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

Preface

Contents

Introduction

Chapter 1

Markov Decision Processes

Definition 1.0.1. (Kernel) $(Y, \mathcal{A}_Y), (X, \mathcal{A}_X)$ measure spaces
$\lambda\colon X \times \mathcal{A}_Y \to \mathbb{R}$ is a *(probability) kernel* $:\iff$ $\lambda(\cdot, A)\colon x \mapsto \lambda(x, A)$ measurable
$\lambda(x, \cdot)\colon A \mapsto \lambda(x, A)$ a (prob.) measure
Since we will interpret probability kernels as distributions over Y given a certain condition X, the notation $\lambda(\cdot \mid x) \coloneqq \lambda(x, \cdot)$ helps this intuition.

Definition 1.0.2. (Markov Decision Process - MDP)
$\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$, with:

- $\mathcal{X}$ countable (finite) set of states
- $\mathcal{A}$ countable (finite) set of actions

$$\begin{cases} \mathcal{X} \times \mathcal{A} \to \mathscr{P}(\mathcal{X} \times \mathbb{R}) \\ (x, a) \mapsto \mathcal{P}_0(\cdot \mid x, a) \end{cases}$$

transition probability kernel
$\mathscr{P}(\mathcal{X} \times \mathbb{R})$ the set of probability measures on $\mathcal{X} \times \mathbb{R}$,
$\mathcal{X}$ represents the next states,
$\mathbb{R}$ the payoffs

is a *(finite) Markov Decision Process.*
Together with a discount factor $\gamma \in [0, 1]$ it is a:

discounted reward MDP	$\gamma < 1$
undiscounted reward MDP	$\gamma = 1$

For $(Y_{(x,a)}, R_{(x,a)}) \sim \mathcal{P}_0(\cdot \mid x, a)$ a random variable, is

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

An MDP is *evaluated* as follows:
1. Select the initial state X_0 an $\mathcal{X}$-valued random variable.
2. $(A_t, t \in \mathbb{N})$ action selection rules (behaviors) will be discussed later, for now simply assume $\mathcal{A}$-valued random variables.
3. Select inductively: $(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$ with the markov property, i.e.:

$$\begin{aligned} &\mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t) = (x_t, a_t), \dots, (X_0, A_0) = (x_0, a_0)] \\ &= \mathbb{P}[(X_{t+1}, R_{t+1}) = (x, r) \mid (X_t, A_t) = (x_t, a_t)] \end{aligned}$$

resulting in the stochastic process $((X_t, A_t, R_{t+1}), t \geq 0)$, which allows to define the *return*:

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

Remark 1.0.3. $(X_{t+1}, R_{t+1}) \sim \mathcal{P}_0(\cdot \mid X_t, A_t)$ with the markov property is well defined, i.e.:

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

Proof. □

Remark 1.0.4.

1. From now on we assume that $\forall (x, a) \in \mathcal{X} \times \mathcal{A} : |R_{(x,a)}| \leq R \in \mathbb{R}$ almost surely. This also implies: $\|r\|_\infty = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |\mathbb{E}[R_{(x,a)}]| \leq R$
$$|\mathcal{R}| \leq \sum_{t=0}^{\infty} \gamma^t |R_{t+1}| \leq \frac{R}{1-\gamma} \text{ a.s.}$$

2. 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 $\mathcal{X} \to \mathcal{P}(\mathcal{A})$, or alternatively define a (binary) relation on $\mathcal{X} \times \mathcal{A}$.

3. If there is just one admissible action in every state, the MDP is equivalent to a normal Markov Process.

4. 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: $(X_{t+1}, R_{t+1}) = f(X_t, A_t, D_t)$

Definition 1.0.5. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ 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 1.0.6.

- 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

Definition 1.0.7. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ a MDP
An A_t selection-rule $\pi = (\pi_t, t \in \mathbb{N}_0)$ is called *behavior*, where

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

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

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

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

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

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

Π_{stat} is the *set of (stoch.) stationary policies*,
$\Pi_{\text{stat}}^{\text{det}}$ is the *set of deterministic stationary policies* (note $\Pi_{\text{stat}}^{\text{det}} \subseteq \Pi_{\text{stat}}$)

Remark 1.0.8. A stationary policy induces a *time-homogenous* markov chain.

Definition 1.0.9. (Markov Reward Process - MRP)

1.1 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 1.1.1. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP, π Behavior
Select X_0 such that $\forall x \in \mathcal{X} : \mathbb{P}(X_0 = x) > 0$ and evaluate the MDP with $((X_t, A_t, R_{t+1}), t \in \mathbb{N}_0)$ the resulting stoch. process.

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

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

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

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

π is *optimal* $:\iff V^* = V^\pi$

Remark 1.1.2. 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.

Definition 1.1.3. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP
Sometimes we don't care about the probability distribution of the reward, so we define:

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

And use the notation $p(y \mid x,a) := p(\{y\} \mid x,a)$ with $(x,a,y) \in \mathcal{X} \times \mathcal{A} \times \mathcal{X}$

Proposition 1.1.4. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ *MDP,* $\pi \in \Pi^{\text{det}}_{\text{stat}}$

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

[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

Proof.

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

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

Corollary 1.1.5. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ *MDP,* $\pi \in \Pi^{\text{det}}_{\text{stat}}$

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

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 1.1.4 □

Definition 1.1.6. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP, $\pi \in \Pi^{\text{det}}_{\text{stat}}$
The mapping $T^\pi \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ with:

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

is called the *Bellman Operator*

Remark 1.1.7.

1. $\forall \pi \in \Pi^{\text{det}}_{\text{stat}} : T^\pi V^\pi = V^\pi$ (c.f. 1.1.5)
2. T^π meets the requirements of the Banach fixed-point theorem 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

3. T^π is an affine operator

4. $W_1, W_2 \in \mathbb{R}^{\mathcal{X}}$, 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. 2. $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_\infty)$ is a non-empty, complete metric space and the mapping maps onto itself. It is left to show, that T^π is a contraction. Be $V, W \in \mathbb{R}^{\mathcal{X}}$:

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

4. Be $W_1, W_2 \in \mathbb{R}^{\mathcal{X}}$, $W_1 \leq W_2$ and $x \in \mathcal{X}$:

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

$\square$

Definition 1.1.8. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP

$$\begin{aligned}
\tilde{V}(x) &\coloneqq \sup_{\pi \in \Pi_{\text{stat}}^{\text{det}}} V^\pi(x) \\
\tilde{Q}(x, a) &\coloneqq \sup_{\pi \in \Pi_{\text{stat}}^{\text{det}}} Q^\pi(x, a)
\end{aligned}$$

Definition 1.1.9. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ MDP
The mapping $T^* \colon \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ with:

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

is called the *Bellman Optimality Operator*

Lemma 1.1.10. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ *MDP*

(i) $\tilde{V}(x) = \sup_{a \in \mathcal{A}} \tilde{Q}(x, a)$

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

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

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

Proof. **(i)**, **(ii)** By 1.1.5 we know $V^\pi(x) = Q^\pi(x,\pi(x))$ thus:

$$\tilde{V}(x) = \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} V^\pi(x) \leq \sup_{a\in\mathcal{A}} \sup_{\pi\in\Pi_{\text{stat}}^{\text{det}}} Q^\pi(x,a) = \sup_{a\in\mathcal{A}} \tilde{Q}(x,a)$$

Because of 1.1.4 we know:

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

□

Corollary 1.1.11.

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

Proof.

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

$\tilde{V}$ analogous □

Theorem 1.1.12. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ *MDP*
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 1.1.13. *(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). □

Proposition 1.1.14. $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ *MDP*
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) $\forall x \in \mathcal{X} : \pi = \arg \max_{\pi \in \Pi_{\text{stat}}} \sum_{a \in \mathcal{A}} \pi(a \mid x) Q^*(x,a)$

(iv) $\pi(a \mid x) > 0 \iff Q^*(x,a) = V^*(x) = \sup_{b \in \mathcal{A}} Q * (x,b)$
"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 1.1.15. $Q \colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ 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 1.1.16.

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

Chapter 2

Title Chapter 2

Bibliography