University of Mannheim

Bachelor Thesis

Markov-Decision Processes

by

Felix Benning

born on the 27.11.1996 in Nürtingen
matriculation number 1501817

in the
Fakulty for Mathematics in Business and Economics
Supervisor: Prof. Dr. Leif Döring

Due Date: ???

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.

This thesis was not previously presented to another examination board and has not been published.

City, Date Signature

Contents

Chapter 1

Markov Decision Processes

1.1 Introduction

A Markov Process is a random process in a state space with no memory of where it was, that is, only the current state influences where it goes next. While Markov Processes allow to model random phenomena evolving over time and make predictions about certain events (e.g. terminal states), they are unable to model the interaction of an actor with such a processes. *Markov Decision Processes* (MDPs) introduce *actions* and *rewards* to the state space and transition probabilities of Markov Processces, and shift the focus from *describing* terminal distributions, absorption times, etc. towards *finding* the optimal action(s) to take in each state (if such an action exists).

The MDP model inherits the restriction of Markov Chains to have no memory of past states. We will also not consider changing transition probabilities over time. Rather the transition probabilities will only be influenced by the state and the action.
Both of these limitations could in principle be circumvented by including the time in the state space at the expense of a larger state space. Although it is questionable whether such a construct would yield any interesting results, as then no state is visited twice. So it is of no use to an actor to learn the value of an action in a certain state without further assumptions.

To illustrate the uses of such a framework, I have selected a few examples from White (1985):

1. Resource Management: The state is the resource level
    - Inventory Management: The resource is the inventory, the possible action is to order resupply, influencing the inventory (state) together with the stochastic demand, and the reward is the profit. The essential trade-off is the cost of storage versus lost sales from a stock-out.
    - Fishing: The resource is the amount of fish, the action is the amount fished, the reward is directly proportional to the amount fished, and

the repopulation is the random element.

- Pumped storage Hydro-power: The state is the amount of water in the higher reservoir and the electricity price, the action is to use water to generate electricity or wait for higher prices.
- Beds in a hospital: How many empty beds are needed for emergencies?

2. Stock trading: The state is the price level and stock and liquidity owned.
3. Maintenance: When does a car/road become too expensive to repair?
4. Evacuation in response to flood forecasts

1.2 Model Formulation

Most of the definitions in this chapter are adaptions from Szepesvári (2010). But to properly define the transition probabilities given an action in a certain state, let us define a probability kernel first.

Definition 1.2.1. (Kernel) Let $(Y, \sigma_Y), (X, \sigma_X)$ be measure spaces.

$$\begin{aligned} &\lambda\colon X \times \sigma_Y \to \mathbb{R} \text{ is a } \textit{(probability) kernel} \\ :\iff\ &\lambda(\cdot, A)\colon x \mapsto \lambda(x, A) \text{ measurable} \\ &\lambda(x, \cdot)\colon A \mapsto \lambda(x, A) \text{ a (probability) measure} \end{aligned}$$

Since we will interpret probability kernels as distributions over Y given a certain condition $x \in X$, the notation $\lambda(\cdot \mid x) \coloneqq \lambda(x, \cdot)$ helps this intuition.

Definition 1.2.2.
$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ is called a *(finite) Markov Decision Process* (MDP). Where:
$\mathcal{X}$ is a countable (finite) set of states.
$\mathcal{A}$ is a countable (finite) set of actions.
And $\mathcal{P}_0\colon (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X}\times\mathbb{R}} \to \mathbb{R}$ is a probability kernel.

$\mathcal{X} \times \mathbb{R}$ represents the next state and the reward. So $\mathcal{P}_0(\cdot \mid x, a)$ represents the probability distribution over the next states and rewards given an action a in the state x.

$\mathcal{P}_0$ is called the *transition probability kernel* or in short, transition kernel.

Remark 1.2.3. Instead of a transition probability kernel $\mathcal{P}_0$, sometimes a *transition function* f with a and an exogenous random element D_t (e.g. Demand) is used to define the next state and reward: $(X_{t+1}, R_{t+1}) = f(X_t, A_t, D_t)$

Some authors include a Time set T in the tuple (e.g. Puterman 2014) this allows for finite horizons but not for continuous time, since the transition kernel

is defined for discrete steps. Most authors split the transition kernel into a state transition kernel and a reward kernel (e.g. Puterman 2014). But since it is easier to define a marginal distribution from a joint distribution than vice versa, and since this notation is more compact I will stick to the definition from Szepesvári (2010).

According to Puterman (2014) some authors call this tuple a Markov Decision Problem instead of Markov Decision Process, presumably to reserve the term Markov Decision Process for the resulting sequence of states, actions and rewards $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$, aligning the Definition with the definition of a Markov process. Although this does not appear to be common practice.

I can find no explanation for this deviation from the notation of Markov processes. So I offer my own interpretation:

The objective of the theory of MDPs is to find an optimal action selection rule (behavior). And without a fixed behavior the sequence $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$ is undefined, since the $(A_t, t \in \mathbb{N}_0)$ are not defined. But fixing the behavior defeats the purpose of modeling decisions. As it would not make sense to talk about optimal behaviors in an MDP if every behavior creates its own MDP.

Nevertheless we still need to construct a stochastic process from the MDP when we have an action selection rule.

First we need to select the random variable X_0 of the initial state. The initial state is not included in the definition of an MDP because later objects will be defined conditional on the current state. They are thus invariant to different starting distributions, as long as $\mathbb{P}(X_0 = x) > 0$ holds for all $x \in \mathcal{X}$ ensuring that conditioning on every state is possible.

To inductively define a stochastic process we need an action selection rule, more formally:

Definition 1.2.4. An A_t selection-rule $\pi = (\pi_t, t \in \mathbb{N}_0)$ is called *behavior*, where

$$\pi_t \colon \begin{cases} [(\mathcal{X} \times \mathcal{A} \times \mathbb{R})^t \times \mathcal{X}] \times \sigma_{\mathcal{A}} \to \mathbb{R} \\ (y, A) \mapsto \pi_t(A \mid y) \end{cases} \quad \text{is a probability kernel,}$$

and $A_t \sim \pi_t(\cdot \mid (X_0, A_0, R_1), \ldots, (X_{t-1}, A_{t-1}, R_t), X_t))$.
Special cases:

1. *Determinisitic stationary policies* specified with some abuse of notation:

$$\pi \colon \mathcal{X} \to \mathcal{A} \quad \text{with } A_t = \pi(X_t)$$

2. *(Stochastic) stationary policies* specified by:

$$\pi \colon \begin{cases} \mathcal{X} \times \sigma_{\mathcal{A}} \to \mathbb{R} \\ (x, A) \mapsto \pi(A \mid x) \end{cases} \quad \text{with } A_t \sim \pi(\cdot \mid X_t)$$

A_t is here selected such that it has the markov property (well defined c.f. 1.2.5), i.e.:

$$\mathbb{P}[A_t = a \mid X_t] = \mathbb{P}[A_t = a \mid X_t, (X_{t-1}, A_{t-1}, R_t), \dots (X_0, A_0, R_1)]$$

Π is the set of behaviors,
Π_{stat} is the set of (stochastic) stationary policies,
$\Pi^{\text{det}}_{\text{stat}}$ is the set of deterministic stationary policies (note $\Pi^{\text{det}}_{\text{stat}} \subseteq \Pi_{\text{stat}} \subseteq \Pi$)

Now we define inductively: $(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$ with the Markov property (well defined c.f. 1.2.5), i.e.:

$$\begin{aligned} &\mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t)] \\ &= \mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t), (X_{t-1}, A_{t-1}, R_t), \dots (X_0, A_0, R_1)] \end{aligned} \tag{1.1}$$

resulting in the stochastic process $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$

Remark 1.2.5. $(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$ with the Markov property, is well defined i.e.:

$$\begin{aligned} &\exists (X_{t+1}, R_{t+1})\ \mathcal{X} \times \mathbb{R}\text{-valued random variable :} \\ &(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t) \text{ and satisfies the Markov property} \end{aligned}$$

(analogous A_t well defined)

Proof. TODO □

Remark 1.2.6. A stationary policy π induces a *time-homogeneous* Markov chain $(X_t, A_t, R_{t+1}, t \in \mathbb{N}_0)$.

Proof. $\mathcal{H}^t_s \coloneqq \{(X_t, A_t, R_{t+1}) \in H_t, \dots, (X_s, A_s, R_{s+1}) \in H_s\},\ H_i \in \sigma_{\mathcal{X} \times \mathcal{A} \times \mathbb{R}}$
Because of

$$\mathbb{P}(A \cap B \mid C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} = \mathbb{P}(A \mid B \cap C)\mathbb{P}(B \mid C)$$

we can show:

$$\begin{aligned} &\mathbb{P}\left[(X_t, A_t, R_{t+1}) \in \{(x, a)\} \times U \mid \mathcal{H}^{t-1}_0\right] \\ &= \mathbb{P}\left[R_{t+1} \in U \mid (X_t, A_t) = (x, a), \mathcal{H}^{t-1}_0\right] \mathbb{P}\left[(X_t, A_t) = (x, a) \mid \mathcal{H}^{t-1}_0\right] \\ &\overset{(1.1)}{=} \mathbb{P}\left[R_{t+1} \in U \mid (X_t, A_t) = (x, a)\right] \underbrace{\mathbb{P}\left[A_t = a \mid X_t = x\right]}_{=\pi(a|x)} \mathbb{P}\left[X_t = x \mid \mathcal{H}^{t-1}_{t-1}\right] \\ &\overset{(*)}{=} \mathbb{P}\left[R_{t+1} \in U \mid (X_t, A_t) = (x, a), \mathcal{H}^{t-1}_{t-1}\right] \mathbb{P}\left[(X_t, A_t) = (x, a) \mid \mathcal{H}^{t-1}_{t-1}\right] \\ &= \mathbb{P}\left[(X_t, A_t, R_{t+1}) \in \{(x, a)\} \times U \mid \mathcal{H}^{t-1}_{t-1}\right] \end{aligned}$$

$(*)$ *Some* of the History is irrelevant if all of the History is irrelevant. (appendix?, TODO: time-homogeneous) □

Remark 1.2.7. Sometimes not all actions make sense in all states. A simple fix would be to set the immediate reward functions for those actions very low, or (if possible) redirect them to the closest possible action.
A more formal approach would be to introduce an additional mapping, which assigns the set of admissible actions to each state $\mathcal{X} \to \mathcal{P}(\mathcal{A})$ and define $\mathcal{X} \times \mathcal{A} \coloneqq \{(x, a) : x \in \mathcal{X}, a \in f(x)\}$.

If there is just one admissible action in every state, the MDP is equivalent to a normal Markov Process. Since then there is a mapping $f \colon \mathcal{X} \to \mathcal{A}$ mapping the state to the only admissible action. Which implies that $A_t = f(X_t)$ which forces every behavior to be equal to f. And since f is a deterministic stationary behavior, 1.2.6 applies.

Definition 1.2.8. An MDP together with a discount factor $\gamma \in [0, 1]$ is a
discounted reward MDP for $\gamma < 1$
undiscounted reward MDP for $\gamma = 1$
This allows us to define the *return*:

$$\mathcal{R} \coloneqq \sum_{t=0}^{\infty} \gamma^t R_{t+1}$$

Definition 1.2.9. Let $(Y_{(x,a)}, R_{(x,a)}) \sim \mathcal{P}_0(\cdot \mid x, a)$ be a random variable.

$$r(x, a) \coloneqq \mathbb{E}[R_{(x,a)}] \quad \text{is the } \textit{immediate reward function.}$$

Definition 1.2.10. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ a MDP
$x \in \mathcal{X}$ is a *terminal (absorbing)* state $:\iff \forall s \in \mathbb{N} : \mathbb{P}(X_{t+s} = x \mid X_t = x) = 1$
An MDP with such states is called *episodic*.
An *episode* is the random time period $(1, \dots, T)$ until a terminal state is reached.

Remark 1.2.11.

- The reward in a terminal state is by convention zero, i.e. x terminal state implies $\forall a \in \mathcal{A} : R_{(x,a)} = 0$.
- Episodic MDPs are often undiscounted

To avoid clutter we will from now on we assume an underlying MDP with the accompanying definitions and notation.

1.3 Value functions

The goal in this section is to

- define Value functions which assign states (and actions) a value, which allow the agent to make a more nuanced decisions than comparing immediate rewards of different actions

- explore the relation of different value functions
- show uniqueness of optimal value functions with the Banach fixpoint theorem, yielding a simple approximation methode along the way
- demonstrate that in MDPs deterministic stationary policies are generally a large enough set of policies to choose from

Definition 1.3.1. Let π be a behavior. Select X_0 such that $\forall x \in \mathcal{X} : \mathbb{P}(X_0 = x) > 0$, be $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$ the resulting stoch. process.

$$V^\pi \colon \begin{cases} \mathcal{X} \to \mathbb{R} \\ x \mapsto \mathbb{E}[\mathcal{R} \mid X_0 = x] \end{cases} \quad \text{is the } \textit{value function} \text{ for } \pi^1$$

$$Q^\pi \colon \begin{cases} \mathcal{X} \times \mathcal{A} \to \mathbb{R} \\ (x, a) \mapsto \mathbb{E}[\mathcal{R} \mid X_0 = x, A_0 = a] \end{cases} \quad \text{is the } \textit{action value function} \text{ for } \pi^2$$

$$V^* \colon \begin{cases} \mathcal{X} \to \mathbb{R} \\ x \mapsto \sup\limits_{\pi \in \Pi} V^\pi(x) \end{cases} \quad \text{is the } \textit{optimal value function}$$

$$Q^* \colon \begin{cases} \mathcal{X} \times \mathcal{A} \to \mathbb{R} \\ (x, a) \mapsto \sup\limits_{\pi \in \Pi} Q^\pi(x, a) \end{cases} \quad \text{is the } \textit{optimal action value function}$$

π is *optimal* $:\iff V^* = V^\pi$

Remark 1.3.2. With the distribution of X_0 set (or X_0 being realized with a fixed value x), the distribution of X_t, A_t, R_{t+1} is determined for all $t \in \mathbb{N}_0$. The conditional expectation is thus unique for a given $X_0 = x$, for all possible realizations of the MDP with a given behavior.
This means V^π, Q^π are well defined.

As mentioned previously, the marginal probability distribution of the state instead of the joint distribution with the reward, will now makes some notation shorter.

Definition 1.3.3.

$$p \colon \begin{cases} (\mathcal{X} \times \mathcal{A}) \times \sigma_{\mathcal{X}} \to \mathbb{R} \\ (x, a, Y) \mapsto \mathcal{P}_0(Y \times \mathbb{R} \mid x, a) \end{cases} \quad \text{is the } \underline{\textit{state}} \textit{ transition kernel.}$$

Notation: $p(y \mid x, a) \coloneqq p(\{y\} \mid x, a)$ with $(x, a, y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$

Now we can start to explore the relation of V^π and Q^π.

[1] Well defined because $\mathbb{P}(X_0 = x) > 0$
[2] Well defined because $A_1 \sim \pi_1(\cdot \mid (x, a, r_0), x_1)$ is defined for all a regardless of π_0

Proposition 1.3.4. *Let $\pi \in \Pi^{\text{det}}_{\text{stat}}$ be a deterministic stationary behavior, then:*

$$Q^\pi(x,a) = r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x,a) V^\pi(y)$$

Proof. split equation or accept eq sticking out a bit?

$$\begin{aligned}
Q^\pi(x,a) &= \mathbb{E}[\mathcal{R}(\pi) \mid X_0 = x, A_0 = a] \\
&= \mathbb{E}[R_1(\pi) \mid X_0 = x, A_0 = a] + \gamma \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \middle| X_0 = x, A_0 = a\right] \\
&= \mathbb{E}[R_{(x,a)}] + \gamma \sum_{y \in \mathcal{X}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \middle| X_0 = x, A_0 = a, X_1 = y\right] p(y \mid x,a) \\
&\overset{\text{Markov}}{=} r(x,a) + \gamma \sum_{y \in \mathcal{X}} \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \middle| X_1 = y, A_1 = \pi(y)\right]}_{= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+2}(\pi) \middle| X_1 = y\right]} p(y \mid x,a) \\
&\overset{(*)}{=} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \tilde{R}_{t+1}(\pi) \middle| \tilde{X}_0 = y\right] = V^\pi(y)
\end{aligned}$$

$(*)$ Rename: $\tilde{X}_t := X_{t+1}, \tilde{A}_t := A_{t+1}, \tilde{R}_t := R_{t+1}$, then $(\tilde{X}_t, \tilde{A}_t, \tilde{R}_{t+1}, t \in \mathbb{N}_0)$ is an "evaluation"/ "Markov Action Process" of the MDP with the (stationary!) policy π. □

Corollary 1.3.5. *For $\pi \in \Pi^{\text{det}}_{\text{stat}}$ this fix-point equation holds:*

$$\begin{aligned}
V^\pi(x) &= Q^\pi(x, \pi(x)) \\
&= r(x, \pi(x)) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi(x)) V^\pi(y)
\end{aligned}$$

and this fix-point equation:

$$Q^\pi(x,a) = r(x,a) + \gamma \sum_{y \in \mathcal{X}} p(y \mid x,a) Q^\pi(x, \pi(x))$$

Proof. Since π is a deterministic stationary policy:

$$V^\pi(x) = \mathbb{E}[\mathcal{R}(\pi) \mid X_0 = x] = \mathbb{E}[\mathcal{R}(\pi) \mid X_0 = x, A_0 = \pi(x)] = Q^\pi(x, \pi(x))$$

The rest follows from 1.3.4 □

With this relation we can use the Banach fix-point theorem (BFT) for the first time. But to do that we first need to make an assumption.

Assumption 1. $\forall (x,a) \in \mathcal{X} \times \mathcal{A}: \quad \mathbb{E}[|R_{(x,a)}|] \leq R \in \mathbb{R}$

This implies that the immediate reward is bounded:

$$\|r\|_\infty = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} |\mathbb{E}[R_{(x,a)}]| \leq R$$

But more importantly:

$$\mathbb{E}[|\mathcal{R}|] \leq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t |R_{t+1}|\right] \leq \frac{R}{1-\gamma}$$

which implies implies $V^\pi, V^* \in B(\mathcal{X}), Q^\pi, Q^* \in B(\mathcal{X}\times\mathcal{A})$ with $B(M) \coloneqq \{f: M \to \mathbb{R} : \|f\|_\infty < \infty\}$ the set of bounded functions on $\mathcal{X}$, which happens to be a complete metric space which we will need for the Banach fix-point Theorem to work. Bounded Value functions will also be neccessary for a lot of other arguments later on. In finite MDPs this assumption is of course always fulfilled.

Definition 1.3.6. For policy $\pi \in \Pi^{\text{det}}_{\text{stat}}$ is the mapping $T^\pi: B(\mathcal{X}) \to B(\mathcal{X})$ with:

$$T^\pi V(x) \coloneqq r(x,\pi(x)) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x, \pi(x)) V(y) \qquad V \in B(\mathcal{X}), x \in \mathcal{X}$$

called the *Bellman Operator*. With some abuse of notation, define the mapping $T^\pi: B(\mathcal{X}\times\mathcal{A}) \to B(\mathcal{X}\times\mathcal{A})$ with:

$$T^\pi Q(x,a) \coloneqq r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x, a) Q(y, \pi(y)) \quad V \in B(\mathcal{X}), (x,a) \in \mathcal{X}\times\mathcal{A}$$

Remark 1.3.7. $\forall \pi \in \Pi^{\text{det}}_{\text{stat}} : T^\pi V^\pi = V^\pi$ (c.f. 1.3.5)

T^π meets the requirements of the Banach fixed-point theorem appendix? for $\gamma < 1$, this implies that V^π for $\pi \in \Pi^{\text{det}}_{\text{stat}}$ is a *unique* fixpoint and can be approximated with the canonical iteration

Proof. $(B(\mathcal{X}), \|\cdot\|_\infty)$ is a non-empty, complete metric space appendix? and the mapping maps onto itself. It is left to show, that T^π is a contraction. Be $V, W \in B(\mathcal{X})$:

$$\begin{aligned}
\|T^\pi V - T^\pi W\|_\infty &= \|\gamma \sum_{y\in\mathcal{X}} p(y \mid \cdot, \pi(\cdot))(V(y) - W(y))\|_\infty \\
&\leq \gamma \sup_{x\in\mathcal{X}} \left\{ \sum_{y\in\mathcal{X}} p(y \mid x, \pi(x)) \|V - W\|_\infty \right\} \\
&= \gamma \|V - W\|_\infty \sup_{x\in\mathcal{X}} \underbrace{\left\{ \sum_{y\in\mathcal{X}} p(y \mid x, \pi(x)) \right\}}_{=1} \\
&= \gamma \|V - W\|_\infty
\end{aligned}$$

The proof for $T^\pi \colon B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ is analogous. □

Remark 1.3.8. Some observations which will come in useful later:

1. T^π is an affine operator
2. $W_1, W_2 \in B(\mathcal{X})$, write $W_1 \leq W_2$ for $\forall x \in \mathcal{X} : W_1(x) \leq W_2(x)$, then:

$$W_1 \leq W_2 \implies T^\pi W_1 \leq T^\pi W_2$$

Proof. Be $W_1, W_2 \in B(\mathcal{X})$, $W_1 \leq W_2$ and $x \in \mathcal{X}$:

$$T^\pi W_2(x) - T^\pi W_1(x) = \gamma \sum_{y \in \mathcal{X}} p(y \mid x, \pi(x)) \underbrace{(W_2(y) - W_1(y))}_{\geq 0} \geq 0 \qquad \square$$

Now we get to the more interesting but also harder optimal value functions. We will later see that taking the supremum over all behaviors is the same as taking it just over the deterministic stationary behaviors. But for now we need to make the distinction:

Definition 1.3.9.

$$\tilde{V}(x) \coloneqq \sup_{\pi \in \Pi^{\text{det}}_{\text{stat}}} V^\pi(x)$$
$$\tilde{Q}(x, a) \coloneqq \sup_{\pi \in \Pi^{\text{det}}_{\text{stat}}} Q^\pi(x, a)$$

The goal is to show that these two pairs of optimal value functions are actually the same using the uniquness of the fix-point of the BFT.

The intuition for why this should be the case comes from the fact that the MDP is memory-less. So winding up in the same state results in the agent having the exact same decision problem. And if the agent decided for one optimal action in the past, then this action will again be optimal. For this reason the supremum over all behaviors should be equal to the supremum over stationary behaviors.

Deterministic policies are enough, because if an optimal policy randomizes over different actions, then the values of theses actions must all be equally high. Which means that just picking one and sticking with it should be just as good.

But before we get to tackle this problem, we first need to adress another one. Notice how we take the supremum for every state x individually?

$$\tilde{V}(x) = \sup_{\pi \in \Pi^{\text{det}}_{\text{stat}}} V^\pi(x)$$

Since the stationary policy still allows us to condition on the state it is intuitive to assume that we do not need different sequences of policies to apporximate V^* for each state $x \in \mathcal{X}$. This will eventually turn out to be true using statements

about the optimal value functions. We can not show this fact right now, since sequences approximating the different suprema are indexed by the possibly countable state space, what we would need is a single sequence of policies which matches all the policies in this countable set of sequences in every state. This turns out to be an impossible task.

Example 1.3.10. (The genie cubicles) Imagine an infinite (countable) amount of cubicles $\mathbb{N} \subset \mathcal{X}$, with a genie in every cubicle. You can wish for an arbitrary amount of money $\mathcal{A} = \mathbb{N}$, after that you have to leave (end up in the terminal state 0, i.e.: $\mathcal{X} = \mathbb{N}_0$). Then of course $V^*(x) = \infty$ for $x \in \mathbb{N}$, is achived by the seqences of behaviors: $(\pi_x^{(n)}, n \in \mathbb{N})$ for $x \in \mathbb{N}$ with $\pi_x^{(n)}(y) = x + n \quad \forall y \in \mathcal{X}$. Then there is no policy $\pi^{(n)}$ which can match every $\pi_x^{(n)}$ for all $x \in \mathbb{N}$.

Even if $\tilde{V}$ was finite, we could modify the example by cutting gifts off above a certain threshold with behaviors approaching that threshold.

So we can not match an infinite set of behaviors. Which means we will have to settle for a finite version. This version will help us to handle the optimal value functions. With the later facts about optimal value functions we can then define a sequence of policies which approach the optimal value functions uniformly confirming our earlier suspicions, that the suprema can be attained with a single sequence of policies.

Proposition 1.3.11. *Be $n \in \mathbb{N}$ and $\{\pi_1, \dots, \pi_n\} \subseteq \Pi_{\text{stat}}^{\text{det}}$, then:*

$$\exists \hat{\pi} \in \Pi_{\text{stat}}^{\text{det}} : \quad \max_{i=1,\dots,n} V^{\pi}(x) \leq V^{\hat{\pi}}(x) \qquad \forall x \in \mathcal{X}$$

Proof. The idea is to to pick the same action in state x, as the policy which generates the most value out of this state, i.e. $\max_{i=1,\dots,n} V^{\pi_i}(x)$.

$$\hat{\pi}: \begin{cases} \mathcal{X} \to \mathcal{A} \\ x \mapsto \pi_{\underset{i=1,\dots,n}{\text{argmax}\, V^{\pi_i}(x)}}(x) \end{cases}$$

One might be surprised that such an appearingly short sighted policy should surpass every policy in the finite set. At first it might seem that different policies achive their values of their state through different, maybe incompatible strategies. Would a strategy which takes a policy because it changes transition probabilities in a way that leads to a high-payoff state not be sabotaged by switching policies erratically? It is important to realize that if a policy A attaches a high value to a state because it provides easy access to states which can yield a high payoff, then the other policies will either exploit the high payoff later on as well, or the policy A will once again be the maximum. Either way it would result in $\hat{\pi}$ exploiting the high payoff later on.

For the maximum value write: $V(x) \coloneqq \max_{i=1,\dots,n} V^{\pi_i}(x)$ and be $m(x) \coloneqq \arg\max_{i=1,\dots,n} V^{\pi_i}(x)$, then:

$$\begin{aligned} T^{\hat{\pi}}V(x) &= r(x, \pi_{m(x)}(x)) + \gamma \sum y \in \mathcal{X} p(y \mid x, \pi_{m(x)}(x)) V(y) \\ &\overset{V \geq V^{\pi_{m(x)}}}{\geq} r(x, \pi_{m(x)}(x)) + \gamma \sum y \in \mathcal{X} p(y \mid x, \pi_{m(x)}(x)) V^{\pi_{m(x)}}(y) \\ &\overset{1.3.5}{=} V^{\pi_{m(x)}}(x) = V(x) \end{aligned}$$

By using the monotonicity of $T^{\hat{\pi}}$ (1.3.8) inductively with $(T^{\hat{\pi}})^1 V \geq (T^{\hat{\pi}})^0 V$, we get $(T^{\hat{\pi}})^n V \geq (T^{\hat{\pi}})^{n-1} V$ thus:

$$V^{\hat{\pi}}(x) = \lim_{n \to \infty} (T^{\hat{\pi}})^n V(x) \geq V(x) \geq V^{\pi_i}(x) \qquad \forall i = 1, \dots, n, \forall x \in \mathcal{X} \qquad \square$$

To prove that the value functions V^* (Q^*) and $\tilde{V}$ ($\tilde{Q}$) are equal we need to show that they are both fix points of the following mapping. And that this mapping satisfies the requirements of the BFT. The value of stochastic stationary behaviors will be sandwiched in between, thus also equal.

Definition 1.3.12. The mapping $T^* \colon B(\mathcal{X}) \to B(\mathcal{X})$ with:

$$T^*V(x) \coloneqq \sup_{a \in \mathcal{A}} \left\{ r(x,a) + \sum_{y \in \mathcal{X}} p(y \mid x, a) V(y) \right\} \qquad V \in B(\mathcal{X}), x \in \mathcal{X}$$

is called the *Bellman Optimality Operator*. With some more abuse of notation, define the mapping $T^* \colon B(\mathcal{X} \times \mathcal{A}) \to B(\mathcal{X} \times \mathcal{A})$ with:

$$T^*Q(x,a) \coloneqq r(x,a) + \sum_{y \in \mathcal{X}} p(y \mid x, a) \sup_{a' \in \mathcal{A}} Q(y, a') \qquad V \in B(\mathcal{X}), x \in \mathcal{X}, a \in \mathcal{A}$$

We will be able to use the established relation of V^π and Q^π in case of deterministic stationary policies It is quite important to remember that this relation is only defined for these policies though. So we need a different approach for the general behaviors. The advantage of these is, that one can condition on a finer level on previous states which will allow us to swap places of suprema and sums.

Lemma 1.3.13.

(i) $\tilde{Q}(x,a) = r(x,a) + \gamma \sum\limits_{y \in \mathcal{X}} p(y \mid x, a) \tilde{V}(y)$

(ii) $\tilde{V}(x) = \sup\limits_{a \in \mathcal{A}} \tilde{Q}(x,a)$

(iii) $V^*(x) = \sup\limits_{a \in \mathcal{A}} Q^*(x,a)$

(iv) $Q^*(x,a) = r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x,a)V^*(y)$

Proof. **(i)** The smaller or equal part is easy:

$$\begin{aligned}\tilde{Q}(x,a) &= \sup_{\pi\in\Pi^{\text{det}}_{\text{stat}}} \left\{ r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x,a)V^{\pi}(y) \right\} \\ &\leq r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x,a) \underbrace{\sup_{\pi\in\Pi^{\text{det}}_{\text{stat}}} V^{\pi}(y)}_{=\tilde{V}(y)}\end{aligned}$$

For the other direction we need to do a bit more work. Since the $r(x,a)$ and γ, are unaffected by the supremum what is left to show is:

$$\sup_{\pi\in\Pi^{\text{det}}_{\text{stat}}} \sum_{y\in\mathcal{X}} p(y \mid x,a)V^{\pi}(y) \geq \sum_{y\in\mathcal{X}} p(y \mid x,a)\tilde{V}(y)$$

This problem should look familiar. It is the question whether there is a single sequence of policies that can match multiple sequences of policies indexed by the state space $y \in \mathcal{X}$. As we can find a policy which can outmatch a finite set of policies (1.3.11), we will consider only the $y \in \mathcal{X}$ with the largest probability to occur and match these sequences with a single sequence. Since we then have just a single policy in the sum, we can estimate up by taking the outer supremum over deterministic policies. That is the idea, whcih can be executed as follows:

The set $M_\delta \coloneqq \{y \in \mathcal{X} : p(y \mid x,a) > \delta\}$ is finite for all $\delta > 0$, and

$$\begin{aligned}1 = p(\mathcal{X} \mid x,a) &= p\left(\bigcup_{\delta\to 0} M_\delta \mid x,a\right) = \lim_{\delta\to 0} p(M_\delta \mid x,a) \\ &= \lim_{\delta\to 0} \sum_{y\in M_\delta} p(y \mid x,a)\end{aligned}$$

Therefore for all $\varepsilon > 0$ exists a $\delta > 0$ such that:

$$\sum_{y\in M_\delta^{\complement}} p(y \mid x,a) < \frac{\varepsilon/4}{R/(1-\gamma)} \tag{1.2}$$

Be $(\pi_y^{(n)}, n \in \mathbb{N})$ with $V^{\pi_y^{(n)}}(y) \nearrow \tilde{V}(y) \quad (n \to \infty)$. Since M_δ is finite, there exists $N \in \mathbb{N}$ such that:

$$|\tilde{V}(y) - V^{\pi_y^{(n)}}(y)| < \varepsilon/2 \qquad \forall n \geq N, \forall y \in M_\delta^{\complement} \tag{1.3}$$

And also because M_δ is finite and 1.3.11 we know:

$$\exists \hat{\pi}^{(n)} \in \Pi^{\text{det}}_{\text{stat}} : \quad V^{\hat{\pi}^{(n)}} \geq V^{\pi_y^{(n)}} \qquad \forall y \in M_\delta \tag{1.4}$$

This finally implies:

$$\begin{aligned}
&\sum_{y\in\mathcal{X}} p(y \mid x,a)\tilde{V}(y) - \sum_{y\in\mathcal{X}} p(y \mid x,a)V^{\hat{\pi}^{(n)}}(y) \\
&\leq \sum_{y\in M_\delta} p(y \mid x,a)(\tilde{V}(y) - V^{\hat{\pi}^{(n)}}(y)) + \sum_{y\in M_\delta^{\complement}} p(y \mid x,a)\underbrace{|\tilde{V}(y) - V^{\hat{\pi}^{(n)}}(y)|}_{\leq \|\tilde{V}\|_\infty + \|V^{\hat{\pi}^{(n)}}\|_\infty} \\
&\overset{(1.4)}{\leq} \sum_{y\in M_\delta} p(y \mid x,a)\underbrace{(\tilde{V}(y) - V^{\pi_y^{(n)}}(y))}_{<\varepsilon/2 \quad (1.3)} + 2R/(1-\gamma)\underbrace{\sum_{y\in M_\delta^{\complement}} p(y \mid x,a)}_{<\frac{\varepsilon/4}{R/(1-\gamma)} \quad (1.2)} \\
&\leq \varepsilon \qquad \forall n \geq N
\end{aligned}$$

This results in:

$$\begin{aligned}
\sum_{y\in\mathcal{X}} p(y \mid x,a)\tilde{V}(y) &\leq \varepsilon + \sum_{y\in\mathcal{X}} p(y \mid x,a)V^{\hat{\pi}^{(n)}}(y) \\
&\leq \varepsilon + \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} \sum_{y\in\mathcal{X}} p(y \mid x,a)V^{\pi}(y)
\end{aligned}$$

Since this equation holds for all $\varepsilon > 0$ we are finished.
(ii) By 1.3.5 we know $V^\pi(x) = Q^\pi(x,\pi(x))$ thus:

$$\begin{aligned}
\tilde{V}(x) &= \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} V^\pi(x) = \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} Q^\pi(x,\pi(x)) \\
&\leq \sup_{a\in\mathcal{A}} \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} Q^\pi(x,a) = \sup_{a\in\mathcal{A}} \tilde{Q}(x,a) \qquad (1.5) \\
&\overset{\text{(i)}}{=} \sup_{a\in\mathcal{A}} \left\{ r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x,a) \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} V^\pi(y) \right\} \qquad (1.6)
\end{aligned}$$

Assume (1.5) is a true inequality for some $x \in \mathcal{X}$, since the suprema in (1.6) can be arbitrarily closely approximated:

$$\exists\pi, \exists a : \tilde{V}(x) < r(x,a) + \gamma \sum_{y\in\mathcal{X}} p(y \mid x,a)V^\pi(y)$$

Define a slightly changed deterministic policy with this π and a:

$$\hat{\pi}\colon \begin{cases} \mathcal{X} \to \mathcal{A} \\ y \mapsto \begin{cases} \pi(y) & y \neq x \\ a & y = x \end{cases} \end{cases}$$

Define $W_n \coloneqq (T^{\hat{\pi}})^n V^\pi$, then:

$$\begin{aligned} W_1(y) = T^{\hat{\pi}} V^\pi(y) &\overset{y \neq x}{=} T^\pi V^\pi(y) = V^\pi(y) \\ &\overset{y=x}{=} r(x, \hat{\pi}(x)) + \gamma \sum_{z \in \mathcal{X}} p(z \mid x, \hat{\pi}(x)) V^\pi(z) \\ &= r(x,a) + \gamma \sum_{z \in \mathcal{X}} p(z \mid x, a) V^\pi(z) \\ &> \tilde{V}(x) \geq V^\pi(x) \end{aligned}$$

In either case we get $W_1(y) \geq V^\pi(y) = W_0(y)$. By induction with 1.3.8 we get: $W_{n+1} = T^{\hat{\pi}} W_n \geq T^{\hat{\pi}} W_{n-1} = W_n$, thus:

$$\begin{aligned} V^{\hat{\pi}}(x) = \lim_{n \to \infty} (T^{\hat{\pi}})^n V^\pi(x) &= \lim_{n \to \infty} W_n(x) \geq W_1(x) \\ &= r(x,a) + \gamma \sum_{z \in \mathcal{X}} p(z \mid x, a) V^\pi(z) \\ &> \tilde{V}(x) \quad \lightning \quad \hat{\pi} \in \Pi^{\text{det}}_{\text{stat}} \end{aligned}$$

□

Corollary 1.3.14. *All the optimal value functions are fix-points of T^**

$$\begin{aligned} T^* \tilde{V} &= \tilde{V} & T^* \tilde{Q} &= \tilde{Q} \\ T^* V^* &= V^* & T^* Q^* &= Q^* \end{aligned}$$

Proof.

$$V^*(x) \overset{\textbf{(iii)}}{=} \sup_{a \in \mathcal{A}} Q^*(x,a) \overset{\textbf{(iv)}}{=} \sup_{a \in \mathcal{A}} \left\{ r(x,a) + \sum_{y \in \mathcal{X}} p(y \mid x, a) V^*(y) \right\} = T^* V^*(x)$$

The others are analogous □

Theorem 1.3.15. *T^* satisfies the requirements of the Banach fixpoint theorem, in particular:*

$$V^*(x) = \sup_{\pi \in \Pi_{\text{stat}}} V^\pi(x) = \tilde{V}(x)$$

is the unique fixpoint of T^*

Lemma 1.3.16. *(Blackwell's condition for contraction)*

Proof. https://math.stackexchange.com/questions/1087885/blackwells-condition-for-a-contraction-why-is-boundedness-neccessary?rq=1 □

Proof (Theorem). □

Now that we proved the uniqueness of the optimal value functions we need to ask the question whether this supremum can be attained with a single policy. And if not, if it can be approximated over all states by the same sequence of policies.

Let us first consider the case where the supremum can be attained, since this includes the important special case of finite MDPs.

Proposition 1.3.17. *Be $\pi^* \in \Pi_{\text{stat}}$, then the following statements are equivalent:*

(i) $\pi^* \in \Pi_{\text{stat}}$ *is optimal ($V^* = V^{\pi^*}$)*

(ii) $\forall x \in \mathcal{X} : V^*(x) = \sum_{a \in \mathcal{A}} \pi^*(a \mid x) Q^*(x, a)$

(iii) $\forall x \in \mathcal{X} : \pi^*(x) = \arg \max_{\pi^* \in \Pi_{\text{stat}}} \sum_{a \in \mathcal{A}} \pi^*(a \mid x) Q^*(x, a)$

(iv) $\pi^*(a \mid x) > 0 \implies Q^*(x, a) = V^*(x) = \sup_{b \in \mathcal{A}} Q^*(x, b)$
"actions are concentrated on the set of actions that maximize $Q^*(x, \cdot)$"
(this also implies: $Q^*(x, a) < V^*(x) \implies \pi^*(a \mid x) = 0$)

Proof. □

Definition 1.3.18. $Q \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ an action value function, $\tilde{\pi} \colon \mathcal{X} \to \mathcal{A}$ with:

$$\tilde{\pi}(x) \coloneqq \arg \max_{\pi \in \Pi_{\text{stat}}} \sum_{a \in \mathcal{A}} \pi(a \mid x) Q(x, a) \qquad x \in \mathcal{X}$$

$\tilde{\pi}(x)$ is called *greedy* with respect to Q in $x \in \mathcal{X}$
$\tilde{\pi}$ is called *greedy* w.r.t. Q

Remark 1.3.19.

- 1.3.17(iii) implies that greedy w.r.t. Q^* is optimal. This means that knowledge of Q^* is sufficient to select the best action.
- 1.3.13 implies that knowledge of V^*, r, p is sufficient as well.

Bibliography

Puterman, Martin L. (Aug. 28, 2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons. 615 pp. ISBN: 978-1-118-62587-3.

Szepesvári, Csaba (Jan. 1, 2010). "Algorithms for Reinforcement Learning". In: *Synthesis Lectures on Artificial Intelligence and Machine Learning* 4.1, pp. 1–103. ISSN: 1939-4608. DOI: 10.2200/S00268ED1V01Y201005AIM009. URL: https://www.morganclaypool.com/doi/abs/10.2200/S00268ED1V01Y201005AIM009 (visited on 02/06/2019).

White, Douglas J. (Dec. 1985). "Real Applications of Markov Decision Processes". In: *Interfaces* 15.6, pp. 73–83. ISSN: 0092-2102, 1526-551X. DOI: 10.1287/inte.15.6.73. URL: http://pubsonline.informs.org/doi/abs/10.1287/inte.15.6.73 (visited on 02/05/2019).