Geometry of Policy Improvement

Guido Montúfar[1,2] and Johannes Rauh[1]

[1]Max Planck Institute for Mathematics in the Sciences
Inselstraße 22, 04103 Leipzig, Germany
[2]Departments of Mathematics and Statistics, UCLA, CA 90095-1555, USA

Abstract. We investigate the geometry of optimal memoryless time independent decision making in relation to the amount of information that the acting agent has about the state of the system. We show that the expected long term reward, discounted or per time step, is maximized by policies that randomize among at most k actions whenever at most k world states are consistent with the agent's observation. Moreover, we show that the expected reward per time step can be studied in terms of the expected discounted reward. Our main tool is a geometric version of the policy improvement lemma, which identifies a polyhedral cone of policy changes in which the state value function increases for all states.

Keywords: Partially Observable Markov Decision Process, Reinforcement Learning, memoryless stochastic policy, policy gradient theorem

1 Introduction

We are interested in the amount of randomization that is needed in action selection mechanisms in order to maximize the expected value of a long term reward, depending on the uncertainty of the acting agent about the system state.

It is known that in a Markov Decision Process (MDP), the optimal policy may always be chosen deterministic (see, e.g., [5]), in the sense that the action a that the agent chooses is a deterministic function of the world state w the agent observes. This is no longer true in a Partially Observable MDP (POMDP), where the agent does not observe w directly, but only the value s of a sensor. In general, optimal memoryless policies for POMDPs are stochastic. However, the more information the agent has about w, the less stochastic an optimal policy needs to be. As shown in [4], if a particular sensor value s uniquely identifies w, then the optimal policy may be chosen such that, on observing s, the agent always chooses the same action. We generalize this as follows: The agent may choose an optimal policy such that, if a given sensor value s can be observed from at most k world states, then the agent chooses an action probabilistically among a set of at most k actions.

Such characterizations can be used to restrict the search space when searching for an optimal policy. In [1], it was proposed to construct a low-dimensional manifold of policies that contains an optimal policy in its closure and to restrict the learning algorithm to this manifold. In [4], it was shown how to do this in

the POMDP setting when it is known that the optimal policy can be chosen deterministic in certain sensor states. This construction can be generalized and gives manifolds of even smaller dimension when the randomization of the policy can be further restricted.

As in [4], we study the case where at each time step the agent receives a reward that depends on the world state w and the chosen action a. We are interested in the long term reward in either the average or the discounted sense [6]. Discounted rewards are often preferred in theoretical analysis, because of the properties of the dynamic programming operators. In [4], the analysis of average rewards was much more involved than the analysis of discounted rewards. While the case of discounted rewards follows from a policy improvement argument, an elaborate geometric analysis was needed for the case of average rewards.

Various works have compared average and discounted rewards [8, 3, 2]. Here, we develop a tool that allows us to transfer properties of optimal policies from the discounted case to the average case. Namely, the average case can be seen as the limit of the discounted case when the discount factor γ approaches 1. If the Markov chain is irreducible and aperiodic, this limit is uniform, and the optimal policies of the discounted case converge to optimal policies of the average case.

2 Optimal Policies for POMDPs

A (discrete time) partially observable Markov decision process (POMDP) is 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 $R\colon W \times A \to \mathbb{R}$ is a reward signal. We consider stationary (memoryless and time independent) action selection mechanisms, described by Markov kernels of the form $\pi\colon S \to \Delta_A$. We denote the set of stationary policies by $\Delta_{S,A}$. We write $p^\pi(a|w) = \sum_s \beta(s|w)\pi(a|s)$ for the effective world state policy. Standard reference texts are [6, 5].

We assume that the Markov chain starts with a distribution $\mu \in \Delta_W$ and then progresses according to α, β and a fixed policy π. We denote by $\mu_\pi^t \in \Delta_W$ the distribution of the world state at time t. It is well known that the limit $p_\mu^\pi := \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mu_\pi^t$ exists and is a stationary distribution of the Markov chain. The following technical assumption is commonly made:

$(*)$ For all π, the Markov chain over world states is aperiodic and irreducible.

The most important implication of irreducibility is that the limit distribution p_μ^π is independent of μ. If the chain has period s, then $p_\mu^\pi = \lim_{T\to\infty} \frac{1}{s} \sum_{t=1}^{s} \mu_\pi^{T+t}$. In particular, under assumption $(*)$, $\mu_\pi^t \to p_\mu^\pi$ for any μ. (Since we assume finite sets, all notions of convergence of probability distributions are equivalent.)

The objective of learning is to maximize the expected value of a long term reward. The (normalized) discounted reward with discount factor $\gamma \in [0, 1)$ is

$$\mathcal{R}_\mu^\gamma(\pi) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \sum_w \mu_\pi^t(w) \sum_a p^\pi(a|w) R(a, w) = (1-\gamma) \mathbb{E}_{\pi,\mu}\Big[\sum_{t=0}^{\infty} \gamma^t R(a_t, w_t) \Big].$$

The average reward is

$$\mathcal{R}_\mu(\pi) = \sum_w p^\pi_\mu(w) \sum_a p^\pi(a|w) R(a,w).$$

Under assumption $(*)$, $\mathcal{R}_\mu$ is independent of the choice of μ and depends continuously on π, as we show next. Since $\Delta_{S,A}$ is compact, the existence of optimal policies is guaranteed. Without the assumption, optimal policies need not exist. On the other hand, the expected discounted reward $\mathcal{R}^\mu_\gamma$ is always continuous, so that, for this, optimal policies always exist.

Lemma 1. *Under assumption $(*)$, $\mathcal{R}_\mu(\pi)$ is continuous as a function of π.*

Proof. By $(*)$, p^π_μ is the unique solution to a linear system of equations that smoothly depends on π. Thus, $\mathcal{R}_\mu$ is continuous as a function of π. □

Lemma 2. *For fixed μ and $\gamma \in [0,1)$, $\mathcal{R}^\mu_\gamma(\pi)$ is continuous as a function of π.*

Proof. Fix $\epsilon > 0$. There exists $l > 0$ such that $(1-\gamma)\sum_{t=l}^\infty \gamma^t R \leqslant \epsilon/4$, where $R = \max_{a,w} |R(a,w)|$. For each t, the distribution μ^t_π depends continuously on π. For fixed π, let U be a neighborhood of π such that $|\mu^t_\pi(w) - \mu^t_{\pi'}(w)| \leqslant \frac{1}{2|W|R}\epsilon$ for $t = 0, \dots, l-1$ and $\pi' \in U$. Then, for all $\pi' \in U$,

$$|\mathcal{R}^\mu_\gamma(\pi) - \mathcal{R}^\mu_\gamma(\pi')| \leqslant \frac{\epsilon}{2} + (1-\gamma)\sum_{t=0}^{l-1} \gamma^t \sum_w |\mu^t_\pi(w) - \mu^t_{\pi'}(w)| R \leqslant \frac{\epsilon}{2} + \frac{|W|}{2|W|R}\epsilon R = \epsilon.$$

□

The following refinement of the analysis of [4] is our main result.

Theorem 1. *Consider a POMDP $(W, S, A, \alpha, \beta, R)$, and let $\mu \in \Delta_W$ and $\gamma \in [0,1)$. There is a stationary (memoryless, time independent) policy $\pi^* \in \Delta_{S,A}$ with $|\operatorname{supp}(\pi^*(\cdot|s))| \leqslant |\operatorname{supp}(\beta(s|\cdot))|$ for all $s \in S$ and $\mathcal{R}^\gamma_\mu(\pi^*) \geqslant \mathcal{R}^\gamma_\mu(\pi)$ for all $\pi \in \Delta_{S,A}$. Under assumption $(*)$, the same holds true for $\mathcal{R}_\mu$ in place of $\mathcal{R}^\gamma_\mu$.*

We prove the discounted case in Section 3 and the average case in Section 4.

3 Discounted Rewards from Policy Improvement

The state value function V^π of a policy π is defined as the unique solution of the Bellman equation

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

It is useful to write $V^\pi(w) = \sum_a p^\pi(a|w) Q^\pi(w,a)$, where

$$Q^\pi(w,a) = R(w,a) + \gamma \textstyle\sum_{w'} \alpha(w'|w,a) V^\pi(w'), \quad w \in W, a \in A,$$

is the state action value function. Observe that $\mathcal{R}^\gamma_\mu(\pi) = (1-\gamma)\sum_w \mu(w) V^\pi(w)$. If two policies π, π' satisfy $V^{\pi'}(w) \geqslant V^\pi(w)$ for all w, then $\mathcal{R}^\gamma_\mu(\pi') \geqslant \mathcal{R}^\gamma_\mu(\pi)$ for all μ. The following is a more explicit version of a lemma from [4]:

Lemma 3 (Policy improvement lemma). *Let* $\pi, \pi' \in \Delta_{S,A}$ *and* $\epsilon(w) = \sum_a p^{\pi'}(a|w) Q^\pi(w,a) - V^\pi(w)$ *for all* $w \in W$. *Then*

$$V^{\pi'}(w) = V^\pi(w) + \mathbb{E}_{\pi', w_0 = w}\Big[\sum_{t=0}^{\infty} \gamma^t \epsilon(w_t)\Big] \quad \text{for all } w \in W.$$

If $\epsilon(w) \geqslant 0$ *for all* $w \in W$, *then*

$$V^{\pi'}(w) \geqslant V^\pi(w) + d^{\pi'}(w)\epsilon(w) \quad \text{for all } w \in W,$$

where $d^{\pi'}(w) = \sum_{t=0}^{\infty} \gamma^t \Pr(w_t = w | \pi', w_0 = w) \geqslant 1$ *is the discounted expected number of visits to* w.

Proof.
$$\begin{aligned} V^\pi(w) &= \sum_a p^{\pi'}(a|w) Q^\pi(w,a) - \epsilon(w) \\ &= \mathbb{E}_{\pi', w_0 = w}\Big[\Big(R(w_0, a_0) - \epsilon(w_0)\Big) + \gamma V^\pi(w_1)\Big] \\ &= \mathbb{E}_{\pi', w_0 = w}\Big[\Big(R(w_0, a_0) - \epsilon(w_0)\Big) + \gamma\Big(\sum_a p^{\pi'}(a|w_1) Q^\pi(w_1, a) - \epsilon(w_1)\Big)\Big] \\ &= \mathbb{E}_{\pi', w_0 = w}\Big[\sum_{t=0}^{\infty} \gamma^t \Big(R(w_t, a_t) - \epsilon(w_t)\Big)\Big] = V^{\pi'}(w) - \mathbb{E}_{\pi', w_0 = w}\Big[\sum_{t=0}^{\infty} \gamma^t \epsilon(w_t)\Big]. \quad \square \end{aligned}$$

Lemma 3 allows us to find policy changes that increase $V^\pi(w)$ for all $w \in W$ and thereby $\mathcal{R}_\mu^\gamma(\pi)$ for any μ.

Definition 1. *Fix a policy* $\pi \in \Delta_{S,A}$. *For each sensor state* $s \in S$ *consider the set* $\operatorname{supp}(\beta(s|\cdot)) = \{w \in W \colon \beta(s|w) > 0\} = \{w_1^s, \ldots, w_{k_s}^s\}$, *and define the linear forms*

$$l_i^{\pi,s} \colon \Delta_A \to \mathbb{R};\ q \mapsto \sum_a q(a) Q^\pi(w_i^s, a), \quad i = 1, \ldots, k_s.$$

The policy improvement cone *at policy* π *and sensation* s *is*

$$L^{\pi,s} = \big\{q \in \Delta_A : l_i^{\pi,s}(q) \geqslant l_i^{\pi,s}(\pi(\cdot|s)) \text{ for all } i = 1, \ldots, k_s\big\}.$$

The (total) policy improvement cone *at policy* π *is*

$$L^\pi = \big\{\pi' \in \Delta_{S,A} : \pi'(\cdot|s) \in L^{\pi,s} \text{ for all } s \in S\big\}.$$

$L^{\pi,s}$ and L^π are intersections of Δ_A and $\Delta_{S,A}$ with intersections of affine halfspaces (see Fig. 1). Since $\pi \in L^\pi$, the policy improvement cones are never empty.

Lemma 4. *Let* $\pi \in \Delta_{S,A}$ *and* $\pi' \in L^\pi$. *Then, for all* w,

$$V^{\pi'}(w) - V^\pi(w) \geqslant d^{\pi'}(w) \sum_s \beta(s|w) \sum_a (\pi'(a|s) - \pi(a|s)) Q^\pi(w,a) \geqslant 0.$$

Proof. Fix $w \in W$. In the notation from Lemma 3, suppose that $\operatorname{supp}(\beta(\cdot|w)) = \{s_1, \ldots, s_l\}$ and that $w = w_{i_j}^{s_j}$ for $j = 1 \ldots, l$. Then

$$\begin{aligned}\epsilon(w) &= \sum_a p^{\pi'}(a|w) Q^\pi(w,a) - \sum_a p^\pi(a|w) Q^\pi(w,a) \\ &= \sum_{j=1}^{l} \beta(s_j|w) l_{i_j}^{\pi,s_j}(\pi'(\cdot|s_j) - \pi(\cdot|s_j)) \geqslant 0,\end{aligned}$$

since $\pi' \in L^\pi$. The statement now follows from Lemma 3. ▫

Remark 1. Lemma 4 relates to the policy gradient theorem [7], which says that

$$\frac{\partial V^\pi(w)}{\partial \pi(a'|s')} = d^\pi(w) \sum_s \beta(s|w) \sum_a \frac{\partial \pi(a|s)}{\partial \pi(a'|s')} Q^\pi(w,a). \tag{1}$$

Our result adds that, for each w, the value function $V^{\pi'}(w)$ is bounded from below by a linear function of π' that takes value at least $V^\pi(w)$ within the entire policy improvement cone L^π. See Fig. 1.

Now we show that there is an optimal policy with small support.

Lemma 5. *Let P be a polytope, and let $l_1, \ldots, l_k$ be linear forms on P. For any $p \in P$, let $L_{i,+} = \{q \in P \colon l_i(q) \geqslant l_i(p)\}$. Then $\bigcap_{i=1}^k L_{i,+}$ contains an element q that belongs to a face of P of dimension at most $k-1$.*

Proof. The argument is by induction. For $k = 1$, the maximum of l_1 on P is attained at a vertex q of P. Clearly, $l_1(q) \geqslant l_1(p)$, and so $q \in L_{1,+}$.

Now suppose that $k > 1$. Let $P' := P \cap L_{k,+}$. Each face of P' is a subset of a face of P of at most one more dimension. By induction, $\bigcap_{i=1}^{k-1} L_{i,+} \cap P'$ contains an element q that belongs to a face of P' of dimension at most $k-2$. ▫

Proof (of Theorem 1 for discounted rewards). By Lemma 5, each policy improvement cone $L^{\pi,s}$ contains an element q that belongs to a face of Δ_A of dimension at most $(k-1)$ (that is, the support of q has cardinality at most k), where $k = |\operatorname{supp}(\beta(s|\cdot))|$. Putting these together, we find a policy π' in the total policy improvement cone that satisfies $|\operatorname{supp}(\pi(\cdot|s))| \leqslant |\operatorname{supp}(\beta(s|\cdot))|$ for all s. By Lemma 4, $V^{\pi'}(w) \geqslant V^\pi(w)$ for all w, and so $\mathcal{R}_\mu^\gamma(\pi') \geqslant \mathcal{R}_\mu^\gamma(\pi)$. ▫

Remark 2. The $|\operatorname{supp}\beta(s|\cdot)|$ positive probability actions at sensation s do not necessarily correspond to the actions that the agent would choose if she knew the identity of the world state, as shown in our example from Section 5.

4 Average Rewards from Discounted Rewards

The average reward per time step can be written in terms of the discounted reward as $\mathcal{R}(\pi) = \mathcal{R}_{p_\mu^\pi}^\gamma$. However, the hypothesis $V^{\pi'}(w) \geqslant V^\pi(w)$ for all w, does not directly imply any relation between $\mathcal{R}(\pi')$ and $\mathcal{R}(\pi)$, since they compare the value function against different stationary distributions. We show that results for discounted rewards translate nonetheless to results for average rewards.

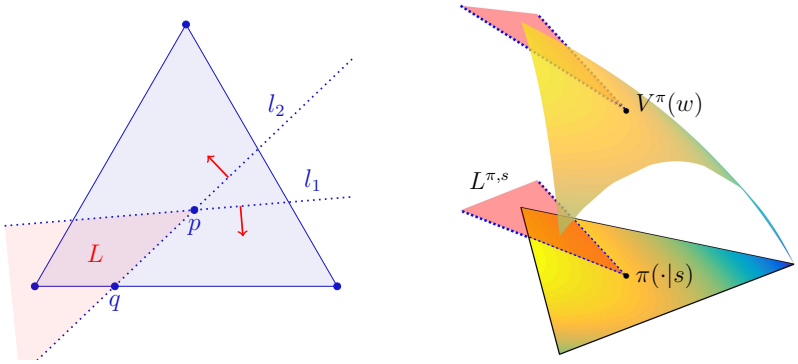

Fig. 1. Left: Illustration of the policy improvement cone. Right: Illustration of the state value function $V^\pi(w)$ for some fixed w, showing the linear lower bound over the policy improvement cone $L^{\pi,s}$. This numerical example is discussed further in Section 5.

Lemma 6. *Let μ be fixed, and assume $(*)$. For any $\epsilon > 0$ there exists $l > 0$ such that for all π and all $t \geqslant l$, $|\mu_\pi^t(w) - p_\mu^\pi(w)| \leqslant \epsilon$ for all w.*

Proof. By $(*)$, the transition matrix of the Markov chain has the eigenvalue one with multiplicity one, with left eigenvector is p_μ^π. Let $p_2, \ldots, p_{|W|}$ be orthonormal left eigenvectors to the other eigenvalues $\lambda_2, \ldots, \lambda_{|W|}$, ordered such that λ_2 has the largest absolute value. There is a unique expansion $\mu = c_1 p_\mu^\pi + c_2 p_2 + \cdots + c_{|W|} p_{|W|}$. Then $\mu_\pi^t = c_1 p_\mu^\pi + \sum_{i=2}^{|W|} c_i \lambda_i^t p_i$. Letting $t \to \infty$, it follows that $c_1 = 1$. By orthonormality, $|c_i|^2 \leqslant \sum_{i=2}^{|W|} c_i^2 \leqslant \|\mu\|_2^2 \leqslant 1$ and $|p_i(w)| \leqslant 1$ for $i = 2, \ldots, |W|$. Therefore, $|\mu_\pi^t(w) - p_\mu^\pi(w)| = |\sum_{i=2}^{|W|} c_i \lambda_i^t p_{|W|}(w)| \leqslant |W||\lambda_2|^t$.

Since $|\lambda_2|$ depends continuously on the transition matrix, which depends continuously on π, $|\lambda_2|$ depends continuously on π. Since $\Delta_{S,A}$ is compact, $|\lambda_2|$ has a maximum d, and $d < 1$ due to $(*)$. Therefore, $|\mu_\pi^t(w) - p_\mu^\pi(w)| \leqslant |W| d^t$ for all π. The statement follows from this. ▫

Proposition 1. *For fixed μ, under assumption $(*)$, $\mathcal{R}_\mu^\gamma(\pi) \to \mathcal{R}_\mu(\pi)$ uniformly in π as $\gamma \to 1$.*

Proof. For fixed μ and ϵ, let l be as in Lemma 6. Let $R = \max_{a,w} |R(a,w)|$. Then

$$\begin{aligned}\mathcal{R}_\mu^\gamma(\pi) &= (1-\gamma) \sum_{k=0}^{l-1} \gamma^k \sum_w \mu_\pi^k(w) \sum_a \pi(a|w) R(a,w) \\ &\quad + (1-\gamma)\gamma^l \sum_{k=0}^{\infty} \gamma^k \sum_w p_\mu^\pi(w) \sum_a \pi(a|w) R(a,w) + O(\epsilon R)(1-\gamma) \sum_{k=0}^{\infty} \gamma^k \\ &= O((1-\gamma) l R) + O(\epsilon R) + \gamma^l \mathcal{R}_\mu(\pi)\end{aligned}$$

for all π. For given $\delta > 0$, we can choose $\epsilon > 0$ such that the term $O(\epsilon R)$ is smaller in absolute value than $\delta/3$. This also fixes $l = l(\epsilon)$. Then, for any $\gamma < 1$ large

enough, the term $O((1-\gamma)lR)$ is smaller than $\delta/3$, and also $|(\gamma^l-1)\mathcal{R}_\mu(\pi)| \leqslant \delta/3$. This shows that for $\gamma < 1$ large enough, $|\mathcal{R}^\gamma_\mu(\pi) - \mathcal{R}_\mu(\pi)| \leqslant \delta$, independent of π. The statement follows since $\delta > 0$ was arbitrary. ▫

Theorem 2. *For any $\gamma \in [0,1)$, let $\hat{\pi}_\gamma$ be a policy that maximizes $\mathcal{R}^\mu_\gamma$. Let $\hat{\pi}$ be a limit point of a convergent subsequence as $\gamma \to 1$. Then $\hat{\pi}$ maximizes $\mathcal{R}_\mu$, and $\lim_{\gamma\to 1} \mathcal{R}^\gamma_\mu(\hat{\pi}_\gamma) = \mathcal{R}_\mu(\hat{\pi})$.*

Proof. For any $\epsilon > 0$, there is $\delta > 0$ such that $\gamma \geqslant 1-\delta$ implies $|\mathcal{R}_\mu(\pi) - \mathcal{R}^\gamma_\mu(\pi)| \leqslant \epsilon$ for all π. Thus $|\max_\pi \mathcal{R}_\mu(\pi) - \max_\pi \mathcal{R}^\gamma_\mu| \leqslant \epsilon$, whence $\lim_{\gamma\to 1} \max_\pi \mathcal{R}^\gamma_\mu(\pi) = \max_\pi \mathcal{R}_\mu(\pi)$. Moreover, $|\max_\pi \mathcal{R}_\mu(\pi) - \mathcal{R}_\mu(\hat{\pi}_\gamma)| \leqslant 2\epsilon + |\max_\pi \mathcal{R}^\gamma_\mu(\pi) - \mathcal{R}^\gamma_\mu(\hat{\pi}_\gamma)| = 2\epsilon$. By continuity, the limit value of $\mathcal{R}_\mu$ applied to a convergent subsequence of the $\hat{\pi}_\gamma$ is the maximum of $\mathcal{R}_\mu$. ▫

Corollary 1. *Fix a world state w, and let $r \geqslant 0$. If there exists for each $\gamma \in [0,1)$ a policy $\hat{\pi}_\gamma$ that is optimal for $\mathcal{R}^\gamma_\mu$ with $|\operatorname{supp}(\pi(\cdot|s))| \leqslant r$, then there exists a policy $\hat{\pi}$ with $|\operatorname{supp}(\pi(\cdot|s))| \leqslant r$ that is optimal for $\mathcal{R}_\mu$.*

Proof. Take a limit point of the family $\hat{\pi}_\gamma$ as $\gamma \to 1$ and apply Theorem 2. ▫

Remark 3. Without $(*)$, one can show that $\mathcal{R}^\gamma_\mu(\pi)$ still converges to $\mathcal{R}_\mu(\pi)$ for each fixed π, but convergence is no longer uniform. Also, $\mathcal{R}_\mu$ need not be continuous in π, and so an optimal policy need not exist.

5 Example

We illustrate our results on an example from [4]. Consider an agent with sensor states $S = \{1,2,3\}$ and actions $A = \{1,2,3\}$. The system has world states $W = \{1,2,3,4\}$ with the transitions and rewards illustrated in Fig. 2. At $w = 1,4$ all actions produce the same outcomes. States $w = 2,3$ are observed as $s = 2$. Hence we can focus on $\pi(\cdot|s=2) \in \Delta_A$. We evaluate 861 evenly spaced policies in this 2-simplex. Fig. 2 shows color maps of the expected reward (interpolated between evaluations), with lighter colors corresponding to higher values. As in Fig. 1, red vectors are the gradients of the linear forms (corresponding to $Q^\pi(w,\cdot)$, $w = 2,3$), and dashed blue lines limit the policy improvement cones $L^{\pi,s=2}$. Stepping into the improvement cone always increases $V^\pi(w) = \mathcal{R}^\gamma_{\mu=\delta_w}(\pi)$ for all $w \in W$. Note that each cone contains a policy at an edge of the simplex, i.e., assigning positive probability to at most two actions. The convergence of $\mathcal{R}^\gamma_\mu$ to $\mathcal{R}_\mu$ as $\gamma \to 1$ is visible. Note also that for $\gamma = 0.6$ the optimal policy requires two positive probability actions, so that our upper bound $|\operatorname{supp}(\pi(\cdot|s))| \leqslant |\operatorname{supp}(\beta(s|\cdot))|$ is attained.

Acknowledgment: We thank Nihat Ay for support and insightful comments.

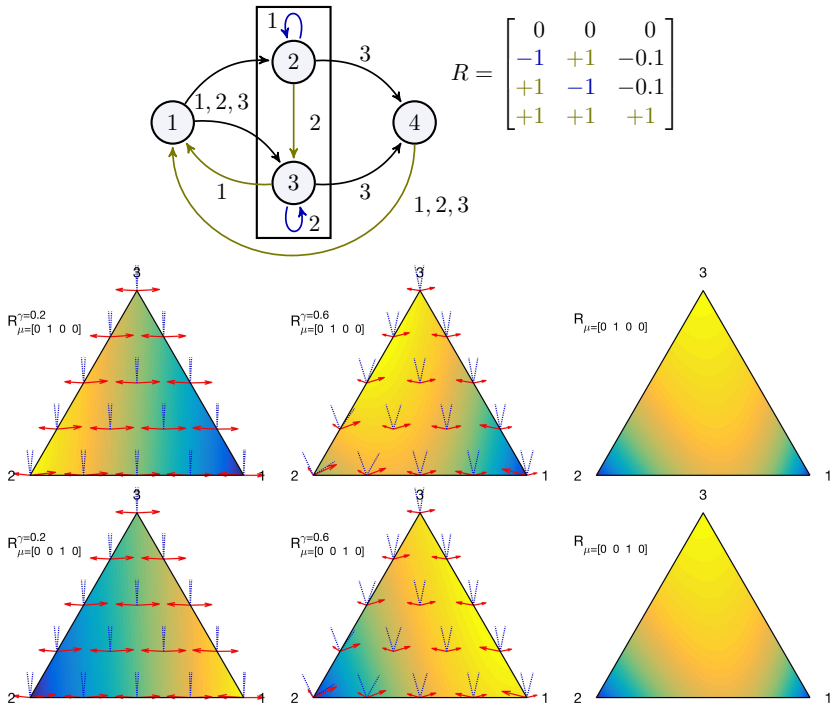

Fig. 2. Illustration of the example form Section 5. Top: State transitions and reward signal. Bottom: Numerical evaluation of the expected long term reward.

References

1. N. Ay, G. Montúfar, and J. Rauh. *Advances in Cognitive Neurodynamics (III)*, chapter Selection Criteria for Neuromanifolds of Stochastic Dynamics, pages 147–154. Springer Netherlands, 2013.
2. M. Hutter. General discounting versus average reward. In *Algorithmic Learning Theory 17*, pages 244–258. Springer Berlin Heidelberg, 2006.
3. S. Kakade. Optimizing average reward using discounted rewards. In *Computational Learning Theory 14*, pages 605–615. Springer Berlin Heidelberg, 2001.
4. G. Montúfar, K. Ghazi-Zahedi, and N. Ay. Geometry and determinism of optimal stationary control in partially observable Markov decision processes. *arXiv:1503.07206*, 2015.
5. S. M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, Inc., 1983.
6. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
7. R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, pages 1057–1063. MIT Press, 2000.
8. J. N. Tsitsiklis and B. Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49(2):179–191, 2002.