \in S_o \setminus \{s\}} \sum_{a'} \eta_{s'a'} \right) \geq 0, \tag{3}$$

  where $S_o := \{s \in \mathcal{S} \mid \beta^+_{os} \neq 0\}$.
- A complete description of the set $\mathcal{N}_\gamma^{\mu,\beta}$ via (in)equalities follows from the description of $\mathcal{N}_\gamma^{\mu}$ via linear (in)equalities given in (1).

1 The Parametrization

2 The Feasible Set

3 Number and Location of Critical Points

4 A few Experiments

Using Theorem 6, Proposition 9 and an upper bound on the algebraic degree of polynomial optimization by [Nie and Ranestad, 2009] yields:

Using Theorem 6, Proposition 9 and an upper bound on the algebraic degree of polynomial optimization by [Nie and Ranestad, 2009] yields:

Theorem 8 (Bound on the number of critical points)

- *Let $\gamma \in (0,1)$, r is generic, $\beta \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}}$ invertible, and $\rho_\gamma^{\pi,\mu}(s) > 0$.*
- *For $I \subseteq \mathcal{A} \times \mathcal{O}$ consider the relative interior of face of $\Delta_{\mathcal{A}}^{\mathcal{O}}$:*

$$\operatorname{int}(F) = \left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} \mid \pi(a|o) = 0 \text{ if and only if } (a,o) \in I\right\}.$$

- *Then, the number of critical points of R_γ^μ on $\operatorname{int}(F)$ is at most*

$$\left(\prod_{o\in O} d_o^{k_o}\right) \cdot \sum_{\sum_{o\in O} i_o = I} \prod_{o\in O} (d_o - 1)^{i_o}, \quad \text{where } I = |\mathcal{S}|(|\mathcal{A}| - 1) - |I|,$$

$O := \{o \in \mathcal{O} \mid (a,o) \in I \text{ f.s. } a\}$, $k_o := |\{a \mid (a,o) \in I\}|$, $d_o := |\{s \mid \beta_{os}^{-1} \neq 0\}|$.

Using Theorem 6, Proposition 9 and an upper bound on the algebraic degree of polynomial optimization by [Nie and Ranestad, 2009] yields:

Theorem 8 (Bound on the number of critical points)

- *Let $\gamma \in (0,1)$, r is generic, $\beta \in \mathbb{R}^{\mathcal{S}\times\mathcal{O}}$ invertible, and $\rho_\gamma^{\pi,\mu}(s) > 0$.*
- *For $I \subseteq \mathcal{A} \times \mathcal{O}$ consider the relative interior of face of $\Delta_{\mathcal{A}}^{\mathcal{O}}$:*

$$\operatorname{int}(F) = \left\{\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} \mid \pi(a|o) = 0 \text{ if and only if } (a,o) \in I\right\}.$$

- *Then, the number of critical points of R_γ^μ on $\operatorname{int}(F)$ is at most*

$$\left(\prod_{o\in O} d_o^{k_o}\right) \cdot \sum_{\sum_{o\in O} i_o = I} \prod_{o\in O} (d_o - 1)^{i_o}, \quad \text{where } I = |\mathcal{S}|(|\mathcal{A}| - 1) - |I|,$$

$O := \{o \in \mathcal{O} \mid (a,o) \in I \text{ f.s. } a\}$, $k_o := |\{a \mid (a,o) \in I\}|$, $d_o := |\{s \mid \beta_{os}^{-1} \neq 0\}|$.

- *If α and μ are generic, this bound can be refined by computing the polar degrees of multi-homogeneous varieties.*
- *A similar bound holds in the mean reward case $\gamma = 1$.*

1. The Parametrization
2. The Feasible Set
3. Number and Location of Critical Points
4. A few Experiments

ROSA

Algorithm 1 **R**eward **O**ptimization in **S**tate-**A**ction space (ROSA)

Require: $\alpha \in \Delta_{\mathcal{S}}^{\mathcal{S}\times\mathcal{A}}$, $\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}}$, $\gamma \in (0,1)$, $\mu \in \Delta_{\mathcal{S}}$

for all $s \in \mathcal{S}$ **do**
  $\ell_s(\eta) \leftarrow \langle \delta_s \otimes \mathbb{1}_{\mathcal{A}} - \gamma\alpha(s|\cdot,\cdot), \eta\rangle_{\mathcal{S}\times\mathcal{A}} - (1-\gamma)\mu_s$ ▷ Define the linear equalities
end for
Compute the defining linear inequalities of $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ ▷ In closed form or algorithmically
Compute the defining polynomial inequalities $p_i(\eta) \geq 0$ of $\mathcal{N}_\beta$ ▷ Can be done according to (3)
$\eta^* \leftarrow \arg\max\langle r, \eta\rangle$ sbj to $\eta \geq 0$, $\ell_s(\eta) = 0$, $p_i(\eta) \geq 0$ ▷ Solve the constrained maximization problem
$R^* \leftarrow \langle r, \eta^*\rangle$ ▷ Evaluate the optimal value
$\tau^* \leftarrow \eta^*(\cdot|\cdot) \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ ▷ Compute an optimal state policy
$\pi^* \leftarrow$ solution of $\beta\pi = \tau^*$ ▷ Compute an optimal observation policy
  return η^*, R^*, π^* ▷ maximizer, optimal value, optimal policy

We use `Ipopt` to solve the constrained problem.

[Müller and Montúfar, 2022b]

Motivation

Policy gradients

- In MDPs, iteration complexity of policy gradient is $O((1-\gamma)^{-\kappa})$, where $\kappa \in \mathbb{N}$ depends on the specific method [Cen et al., 2021].
- Lipschitz constant of $R(\pi)$ is $O((1-\gamma)^{-1})$ [Pirotta et al., 2015].
- Increasingly ill-conditioned problems as $\gamma \to 1$ [Wagner, 2011].
- However, $\gamma \approx 1$ is required to prevent vanishing policy gradients in sparse reward MDPs, where gradients can be as small as $O(2^{-|\mathcal{S}|/2})$ if $\gamma \leq |\mathcal{S}|/(|\mathcal{S}|+1)$ [Agarwal et al., 2021].

Motivation

Optimization in state-action space

- Natural policy gradients or trust region policy optimization; costly.
- Alternatively, solve over state-action frequencies.
  - For MDPs this becomes a linear program [Derman, 1970, Kallenberg, 1994]. This yields a strongly polynomial algorithmic approach, i.e., does not degrade for $\gamma \to 1$ [Post and Ye, 2015].
  - For POMDPs we can use polynomially constrained linear.
- Our experiments show this can yield significant computational savings while remaining numerically stable across values of γ.

Navigation in mazes

- Generate random 2D connected mazes.
- Randomly select a goal state with reward $|\mathcal{S}|$, from which agent transitions to a uniform state.
- For all other states, actions move the agent right, left, up or down.
- Can only observe the 8 neighboring cells, start at a uniform position.

Baselines

Direct Policy Optimization (DPO)

Tabular softmax policies and optimize the parameters for exact discounted reward. We use L-BFGS, which was faster than naive policy gradient.

Bellman Constrained Programming (BCP)

Reformulation of the reward maximization as a quadratically constrained linear program [Amato et al., 2006] (constraint is on the value function)

$$\text{maximize } \langle \mu, v \rangle \quad \text{subject to } \pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} \text{ and } v = \gamma p_\pi v + (1-\gamma) r_\pi, \tag{4}$$

where $p_\pi(s'|s) := \sum_{o,a} \beta(o|s)\pi(a|o)\alpha(s'|s,a)$. We use `Ipopt` to solve (4).

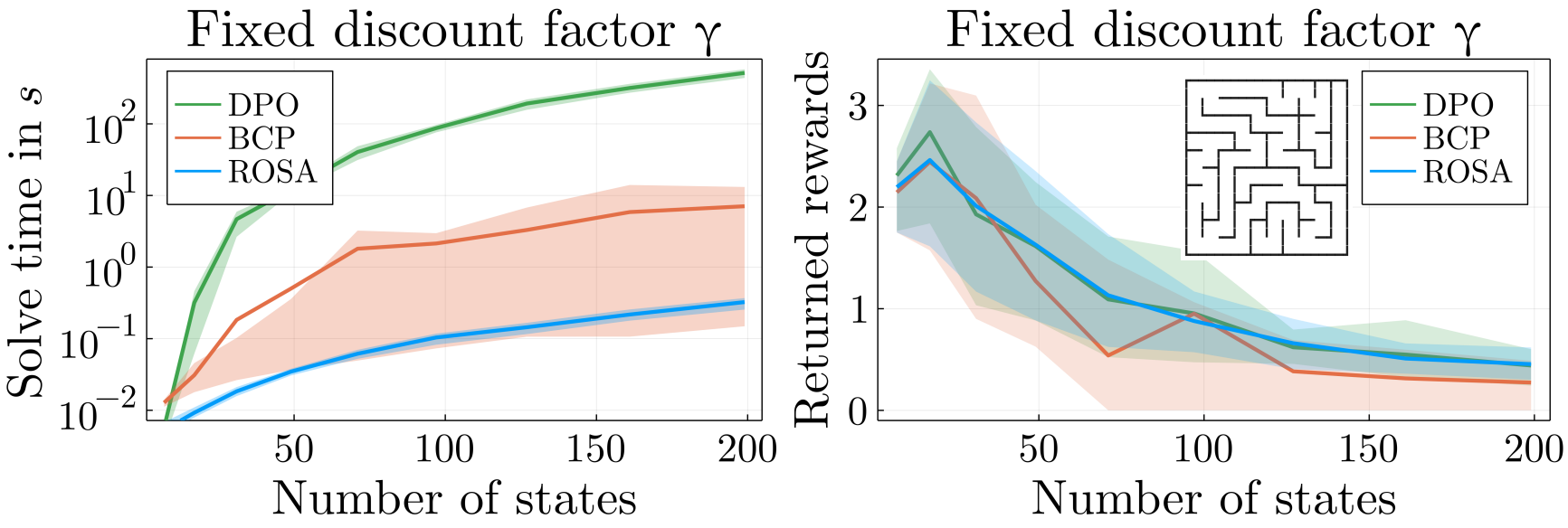


Figure 1: ROSA reaches higher reward while taking less time than other methods.

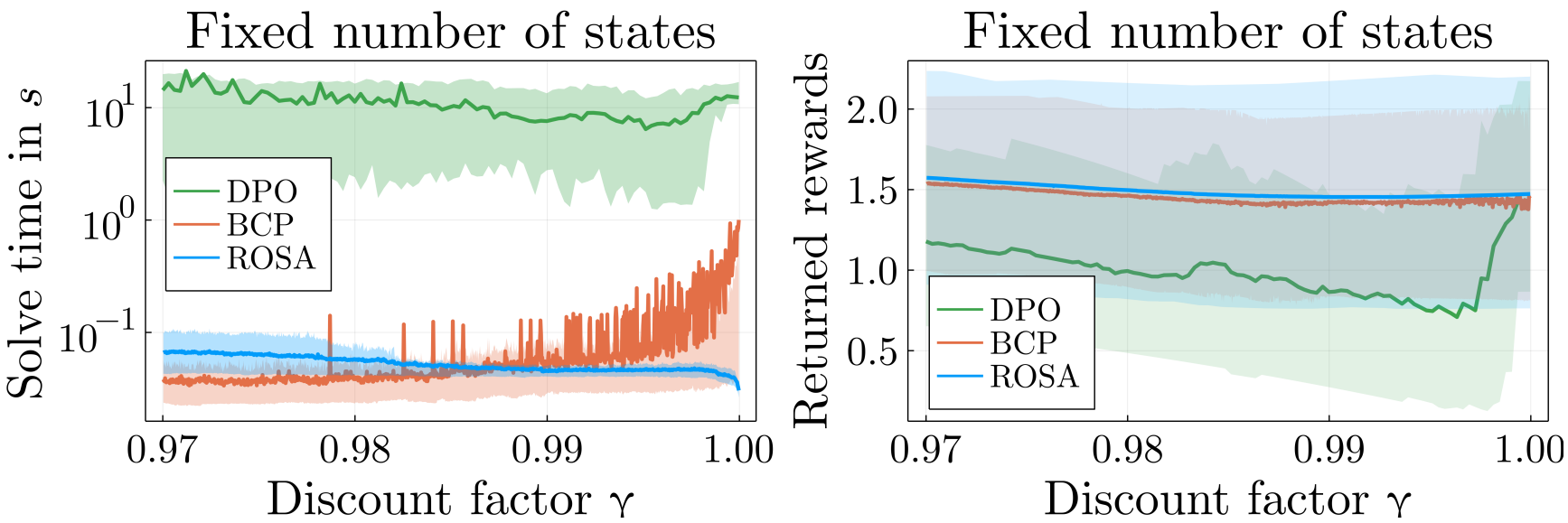


Figure 2: Stability and time improvements from ROSA increase for larger γ.

Conclusion

- Discounted state-action frequencies and expected reward are rational functions of the policy; degree depends on observability
- Feasible state-action frequencies as a basic semialgebraic set
- Polynomial programming formulation of POMDPs
- Bound the number of critical points of the reward
- Implemented and tested algorithms show possible benefits

Future work

- Continuous spaces
- Policies with memory, particularly finite external memory
- Approximate kernels [Rauh et al., 2021] and RL
- Exploit special multi-homogeneous structure of the problem
- Refine policy classes guaranteed to contain maximizers of the reward
- Polynomial optimization methods (with Müller, Rose, Dressler, Garrote Lopez)
- Convergence of natural policy gradients (with Müller)

References I

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. (2021).
On the theory of policy gradient methods: Optimality, approximation, and distribution shift.
Journal of Machine Learning Research, 22(98):1–76.

Amato, C., Bernstein, D. S., and Zilberstein, S. (2006).
Solving pomdps using quadratically constrained linear programs.
In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 341–343.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2021).
Fast global convergence of natural policy gradient methods with entropy regularization.
Operations Research.

References II

Dadashi, R., Taiga, A. A., Le Roux, N., Schuurmans, D., and Bellemare, M. G. (2019).
The value function polytope in reinforcement learning.
In *International Conference on Machine Learning*, pages 1486–1495. PMLR.

Derman, C. (1970).
Finite state Markovian decision processes.
Academic Press.

Kallenberg, L. C. (1994).
Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory.
Zeitschrift für Operations Research, 40(1):1–42.

References III

Montúfar, G., Ghazi-Zahedi, K., and Ay, N. (2015).
Geometry and determinism of optimal stationary control in partially observable Markov decision processes.
arXiv:1503.07206.

Montúfar, G. and Rauh, J. (2017).
Geometry of Policy Improvement.
In *International Conference on Geometric Science of Information*, pages 282–290. Springer.

Montúfar, G., Rauh, J., and Ay, N. (2019).
Task-agnostic constraining in average reward POMDPs.
In *Task-agnostic reinforcement learning Workshop at ICLR 2019*.

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

References IV

Müller, J. and Montúfar, G. (2022a).
The geometry of memoryless stochastic policy optimization in infinite-horizon POMDPs.
In *International Conference on Learning Representations*.

Müller, J. and Montúfar, G. (2022b).
Solving infinite-horizon pomdps with memoryless stochastic policies in state-action space.
In *RLDM 22*.

Nie, J. and Ranestad, K. (2009).
Algebraic Degree of Polynomial Optimization.
SIAM Journal on Optimization, 20(1):485–502.

Pirotta, M., Restelli, M., and Bascetta, L. (2015).
Policy gradient in Lipschitz Markov decision processes.
Machine Learning, 100(2):255–283.

References V

Post, I. and Ye, Y. (2015).
The simplex method is strongly polynomial for deterministic Markov decision processes.
Mathematics of Operations Research, 40(4):859–868.

Rauh, J., Ay, N., and Montúfar, G. (2021).
A continuity result for optimal memoryless planning in pomdps.

Wagner, P. (2011).
A reinterpretation of the policy oscillation phenomenon in approximate policy iteration.
Advances in Neural Information Processing Systems, 24.

POMDP state-action frequencies

Proposition 9 (Correspondence of inequalities in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ and $\Delta_{\mathcal{S}\times\mathcal{A}}$)
Let $\gamma \in (0,1]$, $c \in \mathbb{R}$, $b \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$, $S := \{s \in \mathcal{S} \mid b_{sa} \neq 0 \text{ f.s. } a \in \mathcal{A}\}$. Then

$$\sum_{s,a} b_{sa}\tau_{sa} \geq c \;\; \textit{implies} \;\; \sum_{s\in S}\sum_{a} b_{sa}\eta_{sa} \prod_{s'\in S\setminus\{s\}} \sum_{a'} \eta_{s'a'} - c \prod_{s'\in S}\sum_{a'} \eta_{s'a'} \geq 0.$$

Multi-homogeneous polynom in $(\eta_{sa})_{a\in\mathcal{A}} \in \mathbb{R}^{\mathcal{A}}$ of multi-degree $\mathbb{1}_S \in \mathbb{N}^{\mathcal{S}}$.

If further $\rho_\gamma^{\pi,\mu}(s) > 0$, $s \in \mathcal{S}$, for all π, the inverse implication also holds.

A polynomial $p\colon \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k} \to \mathbb{R}$ is *multi-homogeneous* of *multi-degree* $(d_1, \ldots, d_k) \in \mathbb{N}^k$ if it is homogeneous of degree d_j in the jth variable block, $j = 1, \ldots, k$.

By [Montúfar and Rauh, 2017], we may restrict attention to faces with $k = \sum_{o \in \mathcal{O}} k_o$ active inequalities, where $k_o = \max\{|\mathcal{A}| - l_o, 0\}$.

Exploiting structure of polynomials allows for further improvements:

Example 10 (Number of critical points in a blind controller)

For blind controller with two actions, i.e., $\mathcal{O} = \{o\}$ and $\mathcal{A} = \{a_1, a_2\}$, R_γ^μ has at most $|\mathcal{S}|$ critical points on $\mathrm{int}(\Delta_{\mathcal{A}}^{\mathcal{O}}) \cong (0, 1)$.

Overview

	(In)equalities of state policies	(In)equalities of state-action frequencies
MDPs	$\Delta_{\mathcal{A}}^{\mathcal{S}}$ is described by	$\mathcal{N}_{\gamma}^{\mu}$ is described by
	$\tau(a\|s) \geq 0$ Row normalization: $\sum_a \tau(a\|s) - 1 = 0$ – –	$\eta(s,a) \geq 0$ – Discounted stationarity: $\langle w_{\gamma}^{s}, \eta \rangle - (1-\gamma)\mu(s) = 0$ For $\gamma = 1$: $\sum_{s,a} \eta_{sa} - 1 = 0$
POMDPs	$\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ is described in $\Delta_{\mathcal{A}}^{\mathcal{S}}$ by	$\mathcal{N}_{\gamma}^{\mu,\beta}$ is described in $\mathcal{N}_{\gamma}^{\mu}$ by
	Linear (in)equalities See Section 4 Closed form under Assumption 10: See Theorem 12 Closed form for deterministic observ.: See Remark 57	Polynomial (in)equalities See Section 4, Proposition 14 Closed form under Assumption 10: See Remark 18 for inequalities See Remark 56 for equalities Closed form for deterministic observ.: See Remark 57

Observations

$$\beta \in \Delta_{\mathcal{O}}^{\mathcal{S}} \qquad \beta = \begin{array}{c c} & \begin{array}{cc} o_1 & o_2 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \end{array} & \begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \end{bmatrix} \end{array}$$

Policy

memoryless stochastic action selection mechanism based on observations

$$\pi \in \Delta_{\mathcal{A}}^{\mathcal{O}} \qquad \pi = \begin{array}{c c} & \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{array}{c} o_1 \\ o_2 \end{array} & \begin{bmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \end{bmatrix} \end{array}$$

Effective policy

based on states

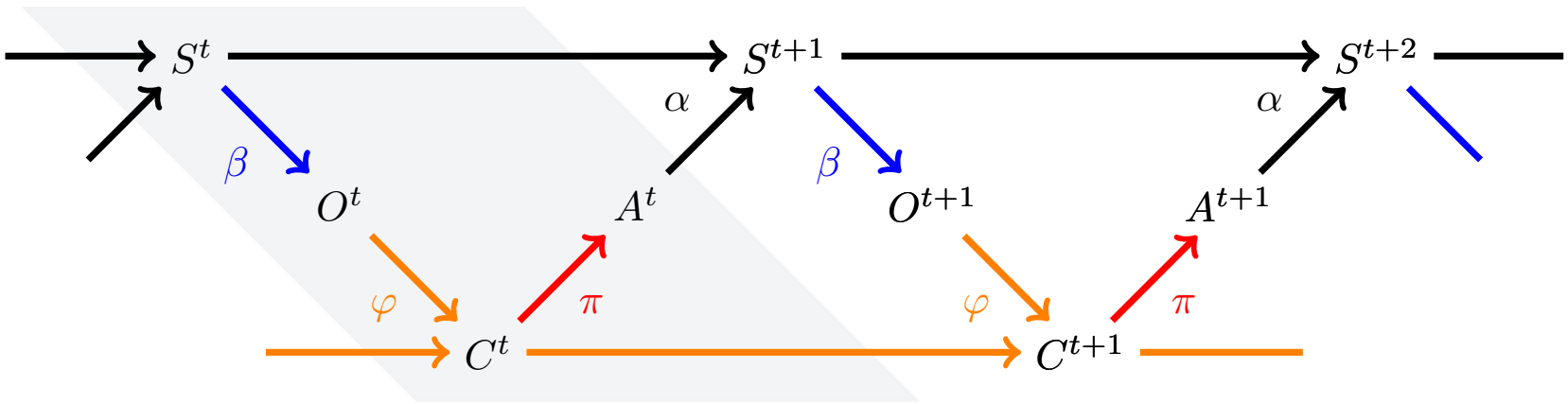
$$\tau = \pi \circ \beta \in \Delta_{\mathcal{A}}^{\mathcal{S}} \qquad \tau = \underbrace{\begin{bmatrix} 1 & 0 \\ 0.5 & 0.5 \end{bmatrix}}_{\beta} \cdot \underbrace{\begin{bmatrix} 0.1 & 0.9 \\ 0.6 & 0.4 \end{bmatrix}}_{\pi} = \begin{array}{c c} & \begin{array}{cc} a_1 & a_2 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \end{array} & \begin{bmatrix} 0.1 & 0.9 \\ 0.35 & 0.65 \end{bmatrix} \end{array}$$

Example 11 (Blind controller)

- For a blind controller the linear equalities defining $\Delta_{\mathcal{A}}^{\mathcal{S},\beta}$ are $\tau_{s_1 a} - \tau_{s_2 a} = \tau(a|s_1) - \tau(a|s_2) = 0$ for all $a \in \mathcal{A}, s_1, s_2 \in \mathcal{S}$.
- These translate to polynomial equalities $\eta_{s_1 a}\rho_{s_2} - \eta_{s_2 a}\rho_{s_1} = 0$.
- If $\mathcal{A} = \{a_1, a_2\}$, we obtain

$$0 = \eta_{s_1 a_1}\eta_{s_2 a_2} - \eta_{s_1 a_2}\eta_{s_2 a_1} \quad \text{for all } s_1, s_2 \in \mathcal{S},$$

which is precisely the condition that all 2×2 minors of η vanish. Hence, $\mathcal{N}_{\gamma}^{\mu,\beta}$ is the set of rank one matrices in $\mathcal{N}_{\gamma}^{\mu}$.



Partially Observable Markov Decision Process with Memory