Uncertainty and Stochasticity of Optimal Policies

Guido Montúfar
University of California, Los Angeles
montufar@math.ucla.edu

Johannes Rauh
MPI for Mathematics in the Sciences
jrauh@mis.mpg.de

Nihat Ay
MPI for Mathematics in the Sciences
Santa Fe Institute
University of Leipzig
nay@mis.mpg.de

Abstract

We are interested in action selection mechanisms, policies, that maximize an expected long term reward. In general, the identity of an optimal policy will depend on the specifics of the problem, including perception and memory limitations of the agent, the system's dynamics, and the reward signal. We discuss results that allow us to use partial descriptions of the observations, state transitions, and reward signal, in order to localize optimal policies to within a subset of all possible policies. These results imply that we can reduce the search space for optimal policies, for all problems that share the same general properties. Moreover, in certain cases of interest, we can identify the policies that produce the same behaviors and the same expected long term rewards, thereby further reducing the search space.

1 Introduction

We study stochasticity of optimal policies in decision making. We want to understand under which conditions a deterministic behaviour is optimal.

In many contexts, optimal policies are deterministic: in each situation, there is an action (not necessarily unique) that can be considered as the optimal action in this situation. Thus, an optimal policy for a decision maker can be implemented algorithmically as a mapping from situations to actions. Examples are Markov Decision Problems (MDPs). In an MDP the variables determining the immediate reward are available to the decision maker. One can show that in an MDP, there always exists a deterministic optimal policy.

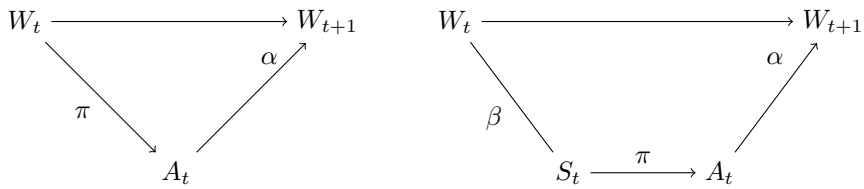

Figure 1: The graphical structure of an MDP and a POMDP.

However, there are situations where optimal policies are not deterministic. The classical example is game theory, as illustrated by the rock-paper-scissors game: when the game is repeated several times, any fixed deterministic strategy can be learned by the opponent, who will then win the game.

Another example is given by Partially Observed Markov Decision Processes (POMDPs). In this case, only a stochastic measurement of the relevant variables is available to the decision maker. In order to be more precise, we now introduce the basic setting and the corresponding notation (see also the graphical representation in Figure 1):

- W_t – world state
- S_t – sensor value
- A_t – chosen action at time t
- α, β – fixed transitions
- π – policy ($\rightarrow$optimize)
- After each step, agent receives reward $R(w_t, a_t)$.

What are the mechanisms that lead to stochastic optimal policies? The game theoretic setting and POMDPs share the following properties:

- **Uncertainty:** The reward R depends on unobserved quantities:
  1. the opponent's strategy
  2. the world state W_t
- **Feedback:** Actions influence the hidden state:
  1. The opponent observes my strategy and adapts.
  2. A_t influence W_{t+1} via α.

Can we control policy stochasticity by controlling uncertainty? One result in this direction is due to [4] (which generalizes [2]). We formulate this kind of problems as localization of optimal policies, and formulate various scenarios in Section 4.

2 Definitions

We consider a POMDP defined by a tuple $(W, S, A, \alpha, \beta, R)$, where W, S, A are finite sets of world states, sensor states, and actions, $\beta\colon W \to \Delta_S$ and $\alpha\colon W \times$

$A \to \Delta_W$ are Markov kernels describing sensor measurements and world state transitions, and $R\colon W \times A \times W \to \mathbb{R}$ is a reward signal depending on the current world state, chosen action, and resulting world state. It is also useful to consider the expected value over resulting world states, $R\colon W \times A \to \mathbb{R}$, with $R(w,a) = \sum_{w' \in W} \alpha(w'|w,a) R(w,a,w')$. A policy is a mechanism for selecting actions. We will focus on stationary (memoryless and time independent) policies, which we simply call *policies*, described by Markov kernels of the form $\pi\colon S \to \Delta_A$. We denote the set of policies by $\Delta_{S,A}$. The deterministic policies map each sensor state to one specific action. They correspond to the vertices of $\Delta_{S,A}$.

The world state is updated at discrete time step by iterating the kernels β, π, α. The objective of learning is to find a policy that maximizes some form of expected long term reward. We focus on the average reward, which for an initial world state distribution $\mu \in \Delta_W$ and a policy $\pi \in \Delta_{S,A}$, is given by

$$\mathcal{R}_\mu(\pi) = \lim_{T\to\infty} \mathbb{E}_{\pi,\mu}\left[\frac{1}{T}\sum_{t=0}^{T-1} R(W_t, A_t, W_{t+1})\right]. \tag{1}$$

We will make the standard assumption that for each fixed policy, the Markov chain of world states is irreducible and aperiodic. This implies that there is a unique stationary limit distribution $p^\pi \in \Delta_W$ of world states. Moreover, this limit distribution is independent of the initial distribution μ. In this case the average reward can be written as

$$\mathcal{R}(\pi) = \sum_w p^\pi(w) \sum_a \sum_s \pi(a|s)\beta(s|w) \sum_{w'} R(w,a,w')\alpha(w'|w,a). \tag{2}$$

Although we restrict the exposition to average rewards as defined above, we note that some of the results hold as well in the setting of discounted rewards, where the rewards in (1) are not weighted uniformly but by a factor γ^t with $\gamma \in (0,1)$.

3 The optimization problem

Before we present the main localization results of this paper, we first explore the structure of the optimization problem in terms of simple examples. They highlight the geometry that underlies the stochasticity of optimal policies.

The objective function (2) reduces to

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

if we introduce the effective state policy $p^\pi(a|w) = \sum_s \beta(s|w)\pi(a|s)$. The stationary state distribution $p^\pi(w)$ is the solution of

$$(T^{(\alpha,\beta,\pi)} - I)p = 0 \quad \text{and} \quad \sum_w p_w = 1 \quad \text{and} \quad p_w \geq 0, \tag{4}$$

where we have defined the state transition matrix

$$T^{(\alpha,\beta,\pi)} = [p^\pi(w'|w)]_{w',w} = \left[\sum_a \alpha(w'|w,a)p^\pi(a|w)\right]_{w',w}. \tag{5}$$

In the following example we give (3) in an explicit form, involving only algebraic operations on π (note that we also use the somewhat more suggestive index notation, for instance $\pi_{1|2}$ instead of $\pi(1|2)$). This gives us a sense of the structure of the optimization problem.

Example 1 (Optimization for two states, two actions, two observations)**.** Let $W = \{1,2\}$, $S = \{1,2\}$, $A = \{1,2\}$. In this case we have

$$\mathcal{R}(\pi) = ((R_{11} - R_{12})\xi_{1|1} + R_{12})p_1 + ((R_{21} - R_{22})\xi_{1|2} + R_{22})(1 - p_1), \tag{6}$$

where the effective world state policy

$$\xi_{1|1} = \beta_{1|1}\pi_{1|1} + (1 - \beta_{1|1})\pi_{1|2} \quad \text{and} \quad \xi_{1|2} = \beta_{1|2}\pi_{1|1} + (1 - \beta_{1|2})\pi_{1|2}, \tag{7}$$

and the stationary state distribution obtained by solving (4)

$$p_1 = \frac{(\alpha_{1|21} - \alpha_{1|22})\xi_{1|2} + \alpha_{1|22}}{(\alpha_{1|21} - \alpha_{1|22})\xi_{1|2} + \alpha_{1|22} + (\alpha_{2|11} - \alpha_{2|12})\xi_{1|1} + \alpha_{2|12}}. \tag{8}$$

Note that the denominator can only vanish if the nominator also vanishes.

So, the objective function $\mathcal{R}(\pi)$ is a rational function of degree 2 in $(\pi_{1|1}, \pi_{1|2}) \in [0,1]^2$, with coefficients depending on α, β, and R. We can plot $\mathcal{R}$ over $\Delta_{S,A} \cong [0,1]^2$. Examples are shown in Fig. 2.

Example 2 (Optimization for two states, two actions, blind agent)**.** Let $W = \{1,2\}$, $S = \{1\}$, $A = \{1,2\}$. We set $\beta(s = 1|w = 1) = 1$, $\beta(s = 1|w = 2) = 1$, and $\pi(a|s) = \pi(a)$, in eq. (6), and solve $\nabla_\pi \mathcal{R}(\pi) = 0$. Using the symbolic mathematics library `SymPy`, we find

$$\pi_1 = \frac{-(\alpha_{122} + \alpha_{212})C \pm \sqrt{(\alpha_{121}\alpha_{212} - \alpha_{122}\alpha_{211})CD}}{C(\alpha_{121} - \alpha_{122} + \alpha_{211} - \alpha_{212})}, \tag{9}$$

where

$$C = ((R_{11} - R_{12})(\alpha_{121} - \alpha_{122}) + (R_{21} - R_{22})(\alpha_{211} - \alpha_{212})) \tag{10}$$

$$D = ((R_{11} - R_{21})(\alpha_{122} + \alpha_{212}) - (R_{12} - R_{22})(\alpha_{121} + \alpha_{211})). \tag{11}$$

These are critical points of the objective function, which might be negative or larger than 1.

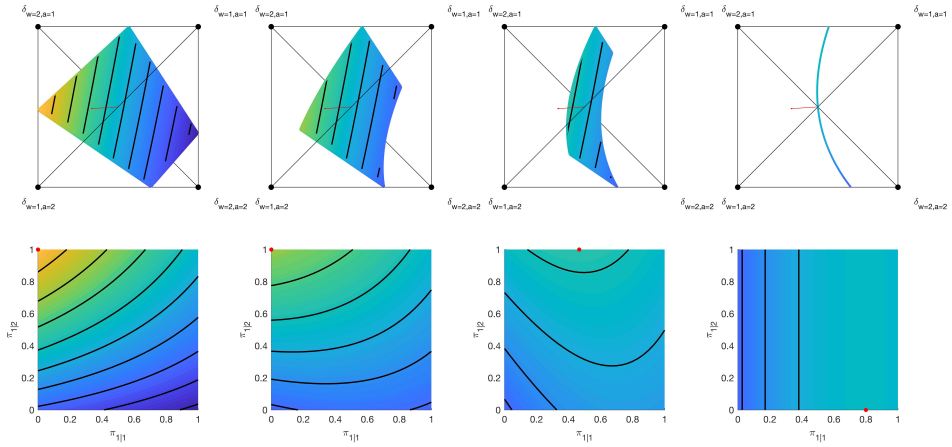

Figure 2: A random choice of α, R, and four choices of $\beta_{s|w}$ going from $[0,1;1,0]$ (fully observable) to $[1,1;0,0]$ (blind). Top row shows the joint distributions $p^\pi(w,a)$. Bottom row shows the policies $(\pi_{1|1},\pi_{1|2})$. Color codes the expected reward $\mathcal{R}(\pi)=\sum_{w,a}p^\pi(w,a)R_{w,a}$. The reward is linear in the space of joint distributions over w and a. The kernel α defines a slice in that space, and β certain inequality constraints. Note that, even in the fully observable case, the expected reward is not a linear function of the policy.

4 Localization of optimal policies

We are interested in saying something about the location of optimal policies. Let $\mathcal{M}$ be a subset of all possible policies. Let $\boldsymbol{A}$ be a subset of all possible state transition kernels $W\times A\to\Delta_W$. Let $\boldsymbol{B}$ be a subset of all possible observation kernels $W\to\Delta_S$. Let $\boldsymbol{R}$ be a subset of all possible reward functions $W\times A\to\mathbb{R}$.

Problem 3. Given $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{R}$, find a subset $\mathcal{N}=\mathcal{N}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{R},\mathcal{M})\subseteq\mathcal{M}$ such that, for any POMDP (W,S,A,α,β,R) with $\alpha\in\boldsymbol{A}$, $\beta\in\boldsymbol{B}$, $R\in\boldsymbol{R}$, there is a policy $\pi^*\in\mathcal{N}$ that is optimal among all policies in $\mathcal{M}$. Ideally, the set $\mathcal{N}$ should be minimal.

We call $\mathcal{N}(\boldsymbol{A},\boldsymbol{B},\boldsymbol{R},\mathcal{M})$ a (*minimal*) *solution set* over $\mathcal{M}$ for the POMDP class $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{R})$. In this language, Theorem 4 below shows that $\mathcal{N}=\{\pi\in\Delta_{S,A}\colon|\operatorname{supp}(\pi(\cdot|s))|\leq k_s\}$ is a minimal solution set over $\mathcal{M}=\Delta_{S,A}$, for the POMDPs with observation kernels from $\boldsymbol{B}=\{\beta\in\Delta_{W,S}\colon|\operatorname{supp}(\beta(s|\cdot))|\leq k_s\}$.

Knowing that there is a set $\mathcal{N}\subseteq\Delta_{S,A}$ which contains an optimal policy allows us to focus the search for optimal policies to the set $\mathcal{N}$. As proposed in [1, 2], this can be used to define a suitable policy model with a reduced number of parameters, without compromising our ability to maximize the average reward.

Observation model

When the world state can be fully recovered from the sensor value, there is an optimal policy within the set of deterministic policies (see, e.g., [5]). This holds irrespective of the specific reward signal and the world state transition kernel. This is an important and well known result in the theory of Markov decision processes. On the other hand, when the agent is blind, it is not possible to localize an optimal policy without taking other specific properties of the system into account.

The following Theorem 4 is a result from [4] refining results from [2]. It generalizes the above discussion to cases where the agent can partially recover the underlying world state. If an observation identifies the world state, there is an optimal policy that is deterministic on this observation. More generally, if a sensor state can result from at most k world states, then there is an optimal policy that randomizes at most k actions at this sensor state. This holds irrespective of the specific reward R and the world state transition kernel α.

Theorem 4. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$. Then there is a policy $\pi^* \in \Delta_{S,A}$ with $|\operatorname{supp}(\pi^*(\cdot|s))| \leq |\operatorname{supp}(\beta(s|\cdot))|$ for all $s \in S$, and $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$. Moreover, there are POMDPs $(W, S, A, \alpha, \beta, R)$ where each policy $\pi^* \in \Delta_{S,A}$ that is optimal among all policies satisfies $|\operatorname{supp}(\pi(\cdot|s))| \geq |\operatorname{supp}(\beta(s|\cdot))|$.*

One might think that the randomization of actions in a POMDP simply allows the agent to assign weights to the optimal deterministic actions that he would choose if he knew the underlying world state. However, the situation is more subtle: being uncertain about the underlying world state, the agent might need to take distance from actions that have the potential of causing a catastrophic outcome when performed in the wrong state, causing him to choose totally different, more conservative, actions.

Transition model

It is also interesting whether we can localize optimal policies given some information about the world state transition kernel α. For instance, [6] studied blind policies for POMDPs and showed that, if each kernel $\alpha(\cdot|\cdot, a)\colon W \to \Delta_W$, $a \in A$, is symmetric and the reward R is proportional to the starting distribution, then there exists a deterministic optimal blind policy. This result can be slightly generalized as follows, dropping the condition on the reward signal.

Theorem 5. *If all kernels $\alpha(\cdot|\cdot, a)\colon W \to \Delta_W$, $a \in A$, have the same stationary distribution and $\beta(\cdot|w) \in \Delta_S$ is independent of $w \in W$, then there is a deterministic policy $\pi^* \in \Delta_{S,A}$ with $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$.*

As an example for the first assumption one can consider doubly stochastic matrices, which all have the uniform distribution as stationary distribution. The second assumption means for practical purposes that the agent is blind.

Let us briefly discuss this result in relation to Theorem 4. Assuming that the sensor kernel β always outputs the same sensor state s, the bound of Theorem 4 is

trivial since it upper bounds the number of actions by $\min\{|W|, |A|\}$. Nonetheless, under the above assumptions, Theorem 5 guarantees the existence of an optimal policy that is deterministic. This shows that incorporating properties of the transitions α, in addition to the observation kernels β, allows us to localize optimal policies even further. The next Theorem 6 extends this line of reasoning by taking some of the structure of the reward R into account.

Reward signal

We consider the case where the reward signal takes the form $R(w, a, w') = R(w, w')$, that is, depending on the sequences of world states but not on the specific actions taken by the agent. For instance, one may be interested in rewarding the motion of a robotic arm, without regard of the specific torques applied to the articulations in order to obtain this motion.

For this type of reward signal, the system can be studied in terms of the world state transitions. Each policy π corresponds to a world state transition $p^\pi \colon W \to \Delta_W$ with $p^\pi(w'|w) = \sum_a \sum_s \beta(s|w)\pi(a|s)\alpha(w'|w, a)$. Since the reward only depends on these transitions, any two policies that represent the same world state transition can be regarded as being equivalent. This allows us to restrict the search for optimal policies to a set of unique representatives.

In [3] it is shown that any feasible world state transition kernel can be represented in terms of a policy π with $|\operatorname{supp}(\pi)| \leq |S| + d_{\alpha,\beta}$. Here $d_{\alpha,\beta}$ is the dimension spanned by the vectors $(\beta(s|w)(\alpha(w'|w, a_0) - \alpha(w'|w, a)))_{w \in W, w' \in W} \in \mathbb{R}^{W \times W}$, for $s \in S$, $a \in A \setminus \{a_0\}$, for any fixed $a_0 \in A$. This is the rank of the linear map from policies to world state transition kernels. In particular, there is an optimal policy that satisfies this property, which implies the following result.

Theorem 6. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$ with $R(w, a, w') = R(w, w')$. Then there is a policy $\pi^* \in \Delta_{S,A}$ with $|\operatorname{supp}(\pi)| \leq |S| + d_{\alpha,\beta}$ and $\mathcal{R}(\pi^*) \geq \mathcal{R}(\pi)$ for all $\pi \in \Delta_{S,A}$.*

Let us comment this result in intuitive terms. It states that if the reward R is not sensitive to the inner working of the control, such that $R(w, a, w')$ is not dependent on a but only on the outcome w' as result of a, then a number of actions can be ignored when maximizing the expected long term reward. The required number of actions essentially involves the dimensionality $d_{\alpha,\beta}$ of the effect that the control has in the world. Note that this number involves α and β in an entangled way.

5 Conclusion

Given a set of policies and a description of the system, we searched for a smallest set of policies that is guaranteed to contain an optimal policy. In particular, we have discussed how the randomization that is needed in optimal actions is related to the amount of information available to an agent at the moment of deciding on the actions. The results presented here can be summarized as follows:

- If the world state is observable, then there is an optimal policy that is deterministic.
- If an observation uniquely identifies the world state, then there is an optimal policy that is deterministic on this observation.
- More generally, if an observation results from at most k world states, then there is an optimal policy which randomizes at most k actions on this observation.
- If all world transition kernels (indexed by the actions) have the same stationary distribution and the observation kernel is independent of the world state, then there is an optimal policy that is deterministic.
- If the reward signal depends on the current and the future world states but not on the specific actions taken, then there is an optimal policy within a low-dimensional face of the set of all possible policies.

References

[1] N. Ay, G. Montúfar, and J. Rauh. Selection criteria for neuromanifolds of stochastic dynamics. In Y. Yamaguchi, editor, *Advances in Cognitive Neurodynamics (III)*, pages 147–154. Springer, 2013.

[2] G. Montúfar, K. Ghazi-Zahedi, and N. Ay. Geometry and determinism of optimal stationary control in POMDPs. *arXiv:1503.07206*, 2015.

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

[4] G. Montúfar and J. Rauh. Geometry of policy improvement. In *Geometric Science of Information, LNCS 10589*, pages 282–290. Springer, 2017.

[5] S. M. Ross. *Introduction to Stochastic Dynamic Programming: Probability and Mathematical*. Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press, Inc., 1983.

[6] N. Vlassis, M. L. Littman, and D. Barber. On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory*, 4(4):12:1–12:8, 2012.