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 $\lambda(\cdot \mid x) := \lambda(x, \cdot)$ 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

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

transition probability kernel
$\mu P(\mathcal{X} \times \mathbb{R})$ 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

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

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} denotes the *set of (stoch.) stationary policies* (note that the deterministic policies are a subset of the stochastic policies)

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

Chapter 2

Title Chapter 2

Bibliography