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

Guido Montúfar
montufar@mis.mpg.de
Max Planck Institute MIS

Keyan Ghazi-Zahedi
zahedi@mis.mpg.de
Max Planck Institute MIS

Nihat Ay
nay@mis.mpg.de
Max Planck Institute MIS,
Leipzig University,
Santa Fe Institute

Abstract

It is well known that any finite state Markov decision process (MDP) has a deterministic memoryless policy that maximizes the discounted long-term expected reward. Hence for such MDPs the optimal control problem can be solved over the set of memoryless deterministic policies. In the case of partially observable Markov decision processes (POMDPs), where there is uncertainty about the world state, optimal policies must generally be stochastic, if no additional information is presented, e.g., the observation history. In the context of embodied artificial intelligence and systems design an agent’s policy underlies hard physical constraints and must be as efficient as possible. Having this in mind, we focus on memoryless POMDPs. We cast the optimization problem 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 means that we can give an upper bound to the number of deterministic policies that need to mixed to obtain an optimal stationary policy, regardless of the specific reward function.

1 INTRODUCTION

Reinforcement learning is a class of learning problems, in which an agent takes an action depending on its current state. As a result of its action, the agent receives a sparse reward signal from the environment, which is used to learn an optimal policy (Sutton and Barto 1998). Theoretical results for reinforcement learning problems are mainly discussed in the context of Markov decision processes (MDPs), which mean that the agent’s state contains all necessary information. For MDPs it is well known that an optimal policy can be chosen from the set of stationary deterministic policies. We are interested in understanding the structure of optimal policies in the context of embodied artificial intelligence and systems design (Pfeifer and Bongard 2006). Here the question is: given a specific morphology, i.e., a body including its physical parameters, sensor, and actuator configuration, what can we say about the structure of optimal policies? This case differs from the MDP setting, as the embodied agent can never access the full information about its environment. This means that the policy has to generate an action based only on partial information about the environment. This setting is also known as partially observable Markov decision processes (POMDPs), for which there are several learning strategies (e.g., Chrisman 1992; Littman et al. 1995; McCallum 1996; Parr and Russell 1995). In the context of POMDPs, optimal policies are generally stochastic. The amount of stochasticity can be reduced with the amount of information stored internally by the agent. One example is to equip the agent with a memory that stores the entire sensor history. This setting leads to nice theoretical results (Still 2009) but is not realistic, because we have to deal with hard constraints in embodied agents. In artificial systems, the on-board computation sets limits to the complexity of the controller with respect to both, memory and computational cost. The same also holds true for natural systems, which have to be energy efficient in order to survive. Efficient, i.e., smaller brains are evolutionary beneficial (e.g., Sol et al. 2010). In this work we focus on memoryless POMDPs.

A stochastic policy can be treated as a mixture of deterministic policies. We can now rephrase the question posed above in the following way: Given a POMDP reinforcement learning problem, what is the maximal number of deterministic policies that we need to mix for an optimal solution, regardless of the particular reward function?

We follow a geometric approach to analyzing the structure of optimal POMDP policies. To understand the general idea, we first look at the structure of optimal policies in the context of MDPs. For these, the existence of stationary deterministic optimal policies can be traced back to the structure of the objective function. The objective is a
linear functional of a convex set of policies. In turn, there always exists an optimizer which is an extreme point of the set of policies. The extreme points of the set of policies are the deterministic policies (which cannot be written as convex combinations of other policies). For POMDPs, the objective is also a linear functional; however, the feasible domain is in general not convex. Nonetheless, the feasible domain can be decomposed into convex pieces. In turn, an optimizer can be chosen as an extreme point of one of these pieces. This is the geometric view that we present in this paper. In general, the more extreme a policy, the more deterministic. Depending on the dimension of the convex pieces, their extreme points are more or less extreme in the set of all policies.

This paper is organized as follows. In Section 2, we briefly comment on MDPs. In Section 3, we comment on POMDPs and specify our settings. In Section 4, we discuss the optimization problem as a constrained linear optimization problem with two types of constraints. Before going into the details of these constraints, in Section 5, we prepare some tools regarding convex sets of probability distributions. In Section 6, we discuss the first constraint, which is about the representability of sets of policies in the underlying MDP. In Section 7, we discuss the second constraint, which is about the stationarity of the world state distribution. In Section 8, we decompose the feasible domain into convex pieces and show that any POMDP has an optimal stationary policy of limited stochasticity. In Section 9, we discuss an example. Section 10 offers a conclusion.

2 MARKOV DECISION PROCESSES

We consider a discrete time Markov decision process (MDP), which is defined by a tuple \((W, A, R, \gamma)\), where \(W\) is a finite set of states, \(A\) is a finite set of actions, \(R: W \times A \rightarrow \mathbb{R}\) is a reward function, and \(\gamma \in (0, 1)\) is a discount factor. A policy \(\pi\) is a mechanism for selecting actions at each time step, depending on the history of world states and actions. The goal is to maximize the discounted long-term expected reward

\[
\lim_{T \rightarrow \infty} \mathbb{E}_{\Pr \{(w_0, a_0, \ldots)_{t=0}^T \mid \pi\}} \left[ \sum_{t=0}^T \gamma^t R(w_t, a_t) \right],
\]

where \(\Pr \{(w_0, a_0, \ldots)_{t=0}^T \mid \pi\}\) is the probability of the sequence \(w_0, a_0, w_1, a_1, \ldots\), given that actions are selected according to the policy \(\pi\). Here we assume some prior initial distribution over the world state at time \(t = 0\). It is well known that, without loss of optimality, this problem can be solved over the set of stationary deterministic policies [Howard 1960]; that is, policies that at any given time step choose an action deterministically based only on the current world state.

3 PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

A discrete time partially observable Markov decision process (POMDP) is defined by a tuple \((W, S, A, \alpha, \beta, R, \gamma)\). Here, in addition to the MDP quadruple \((W, A, R, \gamma)\), we also have a finite set \(S\) of sensor states and a Markov kernel \(\beta: W \rightarrow \Delta_S\), which describes a sensor measurement of the world state. In this setting, the policy is a mechanism for selecting actions based only on the history of sensor states and actions. In general the sensor state conveys only partial information about the world state. The goal, again, is to maximize the discounted long-term expected reward.

In a POMDP, it may happen that none of the optimal policies is memoryless. Furthermore, it may happen that none of the optimal memoryless policies is deterministic (see, e.g., Singh et al. 1994). The intuitive reason is simple: In a POMDP, there may exist a set of world states that cannot be distinguished through the sensor measurement. When the sensor state is ambiguous about the world state, the policy cannot select any one specific action but it rather has to select a combination of the actions that would be optimal for all possibly underlying world states. The history of sensor measurements may remove this ambiguity to some extent. This illustrates why an optimal policy may need to take the entire history of sensor states into account and also why it may need to choose actions stochastically, depending on the degree of uncertainty about the current world state.

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 sensor states are replaced by Bayesian beliefs about the current world state, which effectively encode the history of sensor states. Clearly, this is associated with a costly internal computation from the side of the acting agent. Our emphasis lies on the constraints the agent is subject to, which, besides the perception limitations, include computational and storage limitations. Therefore, in this paper we consider only stationary policies, which select actions based only on the current sensor measurement; that is, policies of the form \(\pi: S \rightarrow \Delta_A\). We denote by \(\Delta_{S,A}\) the set of all such kernels.

The question that motivated our analysis is the following: Given that an MDP has a stationary deterministic optimal policy, how stochastic is the less stochastic optimal stationary policy of a POMDP?

We will assume that the POMDPs have an underlying MDP that is ergodic for every stationary policy. This means that there is a stationary distribution of the world state for each fixed choice of the policy.

\[\text{The main concepts of this paper can be transferred to more general settings involving internal world state representations. That is a natural next step for future considerations.}\]
We consider the following objective function
\[ R_R(\pi) = \sum_w p^\pi(w) \sum_a p^\pi(a|w) R(w, a), \] (2)
where \( p^\pi(w) \in \Delta_W \) is the stationary distribution over the world state when running a stationary policy \( \pi \in \Delta_{S,A} \), and \( p^\pi(a|w) = \sum_s \pi(a|s) \beta(s|w) \). Note that an MDP is the special case where \( S = W \) and \( \beta(s|w) = \delta_w(s) \), such that \( p^\pi(a|w) = \pi(a|w) \). Maximizing this objective function is the same as maximizing the discounted long-term expected reward. See Appendix A.

We claim that there is a set \( M \subseteq \Delta_{S,A} \) that contains an optimal stationary policy, regardless of the choice of \( R \), such that
\[ \max_{\pi \in \Delta_{S,A}} R_R(\pi) = \max_{\pi \in M} R_R(\pi), \quad \text{for all } R \in \mathbb{R}^{W \times A}. \] (3)
We are interested in the possibility that the set \( M \) is much smaller than \( \Delta_{S,A} \).

As mentioned in the introduction, in the case of MDPs, the set \( M \) can be chosen as the set \( C_0 \) of deterministic policies, defined as
\[ C_0 := \{ \pi^f(a|s) = \delta_{f(s)}(a): f \in A^S \}. \] (4)
The policy \( \pi^f \) corresponds to the deterministic function \( f: S \to A \). This is a finite set of cardinality \( |A|^{|S|} \). In contrast, \( \Delta_{S,A} \) is a polytope of dimension \( |S|(|A| - 1) \).

As candidate optimization sets for POMDPs we consider the sets of policies that can be written as convex combinations of \( d + 1 \) or fewer deterministic policies. For any \( 0 \leq d \leq |S|(|A| - 1) \), this set is defined by
\[ C_d := \left\{ \pi \in \Delta_{S,A}: \pi = \sum_{i=1}^{d+1} \lambda_i \pi^f_i, f_i \in A^S \right\}. \] (5)
Here the mixture weights \( \lambda_i \) are non-negative numbers adding to one. Note that \( C_d \) contains all \( d \)-dimensional faces of the polytope \( \Delta_{S,A} \). In particular, we have that \( C_{|S|(|A| - 1)} = \Delta_{S,A} \).

4 A CONSTRAINED LINEAR OPTIMIZATION PROBLEM

In this section we elaborate on the observation that the discounted long-term expected reward is a linear functional of the stationary policy, subject to certain constraints.

The expression \( \sum_w p(w) \sum_a p(a|w) R(w, a) \) that appears in the expected reward 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. However, optimization is with respect to the policy \( \pi \) and, furthermore, the world state distribution \( p(w) \) is a stationary distribution of the underlying Markov process, depending on the policy. This implies that not all joint distributions \( p(w, a) \) are feasible. The feasible set of joint distributions is the subset of \( \Delta_W \times A \) delimited by the following two conditions.

- Representability in terms of the policy:
  \[ p(a|w) = \sum_s \pi(a|s) \beta(s|w), \quad \text{for some } \pi \in \Delta_{S,A}. \] (6)
- Stationarity of the world state distribution:
  \[ p(w, w') = \sum_a p(w, a) \alpha(a'|w, a) \in \Xi, \] (7)
  where \( \Xi \subseteq \Delta_W \times W \) is the polytope of distributions with equal first and second marginals. This means that \( p(w) \) is a stationary distribution of the Markov transition kernel \( p(a'|w, a) \).

Restriction \( \rho \) is that the conditional distribution \( p(a|w) \) belongs to the polytope \( \Xi \subseteq \Delta_W \) defined as the image of \( \Delta_{S,A} \) by the linear map \( \beta: \pi(a|s) \to \sum_s \pi(a|s) \beta(s|w) \). This means that the joint distribution \( p(w, a) \) belongs to the set \( F \subseteq \Delta_W \times A \) of joint distributions with conditionals \( p(a|w) \) contained in \( G \). In general the set \( F \) is not convex, but it is convex in the marginals \( p(w) \) when fixing the conditionals \( p(a|w) \), vice versa. We discuss the details of this constraint in Section 6.

Restriction \( \rho \) is that \( p(w, a) \) is mapped to a point in \( \Xi \) by the linear map \( \alpha: p(w, a) \to \sum_a p(w, a) \alpha(a'|w, a) \). That is, that \( p(w, a) \) belongs to the polytope \( J \) defined as the preimage of \( \Xi \) by \( \alpha \). We discuss the details in Section 7.

Summarizing, the objective function is linear on \( \Delta_W \times A \). At the same time, optimization takes place over a feasible set \( F \cap J \subseteq \Delta_W \times A \), where \( J \) is convex, but \( F \) is not necessarily convex.

5 CONVEXITY OF STATE-ACTION DISTRIBUTIONS

As mentioned in the previous section, the set \( F \) of representable joint distributions over world states and actions is not necessarily convex. This set arises as the set of all joint distributions which have conditionals distributions from a set \( G = f_\beta(\Delta_{S,A}) \). The goal of this section is to describe largest possible convex subsets of \( F \), depending on the properties of \( G \), which we discuss in the next section.

We will use the following definitions:

Definition 1.

- Given a set of distributions \( \mathcal{P} \subseteq \Delta_W \) and a set of Markov kernels \( \mathcal{G} \subseteq \Delta_W \), let\( \mathcal{P} \ast \mathcal{G} := \left\{ q(w, a) = p(w)g(a|w): p \in \mathcal{P}, g \in \mathcal{G} \right\} \)
denote the set of joint distributions in $\Delta_{W \times A}$ with first marginals in $P$ and conditional distributions in $G$.

- For any $V \subseteq W$ let
  \[
  \Delta_W(V) := \left\{ p \in \Delta_W : \supp(p) := \{ w \in W : p(w) > 0 \} \subseteq V \right\}
  \]
  denote the distributions in $\Delta_W$ with support in $V$.

- Given a subset $V \subseteq W$ and a set of kernels $G \subseteq \Delta_{W,A}$, let
  \[
  G|_V := \left\{ h \in \Delta_{V,A} : h(w|a) = g(w|a) \text{ for all } w \in V, \text{for some } g \in G \right\}
  \]
  denote the set of restrictions of elements of $G$ to $V$.

Note that the set of Markov kernels $\Delta_{W,A}$ is a Cartesian product
\[
\Delta_{W,A} = \times_{w \in W} \Delta_A,
\]
since each $g \in \Delta_{W,A}$ is a tuple of probability distributions $g(w|a) \in \Delta_A, w \in W$.

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 marginal 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 1.** Let $V \subseteq W$. Let $P \subseteq \Delta_W(V)$ be a convex set. Let $G \subseteq \Delta_{W,A}$ satisfy $G|_V = \times_{w \in V} G_w \subseteq \Delta_{V,A}$, where $G_w \subseteq \Delta_A$ is a convex set for all $w \in V$. Then $P \ast G \subseteq \Delta_{W \times A}$ is convex.

**Proof.** We need to show that, given any two distributions $g'$ and $g''$ in $P \ast G$, and any $\lambda \in [0, 1]$, the convex combination $q = \lambda g' + (1 - \lambda) g''$ lies in $P \ast G$. This is the case when $q(w, a) = p(w) g'(a|w)$ for some $p \in P$ and some $g \in \Delta_{W,A}$ with $g|_V \in G|_V$. We have
\[
g(w, a) = \lambda g'(w, a) + (1 - \lambda) g''(w, a)
= \lambda p'(w) g'(a|w) + (1 - \lambda) p''(w) g''(a|w)
= (\lambda p'(w) + (1 - \lambda) p''(w))
\times \left( \frac{\lambda g'(w)}{\lambda p'(w) + (1 - \lambda) p''(w)} g'(a|w)
+ \frac{(1 - \lambda) g''(w)}{\lambda p'(w) + (1 - \lambda) p''(w)} g''(a|w) \right).
\]
This shows that $q(w, a) = p(w) g(a|w)$, where $p(w) = \lambda p'(w) + (1 - \lambda) p''(w) \in P$ and $g(a|w) = \lambda g'(w) + (1 - \lambda) g''(w) \in G_w$. Hence $g(a|w)|_V \in G|_V$ and $q \in P \ast G$. \qed

Figure 1: Decomposition of $G = f_\beta(\Delta_{S,A}) \subseteq \Delta_{W,A}$ into a collection of Cartesian products of convex sets, $G_\theta$, $\theta \in [0, 1]$. In this example $W = \{0, 1, 2\}$ and $S = \{0, 1\}$. The sensor measurement $\beta$ fails to distinguish between $w = 0$ and $w = 1$; both are mapped to $s = 0$, while it distinguishes $w = 2$ as $s = 1$.

Note that any set $G \subseteq \Delta_{W,A}$ can be written as a union of Cartesian products of convex sets, as
\[
G = \bigcup_{\theta \in \Theta} G_\theta,
G_\theta = \times_{w \in W} G_{\theta,w},
G_{\theta,w} \subseteq \Delta_A \text{ convex},
\]
for all $w \in W$, for all $\theta \in \Theta$, for some index set $\Theta$. The existence of such a decomposition is clear, since one can always choose $\Theta = G$ and $G_{\theta=a} = \{g\}$ for all $g \in G$.

The situation is illustrated in Figure 1.

6 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
\[
f_\beta : \Delta_{S,A} \to \Delta_{W,A}; \pi(s,a) \mapsto \sum_s \beta(s|w) \pi(a|s).
\]

We are interested in the properties of this set, depending on the properties of the measurement $\beta$.

Consider a deterministic kernel $\beta^b$, defined by $\beta^b(s|w) = \delta_{b(w)}(s)$, for some function $b: W \to S$. Then
\[
f_{\beta^b}(\Delta_{S,A}) = \times_{s \in S} \text{sym} \Delta_{b^{-1}(s), A}.
\]
where $b^{-1}(s)$ is the set of all $w \in W$ that map to the same observation $s$, and $\Sigma \Delta_{B,A} := \{g \in \Delta_{B,A} : g(a | w) = p(a) \text{ for all } w \in B, \text{ for some } p \in \Delta_A \}$ is the symmetric part of $\Delta_{B,A}$. In particular, the set $f_{\beta}(\Delta_{S,A})$ always contains $\Sigma \Delta_{W,A}$. Also, this can be written as

$$f_{\beta}(\Delta_{S,A}) = \bigcup_{\theta \in \Theta} \left( \bigotimes_{s \in U} \left( \bigotimes_{w \in b^{-1}(s)} g_{\theta,s,w} \right) \right) \otimes \Delta_A,$$

where $U := \{s \in S : |b^{-1}(s)| > 1\}$ is the set of sensor states that can come from more than one world state, $\Theta \equiv \Delta_{U,A}$ and $g_{\theta,s,w} = \theta_s$ is a parametrization of $\Delta_A$ by $\theta_s$.

Now, an arbitrary $\beta \in \Delta_{W,S}$ is a convex combination of deterministic kernels, $\beta = \sum_b \lambda(b) \beta^b$. In turn, $f_{\beta}(\pi) = \sum_b \lambda(b) f_{\beta^b}(\pi)$ for all $\pi \in \Delta_{S,A}$. With some abuse of notation we can write the image as $f_{\beta}(\Delta_{S,A}) = \sum_b \lambda(b) \Sigma \Delta_{W,A}$. Where

In the special case of an MDP, we have the trivial factorization $G = \Delta_{W,A} = \times_w \Delta_A$.

The decomposition of $G$ into Cartesian products of convex sets is illustrated in Figure 1.

### 7 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 world state transition kernel, and not some arbitrary distribution stochastically only for those sensor states that are ambiguous about the world state.

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 $\Xi$ has been studied by [Weis (2010)]. The vertices of $\Xi$ are the joint distributions of the following form. For any non-empty subset $W \subseteq W$ and a cyclic permutation $\sigma : W \rightarrow W$, there is a vertex defined by

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

The dimension is given by $\dim(\Xi) = |W|(|W| - 1)$. 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 corresponds to the relative interior of $\Xi$. This has dimension $|W|(|W| - 1)$. The polytope $\Xi$ is illustrated in Figure 2.

### 8 DETERMINISM OF OPTIMAL STATIONARY POLICIES

We have the following result, which states that any POMDP has an optimal stationary policy of limited stochasticity. The statement reflects the natural expectation that an optimal stationary policy gets along by choosing an action stochastically only for those sensor states that are ambiguous about the world state.

**Theorem 1.** Consider any POMDP $(W,S,A,\alpha,\beta,R)$. Let $U \subseteq S$ be the sensor states that can be obtained from several world states. Then there is a policy $\pi^* \in \Delta_{S,A}$ which maximizes the expected reward $R(\pi)$ and which can be written as a convex combination of $|U|(|A| - 1) + 1$ or fewer deterministic policies.

**Proof.** The polytope $J$ has dimension $\dim(J) = |W|(|A| - 1)$. Consider the set $F = \Delta_W \ast G$. Like any subset of $\Delta_{W,A}$, the set $G$ can be decomposed as $G = \bigcup_{\theta \in \Theta} G_{\theta}$, where each $G_{\theta}$ is a product of convex sets, $G_{\theta} = \times_{w \in W} G_{\theta,w}$, $G_{\theta,w} \subseteq \Delta_A$ convex, and $\Theta$ is an index set. Let $d_G$ denote the smallest possible $\dim(\Theta)$. If $U$ denotes the set of sensor states that can be obtained from several world states, then $d_G \leq |U|(|A| - 1) - 1$.

We have a decomposition $F = \bigcup_{\theta \in \Theta} F_{\theta}$, where each $F_{\theta} = \Delta_W \ast G_{\theta}$ is a convex set of dimension $\dim(F_{\theta}) = (|W| - 1) + \dim(G_{\theta})$. The convexity of $F_{\theta}$ follows from the product structure of $G_{\theta}$.

Over each of the polytopes $F_{\theta} \cap J$, the objective is maximized at an extreme point. The extreme points of $F_{\theta} \cap J$ lie at faces of $F_{\theta}$ of dimension at most $\dim(F_{\theta}) - \dim(F_{\theta} \cap$
in cell 3, then only the actions east, south, and west lead to a change of its state, whereas north has no effect.

In an MDP setting, the agent would have full knowledge of its state, i.e., in every time step the agent would know its absolute position in the maze. Hence, a deterministic policy can be constructed that leads to a maximal reward (see Fig. 5). This is very different in a POMDP setting. Let us assume that the agent can only sense the configuration of its immediate surroundings, i.e., it only sees the eight surroundings cells (it cannot distinguish between teleporting and non-teleporting cells). The sensor values in each of the 13 cells is depicted in Figure 5b. Any stationary deterministic policy fails in this setting; for the following reason. The agent’s sensation of the environment renders cells 3 and 9 indistinguishable. Hence, for a stationary deterministic policy, the agent would always pick the same action in both cells. Let us assume that the deterministically chosen action for this type of sensation is east. In this case, the agent would successfully avoid the teleporting cell 5, because it would go east when it reaches cell 3. It would navigate to the teleporting cell 11, when it reaches cell 9 for the same reason. Analogously, if the agent deterministically chooses west for both cells 3 and 9, it would always end up in teleporting cell 5. There is one exception, because the agent can distinguish between cells 6 and 10. This does not violate the example, because a policy that deterministically chooses east in cells 3 and 9 either leads to the teleporting cell 11 or to an oscillation between cells 9 and 10 or cells 3 and e.g., 6. In both cases, the long-term reward is zero.
The optimal policy in this case is one that randomly chooses between east and west whenever it is in cell 3 or 9. The expected reward for this stochastic policy is $\frac{1}{2}$ per cycle, as opposed to zero for any deterministic policy.

In this example, the set $U$ from Theorem 1 consists of the sensor state that the agent measures in cells 3 and 9. The theorem states that, regardless of the specific reward function $R$, there is an optimal stationary policy which is a mixture of not more than $|U|(|A| - 1) + 1 = 4$ stationary deterministic policies. For the specific $R$ from the example, a mixture of 2 stationary deterministic policies suffices.

10 CONCLUSIONS

We have developed a geometric view on the problem of finding an optimal stationary policy for the control of a partially observable Markov decision process. We cast the maximization of the discounted long-term expected reward as a constrained linear optimization problem. For this, we presented two types of constraints; one arising from the partial observability of the world state, and another arising from the stationarity of the world state distribution. We described a convex decomposition of the feasible linear optimization domain. We used these tools to show that every POMDP has an optimal stationary policy of limited stochasticity. The necessary level of stochasticity can be bounded from above, depending on the ambiguity of the sensor measurement, independently of the specific reward function. This generalizes well known observations about the existence of deterministic stationary optimal policies for MDPs to the case of POMDPs.

Acknowledgments

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

References

L. 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.

R. Howard. Dynamic Programming and Markov Processes. Published jointly by the Technology Press of the Massachusetts Institute of Technology and, 1960.

M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially observable environments: Scaling up. In A. Prieditis and S. J. Russell, editors, Machine Learning, Proceedings of the Twelfth International Conference on Machine Learning. Tahoe City, California, USA, July 9-12, 1995, pages 362–370. Morgan Kaufmann, 1995.

A. K. Mccallum. Reinforcement Learning with Selective Perception and Hidden State. PhD thesis, 1996.
R. Parr and S. Russell. Approximating optimal policies for partially observable stochastic domains. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI’95, pages 1088–1094, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

R. Pfeifer and J. C. Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. The MIT Press, 2006.

S. S. Singh, T. Jaakkola, and M. I. Jordan. Learning without state-estimation in partially observable markovian decision processes. In *ICML*, pages 284–292, 1994.

D. Sol, N. Garcia, A. Iwaniuk, K. Davis, A. Meade, W. A. Boyle, and T. Székely. Evolutionary divergence in brain size between migratory and resident birds. *PLoS ONE*, 5(3):e9617, 03 2010.

S. Still. Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85(2):28005, 2009.

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

S. Weis. *Exponential Families with Incompatible Statistics and Their Entropy Distance*. PhD thesis, 2010.

### A THE OBJECTIVE FUNCTION

In the following we motivate the definition of the objective function.

First, note that the objective is the average payoff per time step:

\[
\mathcal{R}(\pi) = \sum_w p^\pi(w) \sum_a p^\pi(a|w) R(w, a) \tag{9}
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_{\pi} \left[ \sum_{t=0}^T R(w_t, a_t) \right]. \tag{10}
\]

The value of a world state \(w\) under a policy \(\pi\) is defined as

\[
V^\pi(w) = \lim_{T \to \infty} \mathbb{E}_{\pi} \left[ \sum_{t=0}^T \gamma^t R(w_t, a_t) \right]_{w_0 = w} \tag{11}
\]

This can be written recursively as

\[
V^\pi(w) = \sum_a p^\pi(a|w) \left[ R(w, a) + \gamma \sum_{w'} \alpha(w'|w, a) V^\pi(w') \right]. \tag{12}
\]

This definition makes sense both for MDPs and POMDPs. However, while for MDPs there is an optimal policy that maximizes the value of each world state simultaneously, for POMDPs the same is not true.

One can define a value function for the sensor states as

\[
V^\pi(s) = \sum_w p^\pi(w|s) V^\pi(w), \tag{13}
\]

where

\[
p^\pi(w|s) = \frac{p(s|w)p^\pi(w)}{p^\pi(s)} = \sum_w p(s|w')p^\pi(w') \tag{14}
\]

is the asymptotic occupancy probability distribution. Here \(p^\pi(w)\) is the limit distribution over the world state when running policy \(\pi\).

It has been shown [Singh et al., 1994] Fact 7) that

\[
\sum_w p^\pi(s) V^\pi(s) = \frac{\mathcal{R}(\pi)}{1 - \gamma}. \tag{15}
\]

Hence maximizing the mean of the sensor state value function is the same as maximizing the average payoff per time step.

### B OPTIMAL POLICIES IN MDPS

An optimal policy is a greedy policy of the form

\[
\pi^*(a|w) = \arg\max_a Q(w, a), \tag{16}
\]

where the \(Q\)-function is the value of executing action \(a\) at state \(w\), and then following the optimal policy. This is given by

\[
Q(w, a) = R(w, a) + \gamma \sum_{w'} \alpha(w'|w, a) V^*(w'). \tag{17}
\]

where the optimal value function is

\[
V^*(w) = \max_a \left\{ R(w, a) + \gamma \sum_{w'} \alpha(w'|w, a) V^*(w') \right\}. \tag{18}
\]