Characterizing Optimal Policies in Switched MDPs

Mihir Dhanakshirur

Supervised by: Prof. Gugan Thoppe

Abstract

We consider a two action, two environment switched MDP in the discounted reward setting. We restrict our analysis to the case where the Markov chain dictating the transitions between environments is action-independent. We present a function which, for each state, approximates the difference between Q-values for the two actions. We use properties of this approximated function to determine if a locally optimal policy is also globally optimal in the switched MDP.

1 Introduction

The problem of a system operating in switching environments has been examined across multiple domains [1]-[3]. Usually, the transition from one environment to another is determined by a Markov transition matrix. Contradictory phenomena are sometimes observed when we transition from a stationary environment to a *switched* environment. [4] considers the setting where there are two environments that are both favorable to the same species, but random switching between these two environments can lead to contrasting results: depending on the parameters of the transition matrix, coexistence of the two species, or extinction of the species favored by the environments can occur. Explaining such counter-intuitive behavior is the impetus for our work, in which we restrict our environments to Markov Decision Processes (MDPs). Switched Markov Decision Processes (SMDPs) accurately model real life scenarios in which an agent sequentially takes decisions in a time-varying environment. Our environment influences the transitions from one state to another (probability transition functions) and the rewards obtained after taking a particular action (reward functions). The agent has knowledge of locally optimal policies, i.e., policies that would, in a certain sense, maximise the long term reward obtained by the agent in each fixed environment. However, playing locally optimal actions may be sub-optimal in the SMDP (or global) setting. Consider a two environment case, where the transitions from one environment to the other are action-independent and extremely rapid. While a locally optimal policy maximises the **long-term** reward the agent receives, in a highly fluctuating environment, it may be wiser to maximise **short-term** and immediate rewards. In practice, there are numerous scenarios in which a locally optimal policy is outperformed by a globally optimal policy. It is thus imperative to characterize optimal policies under the switched MDP framework.

2 Setup and Main Result

2.1 Setup

In our work, we consider two Markov Decision Processes (MDPs) $\mathcal{M}_1$ and $\mathcal{M}_2$ in the discounted reward setting (discount factor γ). The MDPs have common state spaces $\mathcal{S}$ and common action spaces $\mathcal{A}$, but different reward functions $\mathcal{R}_1, \mathcal{R}_2 : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and different probability transition functions $\mathcal{P}_1, \mathcal{P}_2 : (\mathcal{S} \times \mathcal{A}) \times \mathcal{S} \to \mathbb{R}$. We restrict ourselves to the case where an agent can choose only one of two actions at each state, i.e., $|\mathcal{A}| = 2$. Using MDP $\mathcal{M}_1$ and $\mathcal{M}_2$ we can define a switched MDP $\mathcal{M}_s$. The action space of $\mathcal{M}_s$ is $\mathcal{A}$, identical to that of $\mathcal{M}_1$ and $\mathcal{M}_2$. Its state space $\mathcal{S}_s := \mathcal{S} \times \{1, 2\}$ however is twice the size of $\mathcal{S}$. Thus every state in $\mathcal{M}_s$ is a pair where the first entry describes the state the agent is at, and the second entry defines the corresponding probability transition matrix and reward functions involved after the agent takes an action. To define the probability transition matrix for the switched MDP $\mathcal{M}_s$ we require two additional parameters α and β. α and β are the action independent probabilities of staying in MDP $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. In other words, with probability $1 - \alpha$ (or $1 - \beta$), the agent transitions from MDP $\mathcal{M}_1$ to MDP $\mathcal{M}_2$

(or $\mathcal{M}_2$ to $\mathcal{M}_1$), independent of the action taken.

2.2 Main Result

We define the Q-function $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for each stationary policy π as:

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}^{\pi}(s_{t+1}) | s_0 = s, a_0 = a\right]$$

The optimal Q-function Q^* satisfies $Q^*(s,a) := Q^{\pi^*}(s,a) \geq Q^{\pi}(s,a)$ for all π and all s, a. For a state (s,i) in the switched MDP, denote $Q^{\pi_l^*}((s,i),\pi_i^*(s)) - Q^{\pi_l^*}((s,i),a)$ by $\Delta_{s,i}$, where $\pi_l^* := (\pi_1^*, \pi_2^*)$ is the locally optimal policy. Our main results are stated below:

1. $\Delta_{s,i}$ can be approximated *reasonably well* by a second order Taylor expansion, $\widetilde{\Delta}_{s,i}$, that is quadratic in α and linear in β.
2. The error in the approximation, $\left|\Delta_{s,i} - \widetilde{\Delta}_{s,i}\right|$, is bounded by $\frac{6\gamma^3}{(1-\gamma)^4}K$, where K is a constant depending on the reward functions.
3. Let $m_{s,i}$ denote the minimum value of the function $\widetilde{\Delta}_{s,i}$. For each $(s,i) \in \mathcal{S} \times \{1,2\}$, if $m_{s,i} - \frac{6\gamma^3}{(1-\gamma)^4}K \geq 0$, then the locally optimal policy is also globally optimal.

3 Analysis

At state (s,i) in the switched MDP, after taking action a, the agent receives a reward $\mathcal{R}_i(s,a)$. The transition from state (s,i) to state (s',j), having taken action a, is given by $P_s^{\alpha,\beta}((s',j)|(s,i),a) = P_i(s'|s,a)P(j|i)$ where:

$$P(j|i) := \begin{cases} \alpha & \text{if } i = j = 1 \\ 1-\alpha & \text{if } i = 1 \text{ and } j = 2 \\ \beta & \text{if } i = j = 2 \\ 1-\beta & \text{if } i = 2 \text{ and } j = 1 \end{cases}$$

Defining the probability transition matrix $\mathcal{P}_s^{\alpha,\beta}$ in such a way allows us to make use of the action independent, multiplicative property and rewrite $\mathcal{P}_s^{\alpha,\beta}$ as:

$$\mathcal{P}_s^{\alpha,\beta} = \begin{pmatrix} \alpha\mathcal{P}_1 & (1-\alpha)\mathcal{P}_1 \\ (1-\beta)\mathcal{P}_2 & \beta\mathcal{P}_2 \end{pmatrix}$$

where we have identified the probability transition matrices $\mathcal{P}_1$ and $\mathcal{P}_2$ as $(2|\mathcal{S}|, |\mathcal{S}|)$-dimensional matrices.

For each MDP $\mathcal{M}_i = (\mathcal{S}, \mathcal{A}, \mathcal{P}_i, \mathcal{R}_i)$ there exists a corresponding (not necessarily unique) optimal policy $\pi_i^* : \mathcal{S} \times \mathcal{A} \to [0,1]$. A policy π is considered optimal if for every policy π', $V^{\pi}(s) \geq V^{\pi'}(s) \, \forall s \in \mathcal{S}$, where V^{π} is the value function of the policy π given by:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}^{\pi}(s_{t+1}) | s_0 = s\right]$$

It also satisfies the Bellman Optimality Equation:

$$\begin{aligned} Q^*(s,a) &= \mathcal{R}(s,a) + \gamma\mathbb{E}\left[\max_{a' \in \mathcal{A}} Q^*(s_1,a') | s_0 = a, a_0 = a\right] \\ &= \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a)(\max_{a' \in \mathcal{A}} Q^*(s_1,a')) \end{aligned}$$

In particular, any deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ such that $\pi(s) \in \arg\max_a Q^*(s,a)$ is an optimal policy and $Q^{\pi} = Q^*$. From the above relation, we obtain the equivalence: a policy π is optimal if and only if $\pi(s) \in \arg\max_a Q^{\pi}(s,a) \; \forall s \in \mathcal{S}$. The Q-function and value function of a policy are related by the following equation:

$$\begin{aligned} Q^{\pi}(s,a) &= \mathcal{R}(s,a) + \gamma\mathbb{E}\left[V^{\pi}(s_1) | s_0 = s, a_0 = a\right] \\ &= \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s,a)V^{\pi}(s') \end{aligned}$$

For a policy π, the induced reward function and transition probability matrix are maps $\mathcal{R}^{\pi} : \mathcal{S} \to \mathbb{R}$ and $\mathcal{P}^{\pi} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ given by $\mathcal{R}^{\pi}(s) = \sum_{a \in \mathcal{A}} \mathcal{R}(s,a)\pi(s,a)$ and $\mathcal{P}^{\pi}(s'|s) = \sum_{a \in \mathcal{A}} P(s'|s,a)\pi(s,a)$ respectively. For the switched MDP $\mathcal{M}_s$, an optimal policy π_g^* is a mapping $\pi_g^* : \mathcal{S} \times \{1,2\} \to \mathcal{A}$, and we refer to it as the **globally optimal policy**. Any pair of local policies (π_1, π_2) for MDPs $\mathcal{M}_1$ and $\mathcal{M}_2$ can be identified with a global policy π_g for the switched MDP $\mathcal{M}_s$ and vice versa. This is through the bijective map $\pi_g((s,i),a) \leftrightarrow \pi_i(s,a)$.

We treat $\mathcal{P}_i$ as $(2N,N)$-dimensional matrices, $\mathcal{R}_i$ as $(2N,1)$-column vectors, $\mathcal{P}_i^{\pi}$ as (N,N)-dimensional matrices and $\mathcal{R}_i^{\pi}$ as $(N,1)$-column vectors for $i = 1,2$, where $|\mathcal{S}| = N$. $\mathcal{P}_i(\cdot|s)$ is a $(1,2N)$-row vector and $\mathcal{P}_i^{\pi}(\cdot|s)$ is a $(1,N)$-row vector. The following result allows us to directly compute the value function given its probability transition matrix and reward function:

Lemma 3.1 *For a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$,*

$$\begin{aligned} V^{\pi} &= (Id - \gamma\mathcal{P}^{\pi})^{-1}\mathcal{R}^{\pi} \\ &= \mathcal{R}^{\pi} + \gamma\mathcal{P}^{\pi}\mathcal{R}^{\pi} + \gamma^2(\mathcal{P}^{\pi})^2\mathcal{R}^{\pi} + \cdots . \end{aligned}$$

The second equation in 3.1 follows from the Neumann Expansion for inverse of matrices whose norm is less than 1. In this case, since $\mathcal{P}^{\pi}$ is row stochastic, i.e., the sum of entries in each row is 1, we can apply the Neumann Expansion. Using 3.1, we can evaluate the Q-function for a policy π using the following equation:

$$Q^{\pi}(s,a) = \mathcal{R}(s,a) + \gamma\mathcal{P}(\cdot|s,a)(\mathrm{Id} - \gamma\mathcal{P}^{\pi})^{-1}\mathcal{R}^{\pi}$$

Then, a policy π is optimal if and only if $\pi(s) \in \arg\max_a Q^{\pi}(s,a) \forall s \in \mathcal{S} \Leftrightarrow \pi(s) \in \arg\max_a(\mathcal{R}(s,a) + \gamma\mathcal{P}(\cdot|s,a)(\mathrm{Id} - \gamma\mathcal{P}^{\pi})^{-1}\mathcal{R}^{\pi})\ \forall s \in \mathcal{S}$. In the two action case, taking $a = \{1,2\} \setminus \pi(s)$, this boils down to checking if:

$$\begin{aligned} & Q^{\pi}(s,\pi(s)) - Q^{\pi}(s,a) = \mathcal{R}(s,\pi(s)) - \mathcal{R}(s,a) \\ & + \gamma(\mathcal{P}(\cdot|s,\pi(s)) - \mathcal{P}(\cdot|s,a))(\mathrm{Id} - \gamma\mathcal{P}^{\pi})^{-1}\mathcal{R}^{\pi} \qquad (1) \\ & \geq 0\ \forall s \in \mathcal{S} \end{aligned}$$

Suppose π_1^* and π_2^* are locally optimal policies for MDP $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively. Then from 1,

$$\begin{aligned} & \mathcal{R}_i(s,\pi_i^*(s)) - \mathcal{R}_i(s,a) \\ & + \gamma(\mathcal{P}_i(\cdot|s,\pi_i^*(s)) - \mathcal{P}_i(\cdot|s,a))(\mathrm{Id} - \gamma\mathcal{P}_i^{\pi_i^*})^{-1}\mathcal{R}_i^{\pi_i^*} \qquad (2) \\ & \geq 0\ \forall s \in \mathcal{S},\ i = 1,2 \end{aligned}$$

The pair (π_1^*, π_2^*) gives us a policy, say π_l^*, for the switched MDP $\mathcal{M}_s$ which we refer to as the **locally optimal** policy. While the locally optimal policy, in general, differs from the globally optimal policy π_g^*, for $\alpha = \beta = 1$ the locally optimal policy is also globally optimal. Intuitively, this addresses the case when an agent remains in the MDP it began in; in such a case, it is natural to expect that the optimal policy for that MDP remains to be optimal even in the switched MDP. To distinguish between locally and globally optimal policies for a broader range of α and β values, we need to understand the behavior of the difference of the Q-functions $Q^{\pi_l^*}((s,i),\pi_i^*(s)) - Q^{\pi_l^*}((s,i),a)$ for each state $(s,i) \in \mathcal{S}_s$, the state space of the switched MDP. As long as the difference remains positive for each state $(s,i) \in \mathcal{S}_s$, the locally optimal policy is also globally optimal.

For any $m \geq 1$, $n \geq 0$, we have the following constants:

- $\rho_{s,i}(\pi(s)) = \mathcal{P}_i(\cdot|s,\pi(s))(1-\gamma)(\mathrm{Id} - \gamma\mathcal{P}_i^{\pi_i^*})^{-1}$
- $\rho_{s,i}(a) = \mathcal{P}_i(\cdot|s,a)(1-\gamma)(\mathrm{Id} - \gamma\mathcal{P}_i^{\pi_i^*})^{-1}$
- $V_1^{\pi_1^*} = (\mathrm{Id} - \gamma\mathcal{P}_1^{\pi_1^*})^{-1}\mathcal{R}_1^{\pi_1^*}$
- $V_2^{\pi_2^*} = (\mathrm{Id} - \gamma\mathcal{P}_2^{\pi_2^*})^{-1}\mathcal{R}_2^{\pi_2^*}$
- $C_{m,n} = \sum_{|\delta|=n}(T - \mathrm{Id})^{\delta_1}(S - \mathrm{Id})(T - \mathrm{Id})^{\delta_2}(S - \mathrm{Id}) \cdots (S - \mathrm{Id})(T - \mathrm{Id})^{\delta_m}$ where $S = (\mathrm{Id} - \gamma\mathcal{P}_i^{\pi_i^*})^{-1}$ and $T = (\mathrm{Id} - \gamma\mathcal{P}_j^{\pi_j^*})^{-1}$

The quantities $\rho_{s,i}(\pi(s))$ and $\rho_{s,i}(a)$ can be interpreted as the vector of *state visitation frequencies* given the initial state distribution $\mathcal{P}_i(\cdot|s,\pi(s))$ and $\mathcal{P}_i(\cdot|s,a)$ respectively. We now obtain expressions for higher order partial derivatives at $(1,1)$.

Theorem 3.2 *Denote $Q^{\pi_l^*}((s,i),\pi_i^*(s)) - Q^{\pi_l^*}((s,i),a)$ by $\Delta_{s,i}$. Let $j = \{1,2\} \setminus i$. We have the following expression for higher order partial derivatives of $\Delta_{s,i}$ with respect to α and β evaluated at $(1,1)$:*

1. *If $m \geq 1$, $n \geq 0$, then:*

$$\left.\frac{\partial^{m+n}\Delta_{s,i}}{\partial\alpha^m\partial\beta^n}\right|_{(1,1)} = m!n!\left(\frac{\gamma}{1-\gamma}\right)(\rho_{s,i}(\pi_i^*(s)) - \rho_{s,i}(a)) \\ C_{m,n}(V_i^{\pi_i^*} - V_j^{\pi_j^*}) \tag{3}$$

2. *If $m = 0$, then $\forall n \geq 1$:*

$$\left.\frac{\partial^{n}\Delta_{s,i}}{\partial\beta^n}\right|_{(1,1)} = 0 \tag{4}$$

The partial derivatives are calculated at $(1,1)$ since the $(2N,2N)$-dimensional matrix $(\mathrm{Id} - \gamma\mathcal{P}_s^{\pi_l^*})^{-1}$ reduces to the block matrix:

$$\begin{pmatrix} \left(\mathrm{Id} - \gamma\mathcal{P}_1^{\pi_1^*}\right)^{-1} & 0 \\ 0 & \left(\mathrm{Id} - \gamma\mathcal{P}_2^{\pi_2^*}\right)^{-1} \end{pmatrix}$$

which is easier to compute. Moreover, it makes sense to approximate $\Delta_{s,i}$ around $(1,1)$ since we can always find a neighbourhood around it where the locally optimal policy is also globally optimal. Using Taylor's Theorem in Several Variables, we obtain the following result,

Theorem 3.3 *We have the second order Taylor series approximation of $\Delta_{s,i}$ for $0 \leq \alpha, \beta \leq 1$:*

$$\begin{aligned}
&\widetilde{\Delta}_{s,i}(\alpha,\beta) \approx \mathcal{R}_i(s,\pi_i^*(s)) - \mathcal{R}_i(s,a) + \\
&\gamma(\rho_{s,i}(\pi(s)) - \rho_{s,i}(a))(S-T) \\
&+ \frac{\gamma}{1-\gamma}(\rho_{s,i}(\pi(s)) - \rho_{s,i}(a))(V_i^{\pi_i^*} - V_j^{\pi_j^*})(\alpha-1) \\
&+ \frac{\gamma}{1-\gamma}(\rho_{s,i}(\pi(s)) - \rho_{s,i}(a))(S-Id)(V_i^{\pi_i^*} - V_j^{\pi_j^*})(\alpha-1)^2 \\
&+ \frac{\gamma}{1-\gamma}(\rho_{s,i}(\pi(s)) - \rho_{s,i}(a))(T-Id)(V_i^{\pi_i^*} - V_j^{\pi_j^*})(\alpha-1) \\
&(\beta-1)
\end{aligned} \tag{5}$$

The next result bounds the error between $\Delta_{s,i}$ and its second order Taylor approximation, $\widetilde{\Delta}_{s,i}$:

Theorem 3.4 *For all $\alpha, \beta \in [0,1]$:*

$$\left|\Delta_{s,i} - \widetilde{\Delta}_{s,i}\right| \leq \frac{6\gamma^3}{(1-\gamma)^4} K \tag{6}$$

where

$$\begin{aligned}
K = &\max_s(\mathcal{R}_1^{\pi_1^*}(s)) + \max_s(\mathcal{R}_2^{\pi_2^*}(s)) - \min_s(\mathcal{R}_1^{\pi_1^*}(s)) \\
&- \min_s(\mathcal{R}_2^{\pi_2^*}(s))
\end{aligned}$$

Let $m_{s,i}$ denote the minimum value of the function $\widetilde{\Delta}_{s,i}$. We now present a test to discern if the locally optimal policy is globally optimal:

Theorem 3.5 *For each $(s,i) \in \mathcal{S} \times \{1,2\}$, if $m_{s,i} - \frac{6\gamma^3}{(1-\gamma)^4}K \geq 0$, then $\pi_l^* = (\pi_1^*, \pi_2^*)$, the locally optimal policy is also globally optimal.*

4 Empirical Results

For a 3 state, 2 action switched MDP, we plot $\Delta_{s,i}$ and $\widetilde{\Delta}_{s,i}$ for $\alpha, \beta \in [0,1]$ and different discount factors γ. Additionally, we randomly sample probability transition functions and sample rewards uniformly from $[0, 100]$. Empirically, we observe that $\widetilde{\Delta}_{s,i}$ approximates $\Delta_{s,i}$ well for a range of γ values. Moreover, in these cases, it is sufficient to check that $\widetilde{\Delta}_{s,i} \geq 0$ to conclude that $\Delta_{s,i} \geq 0$.

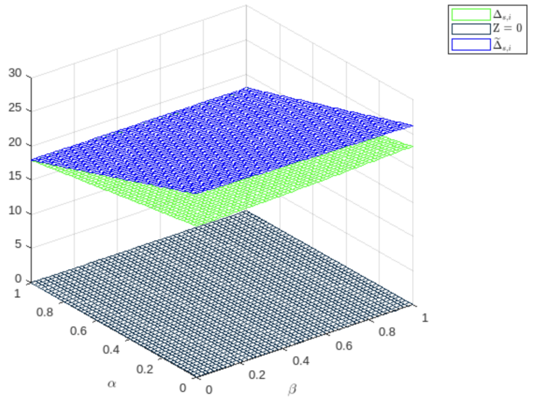

Figure 1: $\gamma = 0.6$

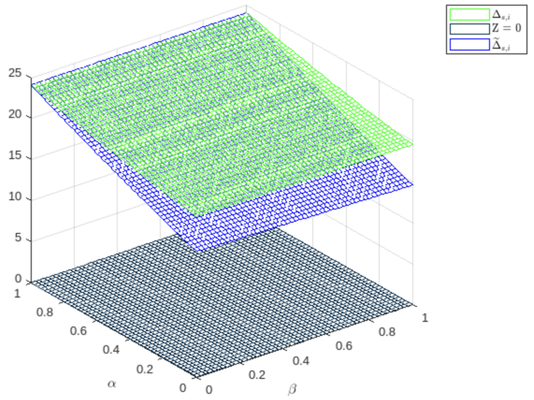

Figure 2: $\gamma = 0.4$

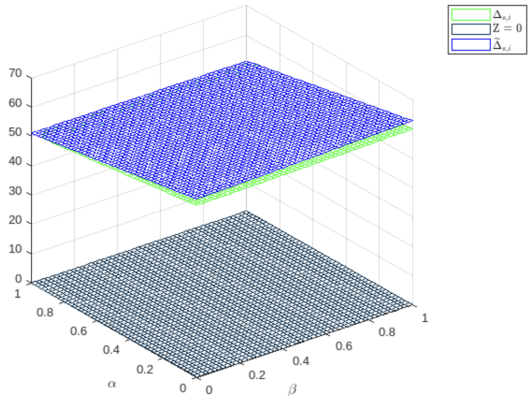

Figure 3: $\gamma = 0.2$

References

[1] G. Ackerson and K. Fu, "On state estimation in switching environments," in IEEE Transactions on Automatic Control, vol. 15, no. 1, pp. 10-17, February 1970, doi: 10.1109/TAC.1970.1099359.

[2] Thompson Sampling in Switching Environments with Bayesian Online Change Detection Joseph Mellor, Jonathan Shapiro Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, PMLR 31:442-450, 2013.

[3] J. Müller, B.A. Hense, T.M. Fuchs, M. Utz, Ch. Pötzsche, Bet-hedging in stochastically switching environments, Journal of Theoretical Biology,Volume 336, 2013, Pages 144-157, ISSN 0022-5193

[4] Benaïm, M., Lobry, C. (2016). LOTKA-VOLTERRA WITH RANDOMLY FLUCTUATING ENVIRONMENTS OR "HOW SWITCHING BETWEEN BENEFICIAL ENVIRONMENTS CAN MAKE SURVIVAL HARDER." The Annals of Applied Probability, 26(6), 3754–3785.