Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Geometry and Determinism of Optimal Stationary Control in Partially Observable Markov Decision Processes

by

Guido Montúfar, Keyan Ghazi-Zahedi, and Nihat Ay

Preprint no.: 22 2016

Geometry and Determinism of Optimal Stationary Control in Partially Observable Markov Decision Processes

Guido Montúfar MONTUFAR@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Keyan Ghazi-Zahedi ZAHEDI@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Nihat Ay NAY@MIS.MPG.DE
Max Planck Institute for Mathematics in the Sciences
04103 Leipzig, Germany

Faculty of Mathematics and Computer Science
Leipzig University
04009 Leipzig, Germany

Santa Fe Institute
Santa Fe, NM 87501, USA

Abstract

It is well known that for any finite state Markov decision process (MDP) there is a memoryless deterministic policy that maximizes the expected reward. For partially observable Markov decision processes (POMDPs), optimal memoryless policies are generally stochastic. We study the expected reward optimization problem over the set of memoryless stochastic policies. We formulate this as a constrained linear optimization problem and develop a corresponding geometric framework. We show that any POMDP has an optimal memoryless policy of limited stochasticity, which allows us to reduce the dimensionality of the search space. Experiments demonstrate that this approach enables better and faster convergence of the policy gradient on the evaluated systems.

Keywords: MDP, POMDP, partial observability, memoryless stochastic policy, average reward, policy gradient, reinforcement learning

1. Introduction

The field of reinforcement learning addresses a broad class of problems where an agent has to learn how to act in order to maximize some form of cumulative reward. On choosing action a at some world state w the world undergoes a transition to state w' with probability $\alpha(w'|w,a)$ and the agent receives a reward signal $R(w,a,w')$. A policy is a rule for selecting actions based on the information that is available to the agent at each time step. In the simplest case, the Markov decision process (MDP), the full world state is available to the agent at each time step. A key result in this context shows the existence of optimal policies which are memoryless and deterministic (see Ross, 1983). In other words, the agent performs optimally by choosing one specific action at each time step based on the current world state. The agent does not need to take the history of world states into account, nor does he need to randomize his actions.

In many cases one has to assume that the agent experiences the world only through noisy sensors and the agent has to choose actions based only the partial information provided by these sensors. More precisely, if the world state is w, the agent only observes a sensor state s with probability $\beta(s|w)$. This setting is known as partially observable Markov decision process (POMDP). Policy optimization for POMDPs has been discussed by several authors (see Sondik, 1978; Chrisman, 1992; Littman et al., 1995; McCallum, 1996; Parr and Russell, 1995). Optimal policies generally need to take the history of sensor states into account. This requires that the agent be equipped with a memory that stores the sensor history or an encoding thereof (e.g., a properly updated belief state) which may require additional computation.

Although in principle possible, in practice it is often too expensive to find or even to store and execute completely general optimal policies. Some form of representation or approximation is needed. In particular, in the context of embodied artificial intelligence and systems design (Pfeifer and Bongard, 2006) the on-board computation sets limits to the complexity of the controller with respect to both, memory and computational cost. We are interested in policies with limited memory (see, e.g., Hansen, 1998). In fact we will focus on memoryless stochastic policies (see Singh et al., 1994; Jaakkola et al., 1995). Memoryless policies may be worse than policies with memory, but they require far fewer parameters and computation. Among other approaches, the GPOMDP algorithm (Baxter and Bartlett, 2001) provides a gradient based method to optimize the expected reward over parametric models of memoryless stochastic policies. For interesting systems, the set of all memoryless stochastic policies can still be very high dimensional and it is important to find good models. In this article we show that each POMDP has an optimal memoryless policy of limited stochasticity, which allows us to construct low-dimensional differentiable policy models with optimality guarantees. The amount of stochasticity can be bounded in terms of the amount of perceptual aliasing, independently of the specific form of the reward signal.

We follow a geometric approach to memoryless policy optimization for POMDPs. The key idea is that the objective function (the expected reward per time step) can be regarded as a linear function over the set of stationary joint distributions over world states and actions. For MDPs this set is a convex polytope and, in turn, there always exists an optimizer which is an extreme point. The extreme points correspond to deterministic policies (which cannot be written as convex combinations of other policies). For POMDPs this set is in general not convex, but it can be decomposed into convex pieces. There exists an optimizer which is an extreme point of one of these pieces. Depending on the dimension of the convex pieces, the optimizer is more or less stochastic.

This paper is organized as follows. In Section 2 we review basics on POMDPs. In Section 3 we discuss the reward optimization problem in POMDPs as a constrained linear optimization problem with two types of constraints. The first constraint is about the types of policies that can be represented in the underlying MDP. The second constraint relates policies with stationary world state distributions. We discuss the details of these constraints in Sections A and B. In Section 4 we use these geometric descriptions to show that any POMDP has an optimal stationary policy of limited stochasticity. In Section 5 we apply the stochasticity bound to define low dimensional policy models with optimality guarantees. In Section 6 we present experiments which demonstrate the usefulness of the proposed models. In Section 7 we offer our conclusions.

2. Partially observable Markov decision processes

A discrete time partially observable Markov decision process (POMDP) is defined by a tuple $(W, S, A, \alpha, \beta, R)$, where W is a finite set of world states, S is a finite set of sensor states, A is a finite set of actions, $\beta\colon W \to \Delta_S$ is a Markov kernel that describes sensor state probabilities given the world state, $\alpha\colon W \times A \to \Delta_W$ is a Markov kernel that describes the probability of transitioning to a world state given the current world state and action, $R\colon W \times A \to \mathbb{R}$ is a reward signal. A Markov decision process (MDP) is the special case where $W = S$ and β is the identity map.

A policy π is a mechanism for selecting actions. In general, at each time step $t \in \mathbb{N}$, a policy is defined by a Markov kernel π_t taking the history $h_t = (s_0, a_0, \ldots, s_t)$ of sensor states and actions to a probability distribution $\pi_t(\cdot|h_t)$ over A. A policy is *deterministic* when at each time step each possible history leads to a single positive probability action. A policy is *memoryless* when the distribution over actions only depends on the current sensor state, $\pi_t(\cdot|h_t) = \pi_t(\cdot|s_t)$. A policy is *stationary* (homogeneous) when it is memoryless and time independent, $\pi_t(\cdot|h_t) = \pi(\cdot|s_t)$ for all t. Stationary policies are represented by kernels of the form $\pi\colon S \to \Delta_A$. We denote the set of all such policies by $\Delta_{S,A}$.

The goal is to find a policy that maximizes some form of expected reward. We consider the long term expected reward per time step (also called average reward)

$$\mathcal{R}_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}_{\Pr\left\{(w_t,a_t)_{t=0}^{T-1}\middle|\pi,\mu\right\}} \left[\frac{1}{T}\sum_{t=0}^{T-1} R(w_t, a_t)\right]. \tag{1}$$

Here $\Pr\left\{(w_t, a_t)_{t=0}^{T-1}\middle|\pi, \mu\right\}$ is the probability of the sequence $w_0, a_0, w_1, a_1, \ldots, w_{T-1}, a_{T-1}$, given that w_0 is distributed according to the start distribution $\mu \in \Delta_W$ and at each time step actions are selected according to the policy π. Another option is to consider a discount factor $\gamma \in (0, 1)$ and the discounted long term expected reward

$$\mathcal{R}^\gamma_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}_{\Pr\left\{(w_t,a_t)_{t=0}^{T-1}\middle|\pi,\mu\right\}} \left[\sum_{t=0}^{T-1} \gamma^t R(w_t, a_t)\right]. \tag{2}$$

In the case of an MDP, it is always possible to find an optimal memoryless deterministic policy. In other words, there is a policy that chooses an action deterministically at each time step, depending only on the current world state, which achieves the same or higher long term expected reward as any other policy. This fact can be regarded as a consequence of the policy improvement theorem (Bellman, 1957; Howard, 1960).

In the case of a POMDP, policies with memory may perform much better than the memoryless policies. Furthermore, within the memoryless policies, stochastic policies may perform much better than the deterministic ones (see Singh et al., 1994). The intuitive reason is simple: Several world states may produce the same sensor state with positive probability (perceptual aliasing). On the basis of such a sensor state alone, the agent cannot discriminate the underlying world state with certainty. On different world states the same action may lead to drastically different outcomes. Sometimes the agent is forced to choose probabilistically between the optimal actions for the possibly underlying world states (see Example 2). Sometimes he is forced to choose suboptimal actions in order to minimize the risk of catastrophic outcomes (see Example 1). On the other hand, the sequence of previous sensor states may help the agent identify the current world state and choose one single

optimal action. This illustrates why in POMDPs optimal policies may need to take the entire history of sensor states into account and also why the optimal memoryless policies may require stochastic action choices.

The set of policies that take the histories of sensor states and actions into account grows extremely fast. A common approach is to transform the POMDP into a belief-state MDP, where the discrete sensor state is replaced by a continuous Bayesian belief about the current world state. Such belief states encode the history of sensor states and allow for representations of optimal policies. However, belief states are associated with costly internal computations from the side of the acting agent. We are interested in agents subject to perceptual, computational, and storage limitations. Here we investigate stationary policies.

We assume that for each stationary policy $\pi \in \Delta_{S,A}$ there is exactly one stationary world state distribution $p^\pi(w) \in \Delta_W$ and that it is attained in the limit of infinite time when running policy π, irrespective of the starting distribution μ. This is a standard assumption that holds true, for instance, whenever the transition kernel α is strictly positive. In this case (1) can be written as

$$\mathcal{R}(\pi) = \sum_w p^\pi(w) \sum_a p^\pi(a|w) R(w,a), \tag{3}$$

where $p^\pi(a|w) = \sum_s \pi(a|s)\beta(s|w)$. An optimal stationary policy is a policy $\pi^* \in \Delta_{S,A}$ with $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$. Note that maximizing (3) over $\Delta_{S,A}$ is the same as maximizing the discounted expected reward (2) over $\Delta_{S,A}$ with $\mu(w) = p^\pi(w)$ (see Singh et al., 1994). The expected reward per time step appears more natural for POMDPs than the discounted expected reward, because, assuming ergodicity, it is independent of the starting distribution, which is not directly accessible to the agent. Our discussion focusses on average rewards, but our main Theorem 7 also covers discounted rewards.

Our analysis is motivated by the following natural question: Given that every MDP has a stationary deterministic optimal policy, does every POMDP have an optimal stationary policy with small stochasticity? Bounding the required amount of stochasticity for a class of POMDPs would allow us to define a policy model $\mathcal{M} \subseteq \Delta_{S,A}$ with

$$\max_{\pi \in \Delta_{S,A}} \mathcal{R}(\pi) = \max_{\pi \in \mathcal{M}} \mathcal{R}(\pi), \tag{4}$$

for every POMDP from that class. We will show that such a model $\mathcal{M}$ can be defined in terms of the number of ambiguous sensor states and actions, such that $\mathcal{M}$ contains optimal stationary policies for all POMDPs with that number of actions and ambiguous sensor states. Depending on this number, $\mathcal{M}$ can be much smaller in dimension than the set of all stationary policies.

The following examples illustrate some cases where optimal stationary control requires stochasticity and some of the intricacies involved in upper bounding the necessary amount of stochasticity.

Example 1. Consider a system with $W = \{1, \dots, n\}$, $S = \{1\}$, and $A = \{1, \dots, n\}$. The reward function $R(w, a)$ is $+1$ on $a = w$ and -1 otherwise. The agent starts at some random state. On state $w = i$ action $a = i$ takes the agent to some random state and all other actions leave the state unchanged. In this case the best stationary policy chooses actions uniformly at random.

Example 2. Consider the grid world illustrated in Figure 1a. The agent has four possible actions, *north*, *east*, *south*, and *west*, which are effective when there is no wall in that direction. On reaching cells 5, 11, and 13 the agent is teleported to cell 1. On 13 he receives a reward of one and otherwise

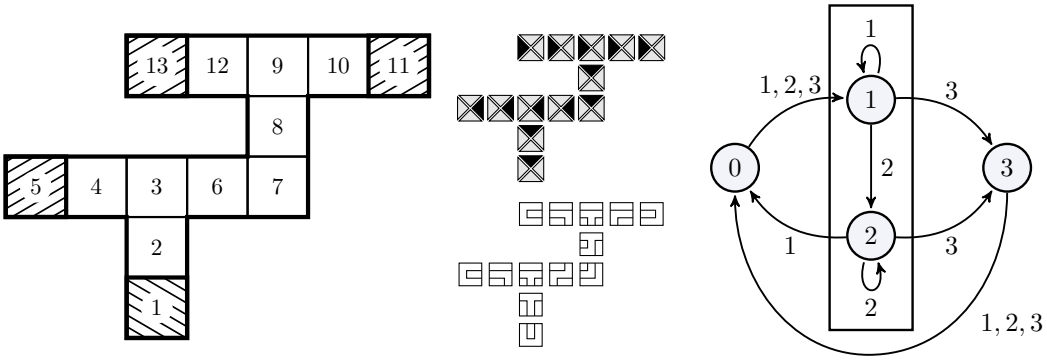

Figure 1: (a) Illustration of the maze Example 2. The left part shows the configuration of world states. The upper right shows an optimal deterministic policy in the MDP setting. At each state, the policy action is in the black direction. The lower right shows the sensor states as observed by the agent in each world state. (b) State transitions from Example 3.

none. In an MDP setting, the agent knows its absolute position in the maze. A deterministic policy can be easily constructed that leads to a maximal reward, as depicted in the upper right. In a POMDP setting the agent may only sense the configuration of its immediate surrounding, as depicted in the lower right. In this case any memoryless deterministic policy fails. Cells 3 and 9 look the same to the agent. Always choosing the same action on this sensation will cause the agent to loop around never reaching the reward cell 13. Optimally, the agent should choose probabilistically between *east* and *west*. The reader might want to have a look at the experiments treating this example in Section 6.

Example 3. Consider the system illustrated in Figure 1b. Each node corresponds to a world state $W = \{0, 1, 2, 3\}$. The sensor states are $S = \{0, 1, 3\}$, whereby $1, 2$ are sensed as 1. The actions are $A = \{1, 2, 3\}$. Choosing action 1 in state 1 and action 2 in state 2 has a large negative reward. Choosing action 2 in state 1 and action 1 in state 2 has a large positive reward. Choosing action 3 in $1, 2$ has a moderate negative reward and takes the agent to state 3. From state 3 each action has a large positive reward and takes the agent to 0. From state 0 any action takes the agent to 1 or 2 with equal probability. In an MDP setting the optimal policy will choose action 2 on 1 and action 1 on 2. In a POMDP setting the optimal policy chooses action 3 on 1. This shows that the optimal actions in a POMDP do not necessarily correspond to the optimal actions in the underlying MDP. Similar examples can be constructed where on a given sensor state it may be necessary to choose from a large set of actions at random, larger than the set of actions that would be chosen on all possibly underlying world states, were they directly observed.

3. Average reward maximization as a constrained linear optimization problem

The expression $\sum_w p(w) \sum_a p(a|w) R(w, a)$ that appears in the expected reward (3) is linear in the joint distribution $p(w, a) = p(w)p(a|w) \in \Delta_{W \times A}$. We want to exploit this linearity. The difficulty is that the optimization problem is with respect to the policy π, not the joint distribution,

and the stationary world state distribution $p^\pi(w)$ depends on the policy. This implies that not all joint distributions $p(w,a)$ are feasible. The feasible set is delimited by the following two conditions.

- Representability in terms of the policy:

$$p(a|w) = \sum_s \pi(a|s)\beta(s|w), \quad \text{for some } \pi \in \Delta_{S,A}. \tag{5}$$

  The geometric interpretation is that the conditional distribution $p(a|w)$ belongs to the polytope $G \subseteq \Delta_{W,A}$ defined as the image of $\Delta_{S,A}$ by the linear map

$$f_\beta \colon \pi(a|s) \mapsto \sum_s \pi(a|s)\beta(s|w). \tag{6}$$

  In turn, the joint distribution $p(w,a)$ belongs to the set $F \subseteq \Delta_{W\times A}$ of joint distributions with conditionals $p(a|w)$ from the set G. In general the set F is not convex, although it is convex in the marginals $p(w)$ when fixing the conditionals $p(a|w)$, and vice versa. We discuss the details of this constraint in Section A.

- Stationarity of the world state distribution:

$$\sum_a p(w,a)\alpha(w'|w,a) \in \Xi, \tag{7}$$

  where $\Xi \subseteq \Delta_{W\times W}$ is the polytope of distributions $p(w,w')$ with equal first and second marginals, $\sum_w p(w,\cdot) = \sum_{w'} p(\cdot,w')$. This means that $p(w)$ is a stationary distribution of the Markov transition kernel $p(w'|w)$. The geometric interpretation is that $p(w,a)$ belongs to the polytope $J := f_\alpha^{-1}(\Xi) \subseteq \Delta_{W\times A}$ defined as the preimage of Ξ by the linear map

$$f_\alpha \colon p(w,a) \mapsto \sum_a p(w,a)\alpha(w'|w,a). \tag{8}$$

  We discuss the details of this constraint in Section B.

Summarizing, the objective function $\mathcal{R} \colon \pi \mapsto \sum_w p^\pi(w) \sum_a p^\pi(a|w) R(w,a)$ is the restriction of the linear function $p(w,a) \mapsto \sum_{w,a} p(w,a)R(w,a)$ to a feasible domain of the form $F \cap J \subseteq \Delta_{W\times A}$, where F is the set of joint distributions with conditionals from a convex polytope G, and J is a convex polytope. We illustrate these notions in the next example.

Example 4. Consider the system illustrated at the top of Figure 2. There are two world states $W = \{1,2\}$, two sensor states $S = \{1,2\}$, and two possible actions $A = \{1,2\}$. The sensor and transition probabilities are given by

$$\beta = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}, \quad \alpha(\cdot|w=1,\cdot) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \alpha(\cdot|w=2,\cdot) = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}.$$

In the following we discuss the feasible set of joint distributions. The policy polytope $\Delta_{S,A}$ is a square. The set of realizable conditional distributions of world states given actions is the line

$$G = f_\beta(\Delta_{S,A}) = \operatorname{conv}\left\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} \right\}$$

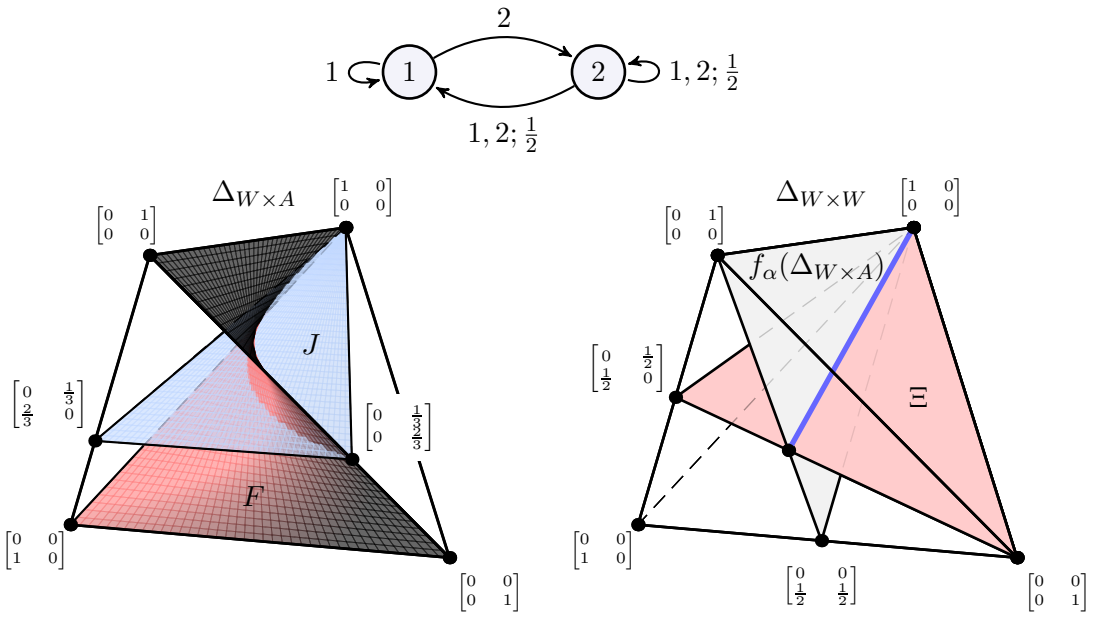

Figure 2: Illustration of Example 4. The world state transitions are shown in the upper part. They are deterministic from $w = 1$ and random from $w = 2$. The lower left shows, inside of $\Delta_{W\times A}$, the set F defined by the representability constraint (5) and the polytope $J = f_\alpha^{-1}(\Xi)$ defined by the stationarity constraint (7). The lower right shows, inside of $\Delta_{W\times W}$, the polytopes $f_\alpha(\Delta_{W\times A})$ and Ξ.

inside of the square $\Delta_{W,A}$. The set F of joint distributions with conditionals from G is a twisted surface. This set has one copy of G for every world state distribution $p(w)$. See the lower left of Figure 2. The set J of joint distributions over world states and actions that satisfy the stationarity constraint (7) is the subset of $\Delta_{W\times A}$ that f_α maps to the polytope Ξ shown in the lower right of Figure 2. This is the triangle

$$J = f_\alpha^{-1}(\Xi) = \operatorname{conv}\left\{\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}, \begin{bmatrix}0 & 1/3\\ 0 & 2/3\end{bmatrix}, \begin{bmatrix}0 & 1/3\\ 2/3 & 0\end{bmatrix}\right\}.$$

As we will show in Lemma 6, the extreme points of J can always be written in terms of extreme points of $\Delta_{W,A}$; in the present example, in terms of $\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$ (or $\left[\begin{smallmatrix}1 & 0\\ 1 & 0\end{smallmatrix}\right]$), $\left[\begin{smallmatrix}0 & 1\\ 0 & 1\end{smallmatrix}\right]$, $\left[\begin{smallmatrix}0 & 1\\ 1 & 0\end{smallmatrix}\right]$. The set $F \cap J$ is a curve. This is the feasible domain of the expected reward $\mathcal{R}$, viewed as a function of joint distributions over world states and actions.

4. Determinism of optimal stationary policies

In this section we discuss the minimal stochasticity of optimal stationary policies. In order to illustrate our geometric approach we first consider MDPs and then the more general case of POMDPs.

Theorem 5 (MDPs)**.** *Consider an MDP* (W, A, α, R)*. Then there is a deterministic optimal stationary policy.*

Proof of Theorem 5. The objective function $\mathcal{R}$ defined in Equation (3) can be regarded as the restriction of a linear function over $\Delta_{W\times A}$ to the feasible set J defined in Equation (8). Since J is a convex polytope, the objective function is maximized at one of its extreme points. By Lemma 6, all extreme points of J can be realized by extreme points of $\Delta_{W,A}$, that is, deterministic policies. □

Lemma 6. *Each extreme point of J can be written as $p(w,a) = p(w)p(a|w)$, where $p(w) \in \Delta_W$ and $p(a|w)$ is an extreme point of $\Delta_{W,A}$.*

Proof of Lemma 6. We can view the map f_α from Equation (8) as taking pairs $(p(w), p(a|w))$ to pairs $(p(w), p(w'|w))$. Here the marginal distribution is mapped by the identity function $\Delta_W \to \Delta_W; p(w) \mapsto p(w)$ and the conditional distribution by

$$\tilde{f}_\alpha \colon \Delta_{W,A} \to \Delta_{W,W};\ p(a|w) \mapsto \sum_a p(a|w)\alpha(w'|w,a) = p(w'|w).$$

Consider some $W' \subseteq W$ for which J contains a distribution q whose marginal has support W'. For each $w \in W'$ let $A_w = \{a \in A \colon \operatorname{supp}(\alpha(\cdot|w,a)) \subseteq W'\}$ denote the set of actions with transitions that stay in W'. With a slight abuse of notation let us write $\Delta_{W',A'} := \bigtimes_{w\in W'} \Delta_{A_w} \subseteq \Delta_{W',A}$ and $\Delta_{W'\times A'} := \Delta_{W'} * \Delta_{W',A'} = \{p(w,a) \colon p(w) \in \Delta_{W'}, p(a|w) \in \Delta_{W',A'}\} \subseteq \Delta_{W'\times A}$ for the corresponding sets of conditional and joint probability distributions. Note that out of $\Delta_{W\times A}$ only points from $\Delta_{W'\times A'}$ are mapped to points in $\Delta_{W'\times W'}$ and hence $J \cap \Delta_{W'\times A} \subseteq \Delta_{W'\times A'}$. The set $f_\alpha(\Delta_{W'\times A'})$ consists of all joint distributions $p(w,w') = p(w)p(w'|w)$ with $p(w) \in \Delta_{W'}$ and $p(w'|w) \in \tilde{f}_\alpha(\Delta_{W',A'}) \subseteq \Delta_{W',W'}$. Now, for each conditional $p(w'|w) \in \Delta_{W',W'}$ there is at least one marginal $p(w) \in \Delta_{W'}$ such that the joint $p(w,w') \in \Delta_{W'\times W'}$ is an element of Ξ. Hence

$$\dim(f_\alpha(\Delta_{W'\times A'}) \cap \Xi) \geq \dim(\tilde{f}_\alpha(\Delta_{W',A'})).$$

The set $J \cap \Delta_{W'\times A}$ is the union of the fibers of all points in $f_\alpha(\Delta_{W'\times A'}) \cap \Xi$. Hence

$$\begin{aligned}\dim(J \cap \Delta_{W'\times A}) &\geq \dim(f_\alpha(\Delta_{W'\times A'}) \cap \Xi) + \Big(\dim(\Delta_{W',A'}) - \dim(\tilde{f}_\alpha(\Delta_{W',A'}))\Big)\\ &\geq \dim(\tilde{f}_\alpha(\Delta_{W',A'})) + \dim(\Delta_{W',A'}) - \dim(\tilde{f}_\alpha(\Delta_{W',A'}))\\ &= \dim(\Delta_{W',A'}).\end{aligned}$$

Let us now consider some extreme point q of J. Suppose that the marginal of q has support W'. By the previous discussion, we know that q is an extreme point of the polytope $J \cap \Delta_{W'\times A'}$. Furthermore, $J \cap \Delta_{W'\times A'}$ is the d-dimensional intersection of an affine space and $\Delta_{W'\times A'}$, where $d \geq \dim(\Delta_{W',A'})$. This implies that q lies at the intersection of d facets of $\Delta_{W'\times A'}$. In turn $|\operatorname{supp}(q(w,\cdot))| = 1$, for all $w \in W'$. This shows that $q(w,a) = p(w)p(w,a)$, where $p(w) \in \Delta_{W'}$ and $p(a|w)$ is an extreme point of $\Delta_{W',A'}$. We can extend this conditional arbitrarily on $w \in W \setminus W'$ to obtain a conditional that is an extreme point of $\Delta_{W,A}$. □

Now we discuss the minimal stochasticity of optimal stationary policies for POMDPs. A policy $\pi \in \Delta_{S,A}$ is called *m-stochastic* if it is contained in an m-dimensional face of $\Delta_{S,A}$. This means that at most $|S| + m$ entries $\pi(a|s)$ are non-zero and, in particular, that π is a convex combination of at most $m+1$ deterministic policies. For instance, a deterministic policy is 0-stochastic and has exactly $|S|$ non-zero entries. The following result holds both in the average reward and in the discounted reward settings.

Theorem 7 (POMDPs)**.** *Consider a POMDP* $(W, S, A, \alpha, \beta, R)$*. Let* $U = \{s \in S\colon |\operatorname{supp}(\beta(s|\cdot))| > 1\}$*. Then there is a* $|U|(|A|-1)$*-stochastic optimal stationary policy. Furthermore, for any* W, S, A *there are* α, β, R *such that every optimal stationary policy is at least* $|U|(|A|-1)$*-stochastic.*

Proof of Theorem 7. Here we prove the statement for the average reward setting using the geometric descriptions from Section 3. We cover the discounted setting in Section C using value functions and a policy improvement argument.

Consider the sets $G = f_\beta(\Delta_{S,A}) \subseteq \Delta_{W,A}$ and $F = \Delta_W * G \subseteq \Delta_{W\times A}$ from Equation (6). We can write G as a union of Cartesian products of convex sets, as $G = \bigcup_{\theta\in\Theta} G_\theta$, with $\dim(G) - \dim(G_\theta) \leq \dim(\Theta) = |U|(|A|-1)$. See Proposition 9 for details. In turn, we can write $F = \bigcup_{\theta\in\Theta} F_\theta$, where each $F_\theta = \Delta_W * G_\theta$ is a convex set of dimension $\dim(F_\theta) = \dim(\Delta_W) + \dim(G_\theta)$. See Proposition 12 for details.

The objective function $\mathcal{R}$ is linear over each polytope $F_\theta \cap J$ and is maximized at an extreme point of one of these polytopes. If $F_\theta \cap J \neq \emptyset$, then each extreme point of $F_\theta \cap J$ can be written as $p(w,a) = p(w)p(a|w)$, where $p(a|w)$ is an extreme point of G_θ. To see this, note that the arguments of Lemma 6 still hold when we replace J by $F_\theta \cap J$ and $\Delta_{W,A}$ by G_θ. Each extreme point of G_θ lies at a face of G of dimension at most $|U|(|A|-1)$. See Proposition 9 for details. Now, since f_β is a linear map, the points in the m-dimensional faces of G have preimages by f_β in m-dimensional faces of $\Delta_{S,A}$. Thus, there is a maximizer of $\mathcal{R}$ that is contained in a $|U|(|A|-1)$ face of $\Delta_{S,A}$.

The second statement, regarding the optimality of the stochasticity bound, follows from Proposition 23, which computes the optimal stationary policies of a class of POMDPs analytically. □

Remark 8.

- Our Theorem 7 also has an interpretation for non-ergodic systems: Among all pairs $(\pi, p^\pi(w))$ of stationary policies and associated stationary world state distributions, the highest value of $\sum_w p^\pi(w) \sum_a p^\pi(a|w) R(w,a)$ is attained by a pair where the policy π is $|U|(|A|-1)$-stochastic. However, this optimal stationary average reward is only equal to (1) for start distributions μ that converge to $p^\pi(w)$.
- For MDPs the set U is empty and the statement of Theorem 7 recovers Theorem 5.
- In a reinforcement learning setting the agent does not know anything about the world state transitions α nor the observation model β a priori, beside from the sets S and A. In particular, he does not know the set U (nor its cardinality). Nonetheless, he can build a hypothesis about U on the basis of observed sensor states, actions, and rewards. This can be done using a suitable variant of the Baum-Welch algorithm or inexpensive heuristics, without estimating the full kernels α and β.

5. Application to defining low dimensional policy models

By Theorem 7, there always exists an optimal stationary policy in a $|U|(|A|-1)$-dimensional face of the policy polytope $\Delta_{S,A}$. Instead of optimizing over the entire set $\Delta_{S,A}$, we can optimize over a lower dimensional subset that contains the $|U|(|A|-1)$-dimensional faces. In the following we discuss various ways of defining a differentiable policy model with this property.

We denote the set of policies in m-dimensional faces of the polytope $\Delta_{S,A}$ by

$$C_m := \{\pi \in \Delta_{S,A}\colon \operatorname{supp}(\pi) \leq |S| + m\}.$$

Note that each policy in C_m can be written as the convex combination of $m+1$ or fewer deterministic policies. For example, $C_0 = \{\pi^f(a|s) = \delta_{f(s)}(a) \colon f \in A^S\}$ is the set of deterministic policies, and $C_{|S|(|A|-1)} = \Delta_{S,A}$ is the entire set of stationary policies.

Conditional exponential families An exponential policy family is a set of policies of the form

$$\pi_\theta(a|s) = \frac{\exp(\theta^\top F(s,a))}{\sum_{a'} \exp(\theta^\top F(s,a'))},$$

where $F \colon S \times A \to \mathbb{R}^d$ is a vector of sufficient statistics and $\theta \in \mathbb{R}^d$ is a vector of parameters. We can choose F suitably, such that the closure of the exponential family contains C_m.

The *k-interaction model* is defined by the sufficient statistics

$$F_\lambda(x) = \prod_{i \in \lambda} (-1)^{x_i}, \quad x \in \{0,1\}^n, \quad \lambda \subseteq \{1,\dots,n\}, 1 \leq |\lambda| \leq k.$$

Here we can identify each pair $(s,a) \in S \times A$ with a length-n binary vector $x \in \{0,1\}^n$, $n = \lceil \log_2(|S||A|) \rceil$. Since we do not need to model the marginal distribution over S, we can remove all λ for which $F_\lambda(s,\cdot)$ is constant for all s. The k-interaction model is (2^k-1)-*neighborly* (Kahle, 2010), meaning that, for $2^k - 1 \geq |S| + m$ it contains C_m in its closure. This results in a policy model of dimension at most $\sum_{i=1}^{\lceil \log_2(|S|+m+1) \rceil} \binom{\lceil \log_2(|S||A|) \rceil}{i}$. Note that this is only an upper bound, both on k and the dimension, and usually a smaller model will be sufficient.

An alternative exponential family is defined by taking $F(s,a)$, $(s,a) \in S \times A$, equal to the vertices of a cyclic polytope. The cyclic polytope $C(N,d)$ is the convex hull of $\{x(t_1), \dots, x(t_N)\}$, where $x(t) = [t, t^2, \dots, t^d]^\top$, $t_1 < t_2 < \cdots < t_N$, $N > d \geq 2$. This results in a $\lfloor d/2 \rfloor$-neighborly model. Using this approach yields a policy model of dimension $2(|S| + m)$.

Mixtures of deterministic policies We can consider policy models of the form

$$\pi_\theta(a|s) = \sum_{f \in A^S} \pi^f(a|s) p_\theta(f),$$

where $\pi^f(a|s) = \delta_{f(s)}(a)$ is the deterministic policy defined by the function $f \colon S \to A$ and $p_\theta(f)$ is a model of probability distributions over the set of all such functions. Choosing this as a $(m+1)$-neighborly exponential family yields a policy model which contains C_m and, in fact, all mixtures of $m+1$ deterministic policies. This kind of model was proposed in Ay et al. (2013).

Identifying each $f \in A^S$ with a length-n binary vector, $n \geq \lceil \log_2(|A|^{|S|}) \rceil$, and using a k-interaction model with $2^k - 1 = m + 1$ yields a model of dimension $\sum_{i=1}^{\lceil \log_2(m+2) \rceil} \binom{\lceil \log_2(|A|^{|S|}) \rceil}{i}$.

Alternatively, we can use a cyclic exponential family for p_θ, which yields a policy model of dimension $2(m+1)$. If we are only interested in modeling the deterministic policies, $m = 0$, then this model has dimension two.

Conditional restricted Boltzmann machines A conditional restricted Boltzmann machine (CRBM) is a model of policies of the form

$$\pi_\theta(y|x) = \frac{1}{Z(x)} \sum_{z \in \{0,1\}^{n_{\text{hidden}}}} \exp(z^\top V x + z^\top W y + b^\top y + c^\top z),$$

with parameter $\theta = \{W, V, b, c\}$, $W \in \mathbb{R}^{n_{\text{hidden}} \times n_{\text{out}}}$, $V \in \mathbb{R}^{n_{\text{hidden}} \times n_{\text{in}}}$, $b \in \mathbb{R}^{n_{\text{out}}}$, $c \in \mathbb{R}^{n_{\text{hidden}}}$. Here we identify each $s \in S$ with a vector $x \in \{0,1\}^{n_{\text{in}}}$, $n_{\text{in}} = \lceil \log_2 |S| \rceil$, and each $a \in A$ with a vector $y \in \{0,1\}^{n_{\text{out}}}$, $n_{\text{out}} = \lceil \log_2 |A| \rceil$. There are theoretical results on CRBMs (Montúfar et al., 2015) showing that they can represent every policy from C_m whenever $n_{\text{hidden}} \geq |S| + m - 1$. A sufficient number of parameters is thus $(|S| + m - 1)(\lceil \log_2 |S| \rceil + \lceil \log_2(|A|) \rceil) + \lceil \log_2(|A|) \rceil$.

Each of these models has advantages and disadvantages. The CRBMs can be sampled very efficiently using a Gibbs sampling approach. The mixture models can be very low dimensional, but may have an intricate geometry. The k-interaction models are smooth manifolds.

6. Experiments

We run computer experiments to explore the practical utility of our theoretical results. We consider the maze from Example 2. In this example, the set U of sensor states s with $|\operatorname{supp}(\beta(s|\cdot))| > 1$ has cardinality two. By Theorem 7, there is a $|U|(|A| - 1) = 6$ stochastic optimal stationary policy. As a family of policy models we choose the k-interaction models from Section 5. The number of binary variables is $n = \lceil \log_2(|S||A|) \rceil = 6$. This results in a sufficient statistics matrix with 64 columns, out of which we keep only the first 40, one for each pair (s, a). For $k = 1, \ldots, 5$, the resulting model dimension is $2, 11, 23, 29, 30$. The policy polytope $\Delta_{S,A}$ has dimension $|S|(|A| - 1) = 30$.

We consider the reinforcement learning problem, where the agent does not know W, α, β, R in advance. We use stochastic gradient with an implementation of the GPOMDP algorithm (Baxter and Bartlett, 2001) for estimating the gradient. We fix a constant learning rate of 1, a time window of $T = 1, \ldots, 100$ for each Markov chain gradient and average reward estimation, and perform $10\,000$ gradient iterations on a random parameter initialization.

The results are shown in Figure 3. The first column shows the learning curves for $k = 1, \ldots, 5$, for the first $2\,500$ gradient iterations. Shown is actually the average of the learning curves for 5 repetitions of the experiment. The individual curves are indeed all very similar for each fixed k. The value shown is the estimated average reward, with a running average shown in bold, for better visibility. The second column shows the final policy. The third column gives a detail of the learning curves and shows the reward averaged over the entire learning process.

The independence model, with $k = 1$, performs very poorly, as it learns a fixed distribution of actions for all sensor states. The next model, with $k = 2$, performs better, but still has a very limited expressive power. All the other models have sufficient complexity to learn a (nearly) optimal policy, in principle. However, out of these, the less complex one, with $k = 3$, performs best. This indicates that the least complex model which is able to learn an optimal policy does learn faster. This model has less parameters to explore and is less sensitive to the noise in the stochastic gradient.

7. Conclusions

Policy optimization for partially observable Markov decision processes is a challenging problem. Scaling is a serious difficulty in most algorithms and theoretical results are scarce on approximative methods. This paper develops a geometric view on the problem of finding optimal stationary policies. The maximization of the long term expected reward per time step can be regarded as a constrained linear optimization problem with two constraints. The first one is a quadratic constraint that arises from the partial observability of the world state. The second is a linear constraint that arises from the stationarity of the world state distribution. We can decompose the feasible domain

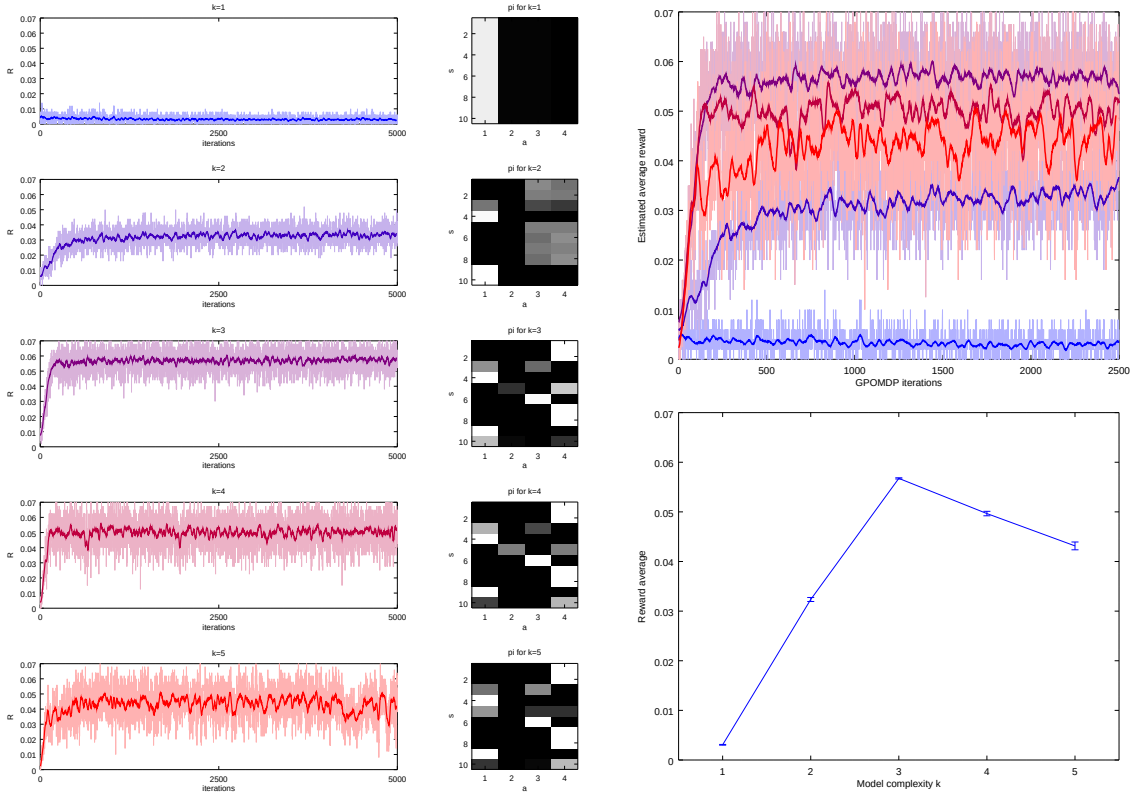

Figure 3: Experimental results on the maze Example 2. The left column shows the average reward learning curves for k-interaction models, with $k = 1, \ldots, 5$ from top to bottom. The second column shows the final policies as matrices of sensor-state action probabilities (white is 1). The right column compares all learning curves and shows the overall average reward for all models. The model with $k = 3$ performs best.

into convex pieces, on which the optimization problem is linear. This analysis sheds light into the complexity of stationary policy optimization for POMDPs and reveals avenues for designing learning algorithms.

We show that every POMDP has an optimal stationary policy of limited stochasticity. The necessary level of stochasticity is bounded above by the number of sensor states that are ambiguous about the underlying world state, independently of the specific reward function. This allows us to define low dimensional models which are guaranteed to contain optimal stationary policies. Our experiments show that the proposed dimensionality reduction does indeed allow to learn better policies faster. Having less parameters, these models are less expensive to train and less sensitive to noise, while at the same time being able to learn best possible stationary policies.

Acknowledgments

We would like to acknowledge support from the DFG Priority Program Autonomous Learning (DFG-SPP 1527).

References

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

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *J. Artif. Int. Res.*, 15(1):319–350, November 2001. URL `http://dl.acm.org/citation.cfm?id=1622845.1622855`.

Richard Bellman. *Dynamic programming*. Princeton University Press, Princeton, NY, 1957.

Lonnie Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *In Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 183–188. AAAI Press, 1992.

Eric Anton Hansen. *Finite-memory Control of Partially Observable Systems*. PhD thesis, 1998.

Ronald A. Howard. *Dynamic Programming and Markov Processes*. The MIT Press, Cambridge, MA, 1960.

Tommi Jaakkola, Satinder P. Singh, and Michael I. Jordan. Reinforcement learning algorithm for partially observable Markov decision problems. In *Advances in Neural Information Processing Systems 7*, pages 345–352. MIT Press, 1995.

Thomas Kahle. Neighborliness of marginal polytopes. *Beiträge zur Algebra und Geometrie*, 51(1): 45–56, 2010. URL `http://eudml.org/doc/224152`.

Michael L. Littman, Anthony R. Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *International Conference on Machine Learning (ICML)*, pages 362–370. Morgan Kaufmann, 1995.

Andrew Kachites McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, 1996.

Guido Montúfar, Nihat Ay, and Keyan Ghazi-Zahedi. Geometry and expressive power of conditional restricted Boltzmann machines. *JMLR*, 16:2405–2436, Dec 2015.

Ronald Parr and Stuart Russell. Approximating optimal policies for partially observable stochastic domains. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, volume 2 of *IJCAI'95*, pages 1088–1094, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

Rolf Pfeifer and Josh C. Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. The MIT Press (Bradford Books), Cambridge, MA, 2006.

Sheldon M. Ross. *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Academic Press, Inc., Orlando, FL, USA, 1983.

Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Learning without state-estimation in partially observable Markovian decision processes. In *ICML*, pages 284–292, 1994.

Edward J. Sondik. The optimal control of partially observable Markov processes over the infinite horizon: Discounted costs. *Operations Research*, 26(2):282–304, 1978. URL `http://www.jstor.org/stable/169635`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

Stephan Weis. *Exponential Families with Incompatible Statistics and Their Entropy Distance*. PhD thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, 2010.

Appendix A. The representability constraint

Here we investigate the set of representable policies in the underlying MDP; that is, the set of kernels of the form $p^\pi(a|w) = \sum_s \beta(s|w)\pi(a|s)$. This set is the image $G = f_\beta(\Delta_{S,A})$ of the linear map

$$f_\beta \colon \Delta_{S,A} \to \Delta_{W,A};\ \pi(a|s) \mapsto \sum_s \beta(s|w)\pi(a|s).$$

We are interested in the properties of this set, depending on the observation kernel β.

Consider first the special case of a deterministic kernel β^b, defined by $\beta^b(s|w) = \delta_{b(w)}(s)$, for some function $b \colon W \to S$. Then

$$f_{\beta^b}(\Delta_{S,A}) = \bigtimes_{s\in S} \operatorname{sym} \Delta_{b^{-1}(s),A},$$

where $b^{-1}(s) \subseteq W$ is the set of world states that b maps to s and $\operatorname{sym} \Delta_{B,A} := \{g \in \Delta_{B,A} \colon g(\cdot|w) = p, p \in \Delta_A\}$ is the set of elements of $\Delta_{B,A}$ that consist of one repeated probability distribution. This set can be written as a union of Cartesian products,

$$f_{\beta^b}(\Delta_{S,A}) = \bigcup_{\theta\in\Delta_{U,A}} \Big[\bigtimes_{s\in U}\Big(\bigtimes_{w\in b^{-1}(s)} \theta(\cdot|s)\Big)\Big] \times \Big[\bigtimes_{s\in S\setminus U} \Delta_A\Big],$$

where $U := \{s \in S \colon |b^{-1}(s)| > 1\}$ is the set of sensor states that can result from several world states. For instance, when β is the identity function we have $G = \Delta_{W,A} = \bigtimes_w \Delta_A$.

Proposition 9. *Consider a measurement $\beta \in \Delta_{W,S}$ and the map $f_\beta \colon \Delta_{S,A} \to \Delta_{W,A};\ \pi(a|s) \mapsto \sum_s \beta(s|w)\pi(a|s)$. Let $U = \{s \in S \colon |\operatorname{supp}(\beta(s|\cdot))| > 1\}$ be the sensor states that can be obtained from several world states. The set $G = f_\beta(\Delta_{S,A})$ can be written as $G = \bigcup_{\theta\in\Theta} G_\theta$, where each G_θ is a Cartesian product of convex sets, $G_\theta = \bigtimes_{w\in W} G_{\theta,w}$, $G_{\theta,w} \subseteq \Delta_A$ convex, and each vertex of G_θ lies in a face of G of dimension at most $|U|(|A|-1)$.*

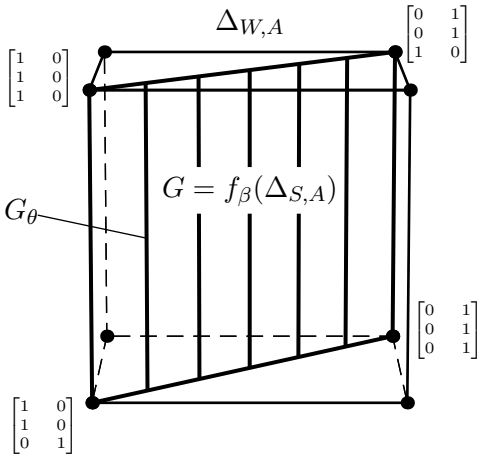

Figure 4: Illustration of Example 10. Shown is a decomposition of $G = f_\beta(\Delta_{S,A}) \subseteq \Delta_{W,A}$ into a collection of Cartesian products of convex sets, G_θ, $\theta \in \Theta$.

Proof of Proposition 9. We use as index set Θ the set of policies $\Delta_{U,A}$. We can write

$$G = \bigcup_{\theta \in \Delta_{U,A}} G_\theta \qquad \text{with} \qquad G_\theta = \bigtimes_{w \in W} \left(\sum_{s \in U} \beta(s|w)\theta(\cdot|s) + \sum_{s \in S \setminus U} \beta(s|w)\Delta_A \right).$$

This proves the first part of the claim. For the second part, note that all G_θ are equal but for addition of a linear projection of $\theta \in \Delta_{U,A}$. □

Example 10. Let $W = \{0, 1, 2\}$, $S = \{0, 1\}$, $A = \{0, 1\}$. Let β map $w = 0$ and $w = 1$ to $s = 0$, and $w = 2$ to $s = 1$, with probability one. Written as a table $(\beta(s|w))_{w,a}$ this is

$$\beta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The policy polytope $\Delta_{S,A}$ is the square with vertices

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The polytope $G = f_\beta(\Delta_{S,A}) \subseteq \Delta_{W,A}$ is the square with vertices

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and can be written as a union of Cartesian products of convex sets, illustrated in Figure 4,

$$G = \bigcup_{\theta \in \Delta_A} G_\theta, \qquad G_\theta = \{\theta\} \times \{\theta\} \times \Delta_A.$$

As mentioned in Section 3, the set $F \subseteq \Delta W \times A$ of joint distributions that are compatible with the representable conditionals $G = f_\beta(\Delta_{S,A}) \subseteq \Delta_{W,A}$, may not be convex. In the following we describe large convex subsets of F, depending on the properties of G. We use the following definitions.

Definition 11.

- Given a set of distributions $\mathcal{P} \subseteq \Delta_W$ and a set of kernels $\mathcal{G} \subseteq \Delta_{W,A}$, let

$$\mathcal{P} * \mathcal{G} := \Big\{ q(w,a) = p(w)g(a|w) \in \Delta_{W\times A} \colon p \in \mathcal{P}, g \in \mathcal{G} \Big\}$$

  denote the set of joint distributions over world states and actions, with world state marginals in $\mathcal{P}$ and conditional distributions in $\mathcal{G}$.

- For any $V \subseteq W$ let

$$\Delta_W(V) := \Big\{ p \in \Delta_W \colon \operatorname{supp}(p) := \{w \in W \colon p(w) > 0\} \subseteq V \Big\}$$

  denote the set of world state distributions with support in V.

- Given a subset $V \subseteq W$ and a set of kernels $\mathcal{G} \subseteq \Delta_{W,A}$, let

$$\mathcal{G}|_V := \Big\{ h \in \Delta_{V,A} \colon h(\cdot|w) = g(\cdot|w) \text{ for all } w \in V, \text{for some } g \in \mathcal{G} \Big\}$$

  denote the set of restrictions of elements of $\mathcal{G}$ to inputs from V.

The following proposition states that a set of Markov kernels which is a Cartesian product of convex sets, with one factor for each input, corresponds to a convex set of joint probability distributions. Furthermore, if the considered input distributions assign zero probability to some of the inputs, then the convex factorization property is only needed for the restriction to the positive-probability inputs.

Proposition 12. *Let $V \subseteq W$. Let $\mathcal{P} \subseteq \Delta_W(V)$ be a convex set. Let $\mathcal{G} \subseteq \Delta_{W,A}$ satisfy $\mathcal{G}|_V = \bigtimes_{w\in V} \mathcal{G}_w \subseteq \Delta_{V,A}$, where $\mathcal{G}_w \subseteq \Delta_A$ is a convex set for all $w \in V$. Then $\mathcal{P} * \mathcal{G} \subseteq \Delta_{W\times A}$ is convex.*

Proof of Proposition 12. We need to show that, given any two distributions q' and q'' in $\mathcal{P} * \mathcal{G}$, and any $\lambda \in [0,1]$, the convex combination $q = \lambda q' + (1-\lambda)q''$ lies in $\mathcal{P} * \mathcal{G}$. This is the case if and only if $q(w,a) = p(w)g(a|w)$ for some $p \in \mathcal{P}$ and some $g \in \Delta_{W,A}$ with $g|_V \in \mathcal{G}|_V$. We have

$$\begin{aligned} q(w,a) &= \lambda q'(w,a) + (1-\lambda)q''(w,a) \\ &= \lambda p'(w)g'(a|w) + (1-\lambda)p''(w)g''(a|w) \\ &= (\lambda p'(w) + (1-\lambda)p''(w)) \\ &\quad \times \Big(\frac{\lambda p'(w)}{\lambda p'(w) + (1-\lambda)p''(w)} g'(a|w) + \frac{(1-\lambda)p''(w)}{\lambda p'(w) + (1-\lambda)p''(w)} g''(a|w) \Big). \end{aligned}$$

This shows that $q(w,a) = p(w)g(a|w)$, where $p(w) = \lambda p'(w) + (1-\lambda)p''(w) \in \mathcal{P}$ and $g(\cdot|w) = \lambda_w g'(\cdot|w) + (1-\lambda_w)g''(\cdot|w) \in \mathcal{G}_w$, $\lambda_w = \frac{\lambda p'(w)}{\lambda p'(w) + (1-\lambda)p''(w)}$, for all $w \in V$. Hence $g(a|w)|_V \in \mathcal{G}|_V$ and $q \in \mathcal{P} * \mathcal{G}$. □

The set of Markov kernels $\Delta_{W,A}$ is a Cartesian product of convex sets $\Delta_{W,A} = \bigtimes_{w\in W} \Delta_A$. The set of joint distributions $\Delta_{W\times A} = \Delta_W * \Delta_{W,A}$ is a simplex, which is a convex set.

A general set $\mathcal{G} \subseteq \Delta_{W,A}$ is not necessarily convex, let alone a Cartesian product of convex sets. However, it can always be written as a union of Cartesian products of convex sets of the form

$$\mathcal{G} = \bigcup_{\theta\in\Theta} \mathcal{G}_\theta, \quad \mathcal{G}_\theta = \bigtimes_{w\in W} \mathcal{G}_{\theta,w}, \quad \mathcal{G}_{\theta,w} \subseteq \Delta_A \text{ convex}.$$

For instance, one can always use $\Theta = \mathcal{G}$, $\mathcal{G}_{\theta=g} = \{g\}$, $\mathcal{G}_{\theta=g,w} = \{g(\cdot|w)\}$. Proposition 12, together with this observation, implies that given any $\mathcal{G} \subseteq \Delta_{W,A}$ and a convex set $\mathcal{P} \subseteq \Delta_W$, the set of joint distributions $\mathcal{F} = \mathcal{P} * \mathcal{G} \subseteq \Delta_{W\times A}$ is a union of convex sets $\mathcal{F}_\theta = \mathcal{P} * \mathcal{G}_\theta$, $\theta \in \Theta$. The situation is illustrated in Example 13.

Example 13. Consider the settings from Example 10. The set $F = \Delta_W * G \subseteq \Delta_{W\times A}$ is the union of following sets:

$$F_\theta = \Delta_W * G_\theta, \quad \theta \in \Delta_A.$$

Each $F_\theta \subseteq F \subseteq \Delta_{W\times A}$ is a polytope with vertices

$$\begin{bmatrix} \theta & (1-\theta) \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ \theta & (1-\theta) \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Appendix B. The stationarity constraint

In the objective function, the marginal distribution over world states is the stationary distribution of the world state transition kernel, and not some arbitrary distribution over world states. The coupling of transition kernels and marginal distributions can be described in terms of the polytope Ξ of joint distributions in $\Delta_{W\times W}$ with equal first and second marginals. This is given by

$$\Xi := \Big\{ p(w,w') \in \Delta_{W\times W} \colon \sum_{w'} p(\cdot,w') = \sum_{w} p(w,\cdot) \Big\}.$$

The second marginal is the result of applying the conditional as a Markov kernel to the first marginal; that is, $\sum_w p(w)p(w'|w) = p(w')$. Hence equality of both marginals means that the marginal is a stationary distribution of the transition $p(w'|w)$.

The polytope Ξ has been studied by Weis (2010) under the name Kirchhoff polytope. The vertices of Ξ are the joint distributions of the following form. For any non-empty subset $\mathcal{W} \subseteq W$ and a cyclic permutation $\sigma \colon \mathcal{W} \to \mathcal{W}$, there is a vertex defined by

$$c_{\mathcal{W},\sigma}(w,w') := \frac{1}{|\mathcal{W}|} \begin{cases} 1, & \text{if } \sigma(w) = w' \\ 0, & \text{otherwise} \end{cases}.$$

The dimension is $\dim(\Xi) = |W|(|W|-1)$. To see this, note that each strictly positive transition $p(w|w)$ is trivially a primitive Markov kernel and hence it has a unique stationary limit distribution. In turn, the set of strictly positive transitions, which has dimension has dimension $|W|(|W|-1)$, corresponds to the relative interior of Ξ.

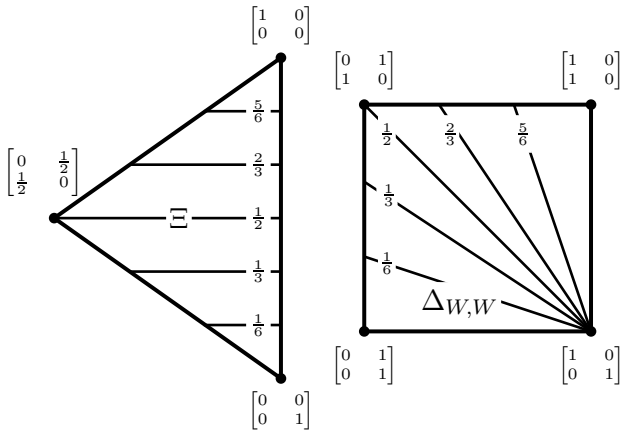

Figure 5: The polytope $\Xi \subseteq \Delta_{W\times W}$, $W = \{0,1\}$, discussed in Example 14. Subsets with first marginals satisfying $p(w=0) = \frac{1}{6}, \dots, \frac{5}{6}$ are highlighted. The right panel shows the corresponding sets of conditional distributions in $\Delta_{W,W}$.

Example 14. Let $W = \{0,1\}$. The non-empty subsets of W are $\mathcal{W} = \{0\}, \{1\}, \{0,1\}$, and the cyclic permutations of these subsets are $0 \to 0$, $1 \to 1$, $(0,1) \to (1,0)$. The Kirchhoff polytope Ξ is the triangle enclosed by the points

$$c_{\{0\},0\to 0} = \begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}, \quad c_{\{1\},1\to 1} = \begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}, \quad c_{\{0,1\},(0,1)\to(1,0)} = \begin{bmatrix}0 & 1/2\\1/2 & 0\end{bmatrix}.$$

Every strictly positive joint distribution $p(w,w')$ corresponds to a marginal $p(w)$ and a conditional distribution $p(w'|w)$. Each point in the interior of Ξ corresponds to a point in the interior of $\Delta_{W,W}$. The situation is illustrated in Figure 5.

Appendix C. Determinism of optimal stationary policies for discounted rewards

Theorem 15. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$, a discount factor $\gamma \in (0,1)$, and a start distribution μ. Then there is a stationary policy $\pi^* \in \Delta_{S,A}$ that is deterministic on each $s \in S$ with $|\operatorname{supp}(\beta(s|\cdot))| \leq 1$ and satisfies $\mathcal{R}^\gamma_\mu(\pi^*) \geq \mathcal{R}^\gamma_\mu(\pi)$ for all $\pi \in \Delta_{S,A}$.*

We will prove Theorem 15 using a policy improvement argument. The world state value function of a policy π is the unique solution of the Bellman equation

$$V^\pi(w) = \sum_a p^\pi(a|w)\Big[R(w,a) + \gamma \sum_{w'} \alpha(w'|w,a) V^\pi(w')\Big].$$

The action value function is given by

$$Q^\pi(w,a) = R(w,a) + \gamma \sum_{w'} \alpha(w'|w,a) V^\pi(w').$$

These definitions make sense both for MDPs and POMDPs. However, while for MDPs there is a stationary policy that maximizes the value of each world state simultaneously, for POMDPs the same is not true in general.

Lemma 16 (POMDP policy improvement)**.** *Let $\pi, \pi' \in \Delta_{S,A}$ be two policies with $\sum_a p^{\pi'}(a|w)Q^{\pi}(w,a) \geq V^{\pi}(w)$ for all w. Then $V^{\pi'}(w) \geq V^{\pi}(w)$ for all w.*

Proof of Lemma 16. The proof follows closely the arguments of the MDP deterministic policy improvement theorem presented by Sutton and Barto (1998).

$$\begin{aligned}
V^{\pi}(w) \leq & \sum_a p^{\pi'}(a|w)Q^{\pi}(w,a) \\
= & \mathbb{E}_{\pi', w_0=w}\Big[R(w_0,a_0) + \gamma V^{\pi}(w_1)\Big] \\
\leq & \mathbb{E}_{\pi', w_0=w}\Big[R(w_0,a_0) + \gamma \sum_a p^{\pi'}(a|w_1)Q^{\pi}(w_1,a)\Big] \\
= & \mathbb{E}_{\pi', w_0=w}\Big[R(w_0,a_0) + \gamma R(w_1,a_1) + \gamma^2 V^{\pi}(w_2)\Big] \\
\leq & \lim_{T\to\infty} \mathbb{E}_{\pi', w_0=w}\Big[\sum_{t=0}^{T-1} \gamma^t R(w_t,a_t)\Big] = V^{\pi'}(w). \qquad \square
\end{aligned}$$

Proof of Theorem 15. Consider any policy $\pi \in \Delta_{S,A}$. Consider some $\tilde{s} \in S$ with $\operatorname{supp}(\beta(\tilde{s}|\cdot)) = \tilde{w}$ and $\tilde{a} \in \operatorname{argmax}_a Q^{\pi}(\tilde{w},a)$. We define an alternative policy by $\pi'(a|s) = \pi(a|s)$, $s \neq \tilde{s}$, and $\pi'(\tilde{a}|\tilde{s}) = 1$. This policy is deterministic on $\tilde{s}$. We have

$$p^{\pi'}(a|w) = \sum_s \beta(s|w)\pi'(a|s) = \sum_{s\neq\tilde{s}} \beta(s|w)\pi'(a|s) = p^{\pi}(a|w), \quad \text{for all } w \neq \tilde{w},$$

and

$$p^{\pi'}(a|\tilde{w}) = \sum_s \beta(s|\tilde{w})\pi'(a|s) = \sum_{s\neq\tilde{s}} \beta(s|\tilde{w})\pi(a|s) + \beta(\tilde{s}|\tilde{w})\delta_{a,\tilde{a}}.$$

In turn,

$$\sum_a p^{\pi'}(a|w)Q^{\pi}(w,a) = \sum_a p^{\pi}(a|w)Q^{\pi}(w,a) = V^{\pi}(w), \quad \text{for all } w \neq \tilde{w},$$

and

$$\begin{aligned}
\sum_a p^{\pi'}(a|\tilde{w})Q^{\pi}(\tilde{w},a) = & \sum_a \Big[\sum_{s\neq\tilde{s}} \beta(s|\tilde{w})\pi(a|s) + \beta(\tilde{s}|\tilde{w})\delta_{a,\tilde{a}}\Big] Q^{\pi}(\tilde{w},a) \\
= & \sum_a \Big[\sum_{s\neq\tilde{s}} \beta(s|\tilde{w})\pi(a|s)\Big] Q^{\pi}(\tilde{w},a) + \sum_a \beta(\tilde{s}|\tilde{w})\delta_{a,\tilde{a}} Q^{\pi}(\tilde{w},a) \\
\geq & \sum_a \Big[\sum_{s\neq\tilde{s}} \beta(s|\tilde{w})\pi(a|s)\Big] Q^{\pi}(\tilde{w},a) + \sum_a \beta(\tilde{s}|\tilde{w})\pi(a|s) Q^{\pi}(\tilde{w},a) \\
= & \sum_a \Big[\sum_{s} \beta(s|\tilde{w})\pi(a|s)\Big] Q^{\pi}(\tilde{w},a) \\
= & \sum_a p^{\pi}(a|\tilde{w})Q^{\pi}(\tilde{w},a) = V^{\pi}(\tilde{w}).
\end{aligned}$$

This shows that $\sum_a p^{\pi'}(a|w)Q^\pi(w,a) \geq V^\pi(w)$, for all w. By Lemma 16 $V^{\pi'}(w) \geq V^\pi(w)$, for all w. Repeating the same arguments, we conclude that any policy π can be replaced by a policy π' which is deterministic on each $s \in S$ with $|\operatorname{supp}\beta(s|\cdot)| = 1$ and which satisfies $V^{\pi'}(w) \geq V^\pi(w)$ for all $w \in W$. Sensor states with $|\operatorname{supp}(\beta(s|\cdot))| = 0$ are never observed and the corresponding policy assignment immaterial. This completes the proof. □

We conclude this section with a few remarks. It is worthwhile to mention the relation

$$\sum_w p^\pi(w)V^\pi(w) = \frac{\mathcal{R}(\pi)}{1-\gamma},$$

which follows from (see Singh et al., 1994, Fact 7)

$$\begin{aligned}
\sum_w p^\pi(w)V^\pi(w) &= \sum_w p^\pi(w) \sum_a p^\pi(a|w) \left[R(w,a) + \gamma \sum_{w'} \alpha(w'|w,a)V^\pi(w') \right] \\
&= \sum_w p^\pi(w) \left[\sum_a p^\pi(a|w)R(w,a) + \gamma \sum_{w'} p^\pi(w'|w)V^\pi(w') \right] \\
&= \mathcal{R}(\pi) + \sum_w p^\pi(w)\gamma \sum_{w'} p^\pi(w'|w)V^\pi(w') \\
&= \mathcal{R}(\pi) + \gamma \sum_{w'} p^\pi(w')V^\pi(w').
\end{aligned}$$

Note that $\mathcal{R}^\gamma_\mu(\pi) = \sum_w \mu(w)V^\pi(w)$. Hence if two policies π, π' satisfy $V^{\pi'}(w) \geq V^\pi(w)$, for all w, then $\mathcal{R}^\gamma_\mu(\pi') \geq \mathcal{R}^\gamma_\mu(\pi)$, for all μ. However, the same hypothesis does not necessarily imply any particular relation between $\mathcal{R}(\pi') = (1-\gamma)\sum_w p^{\pi'}(w)V^{\pi'}(w)$ and $\mathcal{R}(\pi) = (1-\gamma)\sum_w p^\pi(w)V^\pi(w)$.

Appendix D. Examples with analytic solutions

We discuss three examples where it is possible to compute the optimal memoryless policy analytically and show that it has stochasticity equal to the upper bound indicated in Theorem 7. This proves the optimality of the stochasticity bound. The first two examples consider the case $|U| = |S| = 1$ and the third example the case with arbitrarily large $|U|$.

Example 17. Consider a POMDP where the agent has only one sensor state, K possible actions, and the world state transitions are as shown in Figure 6. At each world state only one action takes the agent further to the right, while all other actions take it to $w = 0$ with probability one. At the world state $w = K$, the agent receives a reward and all actions take it to $w = 0$.

We will use the abbreviations $\pi_a = \pi(a|s=1)$ and $p_w = p^\pi(w)$. The world-state transition matrix is given by

$$(p(w'|w))_{w,w'} = \begin{bmatrix}
1-\pi_1 & \pi_1 & & & & \\
1-\pi_2 & & \pi_2 & & & \\
1-\pi_3 & & & \pi_3 & & \\
\vdots & & & & \ddots & \\
1-\pi_K & & & & & \pi_K \\
1 & & & & &
\end{bmatrix}.$$

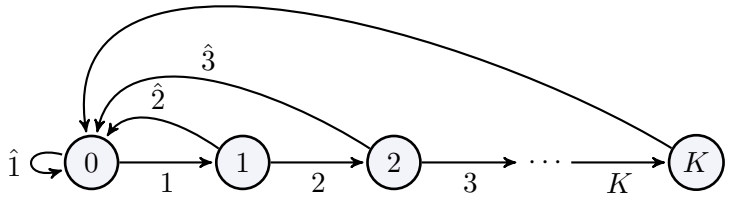

Figure 6: State transitions from Example 17. The number in each node indicates the world state. The sensor state is always $s = 1$. At each world state there is only one action that takes the agent further to the right, while all other actions take it to $w = 0$ with probability one.

For this system the expected reward per time step is $\mathcal{R}(\pi) = p^\pi(w = K)$. The stationary distribution of this transition matrix satisfies

$$\begin{aligned} p_1 &= p_0\pi_1 \\ p_2 &= p_1\pi_2 \\ &\vdots \\ p_K &= p_{K-1}\pi_K. \end{aligned}$$

Using the relation

$$1 = p_0 + p_1 + \cdots + p_K = p_0(1 + \pi_1 + \pi_1\pi_2 + \cdots + \pi_1 \cdots \pi_K),$$

we obtain

$$p_K = \frac{\pi_1 \cdots \pi_K}{1 + \pi_1 + \pi_1\pi_2 + \cdots + \pi_1 \cdots \pi_K}. \tag{9}$$

This is positive if and only if $\pi_1, \ldots, \pi_K$ are all larger than zero. In turn, any optimal memoryless stochastic policy has at least K positive probability actions at the single observation $s = 1$. The next proposition describes the precise form of the optimal memoryless policy (in this case unique).

Proposition 18. *The optimal memoryless policy of the POMDP Example 17 is given by*

$$\begin{aligned} \pi_1 &= c \\ \pi_i &= \pi_{i-1} + c\pi_1 \cdots \pi_{i-1}, \quad i = 2, \ldots, K, \end{aligned}$$

where c is the unique real positive solution of

$$\pi_1 + \cdots + \pi_K = 1. \tag{10}$$

Proof of Proposition 18. The policy that maximizes p_K can be found using the method of Lagrange multipliers. The critical points satisfy $1 - \sum_i \pi_i = 0$ and $\frac{\partial p_K}{\partial \pi_i} - \lambda = 0$ for all $i = 1, \ldots, K$. Computing the derivatives of (9) we find that

$$\frac{\pi_1 \cdots \pi_K}{(1 + \pi_1 + \cdots + \pi_1 \cdots \pi_K)} \times \left(1 - \frac{\pi_1 \cdots \pi_i + \cdots + \pi_1 \cdots \pi_K}{1 + \pi_1 + \cdots + \pi_1 \cdots \pi_K}\right) \frac{1}{\pi_i} = \lambda, \quad \text{for } i = 1, \ldots, K.$$

This implies that

$$\pi_i = c(1 + \pi_1 + \cdots + \pi_1 \cdots \pi_{i-1}), \quad \text{for } i = 1, \ldots, K,$$

where $c = \lambda^{-1} \frac{\pi_1 \cdots \pi_K}{(1+\pi_1+\cdots+\pi_1 \cdots \pi_K)^2}$.

Note that (10), as a polynomial in c, has only positive coefficients, except for the zero degree coefficient, which is -1. By Descartes' rule of sign, a polynomial with coefficients of this form has exactly one real positive root. □

Interestingly, the optimal policy of Example 17 does not assign uniform probabilities to all actions; it satisfies $\pi_1 < \cdots < \pi_K$. The interpretation for this is that, when the agent has covered a longer distance toward the reward position K, the cost of being transported back to the start position before reaching the reward increases. Hence the policy assigns more probability mass to the 'right' actions at positions closer to the reward position. This effect is less pronounced for larger values of K, for which the optimal policy is more uniform. For illustration we solved the polynomial (10) numerically for different values of K. The results are shown in Table 1.

K	$\pi = (\pi_1, \ldots, \pi_K)$	$\mathcal{R} = p_K$
1	(1)	0.5
2	$(0.4142, 0.5858)$	0.1464
3	$(0.2744, 0.3496, 0.3760)$	0.0256
4	$(0.2104, 0.2547, 0.2659, 0.2689)$	0.0030

Table 1: Optimal memoryless policies obtained from Proposition 18 for the POMDP Example 17 for $K = 1, \ldots, 4$.

Example 19. We consider a slight generalization of Example 17. Instead of fully deterministic transitions $p(w'|w, a)$ we now assume that at each $w = 0, \ldots, K-1$ action $a = w + 1$ takes the agent to $w' = w + 1$ with probability $t_{w+1} \in (0, 1]$ and to $w' = 0$ with probability $1 - t_{w+1}$. The world state transition matrix is given by

$$(p(w'|w))_{w,w'} = \begin{bmatrix} 1 - t_1\pi_1 & t_1\pi_1 & & & & \\ 1 - t_2\pi_2 & & t_2\pi_2 & & & \\ 1 - t_3\pi_3 & & & t_3\pi_3 & & \\ \vdots & & & & \ddots & \\ 1 - t_K\pi_K & & & & & t_K\pi_K \\ 1 & & & & & \end{bmatrix}.$$

For this system the expected reward per time step is $\mathcal{R}(\pi) = p^\pi(w = K) = p_K$. Similar to (9) we find that

$$p_K = \frac{t_1\pi_1 \cdots t_K\pi_K}{1 + t_1\pi_1 + \cdots + t_1\pi_1 \cdots t_K\pi_K}.$$

In close analogy to Proposition 18 we obtain the following description of the optimal policies.

Proposition 20. *The optimal memoryless policy of the POMDP Example 19 is given by*

$$\begin{aligned}\pi_1 &= c \\ \pi_i &= \pi_{i-1} + ct_1\pi_1 \cdots t_{i-1}\pi_{i-1}, \quad i = 2, \ldots, K,\end{aligned}$$

where c is the unique real positive solution of

$$\pi_1 + \cdots + \pi_K = 1.$$

The next proposition shows that, in general, for this type of examples, the optimal policy cannot be written as a small convex combination of deterministic policies.

Proposition 21. *There is a choice of $t_1, \ldots, t_K$ for which the optimal memoryless policy of the POMDP Example 19 cannot be written as a convex combination of $K-1$ deterministic policies.*

Proof of Proposition 21. We show that the set of optimal memoryless policies described in Proposition 20, for all $t_1, \ldots, t_K$, is not contained in any finite union of $K-2$ dimensional affine spaces.

Consider the expression $\pi_1 + \cdots + \pi_K$, where $\pi_1 = c$ and $\pi_i = \pi_{i-1} + ct_1\pi_1 \cdots t_{i-1}\pi_{i-1}$. We view this as a polynomial in c with coefficients depending on $t_1, \ldots, t_K$. The derivative with respect to t_{K-1} is non-zero (as soon as $K \geq 2$ and $c \neq 0$). Hence the solution of $\pi_1 + \cdots + \pi_K = 1$ is a non-constant function c of t_{K-1}.

Consider the set of optimal policies for a fixed choice of $t_1, \ldots, t_{K-2}$ and an interval $T \subseteq (0, 1]$ of values of t_{K-1}. This is given by

$$(f_1(c), f_2(c), \ldots, f_{2^{K-1}}(c), f_{2^K}(c)t_{K-1}), \quad \text{for all} t_{K-1} \in T,$$

where c is a non-constant function of t_{K-1}, and f_j is a polynomial of degree j in c with coefficients depending on $t_1, \ldots, t_{K-2}$. The restriction of these vectors to the first $K-1$ coordinates is

$$(f_1(c), f_2(c), \ldots, f_{2^{K-1}}(c)), \quad \text{for all} c \in C,$$

where $C = \{c(t_{K-1})\colon t_{K-1} \in T\} \subset \mathbb{R}$ is an interval with non-empty interior. This set is a linear projection of the interval $\{(c, c^2, \ldots, c^{2^K})\colon c \in C\}$ of the moment curve in 2^K-dimensional Euclidean space, by the matrix M with entry $M_{j,i}$ equal to the degree-i coefficient of f_j, for all $j = 1, \ldots, K-1$ and $i = 1, \ldots, 2^K$. This matrix is full rank, since each of the f_j has different degree.

It is well known that each hyperplane intersects a moment curve at most at finitely many points. Since our linear projection is full rank, the smallest affine space containing infinitely many of its points is equal to the ambient space $\mathbb{R}^{K-1}$. In turn, no finite union of convex hulls of $K-1$ polices contains the set of optimal policies for all $t_{K-1} \in T$. □

Example 22. Consider a POMDP where the agent has U sensor states, K possible actions, and the world state transitions are as shown in Figure 7. The world state transition matrix is given by

$$(p(w'|w))_{w,w'} = \begin{bmatrix} 1 - t_{11}\pi_{11} & t_{11}\pi_{11} & & & & & \\ \vdots & & \ddots & & & & \\ 1 - t_{1K}\pi_{1K} & & & t_{1K}\pi_{1K} & & & \\ 1 - t_{21}\pi_{21} & & & & t_{21}\pi_{21} & & \\ \vdots & & & & & \ddots & \\ 1 - t_{UK}\pi_{UK} & & & & & & t_{UK}\pi_{UK} \\ 1 & & & & & & \end{bmatrix},$$

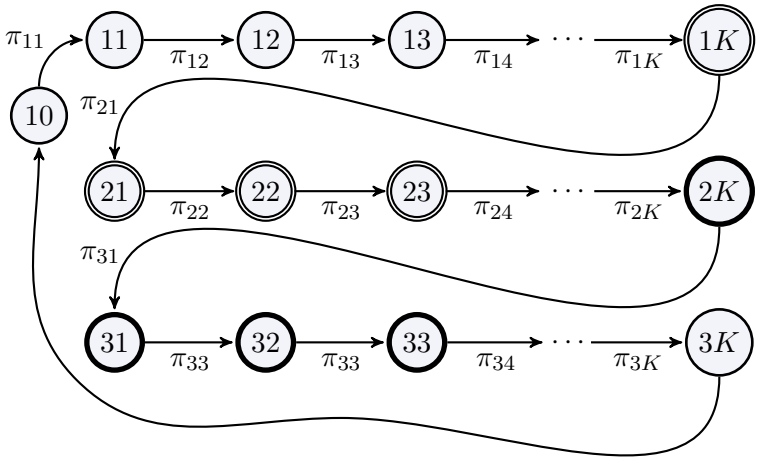

Figure 7: State transitions from Example 22. The number in each node indicates the world state. The type of circle indicates the sensation of the agent (single stroke stands for $s = 1$, double stroke for $s = 2$, etc.). At each world state exactly one action takes the agent further, while all other actions take it back to $w = 0$ (arrows omitted for clarity). At w_{UK} the agent receives a reward of one and is invariably taken to w_{10}.

The next proposition describes the optimal memoryless policy.

Proposition 23. *The optimal memoryless policy is given by*

$$\begin{aligned} \pi_{j1} &= c_j d_j, \\ \pi_{ji} &= \pi_{j,i-1} + c_j e_j t_{j1}\pi_{j1}\cdots t_{j,i-1}\pi_{j,i-1}, \quad i = 2,\dots,K, \end{aligned}$$

where

$$\begin{aligned} d_j &= 1 + t_{11}\pi_{11} + \cdots + t_{11}\pi_{11}\cdots t_{j-1,K}\pi_{j-1,K}, \\ e_j &= t_{11}\pi_{11}\cdots t_{j-1,K}\pi_{j-1,K}, \end{aligned}$$

and c_j is the unique real positive solution of

$$\pi_{j1} + \cdots + \pi_{jK} = 1,$$

for $j = 1,\dots,U$. Here empty products are defined as 1 and empty sums as 0.

Proof of Proposition 23. After some algebra, similar to (9), one finds that the last entry of the stationary world state distribution is given by

$$p_{UK} = \frac{t_{11}\pi_{11}\cdots t_{UK}\pi_{UK}}{1 + t_{11}\pi_{11} + \cdots + t_{11}\pi_{11}\cdots t_{UK}\pi_{UK}}.$$

We can maximize this with respect to π using the method of Lagrange multipliers. This yields the following conditions:

$$\begin{aligned} 1 - \sum_i \pi_{ji} &= 0, \quad \text{for all } j = 1,\dots,U, \\ \frac{\partial p_{UK}}{\partial \pi_{ji}} - \lambda_j &= 0, \quad \text{for all } i = 1,\dots,K \text{ and } j = 1,\dots,U. \end{aligned}$$

From this we obtain

$$\lambda_j = \frac{1}{\pi_{ji}} \frac{t_{11}\pi_{11}\cdots t_{UK}\pi_{UK}}{(1 + t_{11}\pi_{11} + \cdots + t_{11}\pi_{11}\cdots t_{UK}\pi_{UK})} \times \left(1 - \frac{t_{11}\pi_{11}\cdots t_{ji}\pi_{ji} + \cdots + t_{11}\pi_{11}\cdots t_{UK}\pi_{UK}}{1 + t_{11}\pi_{11} + \cdots + t_{11}\pi_{11}\cdots t_{UK}\pi_{UK}}\right), \quad \text{for all } i = 1, \ldots, K \text{ and } j = 1, \ldots, U.$$

This implies

$$\pi_{ji} = c_j(1 + t_{11}\pi_{11} + \cdots + t_{11}\pi_{11}\cdots t_{j,i-1}\pi_{j,i-1}), \quad \text{for all } i = 1, \ldots, K \text{ and } j = 1, \ldots, U,$$

where $c_j = \lambda_j^{-1} \frac{t_{11}\pi_{11}\cdots t_{UK}\pi_{UK}}{(1+t_{11}\pi_{11}+\cdots+t_{11}\pi_{11}\cdots t_{UK}\pi_{UK})^2}$. □

For each sensor state the optimal policy has K positive probability actions. In particular, the smallest face of $\Delta_{S,A}$ which contains the optimal policy has dimension $|U|(|A| - 1)$.